Modeling Collective Anticipation and Response on Wikipedia

Ryota Kobayashi,[1, 2] **Patrick Gildersleve,** [3] **Takeaki Uno,** [4] **Renaud Lambiotte** [3]
[1] The University of Tokyo
[2] JST, PRESTO
[3] University of Oxford
[4] National Institute of Informatics
r-koba@k.u-tokyo.ac.jp, patrick.gildersleve@oii.ox.ac.uk, uno@nii.jp, renaud.lambiotte@maths.ox.ac.uk

Abstract

The dynamics of popularity in online media are driven by a combination of endogenous spreading mechanisms and response to exogenous shocks including news and events. However, little is known about the dependence of temporal patterns of popularity on event-related information, e.g. which types of events trigger long-lasting activity. Here we propose a simple model that describes the dynamics around peaks of popularity by incorporating key features, i.e., the anticipatory growth and the decay of collective attention together with circadian rhythms. The proposed model allows us to develop a new method for predicting the future page view activity and for clustering time series. To validate our methodology, we collect a corpus of page view data from Wikipedia associated to a range of planned events, that are events which we know in advance will have a fixed date in the future, such as elections and sport events. Our methodology is superior to existing models in both prediction and clustering tasks. Furthermore, restricting to Wikipedia pages associated to association football, we observe that the specific realization of the event, in our case which team wins a match or the type of the match, has a significant effect on the response dynamics after the event. Our work demonstrates the importance of appropriately modeling all phases of collective attention, as well as the connection between temporal patterns of attention and characteristic underlying information of the events they represent.

Introduction

In recent years, many aspects of human activities have become increasingly mediated by digital services, leaving electronic footprints as an invaluable resource revealing the forces driving the structure and dynamics of social systems (Backstrom et al. 2006; Crane and Sornette 2008). A central question of computational social science is to understand the mechanisms by which individuals, taken as groups, exhibit collective behaviours (Lazer et al. 2009). This relationship is particularly striking when considering the emergence and subsequent decline in popularity, or success, on the web and in social media (Szabo and Huberman 2010; Goel et al. 2010; Bandari, Asur, and Huberman 2012; Proskurnia et al. 2017; Candia et al. 2019). Take the adoption of certain hashtags, instead of others, when confronted with new social phenomena (Lin et al. 2013); or the fact that certain songs become extremely popular while most remain in the dark (Salganik, Dodds, and Watts 2006). Several studies have looked at large-scale, temporal datasets or performed online experiments in order to identify the mechanisms, or the lack of them, explaining what makes a specific item successful (Hofman, Sharma, and Watts 2017). In this context, it is now understood that a successful outcome arises due to a complex combination of chance, of multiplicative, cascading effects and of intrinsic quality, in different proportions depending on the system under scrutiny (Janosov, Battiston, and Sinatra 2019).

More specifically, the emergence of peaks of popularity in online social systems has attracted much attention (Crane and Sornette 2008; Lehmann et al. 2012). Peaks emerge due to endogenous forces defined as interactions within social media, or as a response to exogenous shocks (e.g., news and disasters) — and most of the time due to a combination of both factors. Peaks may also emerge around planned events, whose date is known ahead of the realisation of the event, where the anticipation of the users plays an important role to shape the dynamics. Representative examples include political elections or sport events. It is the main focus of this article to model the dynamics of collective attention around such planned events.

In this study, we investigate the question of how people respond to planned events in an online setting. While several previous studies have focused on the dynamics of popularity, they have mostly considered specific events (e.g., movie release: Mestyán, Yasseri, and Kertész (2013), elections: Yasseri and Bright (2016), and airplane crashes: García-Gavilanes, Tsvetkova, and Yasseri (2016); García-Gavilanes et al. (2017)) or, looking for universal patterns, have not paid attention to the type of events (Crane and Sornette 2008; Matsubara et al. 2012; Zhao et al. 2015; Kobayashi and Lambiotte 2016; Proskurnia et al. 2017). For these reasons, it remains unclear how event-related information (e.g., category and outcome of an event) influences its peaks of popularity. Addressing this question is important for different reasons. First, it is essential to develop efficient mathematical models of popularity dynamics for the automatic identification of online events (Petrović, Osborne, and Lavrenko 2010; Becker et al. 2012), early detection of emerging stars and viral content on the Internet (Tatar et al. 2014), and

evaluation of the effect of recommender systems (Wu, Rizoiu, and Xie 2019). Second, this understanding of popularity peaks in online systems translates to an insight about the mechanisms of popularity, trends, and fads offline. Previous studies have demonstrated that popularity in online systems is highly correlated with other offline indicators, including stock market volumes (Bordino et al. 2012), movie box office success (Mestyán, Yasseri, and Kertész 2013), the number of people infected with influenza (Hickmann et al. 2015), and election results (Yasseri and Bright 2016).

Here we focus on a range of planned events in order to explore the following two research questions (RQs):

RQ1 What are the essential characteristics of collective attention dynamics towards planned events?

RQ2 How is event-related information (e.g., category and outcome) associated with the dynamics of collective attention?

To answer these questions, we concentrate on page views on Wikipedia and collect data for planned events in a wide range of contexts. We model Wikipedia time series by incorporating the anticipation and response to an event, as well as circadian rhythm. Importantly, the interpretability of model parameters allows us to quantify more robustly the relationship between the collective attention before and after the event, differentiating between their volume and time scale. Based on this model, we develop a Bayesian method for predicting future page view activity and perform a clustering analysis based on time series of collective attention. Furthermore, we restrict our analysis to a data set associated to football matches to investigate how the temporal pattern of collective attention is related to the event outcome.

Related Work

Within the field of computational social science, several works have focused on popularity peaks in social media or on the Web. In Crane and Sornette (2008), the authors analyzed Youtube data and identified two classes in the collective attention dynamics, sudden peak with rapid relaxation, associated to exogenous shocks and gradually growing patterns until the peak, followed by a symmetric relaxation, associated with endogenous effects. Other works have proposed taxonomies of popularity peaks based on their temporal profiles. For instance, in Yang and Leskovec (2011), the authors considered the time-series of popularity of blog posts and news media articles, which they clustered based on a similarity metric invariant to scaling and shifting.

In a related study (Lehmann et al. 2012), the authors analyzed hashtag data on Twitter and identified four classes in the collective attention dynamics: 1) Sudden peak and rapid relaxation, 2) Significant growth before the peak and symmetric relaxation, 3) Significant growth before the peak and sudden decay, and 4) Sudden peak and sudden decay. Furthermore, they conducted semantic analysis using WordNet semantic lexicon. The main differences between our work and this state-of-the-art is the fact that we investigate the association between event-related information (e.g., category and outcome) and collective attention dynamics. Our interpretable model describes all the four classes identified in Lehmann et al. (2012) and allows us to develop a new method for predicting the future page view activity and clustering time series. Another critical difference is that we focus exclusively on planned events. Studies that pay special attention to planned events are more limited. Becker et al. (2012) tackle the task of retrieving online content for planned events across a range of different social media platforms. This work focused on developing the event detection algorithm, but did not investigate the relationship between event-related information and collective attention dynamics, as considered here.

