Forecasting High Tide: Predicting Times of Elevated Activity in Online Social Media

Jimpei Harada[1], David Darmon[2], Michelle Girvan[3], and William Rand[1]

[1]Center for Complexity in Business, University of Maryland, College Park, MD 20740
[2]Department of Mathematics, University of Maryland, College Park, MD 20740
[3]Department of Physics, University of Maryland, College Park, MD 20740

April 3, 2015

Abstract

Social media provides a powerful platform for influencers to broadcast content to a large audience of followers. In order to reach the greatest number of users, an important first step is to identify times when a large portion of a target population is active on social media, which requires modeling the behavior of those individuals. We propose three methods for behavior modeling: a simple seasonality approach based on time-of-day and day-of-week, an autoregressive approach based on aggregate fluctuations from seasonality, and an aggregation-of-individuals approach based on modeling the behavior of individual users. The aggregation-of-individuals approach uses the framework of computational mechanics to automatically infer a state machine that describes the behavior of an individual based on his/her past behavior. We test these methods on data collected from a set of users on Twitter in 2011 and 2012. We find that the performance of the methods at predicting times of high activity depends strongly on the tradeoff between true and false positives, with no method dominating. Our results highlight the challenges and opportunities involved in modeling complex social systems, and demonstrate how influencers interested in forecasting potential user engagement can use complexity modeling to make better decisions.

1 Introduction

For a wide variety of organizations, companies, and individuals there is a growing interest in using social media to get their message out. For instance, brand managers are often tasked with launching promotions that raise the awareness of their brand among users of social media. However, the signal that a brand is trying to convey can easily get lost in the 'noise' produced by other brands, individuals, bots, etc. While good content is important to engage an audience, it is also important to know when users will pay attention to the content in order to increase the chance that the message is spread. Therefore, a brand manager must consider not only what they want to say, but *when* they want to say it.

In order to effectively spread a message on a social media platform, an important first step is to understand the patterns of user engagement. After receiving and becoming aware of information, users on a social media platform then evaluate the content of the information and decide whether it should be retransmitted or not. Previous research has examined different criteria for this decision, including the sender's level of activity and the freshness of the information [31], as well as the user's benefit from spreading the information [34]. In this research, instead of exploring criteria related to evaluation of content and the decision to retransmit, we focus on timing when users on social media are engaged in retransmission behavior. A key assumption of our approach is that information is most likely to be retransmitted during the highest activity periods. In particular, in this paper, we study the task of predicting user engagement on Twitter, and we measure engagement in terms of the number of users actively issuing retweets.

It is well known that user activity on social media services follows both diurnal and weekly patterns [10, 12]. For example, Figure 1 demonstrates the number of users active on Twitter out of a collection of 2145 over a four week period, starting on Mondays at 9am EST. At the daily level, the number of users actively retweeting increases over the course of a day and then decreases at night. However, the times of peak activity also fluctuate from week to week. Such fluctuations in social systems has been attributed to the fact that observed aggregate social behavior, driven by individual human actions, can be described as mixtures of Poisson and non-Poisson processes, where these processes can be seen as modeling individual decision making [2]. Thus, we expect the aggregate behavior of a collection of users who have different decision making processes to exhibit significant temporal fluctuation from seasonality from week to week. In order to effectively reach a large number of activated users, it is therefore important to determine when they are the most engaged by tracking such fluctuations while controlling for the diurnal and weekly seasonality. Moreover, recent work studying the attention of users on Twitter has found that retweets of a given tweet typically occur on the time scale of minutes [15, 14]. Given this observation, it is also important that we track seasonal fluctuations at a fine temporal resolution.

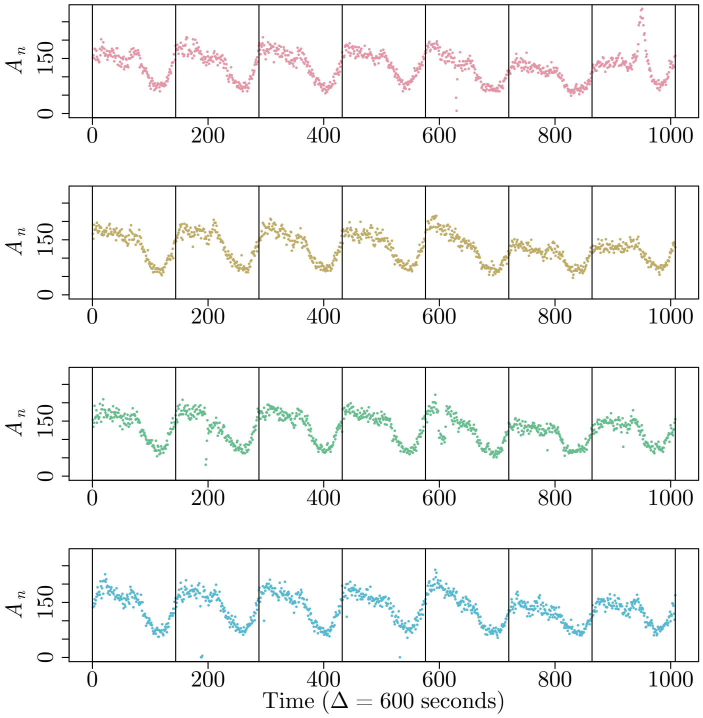

Figure 1: The number of users actively retweeting during disjoint ten minute windows. Each row corresponds to a week, and each column corresponds to a day of the week, starting from Monday.

