University of Vermont Mathematics

Characterizing the Shapes of Collective Attention using Social Media

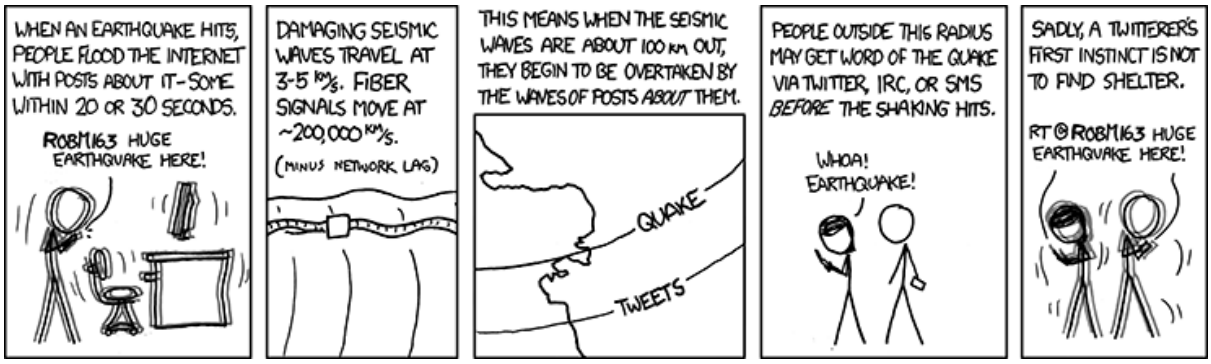
Dilan Kiley Peter Dodds Christopher Danforth
Josh Bongard

May 2, 2014

Special thanks to the members of the Vermont Complex Systems Center and Computational Story Lab at the University of Vermont for their support and input during this project

Abstract

Inspired by the paper "Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System" [2] which analyzed time series of YouTube views, we present an original method for characterizing and visualizing collective public attention using social media. Focussing on 40 billion messages posted to twitter between September 2008 and January 2013, we classify worldwide events into a taxonomy of 5 mathematical shapes. Words corresponding to holidays, political figures, social movements, seasonal trends, celebrities and natural disasters are grouped according to the rate at which their popularity rises and falls in the time series of mentions. In the future, our method will be used to analyze the time series of phrases in many online ecosystems, and quantify and visualize the public response to news events, natural disasters, and policy changes.



[11]

Contents

Chapter 0: Brief Review of *Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System* 4
Chapter 1: Motivation 9
Chapter 2: Taxonomy of Classifications 12
Chapter 3: Redefining the Classification Scheme 14
Chapter 4: Methods 16
Chapter 5: Results 21
Chapter 6: Future Work 34
Chapter 7: Preliminary results of future methods 37
Bibliography 40

List of Figures

1 Christmas and Flu raw time series 10
2 Occupy and Watermelon raw time series 11
3 Google Trends shock event example 12
4 Google Trends anticipated event example 13
5 Google Trends endogenous event example 13
6 Peak fraction classification visualization 14
7 Trend Plot for the “christmas” time series. 23
8 Christmas Time Series, $\rho = 57$ 23
9 Trend Plot for the “flu” time series. 24
10 Flu Time Series, $\rho = 41$ 24
11 Trend Plot for the “occupy” time series. 25
12 Occupy Time Series, $\rho = 99$ 25
13 Trend Plot for the “watermelon” time series. 26
14 Watermelon Time Series, $\rho = 57$ 26
15 Trend Plot for the “christmas” time series (Power Law) 27
16 Christmas Time Series, $\rho = 29$ (Power law fit) 27
17 Trend Plot for the “flu” time series (Power Law) 28
18 Flu Time Series, $\rho = 97$ (Power law fit) 28
19 Trend Plot for the “occupy” time series (Power Law) 29
20 Occupy Time Series, $\rho = 99$ (Power law fit) 29
21 Trend Plot for the “watermelon” time series. (Power Law) 30
22 Watermelon Time Series, $\rho = 93$ (Power law fit) 30
23 Area based Trend Plots 38

Chapter 0: Brief Review of Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System

Here we provide a summary and critique of the inspiring paper by Riley Crane and Didier Sornette. It is the hope that this explanation of the study will aid in illuminating part of the motivation (Section 1) of expanding the techniques of this study to Twitter.

This study by Crane and Sornette examined the possibility of distinct classes which described the shape of views over time on a given YouTube video over time. As mentioned in the original paper, the event of an individual viewing a YouTube video can be influenced, or triggered, in many ways including chance, emails, word of mouth, links on websites, news, and through media sources [2]. To account for this, a complex model must be used to fit an equation to the instantaneous rate of views [2]. The model used in the paper includes two key components, first, an exponential distribution of waiting times, and second, an equation to model the cascading spread of information over a network structure.

To better understand the first component, it is useful to discuss basic theory on how decisions may be made by an individual. In queuing theory, the human thought process is modeled by a dynamic priority queue that helps individuals keep track of various responsibilities, events, and appointments [2, 15]. When the possibility of a new task is introduced, the event is assigned a position in the queue based on its relative priority as determined by the individual [15]. When a task is performed, it is removed from the queue as it no longer requires attention [15]. This can be a complex process as the priority of tasks may change for a variety of reasons, and new events can be added at any moment. The addition of new tasks can be represented with a Poisson random process [15]. In this study, the first component in modeling an instantaneous rate of views is a power law distribution of waiting times used to describe the distribution produced by the queuing process. The exponential decay nature of the distribution means that an individual is more likely to view a video soon after being introduced to it rather than waiting a prolonged period of time [8, 12]. This distribution of delay between cause and action provides the first major component in the model. Here θ is the desired unknown which is determined empirically from the data.

$$\phi(t) \approx \frac{1}{t^{1+\theta}} 0 < \theta < 1 \qquad (\mathbf{1})$$

The second major equation used to produce the dynamic classes is the self-excited Hawkes conditional Poisson process [6]. This equation is used to model the cascading nature of an action, event, or trend over a network. It also is conveniently used to model how an epidemic may spread through a population [2, 6]. The model works well for this study as after individuals view a certain video of interest they may, or may not, impact others to view it through email, text, social media, or in conversation[2].

$$\lambda(t) = V(t) + \sum_{i,t_i \leq t} \mu_i \phi(t - t_i) \tag{2}$$

This "instantaneous rate of views" maintains the rate of views of a particular YouTube video [2]. Here the key parameter, t, represents the current time. Breaking down the equation into pieces, there are two main arguments, the instant turn on views, and the views that are caused by the influence of viewers prior. The first component, $V(t)$, will account for the 'exogenous' views that occur spontaneously in a time series model. The second part of equation 2 models the cascading spread of influence to view a video over time. The summation extends over each individual i who views the video in question at time t_i, where $t_i \leq t$. It is helpful to consider each user as a node in a network, where nodes are linked to other individuals who they may influence to view a video. Thinking in this way, for each i there is a corresponding μ_i which corresponds to its degree, i.e. how many future individuals that individual is potentially capable of influencing [2]. This fluctuating μ_i is then multiplied by the probability that another user has viewed the video in the time interval $(t - t_i)$ which is given from equation 1 described above. In summary, at each time step, each individual who has viewed the topic video becomes an indicator of who may view the video after them which is found by the product of the total number of connections that individual has and the probability that at the time since the original view that a subsequent view will have occurred.

In correspondence to the models outlined above, the study defines distinct 'classes' that will be used to group videos who's views follow a similar shape. In constructing the dynamic classes that will classify each YouTube video that exhibits a burst in activity, two major distinctions are made between a video that exhibits an exogenous burst and a video which exhibits endogenous growth. An exogenous type event has minimal, if any, precursory growth usually due to the unexpected nature of the event [2, 12]. It is characterized by a sudden, sharp spike in growth followed by a rapid decay shortly after the peak is reached. Usually, when speaking of events, these correspond to responses to natural disasters and other unforeseen events. In terms of YouTube videos, these may include highlights of a sporting event in which an unexpected feat was accomplished or simply videos which may become popular once, but don't remain in the public eye. In contrast, endogenous burst are characterized by anticipating precursory growth leading up to a big event, and a somewhat

symmetrical decay afterwards [2, 12] . With the peak usually corresponding to the time when the notable event occurs, there is plenty of time for buildup, and decay is prolonged. This type of burst can be produced, for example, by anticipation for the release of a summer blockbuster, or a presidential election. In terms of YouTube, this may correspond to a movie trailer.

The second distinction that further classifies a video in the study is the notion of criticality [2]. This relates directly to the network analogy of how well connected a network is. If a network is well connected, a trend, information, or an epidemic will spread easier [2]. Conversely, a sparse network will be characterized by minimal spreading as each node is not well connected [2]. In this paper, a network which is "ripe" for spreading (i.e. many connected individuals) is classified as critical, and a network which does not allow for significant cascading influences is classified as subcritical [2]. The following classes were thus determined from data on views of YouTube videos over time to create a model for classes describing collective human behavior using equations 1 and 2.