Our work is relevant to the question of how the popularity of fads rises and falls. Fads, defined as cultural objects including phrases, names, and customs that are very popular for a short period, have been an important research topic in the social sciences (Burgess and Park 1933; Aguirre, Quarantelli, and Mendoza 1988; Abrahamson 1991). However, there has only been a small number of works looking at the temporal property of fads (for exceptions, see Strang and Macy (2001); Rich (2008); Berger and Le Mens (2009); Denrell and Kovács (2015)) compared to the research on the spread of innovation (see references in Rogers (2010)). Fads are considered as two essential processes (Strang and Macy 2001; Rich 2008): the adoption and subsequent abandonment of an object. While most studies on innovation (Rogers 2010) focus on adoption, they pay little attention to the abandonment. In general, it is difficult to track the temporal pattern of popularity of real-world phenomena. Whilst online reactions to news events are not strictly fads, there are certainly parallels between the dynamics of anticipation / adoption and forgetting / abandonment. This link is particularly pertinent when the computational social science literature typically focuses on post-event response and not the anticipation / adoption phase.

Another stream of works related to our contribution concerns the modelling and prediction of time series in social media. Within this question, part of the research focuses on the mechanisms that may explain the growth, or the lack of it, of certain items online. In Lin et al. (2013), for instance, the authors analysed the competition between novel hashtags during the 2012 U.S. presidential debates and investigated mechanisms by which certain hashtag take over. Those mechanisms include preferential attachment and competition driven by the limited amount of attention of users (Weng et al. 2012). From a modelling perspective, various time series models have been proposed to predict the dynamics of popularity in online social systems. Important works include Matsubara et al. (2012), proposing a time series model, SpikeM, that incorporates an exponential rise, power-law decay, and circadian rhythms. In Proskurnia et al. (2017), the authors proposed a time-series model that incorporates reinforcement and circadian rhythms, allowing to predict the popularity dynamics on thepetitionsite.com. While the previous works developed time series models for predicting popularity dynamics, our work additionally exploits a model in order to investigate the relationship between event-related information and popularity dynamics.

Our work also finds connection with the growing research effort on understanding Wikipedia, our main source

of data. Most importantly for our research, several works have shown that popularity on Wikipedia is strongly related to popularity in other online services. User surveys in Singer et al. (2017) show 'media' and 'current events' as key motivations for browsing Wikipedia (30% and 13% of respondents respectively), likely the same driving forces as other online media. To compare platforms more directly; there is a strong correlation between the Wikipedia page view activity and Google search activity (Ratkiewicz, Flammini, and Menczer 2010; Yoshida et al. 2015), despite certain differences, e.g., trends on Twitter and Wikipedia are more ephemeral than on Google, both rising and declining rapidly for newly emerging topics in Althoff et al. (2013). Taken together, these results indicate that Wikipedia page views reflect users' behavior on the Internet in general.

In addition to the consumption of information, a range of works cover the dynamics of the production of knowledge on Wikipedia. Whilst we do not study the relationship between the production and consumption of information, they are intimately connected. The interaction between different editors on the supply side of information can yield rich dynamics, through both collaborative efforts (Keegan, Gergle, and Contractor 2011, 2012; West, Weber, and Castillo 2012), and conflicts (Yasseri et al. 2012), frequently in response to current events. In addition, demand for information on Wikipedia in response to current events is correlated with yet typically precedes, or even drives, supply (as measured by page views and the creation of new articles respectively) (Ciampaglia, Flammini, and Menczer 2015). Thus, understanding the dynamics of collective attention is also key to studying the production of knowledge on Wikipedia and beyond.

Data

We consider the popularity dynamics of events in 5 categories; Elections, Sports events, Association Football matches (also known as soccer and abbreviated to "Football"), Films (with release date), and Holidays. In the early stages of this study, we collected various classes of events over a wider range of categories, including Armed conflicts, Arts, Culture, Business, Economy, Disasters, Accidents, and International relations, among other things. We discarded the other categories after deciding to focus on events that are planned in advance and organized on a single day. Although we limit the scope to this type of event in this study, the proposed model could be generalized to the other types of events. Data for events from each class was obtained by scraping table data from relevant English Wikipedia summary articles. This included the event name (with linked Wikipedia article), date, as well as any class specific auxiliary data such as competition winner, or film director. In the case of football matches, articles for the two teams competing in each match were collected. Events that did not have a corresponding Wikipedia article were removed from analysis. Table 1 summarizes the statistics of this data set[1].

[1]All datasets and code available at github.com/NII-Kobayashi/Collective-Anticipation-and-Response.

We downloaded Wikipedia data dumps for hourly page views (Wikimedia 2020). For each article associated to each event, we collected hourly page views for 10 days before and after the event day (21 days total). Page views towards Wikipedia redirect pages for the articles in question were also included[2]. Whilst the summary articles where the initial events are scraped from act as an important traffic-shaping navigational tool for those browsing within Wikipedia, the majority of attention towards the articles studied comes from sources external to Wikipedia, most commonly Google search (Dimitrov et al. 2018; Gildersleve and Yasseri 2018). As such, we are confident the results are not determined by Wikipedia-specific navigational constraints from our choice of summary page and are representative of wider interest in events from direct information demand. We focus on planned events lasting no more than 1 day and define the peak value as the maximal page view count in a hour within 48 hours from 0:00 UTC on the supplied date of the event. We selected popular events whose peak value is more than 100 views/hour. In this way, 842 events from the initial database were selected for the following analysis: 92, 213, 250, 229, and 58 events for election, sports events, football matches, film, and holiday, respectively.

Modeling Anticipation and Response

Model Description

Here, we propose a simple model for the dynamics of collective anticipation and response, i.e., the number of hourly page views of a Wikipedia article before/after the associated event:

$$f_{\mathrm{peak}}(t) = C(t) D_{\mathrm{peak}}(t), \tag{1}$$

where $C(t) = 1 + \alpha_c \cos(\omega(t - t_c))$ ($\omega = 2\pi/T$) describes the circadian rhythm, α_c ($0 \leq \alpha_c < 1$) is the amplitude, $T = 1$ (day) is the period, and t_c is the peak time in the daily oscillation. The function $D_{\mathrm{peak}}(t)$ describes the anticipation and the response to an event:

$$D_{\mathrm{peak}}(t) = \begin{cases} a_- e^{(t-t_p)/\tau_-} + b_- & (t < t_p) \\ a_+ e^{-(t-t_p)/\tau_+} + b_+ & (t > t_p) \end{cases}, \tag{2}$$

where t_p is the peak time, τ_- (τ_+) is the time constant of the anticipation before the peak (the response after the peak), $a_-(a_+)$ represents the amplitudes, and $b_-(b_+)$ represents the baseline activity before (after) the peak. Note that this model describes activity except for the peak time t_p (hour).