In order to model the number of active retweeters on Twitter at any given time, we propose three approaches: a seasonality model that assumes the overall retweet activity on Twitter is fully explained by the time-of-day and day-of-week, an autoregressive model that explicitly models deviations from the day-to-day seasonality, and an aggregation-of-individuals approach that models the activity patterns of each individual user and then aggregates these models to describe the overall activity pattern.

The seasonality model is based on the assumption that user engagement over time can be explained by seasonal patterns at the daily and weekly level. Thus, in order to predict the time when users are engaged at a certain level using this model, we consider only engagement patterns from the past.

The autoregressive model seeks to describe the population-level fluctuations about the seasonality using a simple linear autoregressive model. We assume that the deviations from seasonality have memory where we can think of this memory in terms of activation / deactivation of the users on Twitter. For example, a certain topic might become popular over the course of several hours, leading to activity greater than expected by the baseline seasonality. Such bursts of activity have been observed on both Twitter and blogging platforms [36]. By noting when and how such bursts occur, we can better predict the number of users active on Twitter compared to using seasonality alone.

The aggregation-of-individuals model explicitly views the overall activity as the accumulation of the activity patterns of all of the users under consideration. In particular, we model each user-to-be-aggregated as a point process with memory [35]. In this case, each individual user can become activated / deactivated, depending on their own previous behavior and the behavior of their inputs. By viewing the user as a computational unit, we can build a predictive model of how they interact with Twitter. This approach has been successfully applied to individual level prediction [7] on Twitter, where many high volume users were found to be well-described by such a model. We can then aggregate these individual level models to produce a global prediction of activity levels that accounts for individual-level activation.

In the rest of this paper, we explore the problem of identifying periods of high activation on a social media platform. We begin by describing our three models and relevant literature. Then we describe the data sets used to test the predictive ability of these models for the proposed problem. Next, we review the predictive ability of the various models, and compare the benefits and tradeoffs of each approach. Finally, we conclude with the limitations of the present work and future directions to extend and improve it.

2 Related Work

A large body of work has investigated the dynamics of technology-mediated human interaction. Relevant to our work, [9] found that human behavior on email services is dominated by bursty-type behavior, with periods of high activity separated by long stretches of inactivity. The authors of [19] found stereotypical temporal patterns in the interaction between blogs and mainstream media news. Studies of Twitter have found similar stereotypical aggregate behavioral patterns for the popularity of particular hashtags over time [19, 18]. More recent work has sought to develop first principle mathematical models explicitly geared towards human behavior on social media [4, 5].

A great deal of work has been done on the problem of predicting the future popularity of individual tweets and hashtags based on their features. As a very recent example, in [33], the authors performed an experiment to investigate how the wording of a tweet impacts whether it is retweeted, controlling for both the author and the topic of the tweet. In [21], the authors predict the volume of tweets about a hashtag day-to-day using features extracted from a corpus of tweets containing that hashtag on previous days. Similar studies can be found in [16, 24, 37, 32]. The problem of predicting individual tweet, hashtag, and topic popularity has been well-studied, and these references are only meant to give a sampling of the much larger literature on the subject.

The problem of predicting the total volume of tweets over time has attracted much less attention from the research community. Notable exceptions include [26, 1, 25]. In [26], the authors build a predictive model for the overall volume of tweets related to a particular hashtag. Similar to one of our approaches, the authors do this by aggregating individual predictive models for a universe of users, where the users were chosen if they previously tweeted on a topic and followed a user who also tweeted on that topic. They then identified predictive models for each user at the resolution of days, where predictions were made based on previous activity of a user and their local network structure. The goal of predicting day-resolution volume from users on a particular topic differs greatly from predicting high volume times from a collection of users determined based on their network properties, which is the goal of this paper. In [1], the authors seek to determine the one hour period in which the followers of a given collection of users are most likely to be active. However, their investigation is purely sociological in nature, in that they make no predictions, and the data used in their analysis only covered a single week of activity. Thus, their approach is not directly applicable to forecasting retweet volume from streaming data. Finally, in [25], the authors use a two state Hidden Markov modeling framework, where the hidden states correspond to when the user is either in an active mode or an inactive state. Using these models, they predict the expected interarrival time for a user given their observed previous behavior by filtering their hidden state, and make predictions based on this time. Thus, this approach is similar in spirit to our aggregation-of-individuals approach. However, they assume a particular hidden state model architecture that is homogeneous across users, while our approach, as we will see, allows for model heterogeneity across users. Moreover, while their approach could be used to predict total retweet volume by aggregating their individual model predictions, they focus on individual level prediction.