Exogenous Sub-Critical:

This class corresponds to the class of activity over a sparsely connected network; in this paper this is defined to be when $\langle \mu_i \rangle < 1$. On a network, this corresponds to a situation where the average degree of each node is less than 1, so there are many isolated nodes. The sub-critical label stipulates that the initial burst of activity, occurring at t_c, that began the burst does not cascade more than a few generations. This model is proportional to equation 1 [2].

$$A_{bare}(t) \approx \frac{1}{(t - t_c)^{1+\theta}} \tag{3}$$

Exogenous Critical:

In this class the network is ripe for a particular video i.e., has a well-connected following of interested individuals, spreading is observed to continue over numerous generations as these individuals continue to influence their contacts to view the topic video. Here $\langle \mu_i \rangle$ is defined to be close to 1, so in almost every case, each individual will have the potential to inspire a neighbor (in the network sense) to view the video [2].

$$A_{ex-c}(t) \approx \frac{1}{(t - t_c)^{1-\theta}} \tag{4}$$

Endogenous Critical:

Here, in addition to a ripe network, the topic video has some notion of expectation, that is to say, it likely pertains to an event that users can anticipate. This precursory growth and word-of-mouth spreading leads to symmetric-like growth and decay around the peak time, t_c. As seen in the model below, the $t - t_c$ term is in absolute value to account for a non-zero number or views before the peak event, corresponding to an endogenous event. The buildup and decay rates surrounding this peak event are, in principle, fairly symmetric around the t_c point [2].

$$A_{en-c}(t) \approx \frac{1}{|t - t_c|^{1-2\theta}} \tag{5}$$

Endogenous Sub-Critical

This last class serves as a catch-all for videos that do not exhibit bursts, but simply fluctuate stochastically. This class is not one of the dynamic classes that were discovered as the fluctuations can be approximated with a general Poisson process [2].

$$A_{en-sc}(t) \approx \eta(t) \qquad \text{where } \eta(t) \text{ is a noise process} \tag{6}$$

Once a time series plot of the total number of views around the burst of activity had been obtained, a peak fraction analysis was performed as a preliminary sort into one of the three classes (equations 3-5) [2] . The peak fraction of view is defined to be the fraction of view that occurred on the day that the rate of views peaked [2]. By the nature of the exogenous subcritical class, there is little growth, and little cascading into following generations, so it is predicted that most of the total views will occur on the defined peak day, accordingly, the peak fraction, $80\% \leq F \leq 100\%$ [2]. For the exogenous critical class, there is

still no preliminary growth, but the views may have influence on a number of generations after the peak, thus the peak fraction will be slightly decreased when compared to the exogenous subcritical; $20\% \leq F \leq 80\%$ [2]. Unlike the exogenous classes, the endogenous class is characterized by substantial precursory growth followed by a slow decay. As a result, the peak fraction for endogenous critical events will be significantly smaller than the exogenous classes; $0\% \leq F \leq 20\%$ [2]. See Figure 6.

Once the video has been sorted into one of the three classes, the decay exponent for each event was determined through a least-squared fit on the logarithm of the data on the videos in that class, as determined by the peak fraction analysis [2].

Results: After examining a large sample of YouTube videos to search for burst-type activity in views that may fit into one of the three dynamic classes only about 10%(about 500,000 videos) fit [2]. After sorting the refined set of videos through the peak fraction analysis described above, the distribution of the exponents was assessed to obtain a value for θ in the class equations. Based on a plot of likely exponents, the following value $\theta = 0.4 \pm 0.1$ was determined [2]. As a result, the exogenous subcritical exponent becomes 1.4, the exogenous critical, 0.6, and the endogenous critical, 0.2.

It is highlighted in the paper that the common value of $\theta = 0.4$ which applies to all three dynamic classes is in agreement with other studies such as one that examined the rank of book sales which found $\theta = 0.3 \pm 0.1$ [2] [12]. It is argued that θ affects the overall persistence of the collective number of views [2]. In this way, a large θ value corresponds to a case where the individual response time to the external, initial introduction of the topic, here a YouTube video, is rapid. Crane and Sornette highlight the paradoxical effect that a larger θ leads to a slower, more persistent response in the collective system [2].

Chapter 1: Motivation

Since its creation in 2006, the number of messages posted on the social media website Twitter has increased to nearly 400 million per day (March, 2013) [14]. Twitter has become a viable resource in the field of data science and analytics [1]. The instantaneous and widespread response of its users provides a wealth of opinions, reactions, and statements about events happening all over the world, yielding ample opportunity to study and quantify the behavior of the activity Twitter users generate.

Twitter provides an ideal environment to study volume around particular words and phrases due to the intrinsic information associated with individual words. In this way, examining a relative frequency plot of a particular word over time can provide very useful information not only about the amount of tweets that occur around the period of a major event, but also are directly tied to the meaning of the word that the tweet frequency represents. A time series, in this study, is described as a data plot that tracks the relative number of tweets containing a particular word over time, i.e the history of that word. In this way, as one can imagine, a word's usage over time may fluctuate greatly, if it is tied to a very opinionated topic, a major event, or a seasonal trend. The potential of tracking search volume and user activity on the Internet has just begun to be realized by the scientific world [1]. The goal of this research project is threefold.

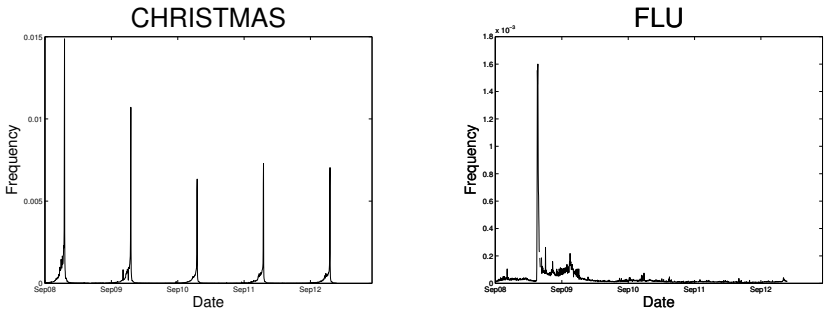
First, the project aims to understand and quantify the behavior of time series in a meaningful way. Given a time series of interest one would hope to achieve the ability to isolate data from a well-fit burst in activity, and extract useful information about the burst that can then be used to better understand the event or topic that a word is associated with. In this way there becomes a straightforward method to quantify a time series and break it down into meaningful events which are then classified according to a defined taxonomy of event shapes.

Second, the project will, through the analysis of carefully selected time series, gain an understanding of the implications of such bursts in monitoring public opinions and understanding human behavior. In particular, we create a taxonomy of the shapes of events observed on Twitter. In addition to identifying and classifying bursts in time series the study aims to develop a method which will determine the best time scale to fit event intervals to. The data for each time series is the relative frequency of a particular word's usage measured every day over the interval of September 2008 to January 2013. This corresponds to 1606 data points for each time series. We classify the event duration by the interval in which the burst is

observed, thus a burst over 21 data points would correspond to a real-world trend of three weeks.

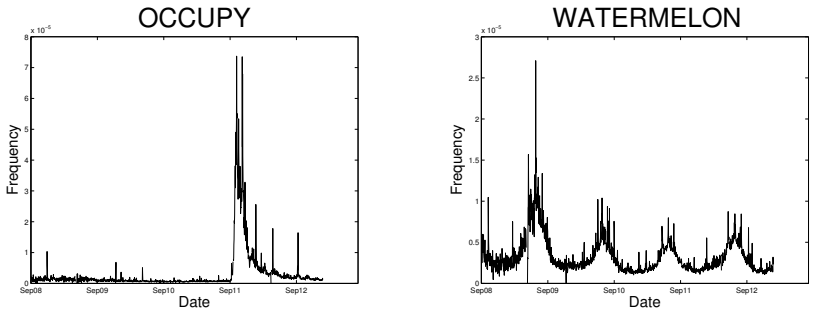
Third, the project will work towards the ability to replicate the study performed by Crane and Sornette on YouTube but instead on Twitter. Of particular interest is determining the θ factor for bursts on Twitter to see how it compares to the value of $\theta = 0.4$ described in the Crane paper [2]. In addition, this study will begin to explore other functional forms, such as the exponential, for the possibility that they provide a better explanation of burst activity on Twitter.

One can imagine that the data for any given word over a large period of time is very noisy and this is exactly the case with most time series we encounter. In particular, this creates a challenge when fitting event intervals as there is a large diversity in the duration of interesting bursts in almost any time series. Many burst may appear and disappear over a period of only a few days. Other trends have a much longer duration. In addition, a large scale trend may contain a wealth of micro-bursts that cumulatively make up the large scale trend. Of course for many time series one would hope to capture the large scale trends with a long duration, while for equally as many other times series one would like to fit very small bursts that only occur for brief periods. This becomes one of the challenges that this study addresses and develops a method to deal with such diversity in burst duration and makeup. Below we exhibit four characteristic time series that demonstrate some of the aforementioned diversity in burst and trend shape.



(a) Christmas raw time series (b) Flu raw times series

Figure 1: (a) Raw time series for mentions of the word "christmas" on Twitter. One notices a period of build up before each Christmas day. Classifications of this time series would hope to find five event intervals which correspond to each Christmas. (b) Raw time series for "flu" on Twitter. Here we exhibit the uncertainty into classifying the number of meaningful event intervals as well as their durations. One clear burst, corresponding to the Swine Flu scare in 2009 [4], is evident as well as a few smaller intervals which have a type of burst growth and decay.



(a) Occupy raw time series (b) Watermelon raw times series

Figure 2: (a) Raw time series for the word "occupy." Here we highlight the notion of a macro-trend, the Occupy movement [13], which itself is composed of small micro-shocks. The analysis aims to recognize the large scale trend of the occupy movement and classify it as one event. (b) Raw time series for mentions of "watermelon". This time series is unique in that the bursts in activity follow a much more relaxed and seasonal pattern. In particular, the bursts are not as extreme, but instead are somewhat symmetrical in build up and decay. Analysis hopes to capture the seasonal fluctuations of watermelon mentions.

Chapter 2: Taxonomy of Classifications

As in the study by Crane and Sornette on YouTube videos we define distinct classes to describe the shape of a given burst interval. Previously defined by Crane and Sornette were four major classes; endogenous (sub-critical / critical) and exogenous (sub-critical / critical) [2]. Here, an endogenous burst would be characterized by a symmetrical, usually slower, build-up and decay around a peak and a exogenous burst would correspond to a sudden shock and characterized by a sharp spike in frequency followed by a decay. The duration of the decay determined the classification of sub-critical or critical. Only the endogenous critical events were considered as the sub-critical type is reserved for time series which don't exhibit any large burst of activity [2]. Thus there are three major classes examined by Crane and Sornette.