The proposed model incorporates the essential features of the peaks in Wikipedia: 1) anticipation, 2) response, and 3) circadian rhythm. Several studies have attempted to model burst activity on the web and in social media (Crane and Sornette 2008; Matsubara et al. 2012; Tsytsarau, Palpanas, and Castellanos 2014; Kobayashi and Lambiotte 2016; Proskurnia et al. 2017; Rizoiu et al. 2017). However, to the best of our knowledge, the proposed model is the first to incorporate all three features. While we consider the gradual growth of popularity before a peak as anticipation and describe it using an exponential function, this growth can be attributable to

[2]See the following for more details wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Traffic/pageviews/Redirects

Category	Summary Articles	# events	Period
Election	[2015-18] national electoral calendar	342	01/07/15[3]–31/12/18
Sports Events	[January*-December] [2015-18] in sports	1,983	01/07/15–31/12/18
Football Matches	2017-18 UEFA [Champions, Europa] League	330	12/09/17–26/05/18[4]
Film	[2017, 2018] in film	547	01/01/17–31/12/18
Holiday	Public holidays in [the United States, the United Kingdom, Australia]	71	01/01/18–31/12/18

Table 1: Statistics of Wikipedia data.

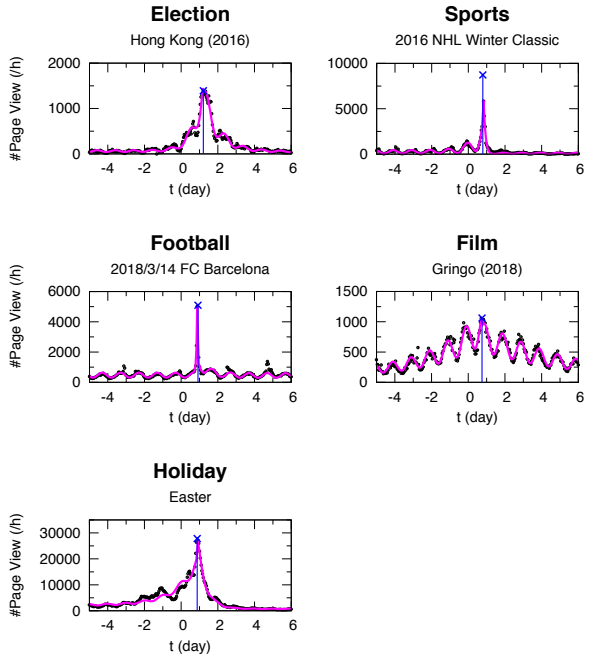

Figure 1: The proposed model reproduces collective attention dynamics to a variety of planned events. Examples of five event categories (election, sports, football matches, film releases, and holidays) are shown. Data points are shown as dots, the peaks are shown as blue crosses, and the fit by the proposed model is shown in magenta lines.

endogenous factors such as word of mouth such as in Crane and Sornette (2008). In this endogenous model, the peak is described as a symmetric power-law function before and after the peak $\propto |t - t_p|^{-\gamma}$. Some of existing models incorporate circadian rhythm (Matsubara et al. 2012; Kobayashi and Lambiotte 2016); however, these models focus on the information cascade after the event and do not consider activity before the peak.

Fitting Accuracy

We examine whether the proposed model can fit the observed dynamics of collective attention. Here, we analyze Wikipedia data related to planned events from five categories (election, sports events, film releases, football matches, and

[3]Page view data made readily available from this date.

[4]Corresponding to the 2017/18 season.

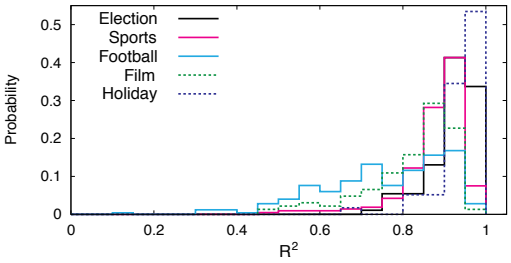

Figure 2: Histogram of Fitting Accuracy. The coefficient of variation R^2 was compared among five categories: election (black), sports (magenta), football (cyan), film (dashed green), and holiday (dashed blue).

holidays). Figure 1 shows that the model accurately fits the temporal patterns in the Wikipedia page view activity for a variety of events. The quality of the model fit is evaluated by the coefficient of determination, defined as

$$R^2 = 1 - \frac{\sum_t (s_t - f_{\mathrm{peak}}(t))^2}{\sum_t (s_t - \langle s \rangle)^2}, \tag{3}$$

where s_t is the time series of page view data, i.e., the number of page views in the last hour, $\langle s \rangle$ is its average over time, and the summation is over all time points except for the peak time t_p. The coefficient R^2 measures the amount of variance described by the model, and is high when the model can fit the data accurately. Note that the maximum value of 1 achieved is only if the model perfectly fits the data: $s_t = f_{\mathrm{peak}}(t)$ for any time t. The model accurately fits the Wikipedia page view data: median of the coefficient R^2 was 0.88, which indicates that our model describes 88% of the variation in data. Figure 2 shows the histogram of the coefficient R^2 for each category. While the model can fit the peaks of election, sports, holiday events very accurately, the case of football events is more delicate. This loss of accuracy could originate from the fact that teams may play another match in a different competition within several days of the studied event, inducing additional peaks. In addition, we compared the fitting accuracy of the proposed model to that of two existing methods, SpikeM (Matsubara et al. 2012) and power-law model (Crane and Sornette 2008). The fitting accuracy of the proposed method was better than both of these methods, with median R^2 of 0.76 and 0.74 for SpikeM and the power-law model, respectively.

	τ_- (hours)	τ_+ (hours)	ρ
Election	6.2 (3.7−14)	19 (14−25)	0.5 (0.4−0.8)
Sports	7.0 (2.5−18)	14 (11−17)	0.7 (0.4−1.0)
Football	1.1 (0.8−5.3)	6.7 (1.3−11)	0.9 (0.8−1.0)
Film	34 (20−56)	87 (55−140)	0.7 (0.6−0.8)
Holiday	11 (7.7−17)	12 (9.0−16)	1.3 (1.1−1.7)

Table 2: Fitted parameter distribution (Median and interquartile ranges): Anticipation and response time constants $\{\tau_-, \tau_+\}$ and the ratio $\rho = S_-/S_+$.