3 Methodology

Here we define our exact problem and the proposed solutions. Consider a set $\mathcal{U} = \{u_1, u_2, \ldots, u_U\}$ of U users. Each user in $\mathcal{U}$ has an individual retweet history. Let Δ be a time interval; here we take $\Delta = 10$ minutes. Then for each user u in the set of users $\mathcal{U}$, we specify their retweet activity during any window of length Δ by

$$X_n(u) = \begin{cases} 1 & : \text{user } u \text{ retweeted between times} \\ & \quad (n-1)\Delta \text{ and } n\Delta \\ 0 & : \text{otherwise} \end{cases}. \tag{1}$$

That is, $\{X_n(u)\}_{n=1}^{N}$ specifies the retweet activity of the user during each of the N time intervals $[0, \Delta), [\Delta, 2\Delta), \ldots, [(N-1)\Delta, N\Delta)$.

The total number of users active during any time interval $[(n-1)\Delta, n\Delta)$ is then given by

$$A_n = \sum_{u \in \mathcal{U}} X_n(u). \tag{2}$$

This is the value we seek to predict.

3.1 Seasonality

For the seasonality model, we assume that retweet activity shows day-to-day variability, but regularity from week-to-week. We assume that the seasonality repeats every T timesteps,

$$s_n = s_{n+jT}, \qquad j = 1, 2, \ldots \tag{3}$$

and that the observed number of users retweeting A_n is given by

$$A_n = s_n + \epsilon_n \tag{4}$$

where ϵ_n can be thought of as the deviation from the seasonality at any given time n. Under the assumption of seasonality, we infer the seasonal component by averaging across W weeks [8],

$$\hat{s}_n = \frac{1}{W} \sum_{j \in \{0,1,\ldots,W-1\}} A_{n+jT}, \qquad n = 1, \ldots, T. \tag{5}$$

Figure 2 shows the aggregate retweet activity across the four weeks from Figure 1 with the estimated seasonality superimposed.

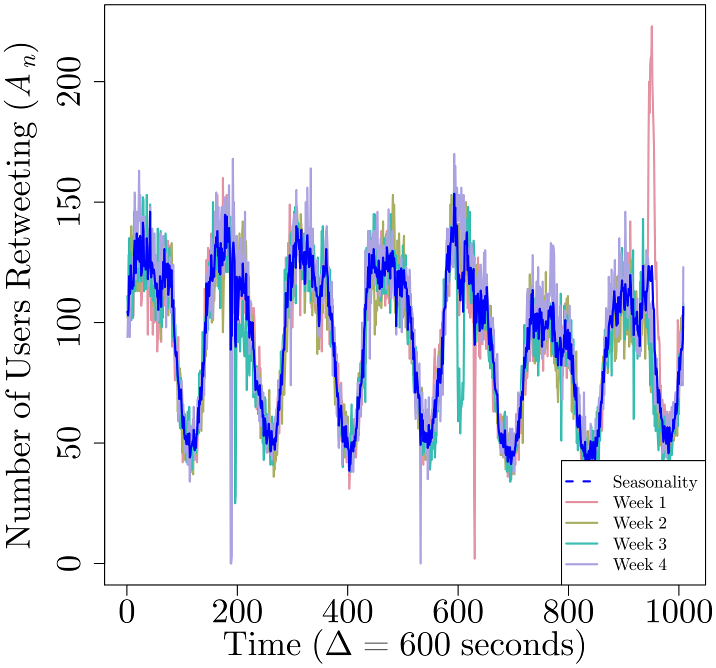

Figure 2: The number of users retweeting A_n over four consecutive weeks. The estimated seasonality $\hat{s}_n$ is shown in blue.

If we assume that $\{\epsilon_n\}_{n=1}^{N}$ is a realization from a white noise process, the optimal predictor under mean-squared loss for A_n is s_n, the seasonality. Thus, we use our estimator for the seasonality as the predictor for the seasonality model,

$$A_n^{\mathrm{S}} = \hat{s}_n. \tag{6}$$

3.2 Aggregate Autoregressive Model

In the seasonality model, we have assumed that the residuals $\{\epsilon_n\}_{n=1}^N$ are white noise. More explicitly, we have assumed that they show no autocorrelation: $E[\epsilon_t \epsilon_s] = \sigma_\epsilon^2 \delta_{st}$, where σ_ϵ^2 is the variance of the white noise process and δ_{st} is the Kronecker delta. A more reasonable model for the residual would incorporate memory, since aggregate social systems are known to exhibit such memory [23]. Thus, a simple refinement of the previous model allows for memory in the deviations from seasonality. More explicitly, we consider the model

$$A_n = s_n + Y_n \tag{7}$$

where we now take $\{Y_n\}_{n=1}^N$ to be a realization from an autoregressive process of order p, an AR(p) model [8]. That is, we consider the dynamics of Y_n to be governed by

$$Y_n = \sum_{j=1}^{p} b_j Y_{n-j} + \epsilon_n \tag{8}$$

where $\{\epsilon_n\}$ is again a white noise process with mean 0 and variance σ_ϵ^2.

The predictor for the aggregate autoregressive model is

$$A_n^{\text{AR}} = \hat{s}_n + \sum_{j=1}^{\hat{p}} \hat{b}_j \hat{Y}_{n-j}, \tag{9}$$

where $\hat{Y}_n = A_n - \hat{s}_n$ is the deviation of the observed aggregate retweeting activity from the estimated seasonality at time n. We choose the autoregressive order $\hat{p}$ by minimizing the the Akaike information criterion on the training set [17].