In this study we expand our classifications from three to five as we examine time series on Twitter. In shifting the medium from the Crane and Sornette study on YouTube [2] to Twitter it quickly becomes apparent that an expansion of the types of events is required. Whereas on YouTube, most burst activity is described by a "shock" type event in which a video becomes viral overnight, on Twitter there are many instances where there is a profound anticipation surrounding a given word, followed by a sharp decline. We expand our classifications to include a distinction between "shock" events characterized by a sharp increase in frequency and corresponding decay and "anticipated" events in which there is a notable buildup to a peak in frequency followed by a rapid decline. In many ways the anticipated bursts behave as mirror images of the shock type events. Just as the decay exponent of events was the focus of the YouTube study [2], in this study we maintain this view and place the same emphasis on the buildup exponent for anticipated events.

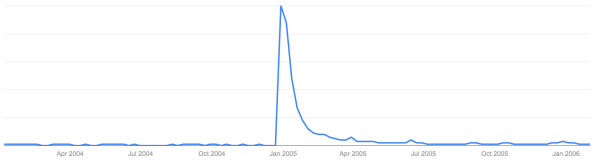


Figure 3: Google Trends [7] produces the following for the search volume of "Tsunami" over the time period of early 2004 to April 2006. This type of event is what is typical of a shock type event. As shown there is almost no build-up precluding a sharp spike. After the peak there is a relaxation. This decay exponent is the interest of the YouTube study as well as this one.

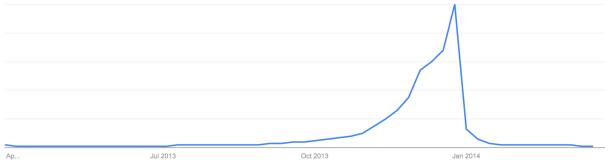


Figure 4: Google Trends [7] search volume for the word "Christmas" for the 2013 holiday season. This is the perfect example of an anticipated event, which we introduce in this study. In contrast to a shock event like a tsunami, search volume, and Twitter mentions of Christmas has a profound buildup and almost immediately drops off. The build-up exponent is what we will be interested in for anticipated events.

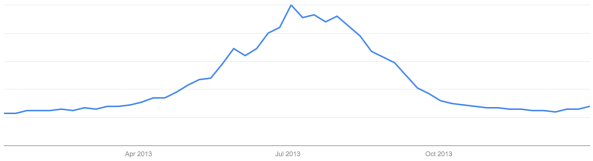


Figure 5: As in the YouTube study ([2], symmetrical events (endogenous), which have similar buildup and decay, are considered in this study. Words like "Watermelon", shown above from Google Trends [7], exhibit this seasonal and symmetric type event.

As depicted in Figures 3-5 we note the three major classifications that we will refer to for the remainder of the report, anticipated, shock, and endogenous. Within the anticipated and shock classifications we also distinguish between sub-critical and critical which corresponds to the relaxation and buildup of the shock and anticipated events respectively.

Chapter 3: Redefining the Classification Scheme

In the study conducted on YouTube [2], events were classified by a peak fraction defined as the number of views on a given video divided by the total number of views for that video (See Figure 6 below from Crane and Sornette paper [2])

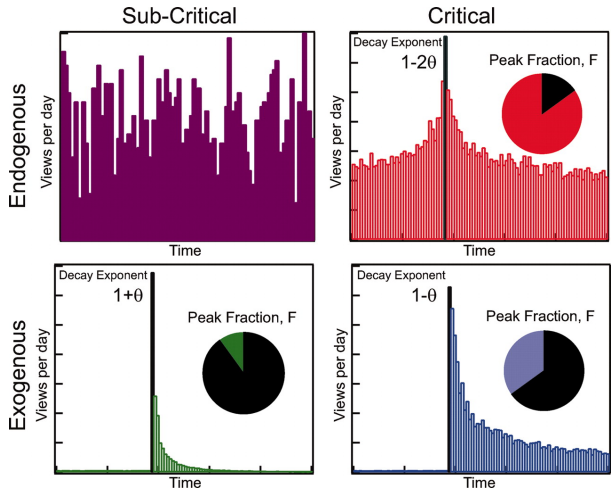


Figure 6: Peak fraction classification visualization. Seen in the image above from the Crane and Sornette paper are distinctions between sub-critical and critical events. [2]

As mentioned above, the notion of a peak fraction is introduced. In the Crane study the classification for an event was determined by a simple metric of the peak fraction [2]. A large peak fraction would therefore correspond to a video that had a large spike and quickly relaxed, exogenous sub-critical. A slightly lower peak fraction would correspond to a similar situation but the exogenous event would exhibit more memory since the initial shock and thus maintain a higher number of views in the relaxation [2]. Lastly, the endogenous critical class would have a relatively low peak fraction as this class is not characterized by significant, sudden bursts but rather a slow and symmetric build-up and relaxation [2].

Here we define a new way to classify a burst of activity in a time series. After determining the event interval around a peak we fix a function to the buildup interval and the decay interval. After both are determined we take the difference of the two exponents (or powers) to classify the event. We call this the "Score" of the event.

$$\text{Score}_i = \alpha_i - \beta_i \tag{7}$$

where α_i and β_i correspond to the exponential, or power, coefficient in the fit to the relaxation and build-up intervals of the i^{th} burst in a time series respectively.

An expected event will have a negative score, whereas a positive score will correspond to a shock event. How positive or negative the score is will classify the event as critical or sub-critical. This translates to how quickly the decay occurs after the peak of a shock event. Endogenous events, such as watermelon, have near zero scores. We classify events according to their score on a scale bracketing zero.

$$\text{Classification}_i = \begin{cases} \text{Anticipated Exogenous Sub-Critical,} & \text{if } -\infty \leq \text{Score}_i \leq -1 \\ \text{Anticipated Exogenous Critical,} & \text{if } -1 \leq \text{Score}_i \leq 0.01 \\ \text{Endogenous,} & \text{if } -0.01 \leq \text{Score}_i \leq 0.01 \\ \text{Shock Exogenous Critical,} & \text{if } 0.01 \leq \text{Score}_i \leq 1 \\ \text{Shock Exogenous Sub-Critical,} & \text{if } 1 \leq \text{Score}_i \leq \infty \end{cases} \tag{8}$$

Chapter 4: Methods

Before detailing the process to produce the best results we take a brief detour to outline prior attempts at determining event intervals in a time series that did not yield as much success but nonetheless lend insight into the series of attempts which led to the current version. While the scheme, described in section 3, of determining the classification for an event interval based on its Score_i remains constant across all methods that will be outlined what will change is the way in which the event intervals are determined.

In hopes to make the proposed method of classifying a time series as transparent as possible we continue by detailing the exact procedure performed on each time series in a "pseudo-code" manner. While the methods described are the most recent iteration of this procedure, the refinement of the process is ongoing and may improve as new approaches are applied to the main method.

Determining the best scale to fit an event interval is indeed a quite challenging problem. As described earlier, any one given time series may contain events that occur on many scales and must be fit accordingly. Here we decide that that an imperfect fit of small event over large duration is preferred to finding an interval too that is not large enough to span the lifetime of a large scale event. That is to say that we would rather overestimate the interval in which an event is defined to occur than underestimate it.

We continue by outlining some attempts on fitting the best interval around a given spike in a time series:

Method 1:

We present the following scheme for determining event intervals around a given spike by monitoring the R-squared statistic for a fit as the length of a fit interval is increased. Given a local maximum in a time series, t_0, we set out to find the best place to begin the event, t_{start} and also end, t_{end}. We propose the following possibility. To determine either t_{start} or t_{end} we set out from the local maximum in one direction, either forwards in time, to determine t_{end}, or backwards, to determine t_{start}. As we expand the bound on our interval we include more and more data points in out interval of interest. Accordingly we continuously fit our functional form to the current interval, determine the value for the parameter of interest, and also record the R-squared statistic for this fit. We

continue to expand our interval in both directions. We then look for the point at which there is a marked drop-off in the R-squared statistic. This would indicate that the fit of the functional form is no longer good and the interval is thus defined. We do this moving outward in both directions from t_0 and yield a value for t_{start} and t_{end}. The event interval for that local maximum is now defined to be $\{t_{start} : t_{end}\}$.

We find that in general, due to the large amount of noise in the data for any given time series that the R-squared statistic can fluctuate greatly as we increase our interval size. This in turn produces another noisy time series to analyze with the hope of determining a critical point where the R-squared statistic no longer is sufficiently large. In particular this method was observed to yield very small event intervals and performed poorly when applied to large scale trends with long event durations.

Method 2

In a similar attempt as in Method 1 here we modify our approach slightly by instead of monitoring the R-squared statistic as we increase the size of the event interval but instead the value of the fit parameter of interest. As in Method 1 after we determine a local maximum at t_0 we set out in both directions to determine t_{start} and t_{end}. For each iteration we increase the size of the fit interval and record the value of parameter of interest (ex. for a power law this value would be the α in the fit $\frac{1}{t^{\alpha}}$). We monitor the value of this parameter over time and look for the point at which the slope of the line connecting two successive values of the parameter undergoes a sharp change. This would indicate that the interval has grown too large and perhaps the interval has collided with a neighboring event interval.

Here, as in Method 1, we find that it is difficult to find the ideal time to stop the fit, and attempts to do so yielded very small event intervals.

Method 3

In contrast to Methods 1 & 2, in Method 3 we explore the ability to smooth a time series to aid in determining the start and end of each interval. Prior studies have explored the avenue of using smoothing to help classify a time series [16]. Here we present an original method which incorporates the use of smoothing by a moving average. We define a smoothing parameter ρ which we will use to determine the best scale to define event intervals on. For each value of ρ between the odd values of 1 and 99 we compute a moving average of the time series in

question where the size of the moving average window is ρ, centered around each point in the time series. This naturally will temper the noise in the data and also will make spikes more moderate. Especially on Twitter data since the resolution is daily, many time series exhibit a data point, corresponding to one day, in which there is a large spike that far exceeds the activity on the day prior and after. This smoothing procedure will produce a subdued version that can then be used to determine new event intervals.