Interpretability of Model Parameters

The proposed model has three types of parameters: anticipation $\{a_-, b_-, \tau_-\}$, response $\{a_+, b_+, \tau_+\}$, and circadian rhythm $\{\alpha_c, t_c\}$. a, b, and τ represent the amplitude, baseline views, and time constant for anticipation and response, and α_c, t_c represent the circadian amplitude and offset. The time constants depend on the event category (Table 2). For example, the anticipation and response of a film release tend to be slow ($>$ 1 day) compared to the other categories. In contrast, a football match demonstrates small time constants ($\leq$ 7 hours). Viewers typically lose interest in the next morning. In addition, we quantify the anticipation-response ratio as $\rho = S_-/S_+$, where $S_- = \int_{-M}^{0} D_{\rm peak}(s)ds$, $S_+ = \int_{0}^{M} D_{\rm peak}(s)ds$ (the area under the anticipation / response curve), and $M = 7$ (days) is the time window. We observe that most events (except for Holidays) are response dominated (Table 2).

The circadian parameters are associated with the regional (time zone) distribution of viewers. To demonstrate this, we infer the ratio of viewers in the United States (US), United Kingdom (UK), and Australia (AU) time zones towards individual articles. Assuming that most viewers come from the US, UK, or AU, the fitted circadian function $C(t)$ can be decomposed into three waves:

$$C(t) = p_{\rm US}C_{\rm US}(t) + p_{\rm UK}C_{\rm UK}(t) + p_{\rm AU}C_{\rm AU}(t), \qquad (4)$$

where $C_{\rm X}(t) = 1 + \bar{\alpha}\cos(\omega(t - t_X))$ is the circadian function in $X \in \{\text{US, UK, AU}\}$, and $\bar{\alpha} = 0.9$ is a constant. The reference time t_X was estimated by fitting the circadian function $C(t)$ from the page view data of public holidays and calculating the average of the peak time t_c for the domestic holidays, such as Flag Day (United States), Ayr Gold Cup, and Melbourne Cup ($\alpha_c > 0.5$). The estimate was $t_X = 20.6$, 16.2, and 5.9, for US, UK, and AU, respectively, which reflects the time difference among these regions (US: UTC-5, AU: UTC+10). The viewer ratio $\{p_{\rm US}, p_{\rm UK}, p_{\rm AU}\}$ was determined using the least squares method. Next, we analyze the Wikipedia data related to the football matches (UEFA Champions league 2017-18) and film releases. The UK dominated the attention of nearly all the football matches (98 %: 244/250), and the US dominated the attention of most of the films (81 %: 186/229). These results indicate that Wikipedia page views reflect the popularity in the real world; all football teams in the UEFA Champions league are based in Europe and the 2018 box office revenue of the US (11.9 billion US dollars) is much greater than that of UK and AU (1.7 and 0.8 billion US dollars, respectively) [5].

Experimental Evaluation

We present two applications of the proposed model:

- predicting the number of page views after the peak;
- clustering page view time series;

Based on the proposed model, we develop a forecasting method and implement a clustering method. Then, these methods are applied to the Wikipedia data set, and their performance is compared to the existing methods.

Predicting the Number of Page Views

Here we formally define the prediction problem. Given the page view time series s_t up to the observation time $t_{\rm obs}$, we seek to predict the page view activity during prediction period $t_{\rm obs} < t \leq M (= 7$ days). Figure 3 shows examples of the prediction results.

Bayesian method for predicting page view time series We develop a method for predicting the future view activity of a Wikipedia article based on the proposed model. First, the circadian parameters $\{\alpha_c, t_c\}$ and anticipation parameters $\{a_-, b_-, \tau_-\}$ are determined by minimizing the least squared error before the peak.

We then employ a Bayesian approach for fitting the remaining parameters (response parameters), $\vec{\theta}_+ = \{a_+, b_+, \tau_+\}$, from short observations. The Gaussian distribution is assumed for the observed data

$$P\left(s_t|\vec{\theta}_+\right) = N(s_t|f_{\rm peak}(t), \sigma_n^2), \qquad (5)$$

where $N(x|\mu, \sigma^2)$ is the normal distribution with the mean μ and the variance σ^2. The log-normal prior distribution is assumed to utilize the categorical information (e.g., election) and the page view activity before the peak.

$$P\left(\vec{\theta}_+\right) = \prod_{q_+ \in \vec{\theta}_+} {\rm LN}(q_+|c_{q,k}\hat{q}_- + d_{q,k}, \sigma_{q,k}^2), \qquad (6)$$

where $q_+ \in \{a_+, b_+, \tau_+\}$ is a response parameter and $\hat{q}_- \in \{\hat{a}_-, \hat{b}_-, \hat{\tau}_-\}$ is the fitted value before the peak, and k represents the event category. ${\rm LN}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma x}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}$ denotes the log-normal distribution. Hyper-parameters $\{c_{q,k}, d_{q,k}\}$ were determined from fitted parameters. The response parameters are determined by maximizing the posterior probability

$$P\left(\vec{\theta}_+|\{s_t\}\right) \propto P\left(\{s_t\}|\vec{\theta}_+\right) P\left(\vec{\theta}_+\right). \qquad (7)$$

Note that the Bayesian method is equivalent to the least squares method when we assume uniform prior distribution: $P\left(\vec{\theta}_+\right) \propto 1$. Finally, the page view activity is predicted by calculating the model (Eq. 1) using the fitted parameters.

[5] www.boxofficemojo.com/year/?area=USA, www.boxofficemojo.com/year/?area=UK, www.boxofficemojo.com/year/?area=AU

Prior	1 day	2 days	3 days
No prior	0.71	0.57	0.49
Anticipation	0.68	0.58	0.50
Anticipation, Category	**0.54**	**0.51**	**0.46**

Table 3: Prediction error of the Bayesian method. Three methods are compared: No prior, Anticipation (only anticipation parameters are utilized, i.e., hyper parameters $c_{q,k}, d_{q,k}$ do not depend on the category k), and Anticipation, Category (anticipation parameters and event category are utilized). The best method is shown in bold. The observation time $t_{\rm obs}$ is varied from 1 day to 3 days.

Method	Complexity	1 day	2 days	3 days
Proposed	8	**0.54**	**0.51**	**0.46**
SpikeM	7	0.81	0.68	0.59
Power-law	4	0.68	0.64	0.62
LR	≥ 192	0.69	0.56	0.54
LSTM	$\approx 160{,}000$	1.4	1.0	0.83

Table 4: Comparison of prediction errors among the proposed method and existing methods. The best method is shown in bold. Complexity represents the number of parameters of each method to describe a peak.