3.3 Aggregation of Causal State Models

Before describing the aggregation procedure, we briefly review computational mechanics, which is our basic modeling approach for the individual-level models. [29] provides a more in-depth introduction to computational mechanics, and [7] describes an application of computational mechanics to modeling individual user activity on Twitter. Computational mechanics provides a framework for describing stationary [11] (and more generally, *conditionally* stationary [6]), discrete-time, discrete-alphabet stochastic processes by linking the observed process to

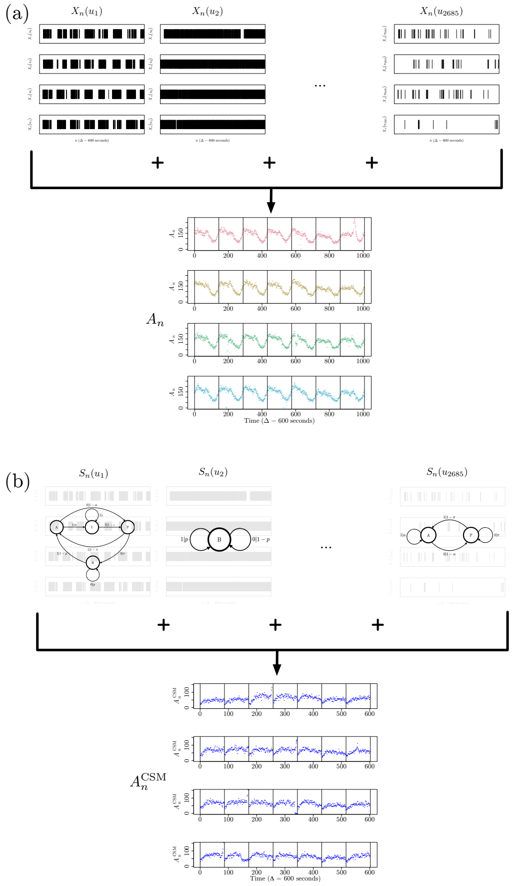

Figure 3: A demonstration of how (a) the retweet volume A_n results from the summation of the individual retweet behavior $\{X_n(u)\}_{u \in \mathcal{U}}$ of the users in $\mathcal{U}$ and (b) the aggregation-of-individuals prediction A_n^{CSM} is formed via filtering through each user u's ϵ-machine.

a hidden state process. In this way, the formalism of computational mechanics is closely related to Hidden Markov Models and other state-based models of discrete-alphabet stochastic processes [20]. In particular, any conditionally stationary stochastic process $\{X_n\}$ naturally induces a hidden state process $\{S_n\}$, where the transition structure of the hidden state process is determined by the predictive distribution of $\{X_n\}$. The hidden state process $\{S_n\}$ is always Markov, and the combination of its Markov chain representation and the state conditional emission probabilities $P(X_n = x | S_{n-1} = s)$ is called the *causal state model* or

ϵ-*machine* for the stochastic process $\{X_n\}$. In the case where the predictive distribution for $\{X_n\}$ is unknown, machine reconstruction algorithms can be used to automatically infer the ϵ-machine that best describes the observed data $\{X_n\}_{n=1}^{N}$. We use the Causal State Splitting Reconstruction (CSSR) algorithm [30] to infer an ϵ-machine for each user's observed retweeting activity.

As with the autoregressive model, the CSSR algorithm requires a maximum history length $L_{\max}$ to look into the past in order reconstruct the ϵ-machine associated with a user u's behavior. While theory exists for choosing the largest $L_{\max}$ such that we can consistently infer the one-step-ahead predictive distributions used in CSSR [22], we take the practical approach of choosing $L_{\max}$ based on cross-validation. In particular, we perform 5-fold cross validation using the log-likelihood of the held out data as our objective function. The form of the log-likelihood associated with a realization from a stochastic process under an ϵ-machine model may be found in [13].

For each user u, we reconstruct their associated ϵ-machine. We then perform prediction as follows: at time $n-1$, we determine the current causal state $S_{n-1}(u)$ for each user u based on their activity pattern $X_1^{n-1}(u) = (X_1(u), X_2(u), \ldots, X_{n-1}(u))$. The causal state $S_{n-1}(u)$ specifies the one-step-ahead predictive distribution for each user, $P(X_n(u) = 1|S_{n-1}(u) = s(u))$. We then aggregate these probabilities to form our prediction for the number of active users at the next time step,

$$A_n^{\text{CSM}} = \sum_{u \in \mathcal{U}} P(X_n(u) = 1|S_{n-1}(u) = s(u)). \tag{10}$$

This can be seen to be the expected number of users active at time n given the causal states of the users at time $n-1$, under the assumption that the behavior of a user u at time n is independent of the causal states of all others users at time $n-1$ given the causal state of u at time $n-1$.