Given a particular smoothing parameter, ρ, we have a smooth version of our original time series. We define an event interval in the following way. After finding the peak location of a local maximum we set out in each direction of t_0. We continue in each direction until a value of a local minimum is reached. Upon reaching the first local minimum the interval expansion is ceased. In this way we have defined the interval $\{t_{start} : t_0\} \cup \{t_0 : t_{end}\}$ where t_0 is a local maximum and t_{start} and t_{end} are local minimums of the smoothed time series.

Given one of the aforementioned methods for determining the individual event intervals we continue by walking through the process of quantifying a time series into a sequence of classified events.

In addition to choosing a method for determining event intervals we also must choose the form of the function in which to fit each buildup and decay interval for each interval. Here we explore two major types of classification, an exponential function, and a power law function. Given an event interval we fit one of the two functions to the intervals $I_{pre} = \{t_{start} : t_0\}$ and $I_{post} = \{t_0 : t_{end}\}$. We define the functional forms of the functions as follows:

Exponential

$$V(t) = \alpha e^{(t-t_0)\gamma} \tag{9}$$

Equation 9 describes the exponential functional form for the decay interval, I_{post}. For I_{pre} we note that we would have a $t_0 - t$ .The parameter γ is the key value which will be used to determine Score_i described in equation 7.

Power Law

$$V(t) = \alpha(t - t_0)^{\gamma} \tag{10}$$

Equation 10, like 9 describes the fit for the decay interval. Likewise the parameter γ is what will be used in equation 7.
We define the following procedure for classifying a given time series.

1. Determine the locations of the local maxima, t_{0_i} defined to be peaks that exceed the average of the span of the time series.
2. Determine a preliminary I_{pre} and I_{post} for each t_{0_i} by finding the closest local minimum on each side of t_{0_i}.
3. Resize the interval for each as t_{0_i} according to the method(1,2,or 3).
4. Once the I_{pre} and I_{post} are determined, fit the corresponding functional form to each interval to determine the parameters α_i and β_i in equation 7 which correspond to the γ's described in equations 9 and 10.
5. Determine the Score$_i$ for each local maxima found in step 1.
6. Classify the event interval $\{t_{start} : t_{end}\}$ according to the scale in equation 8.

For methods 1 and 2 the above method is all that is used to determine the classification of the events in a time series. For method 3 however there is the added caveat that we repeat the above process for each smoothing parameter, $\rho \in 1 - 99 \mid \rho$ is an odd number. For each value of ρ we repeat steps 1-6. We note that as we increase the value of ρ we find fewer local maximums in step 1. We thus have produced 50 classifications for the given time series each fitting the events to a larger time scale.

We now clarify how we decide which of the 50 classifications best describes the time series. For each classification corresponding to a value of ρ:

1. Compute the R-squared value for the fit of both the I_{pre} and I_{post}. For each interval determined around each t_{0_i} we average the two R-squared values. This becomes the R-squared for that event interval.
2. Compute the p-value for the fit of both the I_{pre} and I_{post}. For each interval determined around each t_{0_i} we keep the maximum of the two p-values. This becomes the p-value for that event interval.
3. Average all the R-squared values for all event intervals found, call this R_ρ.
4. Find the maximum of all the p-values for all event intervals found, call this p_ρ.

To determine the best parameter, ρ, for a given time series we find the classification which maximizes the ratio $\dfrac{R_\rho}{p_\rho}$. To visualize this incrementing of the smoothing parameter we develop a "Trend Plot" which plots all classifications for a time series for each value of ρ. This presents a visual representation of the transition of the events in a time series from their micro features to their macro trends.

It is also important to note one additional change to the model as the value of ρ is increased. Given the natural tendency of an event which has a larger lifetime to have similar γ values in equations 9 and 10 we must adjust the initial threshold of the classifications

described in equation 8 very slightly. For this model we choose to decrease the exogenous critical/sub-critical threshold and endogenous threshold by a total of 90% from the original value achieved at the final ρ value. Thus at a ρ value of 99 we will have a new classification scheme:

$$\text{Classification}_i = \begin{cases} \text{Anticipated Exogenous Sub-Critical,} & \text{if } -\infty \leq \text{Score}_i \leq -0.1 \\ \text{Anticipated Exogenous Critical,} & \text{if } -0.1 \leq \text{Score}_i \leq 0.001 \\ \text{Endogenous,} & \text{if } -0.001 \leq \text{Score}_i \leq 0.001 \\ \text{Shock Exogenous Critical,} & \text{if } 0.001 \leq \text{Score}_i \leq 0.1 \\ \text{Shock Exogenous Sub-Critical,} & \text{if } 0.1 \leq \text{Score}_i \leq \infty \end{cases} \quad \textbf{(11)}$$

Chapter 5: Results

Here we present the Trend Plots for each of the four characteristic time series described in Section 1. First we present the Trend Plots for the exponential fit described in equation 9. We then present the same plots for the power law fit described in equation 10. Since we are creating the Trend Plots it is transparent that the method being implemented in the following plots is Method 3.

We think it beneficial for the reader to explain the mechanism of the Trend Plots presented in Figures 7-22. The Trend Plots are a variant of the denodrogram plot in that it presents a visualization of the shapes of events in a time series in a hierarchy of micro to macro scale. It is also similar in the way in which it optimizes the best classification of a time series. We now describe the layout of the Trend Plot.

Along the top of the plot, the raw time series for the given word is plotted in red. Directly below the time series is a series of classifications for the time series with varying ρ values which increases as one moves down the vertical axis. Recall that a lower smoothing parameter, ρ, will result in smaller event intervals and thus more events. For each value of ρ the event intervals found with that corresponding value are plotted as colored bars. The color of each section corresponds to the type of event that interval was classified as. For clarification we define:

$$\text{Color}_i = \begin{cases} \text{Powder Blue} & \rightarrow & \text{Anticipated Exogenous Sub-Critical} \\ \text{Yellow} & \rightarrow & \text{Anticipated Exogenous Critical} \\ \text{Green} & \rightarrow & \text{Endogenous} \\ \text{Red} & \rightarrow & \text{Shock Exogenous Critical} \\ \text{Blue} & \rightarrow & \text{Shock Exogenous Sub-Critical} \end{cases}$$

Note that there is no overlap between the event intervals within one value of ρ. However, if two events of the same type appear adjacent in the time series, there is no distinction plotted on the Trend Plot. This will show up on the subsequent classification of the time series.

Along the Eastern edge of the Trend Plot the average R-squared statistic, R_ρ is plotted as the ρ value is varied. Likewise, p_ρ and the ratio we wish to optimize, $\dfrac{R_\rho}{p_\rho}$ is plotted in a

similar fashion. R_ρ is plotted in a thin red line, p_ρ in a thin blue line, and $\frac{R_\rho}{p_\rho}$ in a thicker green line. The domain for the R-squared statistic and the p-value is naturally [0,1]. Thus we adopt this same scale to plot $\frac{R_\rho}{p_\rho}$ on. To do so we normalize by adjusting all computed ratios of $\frac{R_\rho}{p_\rho}$ by the max value found across all value of ρ. In this way the best classification using this scheme will be denoted when the ratio reaches 1.

To illustrate where this maximum is achieved a thin dashed line is drawn across the main figure at the smoothing parameter which maximized the ratio $\frac{R_\rho}{p_\rho}$. Additionally, the smooth time series which determined the scale of the events is plotted over the top plot of the raw time series in black.

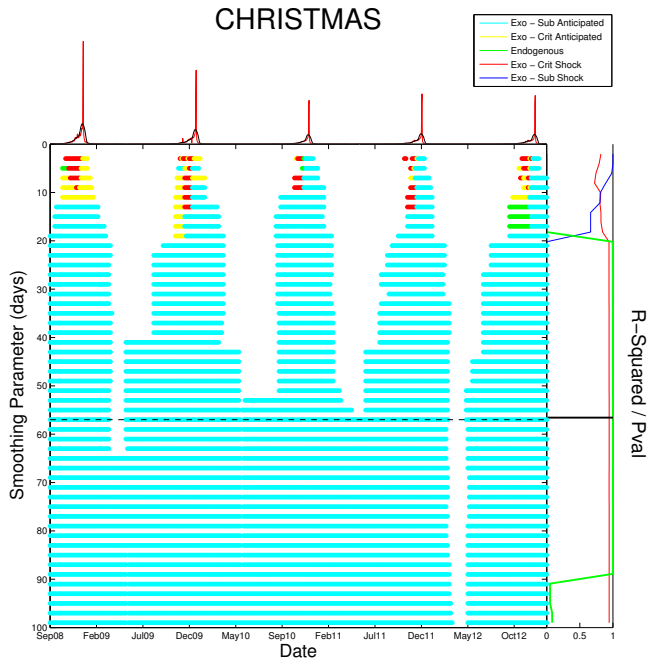


Figure 7: Trend Plot for the "christmas" time series.