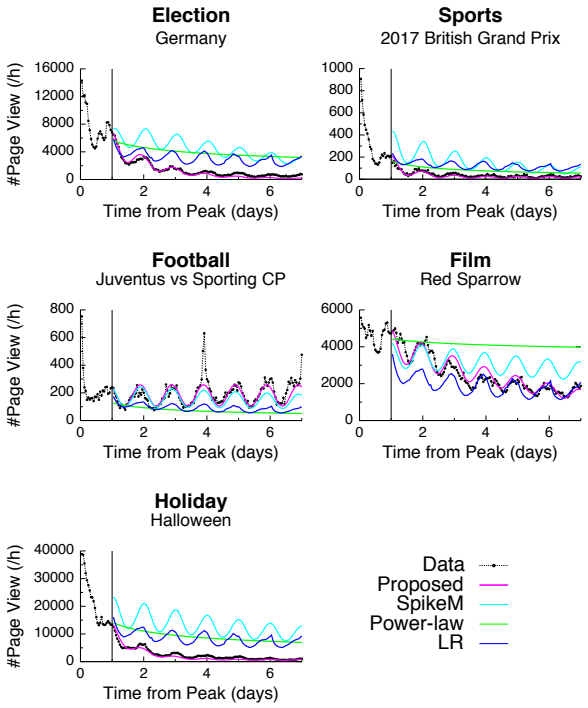

Figure 3: Predicting future page view activity. The proposed method (magenta) outperforms existing methods. Vertical bar represents the observation time $t_{\rm obs} = 1$ (day). Results from the LSTM approach are omitted from the figure as its predictions often lie outside a reasonable plot range.

Evaluation metrics Absolute Percentage Error (APE) of page view time series is used to evaluate the prediction error, $\mathrm{APE} = \sum_{t=t_{\rm obs}+1}^{M} \frac{|s_t - \hat{s}_t|}{N}$, where $N = \sum_{t=t_{\rm obs}+1}^{M} s_t$ is the total number of page view during the prediction period, $\hat{x}$ denotes the predicted value of x.

Prediction results We evaluate prediction performance by analyzing the Wikipedia data set, and compare the performance of the proposed method to state-of-art approaches.

We develop a Bayesian method for fitting the response parameters $\{a_+, b_+, \tau_+\}$, which utilizes the attention dynamics before the peak and event category information. First, we evaluate the prediction performance of the Bayesian methods (Table 3). Incorporating both anticipation parameters and category information substantially improved the accuracy for short observations. In contrast, the utilization of anticipation parameters only did not improve the accuracy. This result indicates that the event category information supplements the insufficiencies of the page view data.

Next, we compare the performance of the proposed method to four existing methods: SpikeM (Matsubara et al. 2012), Power-law model (Crane and Sornette 2008; Tsytsarau, Palpanas, and Castellanos 2014), Linear Regression (LR) (Szabo and Huberman 2010), and LSTM (Hochreiter and Schmidhuber 1997; Mishra, Rizoiu, and Xie 2018). For details see appendix A. Figure 3 shows examples of time series prediction, which demonstrates that the proposed method outperforms the baseline methods. Though SpikeM reproduces the circadian rhythm, it tends to overestimate the page view activity. Furthermore, the proposed method provides the most accurate predictions in terms of the cumulative count and the time series (Table 4). We observe an improvement of around 20% over the two runners up (Power-law and LR) for time-series prediction from 1 day observations. Furthermore, we confirm that the result is qualitatively the same for the prediction performance of the cumulative view count (Appendix B).

Clustering Analysis

We implement a method for clustering time series data based on the proposed model (Eq. 1). The anticipation and response parameters $(a_-, a_+; \tau_-, \tau_+; b_-, b_+)$ are log-transformed before the clustering because they exhibit heavy-tailed distributions. The transformed parameters, along with the circadian parameters (α_c, t_c), are then used as features in clustering with a Gaussian Mixture Model (Reynolds 2009). The number of clusters is determined based on the Bayesian Information Criterion (BIC) (Schwarz 1978), which yields six clusters. The clusters are robust with respect to the initial conditions.

Figure 4 shows the centers for $K = 6$ clusters, and Table 5 shows the distribution of each category. Four clusters (C1, C2, C3, and C4) correspond to a single category: sports events, football matches, and film release, whereas all the categories are mixed in the other clusters (C5 and C6). Cluster C1 exhibits two peaks before/after the main peak due to the circadian rhythm, which corresponds to typical activity for minor sports events. Cluster C2 exhibits a quick rise and decay, which corresponds to typical activity for foot-

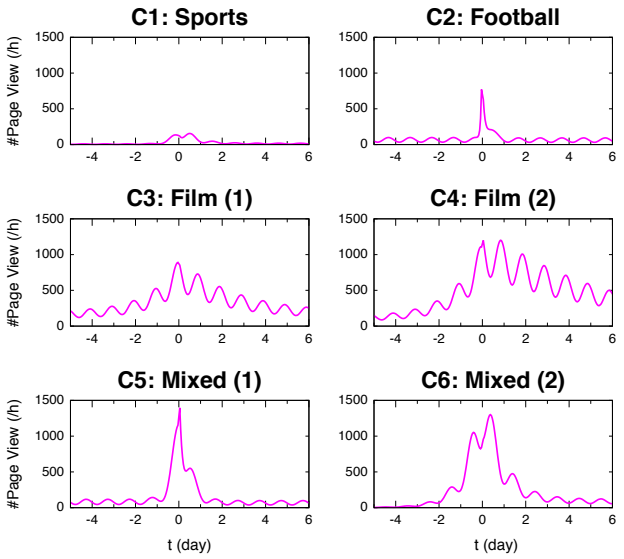

Figure 4: Identified cluster centers based on the parameters of the proposed model. The cluster ID and its corresponding category are shown in top of each panel.

	C1	C2	C3	C4	C5	C6
Election	3	2	4	3	16	64
Sports	**120**	7	1	4	11	70
Football	0	**165**	0	0	5	63
Film	0	0	**140**	**57**	26	6
Holiday	0	0	0	0	14	44

Table 5: Identified clusters (C1, C2, $\cdots$, C6) based on model parameters. Bold indicates that the cluster is dominated by a category ($> 80\%$).

ball matches. The activity for film releases are in clusters C3 and C4, with a slow rise and decay. Cluster C4 represents the activity for films more popular than C3, decaying more slowly. The other two clusters contains all event categories, and cluster C6 shows a slower pattern than C5.

We evaluate the quality of clustering by measuring the similarity to the event category defined by Wikipedia editors. The similarity is quantified by using adjusted mutual information (AMI) (Vinh, Epps, and Bailey 2010) that takes a value from 0 (independent) to 1 (perfectly correlated), and accounts for increased baseline mutual information between partitions with a larger number of clusters. Here we compare the clustering result from our approach against those obtained using parameter sets of alternative models; SpikeM and the power-law model, as before, as well as the method proposed by Lehmann et al. (2012) which is based on the fraction of total views before, during, and after a peak $\{f_-, f_p, f_+\}$. The previously mentioned linear regression model is not considered, owing to its dependence of the number of parameters on the prediction window, likewise the LSTM model with its many parameters. We compare against models with the number of clusters that minimise the respective BIC. We also tested alternative models with six clusters, to compare against our own six cluster model, but performance for each of the alternative models was worse than that of their BIC selected clusterings. Results are displayed in Table 6 and our approach comfortably outperforms the other models, with median AMI of 0.47 over 10,000 different initial conditions.