4 Data Collection and Selection of $\mathcal{U}$

We begin with a collection of 15000 Twitter users whose statuses (Tweet text) were collected over two disjoint five week intervals: from 25 April 2011 to 29 May 2011 and from 1 October 2012 to 5 November 2012. The users are embedded in a 15000 node network collected by performing a breadth-first expansion of the active followers of a random seed user. In particular, the network was constructed by considering the followers of the seed user, and including those followers considered active (i.e. users who tweeted at least once per day over the past one hundred days). The collection of users continued from the followers of these followers, etc., until 15000 users were included. From this network of users, the subset of users $\mathcal{U}$ was chosen to account for 80% of the retweet volume for the first four weeks in the five week period under consideration. That is, we take u_1 to be the user issuing the greatest number of retweets, then u_2 to be the user issuing the second greatest number of retweets, etc., until we reach the user u_U such that the total number of retweets issued by the users in $\mathcal{U}$ account for 80% of the retweet volume. This results in $U = 2145$ users for the 2011 collection and $U = 1610$ users for the 2012 collection. Because we are interested in predicting times of greatest retweet activity, for each day we only consider the retweet activity from 6 AM EST to 10 PM EST. The data used in our analysis can be made available upon request by the corresponding author.

5 Results

In the following results, we use the first four weeks of the five week periods from 2011 and 2012 for inference of the three model types, and leave the last weeks from each year for testing. As described in the methodology section, we choose the parameters of each model as follows. The seasonality model has no tuning parameter, and we use the full four weeks to infer the seasonality component. We choose the model order p of the autoregressive model to maximize the Akaike information criterion on the four week training period. For each causal state model in the aggregation-of-individuals model, we infer the user-specific history length L by 5-fold log-likelihood cross-validation over the 28 days in the training sets.

5.1 Adjustment to the Aggregation-of-Individuals Model

As described in the methodology section, the predictor for the aggregation-of-individuals model (10) is equivalent to the expected number of users in $\mathcal{U}$ who are active at time step n given their causal states at $n - 1$ under a certain independence assumption. In particular, we have taken the predictor to be

$$
\begin{aligned}
A_n^{\text{CSM}} &= E\left[\sum_{u \in \mathcal{U}} X_n(u) \middle| S_{n-1}(u_1), \ldots, S_{n-1}(u_U)\right] &(11)\\
&= \sum_{u \in \mathcal{U}} E[X_n(u) | S_{n-1}(u_1), \ldots, S_{n-1}(u_U)] &(12)\\
&= \sum_{u \in \mathcal{U}} E[X_n(u) | S_{n-1}(u)] &(13)\\
&= \sum_{u \in \mathcal{U}} P(X_n(u) = 1 | S_{n-1}(u)), &(14)
\end{aligned}
$$

where going from (12) to (13) we make the assumption that for all $u \in \mathcal{U}$,

$$X_n(u) \perp\!\!\!\perp \{S_{n-1}(u'), u' \neq u\} | S_{n-1}(u). \qquad (15)$$

That is, we assume that the observed behavior of user u at time n is independent of the causal states of all other users u' at time $n-1$, given the causal state of user u at time $n-1$. While such an independence relationship holds when conditioning on the local causal states of a time-varying random field [28], it need not be true when conditioning on the marginal causal states.

Motivated by the form of the deviation of (10) from the predicted value (see Figure 4), we define the adjusted aggregation-of-individuals predictor as

$$A_n^{\text{CSM}^*} = \beta_0 + \beta_1 A_n^{\text{CSM}}, \qquad (16)$$

where the parameters β_0 and β_1 were estimated by regressing the true values A_n from the training set on the unadjusted aggregation-of-individuals predictions A_n^{CSM} from the training set. We will use this predictor for the remainder of this work, and address alternative corrections in the conclusion.

5.2 Predicting Activation Level at Varying Thresholds

We next present an experiment to test the predictive capability for each of the three proposed models. As mentioned in the introduction, ideally a potential influencer would like to choose the optimal time(s)-of-day to send out a message such that the largest number of users will be active around those times. As a proxy for this goal, we consider the task of identifying whether or not the activity level over an interval of length Δ will fall into the $100p^{\text{th}}$ percentile for that day. As an example, how well can we predict whether the number of activated users falls within the 80^{th} percentile for a given day?

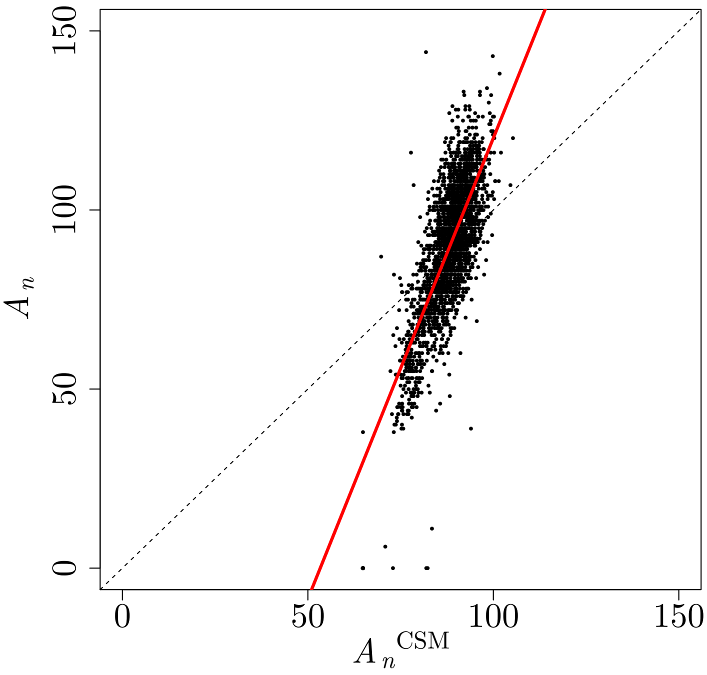

Figure 4: The transformation of the aggregation-of-individuals model used to adjust for associations in user behavior. The red line corresponds to the linear least squares fit from regressing the true values A_n from the training set on the unadjusted aggregation-of-individuals predictions A_n^{CSM} from the 2011 data.