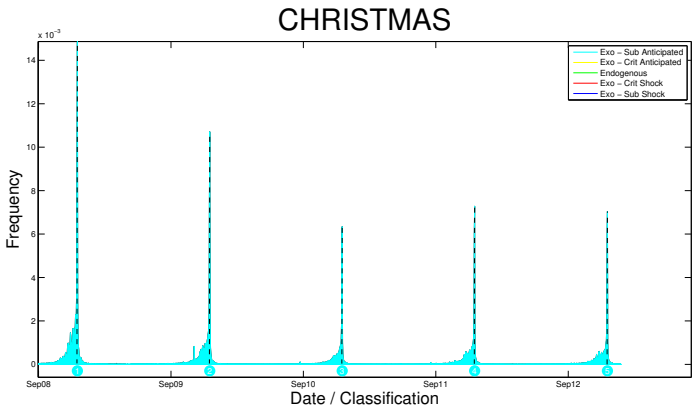


Figure 8: Christmas Time Series, $\rho = 57$

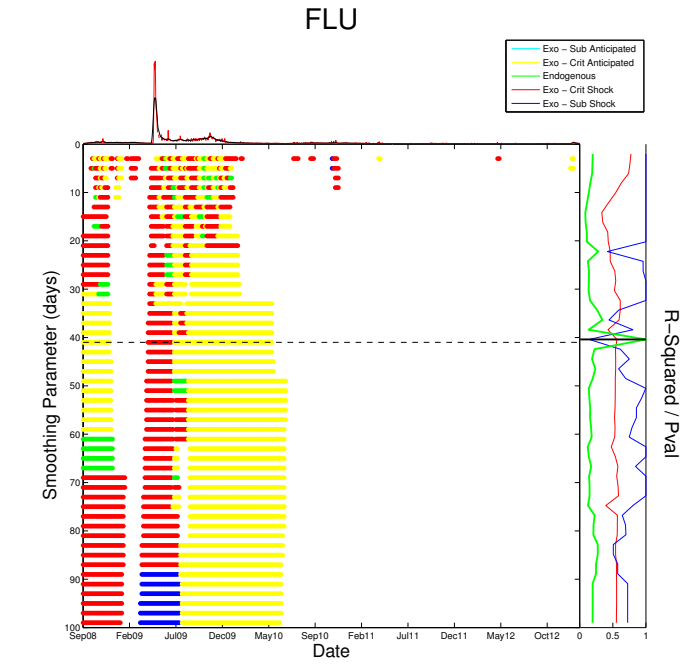


Figure 9: Trend Plot for the "flu" time series.

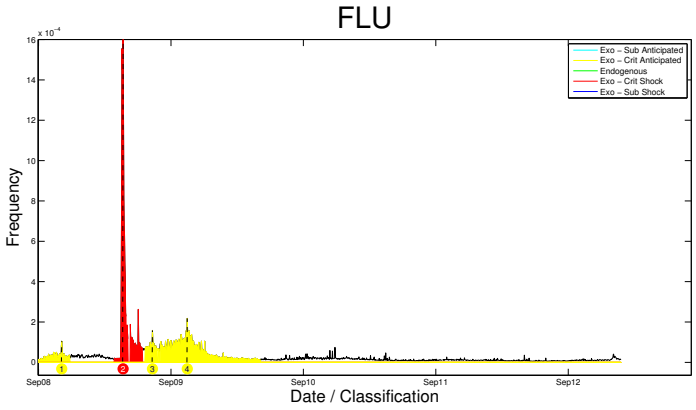


Figure 10: Flu Time Series, $\rho = 41$

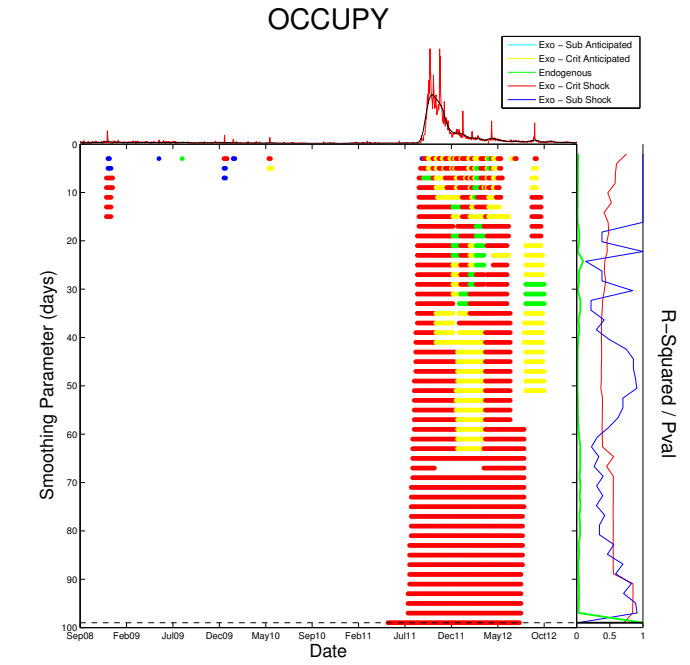


Figure 11: Trend Plot for the "occupy" time series.

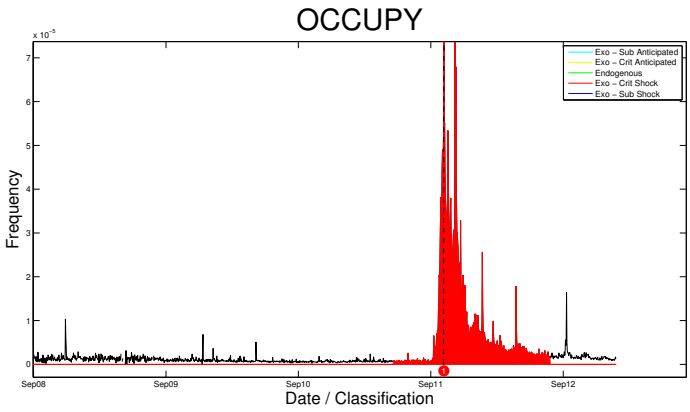


Figure 12: Occupy Time Series, $\rho = 99$

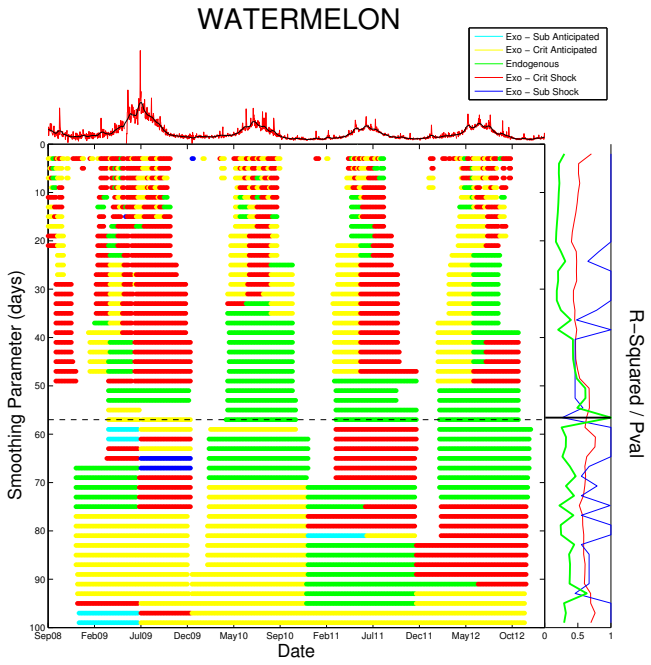


Figure 13: Trend Plot for the "watermelon" time series.

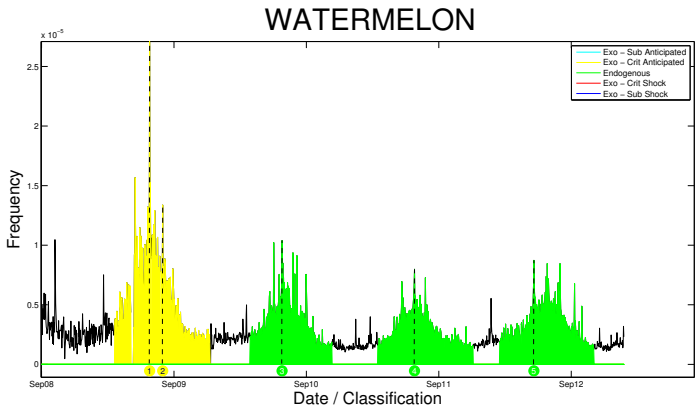


Figure 14: Watermelon Time Series, $\rho = 57$

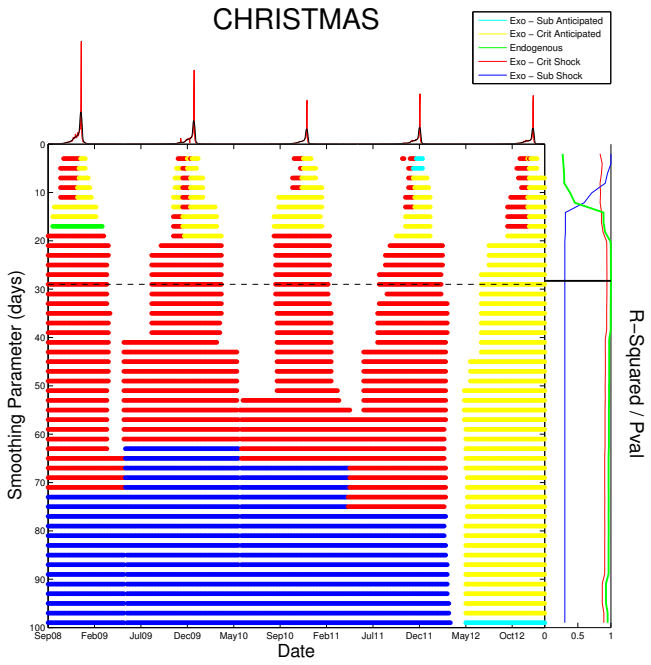


Figure 15: Trend Plot for the "christmas" time series (Power Law)