Method	Clusters	Features	AMI
Proposed	6	8	**0.47 (0.46–0.49)**
SpikeM	8	7	0.35 (0.33–0.38)
Power-law	7	4	0.36 (0.32–0.38)
Fraction	2	3	0.39 (0.39–0.39)

Table 6: Median and interquartile range for AMI over 10,000 initial conditions across the different models. Fraction means the clustering based on the fractions of total views before, during, and after a peak. The best method is shown in bold.

	a_-	a_+
Group	940 (280−2,400)	580 (220−1,500)
Knockout	5,500 (2,300−11,000)	2,200 (1,200−9,100)
Final	27,000	16,000

Table 7: Median and interquartile ranges of parameters for different stages (no range for final).

Relating Event Outcome and Collective Attention Dynamics

The results presented in the previous section suggest that the temporal properties of collective attention (e.g., the parameters of the proposed model) are associated with the event category (e.g., football matches). In this section, we investigate the link between more detailed information about the event and the model parameters. Here, we restrict our analysis to a dataset associated with football matches for easy interpretation of the results.

First, we examine how the model parameters depend on the stage of competition (group stage, knockout stage, and the final match in the Champions League) and the match result (win, draw, and lose). As shown in Table 7, the anticipation and response amplitude (a_-, a_+) depend on the stage of competition, with matches in the latter stages of the competition being more popular. The time constants (τ_-, τ_+) do not depend on the stage (data not shown).

Interestingly, the response parameters (a_+, τ_+) of the losing teams are distinct from the winning ones (Table 8). We can observe two clusters in the response parameters that correspond to excited ($\tau_+ > 2$ hours) and disappointed ($\tau_+ < 2$ hours) response (dashed line in Fig. 5A). The proportion of losing teams in the disappointed class is much higher than that in the excited one: 55 % (50/91) and 31 % (49/159), respectively. Note that the disappointed response cannot be identified based solely on the view counts before and after the peak (Fig. 5B). Ahead of their Champions League semifinal on 2nd May against AS Roma, Liverpool FC manager Jürgen Klopp warned that “no one remembers losers” (Bascombe 2018). Liverpool won the two-legged tie, and our

	a_+	τ_+ (hours)
Win	680 (250−1,300)	8.6 (2.5−12)
Draw	770 (200−1,900)	9.4 (1.0−15)
Lose	1,400 (370−4,600)	1.9 (1.0−7.9)

Table 8: Median and interquartile range of response parameters for different match results.

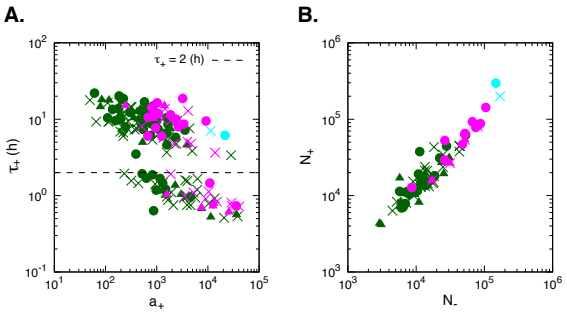

Figure 5: Effect of stage and result of a football match on response dynamics (2017-18 UEFA Champions League). A. Reaction parameters (a_+, τ_+). The peaks with poor fitting accuracy ($R^2 < 0.7$) are omitted from this figure. B. Page view counts before and after the peak (N_-, N_+). Colors represent the stage, i.e., group stage: green, knock Out stage: magenta, final: cyan. Symbols represent the results, i.e., win: circle, draw: triangle, and lose: cross.

results clearly confirm Klopp's statement – with Liverpool FC's $\tau_+ = 12.9$ compared to AS Roma's $\tau_+ = 0.7$ (hours).

To further quantify the association between the model parameters and match result, we infer whether a football team wins, draws, or loses a match based on the page view time series of the team and its opponent. A linear support vector machine classifier[6] (Chang and Lin 2011) implemented in scikit-learn (Pedregosa et al. 2011) is used for the classification task. We consider two kinds of features: model parameters (based on the proposed model, SpikeM, and Power-law model) and a feature set based on the fractions of total views before, during, and after a peak (Lehmann et al. 2012). The classification performance is evaluated by 5-fold cross-validation and also compared against a baseline of always predicting the most frequent result (draw).

Table 9 summarises the classification performance of the selected parameter combinations against alternative models (other combinations with lower performance or equivalent performance and more features omitted). Classification based on the parameters of the proposed model yields better performance than the other approaches. The best performance is achieved by the features of the response parameters $\{a_+, b_+, \tau_+\}$ of both the target and opponent teams. Note that the other parameters, e.g. the anticipation and the circadian parameters, do not improve the performance.

[6]Best performance is achieved with a linear kernel, but the results are qualitatively similar for radial basis function, polynomial, and sigmoid kernels.

Model	Features	Accuracy
Proposed	$\{a_+,b_+,\tau_+\}^{opp}$	**68%**
	$\{a_+,b_+,\tau_+\}$	56%
SpikeM	$\{u(0),\beta,t_b,s_b,\epsilon_0,p_a,p_s\}^{opp}$	57%
Power-law	$\{a_+,\gamma_+\}$	58%
Fraction	$\{f_-,f_p,f_+\}$	55%
Baseline	–	42%

Table 9: Classification accuracy of football match results. By default, only the target team's parameters are included, superscript '*opp*' indicates the opposing team's parameters are also included. The best method is shown in bold.

These results indicate that the parameters of the proposed model are not only able to capture the the different temporal patterns of different event classes, but are also sensitive to detailed event-related information, in this case the stage of competition and match result. Finally, we analyzed another series of football matches (2017-18 UEFA Europa League) and the results are qualitatively similar to those of the Champions League data.

Discussion and limitations

As set out in the introduction, the two objectives of this paper were to investigate the two research questions:

RQ1 What are the essential characteristics of collective attention dynamics towards planned events?

RQ2 How is event-related information (e.g., category and outcome) associated with the dynamics of collective attention?

To address RQ1, we have collected hourly time series of page views for a range of topics. This data set is the first to focus on the hourly page viewing activity observed on relevant Wikipedia pages 10 days before and after each planned event, thus allowing to compare the dynamics of the collective attention to various categories of events. We have then proposed a simple model for characterizing the time series of collective attention, which has helped reveal some of their essential properties:

- Attention dynamics exhibits exponential anticipation and response, and circadian rhythm;
- The time constants of anticipation and response (τ_+, τ_-) are event-related: for example, time constants are shorter for football matches and longer for film releases;
- Circadian rhythm is associated with the geographical distribution of attention.