Let N_Δ be the number of timepoints to predict on in a day ($N_\Delta = 86$ for this analysis). For a given day $d \in \{1, 2, \ldots, 7\}$ in the testing set, the true distribution of the activity levels is given by

$$F_d^{\text{True}}(a) = \frac{1}{N_\Delta} \sum_{n=n_{\text{train}}+N_\Delta(d-1)+1}^{n_{\text{train}}+N_\Delta d} \mathbb{1}\left[A_n \leq a\right]. \tag{17}$$

We then define the historical distribution of the activity levels for a day d in terms of the estimated seasonality for that day from the training set

$$F_d^{\text{Hist}}(a) = \frac{1}{N_\Delta} \sum_{n=N_\Delta(d-1)+1}^{N_\Delta d} \mathbb{1}\left[A_n^S \leq a\right]. \tag{18}$$

We will use $F_d^{\text{Hist}}(\hat{A}_n)$ to predict whether or not a predicted activity level $\hat{A}_n$ exceeds the quantile p^* of activity for a given day, where $\hat{A}_n$ is one of A_n^{S}, A_n^{AR}, or $A_n^{\text{CSM}^*}$. That is, for a threshold p, we predict the indicator for whether the activity at time n will exceed some quantile p^* as

$$\hat{I}_n(p) = \left\{ \begin{array}{ll} 1 & : F_{d(n)}^{\text{Hist}}(\hat{A}_n) > p \\ 0 & : \text{ otherwise} \end{array} \right. . \tag{19}$$

Whether or not the activity at time n exceeded the quantile p^* is then given in terms of the true distribution as

$$I_n = \left\{ \begin{array}{ll} 1 & : F_{d(n)}^{\text{True}}(A_n) > p^* \\ 0 & : \text{ otherwise} \end{array} \right. . \tag{20}$$

As we vary the threshold value p, the true positive rate is given by

$$\text{TPR}(p) = \frac{\sum\limits_{n=n_{\text{train}}+1}^{n_{\text{test}}} \mathbb{1}\left[\hat{I}_n(p) = 1, I_n = 1\right]}{\sum\limits_{n=n_{\text{train}}+1}^{n_{\text{test}}} \mathbb{1}\left[I_n = 1\right]} \tag{21}$$

and the false positive rate is given by

$$\text{FPR}(p) = \frac{\sum\limits_{n=n_{\text{train}}+1}^{n_{\text{test}}} \mathbb{1}\left[\hat{I}_n(p) = 1, I_n = 0\right]}{\sum\limits_{n=n_{\text{train}}+1}^{n_{\text{test}}} \mathbb{1}\left[I_n = 0\right]}. \tag{22}$$

We show the ROC curves associated with the fixed quantile $p^* = 0.8$, along with their AUCs, for the test weeks from 2011 and 2012 in Figure 5 ant Table 1. The true and false positive rates are computed using the last 86 of the 96 time points in each day, since both the autoregressive and aggregation-of-individuals models require up to ten timepoints to begin prediction depending on the model order p or largest history length L, respectively.

Overall, based on the AUC values, the aggregation of individuals model performs best in 2011 and the autoregressive model performs best in 2012. However, inspection of the ROC curves indicates that based on the desired balance between true and false positives, each of the models may outperform the others, with no model strictly dominating. For example, if a high false positive rate is acceptable, the seasonality model achieves the lowest false positive rate to give a 100% true

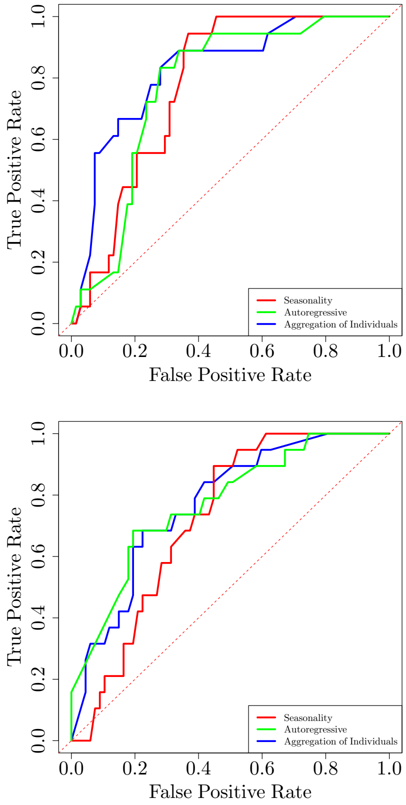

Figure 5: The ROC curves associated with each of the three models for the testing week in 2011 (top) and 2012 (bottom) with p^* fixed at 0.8. The AUC values for the seasonality, autoregressive, and aggregation-of-individuals models for 2011/2012 are 0.778/0.720, 0.773/0.773, and 0.825/0.771.