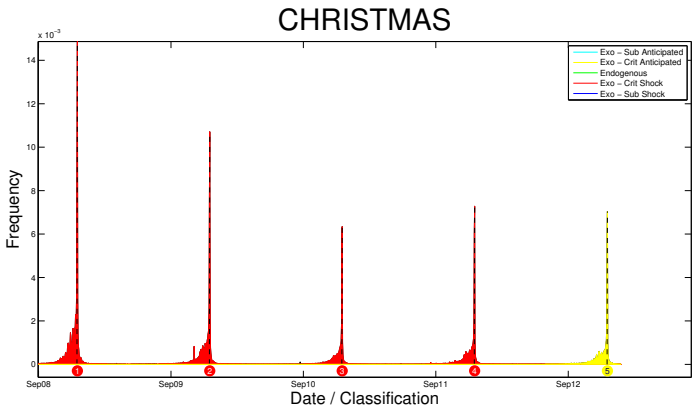


Figure 16: Christmas Time Series, $\rho = 29$ (Power law fit)

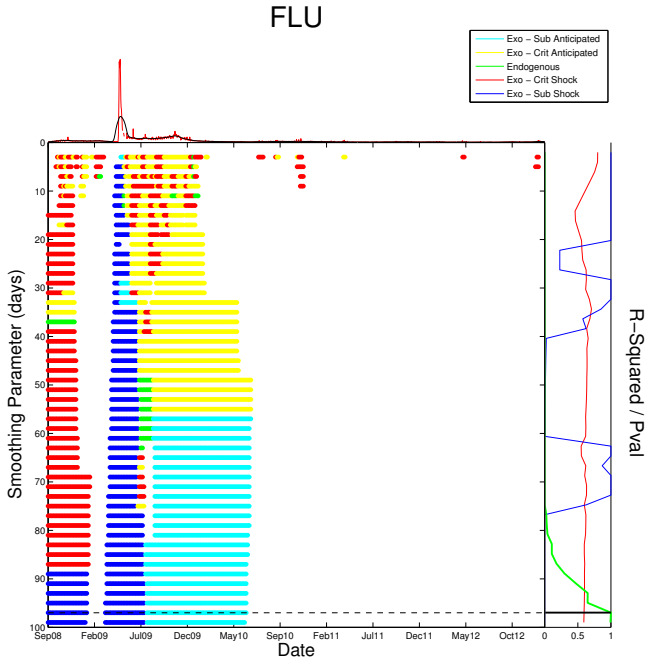


Figure 17: Trend Plot for the "flu" time series (Power Law)

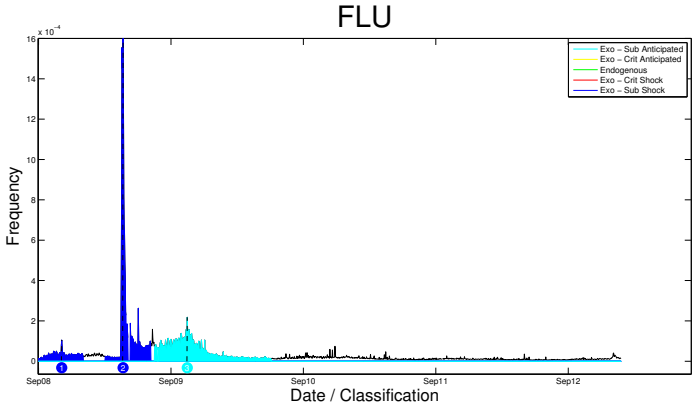


Figure 18: Flu Time Series, $\rho = 97$ (Power law fit)

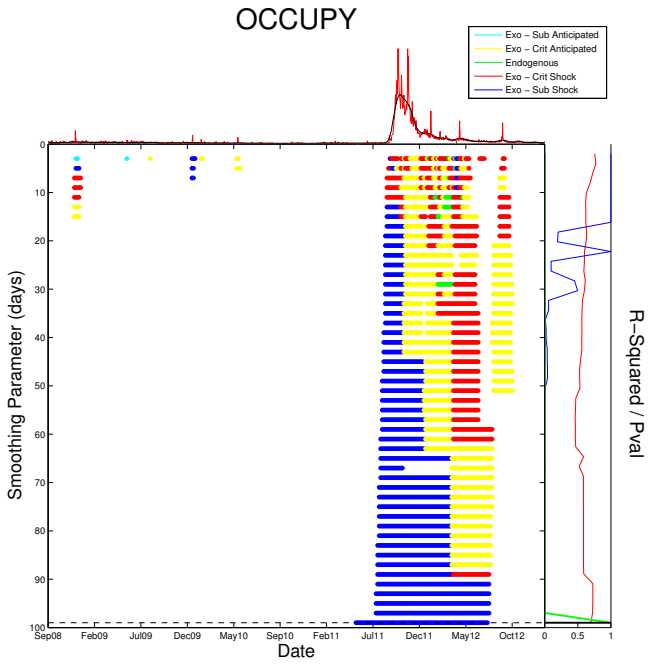


Figure 19: Trend Plot for the "occupy" time series (Power Law)

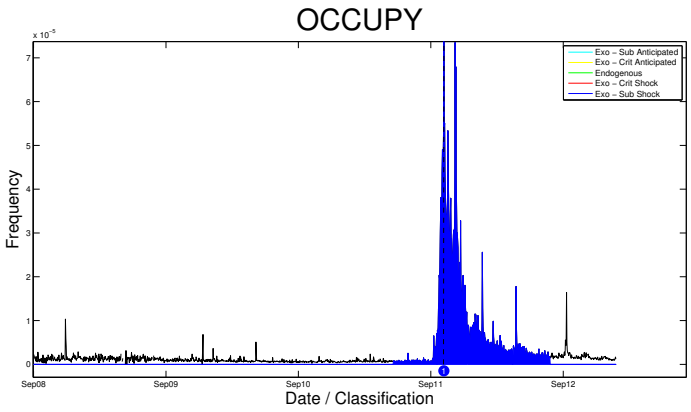


Figure 20: Occupy Time Series, $\rho = 99$ (Power law fit)

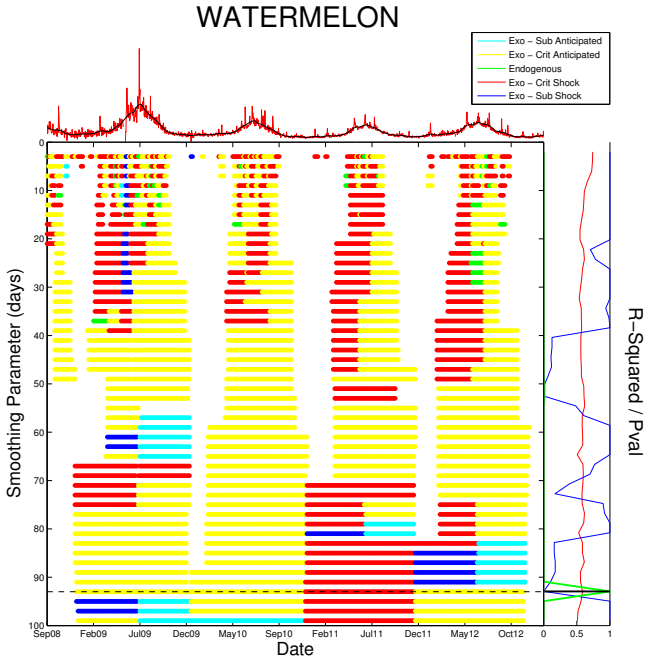


Figure 21: Trend Plot for the "watermelon" time series. (Power Law)

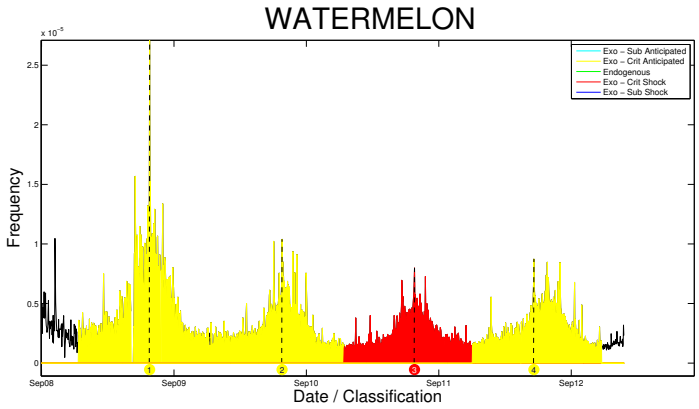


Figure 22: Watermelon Time Series, $\rho = 93$ (Power law fit)

We begin our discussion of Figures 7-22 with the exponentially fit Trend Plots. As seen in Figure 7, the optimized value of $\frac{R_\rho}{p_\rho}$ occurs at $\rho = 57$. However, we see that the classification for every Christmas as a anticipated exogenous sub-critical event is robust for values of $\rho \in [21 : 89]$. This broad plateau of the $\frac{R_\rho}{p_\rho}$ value indicates not only that this is the correct classification for the Christmas time series, but also that the fitting of the model using equation 9 is reasonable. In addition, this successful classification lends merit to the optimization of the $\frac{R_\rho}{p_\rho}$ value as a means of determining the best ρ and hence the best scale to consider the major events in a time series. Although the simplicity of the christmas time series

represents an idealized testing ground, it provides the best example for describing the ideal dynamics of the Trend Plot scheme. We note that for ρ values up to 21 there is a period of transience in which smaller event intervals within the larger scale trend of one Christmas season are identified. As ρ increases, the classification of anticipated exogenous sub-critical remains.

In Figure 8 we explore the Christmas time series with the event regions colored according to their classification. We see that there are only five peaks found, each corresponding to one Christmas. Also shown on this plot is the location of the peaks determined by the model (plotted with dashed lines).