The results on anticipation suggest an exponential distribution of the number of interested viewers before an event. An exponential increase in page-views can be attributed to the characteristics of independent individuals rather than to network effects (e.g., information diffusion on social network). This is because the time constant depends on the category of the event and it is larger (more than 10 hours) than that of information diffusion. Moreover, the technological affordances of Wikipedia as a website are very different from

social networks such as Facebook, Twitter, and YouTube, where interactions between users and 'trending' items are much more emphasised and lead to network popularity phenomena.

We have examined RQ2 through prediction analysis and clustering analysis. We have shown that the event category and the anticipatory pattern before the peak improve the prediction performance for the response pattern, even for short observation windows (Table 3). The clustering results of the attention patterns are akin to the event categories annotated by Wikipedia editors (Table 5). These results indicate that the temporal patterns of collective attention are highly concomitant with the event category (e.g., sports versus elections). In addition, we performed a classification task of the event outcome (win/draw/lose) using a data set of association football matches (UEFA Champions League 2017-18). We found that the realization of an event drives the dynamics of collective attention and demonstrated that the event outcome can be inferred from collective attention patterns.

These results suggest that events should be classified based on whether their *sentimental outcome* is known in advance or not. We define the concept of sentimental outcome according to how positive/negative a user feels after the event. The distribution of all users' feelings is then representative of the sentimental outcome of the event. Events that elicit positive responses such as popular movies or holidays have positive sentimental outcome, and events with negative responses such as disasters or a team losing a football match have negative sentimental outcome. The sentimental outcome is not always known in advance, and can also be bimodal. For example, the sentimental outcome of a football match is largely contingent on its result: the users will be excited (disappointed) if the team they support wins (loses) the match. This differential response is not observed for movies, though the sentimental outcome could vary depending on movie quality/success. Additionally, holidays typically do not have explicit outcome so variation in the response dynamics is not expected. This is confirmed by the relative sizes of the interquartile ranges for τ_+ in table 2. Unfortunately, due to limitations on the structured Wikipedia data, we are unable to test more generally this hypothesis of differential excited/disappointed responses based on the sentimental outcome, as we have done for the football matches. Our results also suggest that certainty of positive sentimental outcome is associated with the anticipatory activity. We found that most of the film releases and holiday events exhibit anticipatory activity, whereas most of election and sports events do not exhibit it (Fig. 1). Future research, using sentiment analysis on user generated content (e.g. data from Twitter (Dodds et al. 2011) or Reddit (Medvedev, Lambiotte, and Delvenne 2017)), could help investigate which type of events exhibit a strong anticipatory pattern by quantifying the sentimental outcome with tools from natural language processing. More generally, we have not fully explored the multitude of possible relations between event content and patterns of collective attention. Enriching our analysis with natural language processing tools in a combination of sources, including news articles and social media, is a promising research direction.

Our model can quantify collective attention patterns of events, and can thus help understand the mechanisms driving the anticipation and response of an event. However, there are some limitations in this study. Events can be classified based on two dimensions: whether or not they are planned in advance, and whether they are scheduled to be held on a single day or over multiple days. This study has focused on a specific type of event, i.e., planned events scheduled for a single day. The proposed model is validated for this restricted type of events, but it could be generalised for more general situations. For instance, the model without anticipation ($a_- = 0$) would be reasonable for unplanned events. Typically, events held over multiple days encompass several *sub-events*, such as in the Olympic games. The collective attention towards such events may be described by the sum of multiple peaks: $\sum_k f_{\rm peak}(t; t_{p,k})$, where $t_{p,k}$ is the time of the k-th sub-event. Further work is required to investigate how to extend our methodology to multi-day events with unknown sub-events and / or continuous stimuli. Moreover the present study has only considered events with a high popularity, owing to their importance online and offline. Although the temporal pattern of less trending events is similar, it is more challenging to determine model parameters and predict the response patterns for such events because of the scarcity of data points. Point process models (Valera and Gomez-Rodriguez 2015; Kobayashi and Lambiotte 2016) could be suitable for such situations. Overall, future studies would thus be required to develop the generalized models and investigate the links between event-related information and attention patterns in a more general setting.

Conclusion

In this paper, we have studied collective attention towards Wikipedia pages associated with planned events, defined as events known to happen at a specific date in the future (in contrast to unexpected, unplanned events such as earthquakes or plane crashes). As a first step, we have collected hourly time series of page views for a range of topics and proposed a simple model by incorporating key features, i.e., the anticipation and the response to an event as well as circadian rhythm. To the best of our knowledge, the proposed model for collective attention is unique in taking all of these factors into account. This model allows us to develop a method for predicting the future time evolution of the popularity and for clustering time series. Our methodology outperforms state-of-the-art methods for prediction and clustering tasks, emphasizing the importance of appropriately modeling the dynamics of collective attention. Interestingly, the event category information (e.g. football match and film release) improved the prediction accuracy from short observations, and the clustering result based on the model parameters was associated with the event categories. More specifically, using football match data, we have demonstrated that the response parameters are associated with the type and the result (win, draw, and lose) of the match. These results suggest that not only the event category but also that more detailed event-related information drives the temporal pattern of popularity, and led us to postulate the notion of sentimental outcome to explain the differences in attention patterns

between events. We believe that the proposed model provides an important contribution towards studying the relationship between the dynamics of collective attention and characteristic information underlying an event.

Acknowledgments

We thank Kazuhiro Kurita and Yuka Takedomi for stimulating discussions. Furthermore, this paper was greatly improved by the comments of anonymous reviewers. R.K. is supported by JSPS KAKENHI Grant Numbers JP17H03279, JP18K11560, and JP19H01133, and JST PRESTO Grant Number JPMJPR1925. T.U. is supported by JSPS KAKENHI Grant Numbers JP19H01133, and JST CREST JPMJCR1401.

Appendix A: Baseline models

We summarize four existing methods for predicting the response dynamics of collective attention.

- **SpikeM (Matsubara et al. 2012):** This model describes the page view count in the the last hour $x(t)$ using a difference equation

$$x(t+1) = p(t+1) \left(u(t) \sum_{k=t_b}^{t} (x(k) + s(k)) f(t+1-k) + \epsilon_0 \right),$$
$$u(t+1) = u(t) - x(t+1),$$

  where $p(t) = 1 - 0.5 p_a \{1 + \sin(\omega(t + p_s))\}$ ($\omega = 2\pi/T$) describes circadian rhythm, $s(t) = s_b$ ($t = t_b$), $s(t) = 0$ (otherwise) represents the shock due to the event, and $f(t) = \beta t^{-1.5}$ is the power-law function. Seven parameters $\{u(0), \beta; t_b, s_b; \epsilon_0; p_a, p_s\}$ are fitted using the least squares method. This model incorporates the power-law relaxation in response activity and the circadian rhythm in human behavior.