Table 1: The AUC for each of the methods on the test weeks from 2011 and 2012.

Year	Seasonality	Autoregressive	Agg. of Individuals
2011	0.778	0.773	**0.825**
2012	0.720	**0.773**	0.771

positive rate on the testing sets in 2011 and 2012. However, the seasonality model generally underperforms when the desired false positive rate is low, in which case either the aggregation of individuals model (in 2011) or the autoregressive model (in 2012) performs better. While we only report on predicting under the condition that $p^* = 0.8$, we find similar results for $p^* \geq 0.7$.

5.3 Utility of Individual Level Models Beyond Aggregate Prediction

Though we do not focus on individual-level prediction in this paper, we wish to highlight some of the possible advantages offered by the aggregation-of-individuals approach not immediately evident from the ROC analysis above. In particular, as demonstrated in Figure 3, the aggregation-of-individuals generates individual level, behavioral models for each user u. These models have the advantage of being interpretable. Consider the four models in Figure 6. The models can be represented as directed graphs, where each vertex corresponds to a causal state, and each arrow corresponds to an allowed emission from that state. The arrows are decorated with the emission symbol $x \in \{0, 1\}$ (*i.e.* user u either retweets or does not during a time interval) and the causal state conditioned emission probability $P(X_n(u) = x | S_{n-1}(u) = s)$ of transitioning from state s while emitting symbol x. That is, each arrow is decorated as $x | P(X_n(u) = x | S_{n-1}(u) = s)$.

These models allowed for user-specific targeting. Consider the model represented by (b). Users of this type tend to retweet in a bursty manner, with an active state A and a passive state P. This corresponds to a simple order-1 Markov model. For such users, it is sufficient to target them when they have recently retweeted. Users exhibiting behavior like models (c) or (d) require more subtle targeting. Model (c) has the same active and passive states as in (b), but with an additional refractory state R that occurs after the user is quiescent while in the passive state P. Depending on the balance between $1 - \rho$ and α, which correspond to the probabilities of retweeting from the active and refractory states, it may be more beneficial to target the user when they are currently active or when they are 'resting' in the refractory state. Model (d) is similar to model (c), with an additional intermediate state I that occurs after the user has issued a retweet from the active state A. Again, depending on the balance between α, $1 - \rho$, and ι, the user can be targeted for when they are most likely to retweet. Many of the users have simple ϵ-machines similar to (a), (b), (c), and (d), which allow for this sort of user-specific targeting.

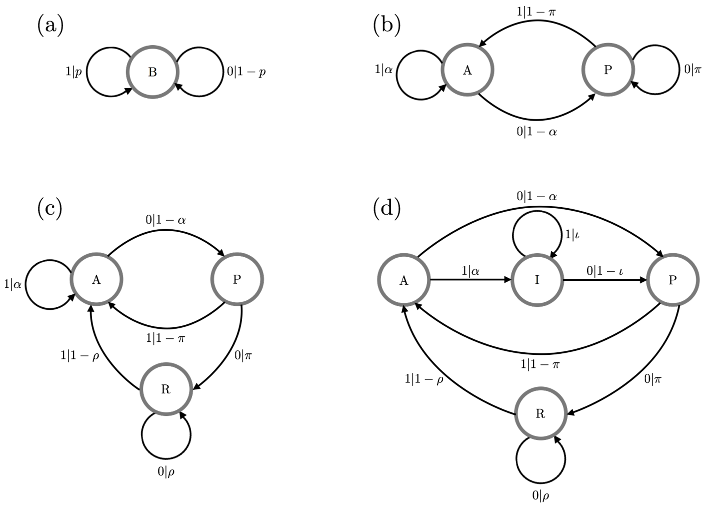

Figure 6: Four example ϵ-machines inferred from the users. (a) A user who retweets at random with bias p. (b) A user who retweets in a bursty manner, with an active state A and a passive state P. (c) A user who retweets in a bursty manner, with a refractory state R. (d) A user who retweets in a bursty manner with both a refractory state R and an intermittent state I.

6 Conclusion

We have found that while user retweet activity clearly exhibits seasonality from day-to-day and week-to-week, seasonality alone does not explain the times of high user activity on social media. Incorporating additional information about either the deviations from seasonality or the behavioral patterns of individual users allows for more accurate prediction of times of high volume, especially when a low false positive rate is desired. Since overexposure to a message may lead to reduced user engagement (content fatigue) due to the repetitive nature of the message, it can be said that having a low false positive is important in this motivating example.

In future work, we will explore more sophisticated models that should provide even greater predictive power. For example, the individual models used in the aggregation-of-individuals method did not incorporate social inputs to the users beyond their own previous behavior. The computational mechanics framework allows for the incorporation of inputs via either dynamic random field-based [28] or transducer-based [27, 3] models of a user's behavior. Such an extension could eliminate the need for the adjustment to the aggregation-of-individuals predictor needed to translate the model's output to a prediction.

This work highlights that in building predictive models for complex social systems, a multi-level view of the system under consideration often leads to improved predictive ability. Thus, in the predictive problem considered in this paper, influencers who track potential user engagement can use complexity modeling to make better informed decisions.