Figure 9 shows a more characteristic Trend Plot of a time series in the dataset. As compared with the Christmas plot we see that there is a larger variety in the type of events in the time series. Here we also note that the value of $\frac{R_\rho}{p_\rho}$ does not show a plateau but instead a sharp peak. Some of the sharpness in the spike is due to the normalization of the $\frac{R_\rho}{p_\rho}$ ratio, but in general we see that the classification performs well in finding a shock exogenous critical event around the large spike. Here we again see a window of resilience for the optimum classification between the ρ values of 33 and 59. In this interval we see only slight variation in the classifications despite the peaked value of $\frac{R_\rho}{p_\rho}$. As ρ increases we see the effects of the variable threshold taking place. In particular, we note that the classification of the main peak changes from critical to sub-critical after ρ exceeds 89.

Figure 10 shows the classification for the flu time series. Of particular note is the main spike where the frequency of the use of the word flu demonstrates a profound peak corresponding to the height of Swine Flu [4]. This main event interval is classified as a shock exogenous critical event whereby a sudden outbreak of Swine Flu in the news prompted a response on the social network of Twitter. The critical classification indicates that this trend did not decay quickly but instead remained in the public eye for some time after the peak. Surrounding the main peak we see two anticipated events. While perhaps the event interval on these is too large considering their relatively low peak value, the large scale trend of the spike at the height of the Swine Flu outbreak is correctly identified and classified.

Figure 11 demonstrates the success of the process in correctly identifying a macro trend comprised of micro shocks as one event. In this case, the Occupy movement provides a perfect example [13]. Several micro shocks are visible which appear on top of the large macro swell of the mentions of Occupy. As ρ increases we see that the number of event intervals found decreases and the micro events blend into one macro trend.

Figure 11 shows the classification for the Occupy time series. We note that the classification as a shock exogenous critical event feels correct give the nature of the Occupy movement [13]. In the case that one would want to examine the micro shocks that make up the macro Occupy movement, a different model for determining the best classification would

be needed.

Last of the exogenous fit examples we examine the watermelon Trend Plot in Figure 13. We note in particular that the optimization ratio $\frac{R_\rho}{p_\rho}$ is fairly unstable, exhibiting several spikes and no plateau like in Figure 7. This seasonal type of trend perhaps does not fit as well into the exogenous model as do more characteristic bursts seen in figures prior. If we consider the ρ values before 57 transient behavior we see that there is not a lot of uniform agreement between the seasonal swells of watermelon. Some of the seasons are split into smaller events and some classifications do not agree with one another across season. After ρ reaches 50 we see that the model settles down with a general agreement that there are 4 major trend regions in the time series. The spike at $\rho = 57$ stands out from the rest of the classifications as it is the last ρ value before the classifications tend to drift away from the desired endogenous classification. The fact that the first watermelon season is classified as an anticipated exogenous critical event is perhaps explained by the changing nature of Twitter. In its early stages, Twitter frequencies were in a heightened period of flux as more users and languages were introduced. Other studies have examined the composition and applications of languages on social media and note that Twitter adoption across countries, which may speak different languages, does not occur at the same rate [10]. This tends to settle down in later years in the time series. This is why, perhaps, the biggest spike in the Christmas time series in Figure 8 occurs during the 2008 Christmas.

Figure 14 demonstrates some of this fluctuation in frequency in the first seasonal swell by classifying the macro seasonal trend as two distinct events. Clearly this is not desirable, but is likely due to the volatile state of Twitter frequencies in the early going [10]. We see that the second, third, and fourth seasons are all correctly found to be endogenous. The fourth trend depicts an interesting dynamic. While it appears that the smoothing of the time series found the second of a triad of micro spikes atop the seasonal watermelon trend, the model then shifts the fitting intervals to split on the first spike. This is due to act that its micro peak is slightly higher than the middle peak. It is interesting that the classification as an endogenous event still holds, and likely would still hold if the middle micro peak was used as the spilt between I_{pre} and I_{post}.

As described above we see that the model in equation 9 for an exponential burst performs quite well in identifying good classifications of the event intervals and optimizing the best value of ρ to describe a time series. We now move to compare these with the Trend Plots that result from using equation 10, the power law fit, to determine intervals.

Figures 15 and 16 show the Trend Plot and classification plot for Christmas using a power law fit. We note from the start that, like Figure 7, there is a large plateau in the ratio $\frac{R_\rho}{p_\rho}$. Despite this, the classifications of the event intervals in the first three Christmases as shock events does not feel correct. In examining the fits of the power laws in those regions, the fits are in fact reasonable and do fit the data well. Where the classification breaks down

however is in using equation 7 to determine the type of event. Due to the fact that the power laws are very sensitive and also depend on a pre factor, the α_i and β_i values are very similar and in fact are the reverse of what one would expect. i.e. one would think that α_i, the relaxation exponent would be greater than β_i since it appears that the volume of mentions of Christmas mentions after the peak dray faster than it grows before the peak.

Figures 17 and 18 depict the flu time series under the power law fit. Here we see that, compared to the exponential counterpart in Figures 9 and 10 that the classifications are the same up to criticality. In addition, the exponential finds a smaller ρ value which in turn results in four total event intervals instead of the three found in the power law fit version.

For the power law classification of the Occupy time series in Figures 19 and 20 we see that again, like in the exponential fit, the best classification for the movement is found to be one macro event. This again feels like the best classification, however the fact that it is found as a sub-critical event does not seem to fit the ideal shape of a trend that becomes very quickly popular but decays very rapidly.

Lastly, in Figures 21 and 22 we see the power law classifications for the watermelon time series. Here we see that compared to the exponential classification that the ρ value is found to be much larger, finding larger event intervals for the seasonal watermelon swells. Additionally, the fit of these event intervals appears to perform worse than exponential classifications. While one could describe the watermelon swells as anticipated events, they do not fit the characteristic spike model intended for this class. The endogenous class is still the desired fit for the watermelon seasons. Also strange about this classification is the fact that the third event interval is found as a shock. This does not fit the model, or desired class, but is conceivable given the fact that the distinction between a shock and an anticipated event is a simple negation of the $Score_i$. In this way, if the seasonal trend for one watermelon cycle lacked build up in comparison to its decay, it could be found to be a shock by the model, and in fact is.

Chapter 6: Future Work

Here we address the potential for further work stemming from the preliminary findings and methods presented in this report. The exploratory methods presented here open the door for many improvements, advances, and new applications. Given the difficulty in determining the best way to partition a diverse population of time series, the procedure presented in Method 3, using an exponential fit, appears to do a decent job in classifying events in a way that a human might. A highlight of the method is the byproduct of the Trend Plot which can be used to provide a micro to macro scale look at the history of a word. However, this may not be the best way to classify the shape of an event in a predetermined event interval. Due to the nuances of curve fitting, perhaps a simpler approach could reveal itself to be more natural in classifying events. This in turn, if it does prove to be a more successful method of classifying event intervals into one of the five classifications, will result in one step back in that we may lose the ability to use the method of maximizing the ratio $\frac{R_\rho}{p_\rho}$.

After observing the results of the current methods, it is the recommendation of this study to continue to use the ratio $\frac{R_\rho}{p_\rho}$ to determine the best scale to fit event intervals but change the way in which events are classified. A simple area comparison of I_{pre} to I_{post} may yield a good starting place to separate shock events from anticipated events. Preliminary attempts are described and presented in section 7.

Given that the base study by Crane and Sornette fit power laws to the intervals of interest and determined the critical $\theta = 0.4$ value corresponding to the video sharing website of YouTube it is proposed that the fits in future work use the power law equation presented in equation 10 [2]. This continuity across studies will allow comparison to the θ value found in the Crane study and other studies cited by Crane and Sornette [2, 12]. Achieving this comparison may yield information regarding the similarities and differences between the cascades and networks pertaining to different social systems.

Given the results presented in Section 5, it may seem counterintuitive to the reader that we here recommend the use of power law fits in the event intervals and not exponentials. The figures presented in Section 5, and the analysis of them, tend to favor a bias for using exponential fits, however much of the discussion regarding the performance of the power law fit versus the exponential in the Trend Plots centered on the discussion on how well the classifications of the event intervals agreed with what one might hope. Given the above recommendation to devise a new, independent and simple method to bin an event interval, it enables the ability to continue to include the power law fit as the fit of choice for revealing robust classes, as in the Crane study (equations 3-5).

We now turn to address some of the future work contingent on the successful

implementation of the above methods. Granted success in the ability to classify a time series by the best ρ value, event intervals, and event classifications we can begin to examine the distributions of the decay exponents of the events in a given class such as the anticipated exogenous critical class. In keeping with the study on YouTube we propose that for shock classified events the exponent of interest be the α_i in equation 7 and β_i for anticipated events [2]. In this way we keep the focus on the relaxation of a shock event, thus we keep the notion of the cascades and influence intact. Likewise for anticipated events we capture the interesting information from a storytelling standpoint, and we are able to speculate on a model for how an event grows to a maximum. For endogenous events, either α_i or β_i should suffice as in principle the event is symmetric.

In addition to determining the corresponding θ value for Twitter, we now address further applications of the successful classification of events in a time series. Of particular interest is the ability to predict how an event will play out given its build up dynamics. Prior studies have examined the predictability of events on Twitter [9]. A confirmation and comparison using our methods and model would be informative. This would seem to be easier for anticipated events as there is more data preceding the spike to draw maps from. The nature of shock events is that they are largely unpredictable. There is currently a lot of interest regarding the ability to predict the future given a large amount of data [1, 5, 9] . Once a large number of event intervals are classified perhaps a macro quality specific to that shape will reveal itself and lend itself towards methods of predicting the shape of anticipated style events in real time.

Another possible application and extension of this study is quantifying the mood of each class of events. Recent research done on Twitter has made advances in quantifying and interpreting the happiness score of a word [3]. With the successful classification of events in a time series in a word one would be able to start to gauge, on a human emotion scale, the sentiment of certain classes of events [3]. For example, the time series for Christmas consists of five event intervals which are all anticipated. Next, one would associate the happiness score for the word Christmas to the classification of anticipated events. On the other hand, a time series like the one for the word occupy contains one large scale event that is a shock. One could then bin the happiness score for Occupy into the shock class. Continuing in this way for all classified time series that demonstrate a uniform classification, i.e. all the event intervals are of the same type, one would be able to address the question whether anticipated or shock events are viewed in a more positive or negative way. Framed in a hypothesis, one could ask, do the events we anticipate have a higher happiness index then shock or surprise events?

We see that just from this short section that there is still much work to be done in improving and expanding the characterization of events in Twitter word frequency time series. This study has explored some techniques that may lend future research on the project valuable insight; in particular, the suggestions made in the preceding paragraphs. We also see that the

potential uses of the successful implementation of the inspiration of this study are vast and exciting [1, 5, 9]. It is the hope of the authors of this project that the outlined improvements and expansions to the current model will be explored and implemented in the near future enabling the ability to begin to answer some very interesting questions, like the happiness question posed in this section, about how social media, in particular Twitter, reflects the user base and how big data in the form of classified event intervals can be applied in the future.

Chapter 7: Preliminary results of future methods

Here we briefly explore the possibility, described in section 6, regarding using a new classification method for event intervals once they are determined. In particular, once we have an event interval in a given time series, instead of classifying by equation 7, we instead classify from an area based analysis. Since we desire to classify first whether an event is of the anticipated or shock variety we preliminarily determine this by comparing the sum of all activity in I_{pre} to I_{post}.

$$\text{Classification}_i = \begin{cases} \text{Anticipated}, & \text{if } \sum_{I_{pre}} f_i > \sum_{I_{post}} f_i \\ \text{Shock}, & \text{if } \sum_{I_{pre}} f_i < \sum_{I_{post}} f_i \end{cases} \text{ where } f_i \text{ is the frequency on day } i \quad (\mathbf{12})$$

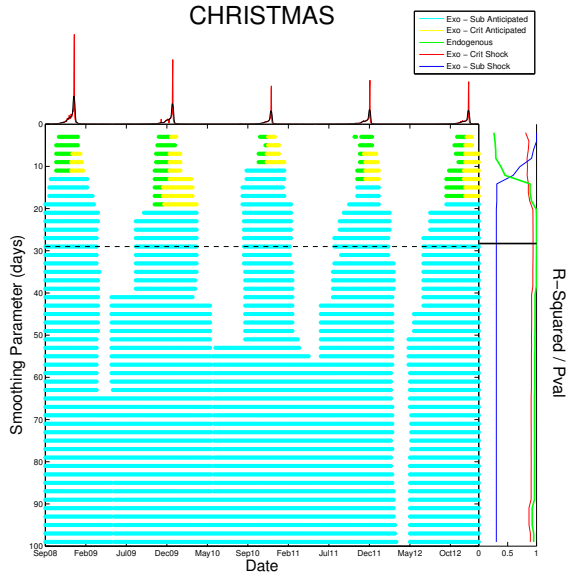
Next we aim to determine criticality of the event interval. To do so we examine the event interval of interest for each event. i.e. for anticipated events we examine I_{pre} and for shock events we examine I_{post}. Here we compare the area of the interval to half the area of the rectangle with height, $h = f_{t_0}$ and length, $l =$ duration of $I_{pre}(I_{post})$. We call this $TriangularArea_i$. Critical events will fill a larger fraction of this area than sub-critical events thus we can define for anticipated events:

$$\text{Criticality}_i = \begin{cases} \text{Sub-Critical}, & \text{if } \dfrac{\sum_{I_{pre}} f_i}{TriangularArea_i} < 0.20 \\ \text{Critical}, & \text{if } \dfrac{\sum_{I_{pre}} f_i}{TriangularArea_i} > 0.20 \end{cases} \text{ where } f_i \text{ is the frequency on day } i \quad (\mathbf{13})$$

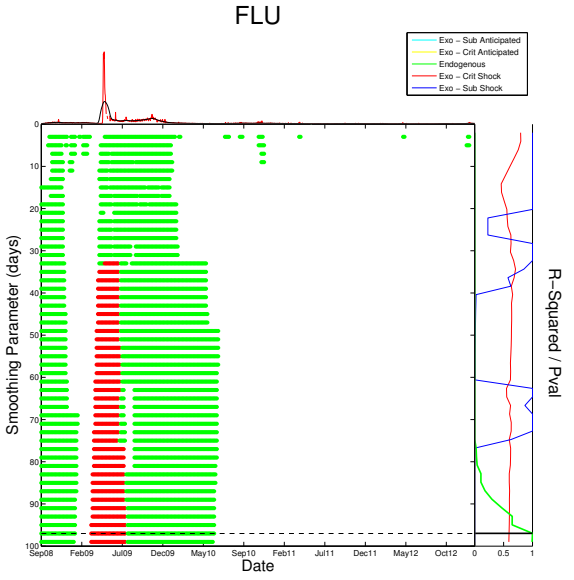
Lastly we have one condition for the possibility of an endogenous event in which we define

$$\text{Classification}_i = \text{Endogenous if } \frac{\sum_{I_{pre}} f_i}{TriangularArea_i} > .30 \quad \text{and} \quad \frac{\sum_{I_{post}} f_i}{TriangularArea_i} > .30 \quad (\mathbf{14})$$

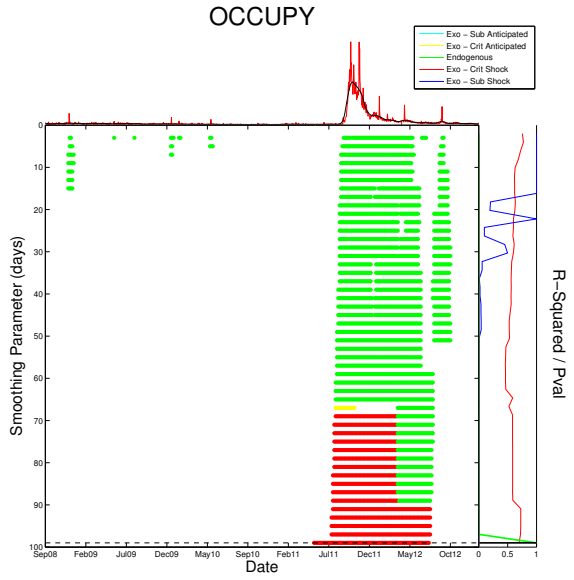
Acknowledging that the following results are preliminary, we generate Trend Plots for the four characteristic time series defined in this study.



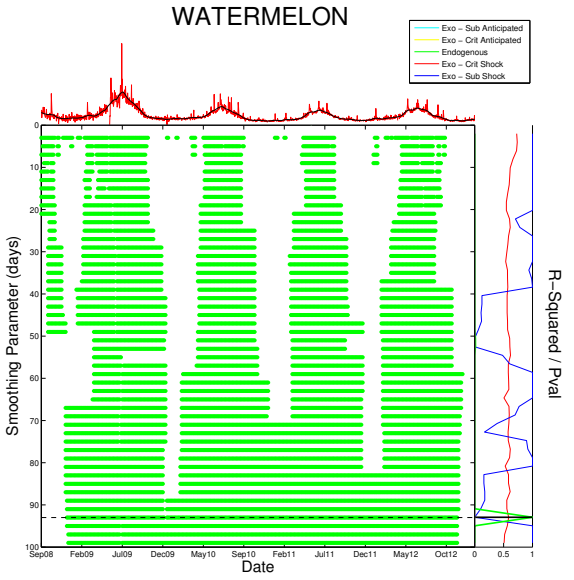
(a) Christmas area Trend Plot



(b) Flu area Trend Plot



(c) Occupy area Trend Plot



(d) Watermelon area Trend Plot

Figure 23: (a) Chistmas, (b) Flu, (c) Occupy, (d) Watermelon area Trend Plots

Bibliography

[1] Asur, S. and B. A. Huberman (2010, March). Predicting the Future with Social Media. *ArXiv e-prints*.

[2] Crane, R. and D. Sornette (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Prod. Nat. Acad. Sci. 105*, 15649–15653.

[3] Dodds, P. S., K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth (2011, 12). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE 6*(12), e26752.

[4] for Disease Control, C. and Prevention (2011, August). 2009 h1n1 flu.

[5] Gruhl, D., R. Guha, R. Kumar, J. Novak, and A. Tomkins (2005). The predictive power of online chatter. In *Proceedings of the Eleventh ACM Conference on Knowledge Discovery and Data Mining (KDD)*, Chicago, Il.

[6] Hawkes, A. G. and D. Oakes (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability 11*(3), pp. 493–503.

[7] http://www.google.com/trends/.

[8] Johansen, A. and D. Sornette (2000, January). Download relaxation dynamics on the WWW following newspaper publication of URL. *Physica A Statistical Mechanics and its Applications 276*, 338–345.

[9] Miotto, J. M. and E. G. Altmann (2014, March). Predictability of extreme events in social media. *ArXiv e-prints*.

[10] Mocanu, D., A. Baronchelli, N. Perra, B. Gonalves, Q. Zhang, and A. Vespignani (2013, 04). The twitter of babel: Mapping world languages through microblogging platforms. *PLoS ONE 8*(4), e61981.

[11] Munroe, R. Seismic waves.

[12] Sornette, D., F. Deschâtres, T. Gilbert, and Y. Ageon (2004, November). Endogenous

Versus Exogenous Shocks in Complex Networks: An Empirical Test Using Book Sale Rankings. *Physical Review Letters 93*(22), 228701.

[13] Staff, T. W. (2011, November). Occupy wall street: A protest timeline.

[14] Tsukayama, H. (2013, March). Twitter turns 7: Users send over 400 million tweets per day. *The Washington Post*.

[15] Vázquez, A., J. a. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási (2006, Mar). Modeling bursts and heavy tails in human dynamics. *Phys. Rev. E 73*, 036127.

[16] Yang, J. and J. Leskovec (2010). Patterns of temporal variation in online media. Technical report.