References

[1] Esam Alwagait and Basit Shahzad. Maximization of tweet's viewership with respect to time. In *Computer Applications & Research (WSCAR), 2014 World Symposium on*, pages 1–5. IEEE, 2014.

[2] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[3] Nix Barnett and James P Crutchfield. Computational mechanics of input-output processes: Structured transformations and the ϵ-transducer. *arXiv preprint arXiv:1412.2690*, 2014.

[4] Christian Bauckhage, Kristian Kersting, and Fabian Hadiji. Mathematical models of fads explain the temporal dynamics of internet memes. In *ICWSM*, 2013.

[5] Christian Bauckhage, Kristian Kersting, and Bashir Rastegarpanah. Collective attention to social media evolves according to diffusion models. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web*, pages 223–224, 2014.

[6] S Caires and JA Ferreira. On the nonparametric prediction of conditionally stationary sequences. *Probability, Networks and Algorithms*, pages 1–32, 2003.

[7] David Darmon, Jared Sylvester, Michelle Girvan, and William Rand. Predictability of user behavior in social media: Bottom-up v. top-down modeling. In *ASE/IEEE Int'l Conf. on Social Computing*, pages 102–107, 2013.

[8] Jianqing Fan and Qiwei Yao. *Nonlinear time series*. Springer, 2002.

[9] K-I Goh and A-L Barabási. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002, 2008.

[10] Scott A Golder and Michael W Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[11] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*, volume 2. Oxford Univ Press, 1992.

[12] Nir Grinberg, Mor Naaman, Blake Shaw, and Gilad Lotan. Extracting diurnal patterns of real world activity from social media. In *ICWSM*, 2013.

[13] Robert Haslinger, Kristina Lisa Klinkner, and Cosma Rohilla Shalizi. The computational structure of spike trains. *Neural Computation*, 22(1):121–157, 2010.

[14] Nathan O Hodas and Kristina Lerman. Attention and visibility in an information-rich world. In *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.

[15] Nathan Oken Hodas and Kristina Lerman. How visibility and divided attention constrain social contagion. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 249–257. IEEE, 2012.

[16] Liangjie Hong, Ovidiu Dan, and Brian D Davison. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 57–58. ACM, 2011.

[17] Rob J Hyndman and Yeasmin Khandakar. Automatic time series for forecasting: the forecast package for r. *Journal of Statistical Software*, 27(3), 2008.

[18] Janette Lehmann, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 251–260. ACM, 2012.

[19] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

[20] Michael L Littman, Richard S Sutton, and Satinder P Singh. Predictive representations of state. In *NIPS*, volume 14, pages 1555–1561, 2001.

[21] Zongyang Ma, Aixin Sun, and Gao Cong. Will this #hashtag be popular tomorrow? In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1173–1174. ACM, 2012.

[22] Katalin Marton and Paul C Shields. Entropy and the consistent estimation of joint distributions. *The Annals of Probability*, pages 960–977, 1994.

[23] Joachim Mathiesen, Luiza Angheluta, Peter TH Ahlgren, and Mogens H Jensen. Excitable human dynamics driven by extrinsic events in massive communities. *Proceedings of the National Academy of Sciences*, 110(43):17259–17262, 2013.

[24] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. In *ICWSM*, 2011.

[25] Vasanthan Raghavan, Greg Ver Steeg, Aram Galstyan, and Alexander G Tartakovsky. Modeling temporal activity patterns in dynamic social networks. *IEEE Transactions on Computational Social Systems*, 2013.

[26] Yiye Ruan, Hemant Purohit, David Fuhry, Srinivasan Parthasarathy, and Amit Sheth. Prediction of topic volume on twitter. *WebSci (short papers)*, 2012.

[27] Cosma Rohilla Shalizi. *Causal architecture, complexity and self-organization in the time series and cellular automata*. PhD thesis, University of Wisconsin–Madison, 2001.

[28] Cosma Rohilla Shalizi. Optimal nonlinear prediction of random fields on networks. *Discrete Mathematics and Theoretical Computer Science*, pages 11–30, 2003.

[29] Cosma Rohilla Shalizi and James P Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3-4):817–879, 2001.

[30] Cosma Rohilla Shalizi and Kristina Lisa Klinkner. Blind construction of optimal nonlinear recursive predictors for discrete sequences. In Max Chickering and Joseph Y. Halpern, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI 2004)*, pages 504–511, Arlington, Virginia, 2004. AUAI Press.

[31] Andrew T Stephen, Yaniv Dover, Lev Muchnik, and Jacob Goldenberg. Fresh is best: The effect of source activity on the decision to retransmit content in social media. *Available at SSRN 1609611*, 2014.

[32] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing, 2010 IEEE Second International Conference on*, pages 177–184. IEEE, 2010.

[33] Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. *arXiv preprint arXiv:1405.1438*, 2014.

[34] Olivier Toubia and Andrew T Stephen. Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter? *Marketing Science*, 32(3):368–392, 2013.

[35] Greg Ver Steeg and Aram Galstyan. Information transfer in social media. In *Proc. of the 21st Inter. WWW Conf.*, pages 509–518. ACM, 2012.

[36] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.

[37] Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. *ICWSM*, 10:355–358, 2010.