- **Power-law model (Crane and Sornette 2008; Tsytsarau, Palpanas, and Castellanos 2014):** Crane and Sornette (2008) showed that the viewing activity of YouTube videos demonstrates power-law relaxation behavior:

$$x(t) = \begin{cases} a_-(t_p - t)^{\gamma_-} & (t < t_p) \\ a_+(t - t_p)^{\gamma_+} & (t > t_p) \end{cases},$$

  where t_p is the peak time. Four parameters $\{a_-, \gamma_-; a_+, \gamma_+\}$ are fitted by least squares method.

- **Linear Regression (LR) (Szabo and Huberman 2010):** This method applies linear regression to the logarithm of the cumulative view counts

$$\log R(t) = \alpha_t + \log R(t_{\rm obs}) + \sigma_t \xi,$$

  where $R(t) = \sum_{k=t_p+1}^{t} x(k)$ is the cumulative view count after the peak, $t_{\rm obs}$ is the observation time, ξ represents Gaussian noise. For each time, the cumulative count is predicted by the unbiased estimator $\hat{R}(t) = R(t_{\rm obs}) \exp(\hat{\alpha}_t + \hat{\sigma}_t^2/2)$, where $\hat{\alpha}_t$ and $\hat{\sigma}_t^2$ are the fitted values obtained using the maximum likelihood method.

- **Long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997):** LSTM is a recurrent neural network model for time series forecasting (Hochreiter and Schmidhuber 1997). With the advent of deep learning methods, LSTM is also getting popular in social media analysis, including popularity prediction (Mishra, Rizoiu, and Xie 2018) and fake news detection (Ruchansky, Seo, and Liu 2017). We used the Matlab Deep Learning Toolbox for the implementation and adopted its default parameters [7]. Specifically, the number of hidden units was 200 and the number of model parameters was 160,800.

Appendix B: Prediction performance of the cumulative view count

We calculate the Absolute Percentage Error (APE) of cumulative page view count, $\mathrm{APE} = \frac{|N - \hat{N}|}{N}$, where $\hat{N}$ denotes the predicted cumulative view count after the observation period. We evaluate the prediction error of the Bayesian methods (Table 10). Then, the best Bayesian method was compared to existing methods (Table 11). The proposed method shows an improvement of around 20 % over the runner up (LR) for the prediction performance.

Prior	1 day	2 days	3 days
No prior	0.65	0.47	0.35
Anticipation	0.62	0.50	0.38
Anticipation, Category	**0.44**	**0.41**	**0.33**

Table 10: Prediction error of the Bayesian method. Three methods are compared: No prior, Anticipation (only anticipation parameters are utilized), and Anticipation, Category (both anticipation parameters and event category are utilized). The best method is shown in bold. The observation time $t_{\rm obs}$ is varied from 1 day to 3 days.

Method	Complexity	1 day	2 days	3 days
Proposed	8	**0.44**	**0.41**	**0.33**
SpikeM	7	0.74	0.60	0.51
Power-law	4	0.62	0.57	0.54
LR	≥ 192	0.54	0.51	0.50
LSTM	$\approx 160{,}000$	1.3	0.90	0.74

Table 11: Comparison of prediction errors among the proposed method and existing methods. The best method is shown in bold. Complexity represents the number of parameters of each method.

References

Abrahamson, E. 1991. Managerial fads and fashions: The diffusion and rejection of innovations. *Academy of management review* 16(3): 586–612.

[7] https://www.mathworks.com/help/deeplearning/ug/time-series-forecasting-using-deep-learning.html

Aguirre, B. E.; Quarantelli, E. L.; and Mendoza, J. L. 1988. The collective behavior of fads: The characteristics, effects, and career of streaking. *American Sociological Review* 569–584.

Althoff, T.; Borth, D.; Hees, J.; and Dengel, A. 2013. Analysis and forecasting of trending topics in online media streams. In *ACM Multimedia*, 907–916.

Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; and Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. In *KDD*, 44–54.

Bandari, R.; Asur, S.; and Huberman, B. A. 2012. The pulse of news in social media: Forecasting popularity. In *ICWSM*.

Bascombe, C. 2018. Jurgen Klopp warns Liverpool players 'no one remembers losers' ahead of Roma second leg. *The Telegraph* URL https://www.telegraph.co.uk/football/2018/05/01/jurgen-klopp-warns-liverpool-players-no-one-remembers-losers/.

Becker, H.; Iter, D.; Naaman, M.; and Gravano, L. 2012. Identifying content for planned events across social media sites. In *WSDM*, 533–542.

Berger, J.; and Le Mens, G. 2009. How adoption speed affects the abandonment of cultural tastes. *Proceedings of the National Academy of Sciences* 106(20): 8146–8150.

Bordino, I.; Battiston, S.; Caldarelli, G.; Cristelli, M.; Ukkonen, A.; and Weber, I. 2012. Web search queries can predict stock market volumes. *PloS one* 7(7): e40014.

Burgess, E. W.; and Park, R. E. 1933. *Introduction to the Science of Sociology*. Chicago University Press.

Candia, C.; Jara-Figueroa, C.; Rodriguez-Sickert, C.; Barabási, A.-L.; and Hidalgo, C. A. 2019. The universal decay of collective memory and attention. *Nature Human Behaviour* 3(1): 82–91. ISSN 2397-3374. doi:10.1038/s41562-018-0474-5. URL https://doi.org/10.1038/s41562-018-0474-5.

Chang, C.-C.; and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2(3): 27.

Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2015. The production of information in the attention economy. *Scientific reports* 5: 9452.

Crane, R.; and Sornette, D. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences* 105(41): 15649–15653.

Denrell, J.; and Kovács, B. 2015. The effect of selection bias in studies of fads and fashions. *PLoS One* 10(4): e0123471.

Dimitrov, D.; Lemmerich, F.; Flöck, F.; and Strohmaier, M. 2018. Query for architecture, click through military: Comparing the roles of search and navigation on wikipedia. In *Web Sci*, 371–380.

Dodds, P. S.; Harris, K. D.; Kloumann, I. M.; Bliss, C. A.; and Danforth, C. M. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one* 6(12): e26752.

García-Gavilanes, R.; Mollgaard, A.; Tsvetkova, M.; and Yasseri, T. 2017. The memory remains: Understanding collective memory in the digital age. *Science advances* 3(4): e1602368.

García-Gavilanes, R.; Tsvetkova, M.; and Yasseri, T. 2016. Dynamics and biases of online attention: the case of aircraft crashes. *Royal Society open science* 3(10): 160460.

Gildersleve, P.; and Yasseri, T. 2018. Inspiration, captivation, and misdirection: Emergent properties in networks of online navigation. In *International Workshop on Complex Networks*, 271–282. Springer.

Goel, S.; Hofman, J. M.; Lahaie, S.; Pennock, D. M.; and Watts, D. J. 2010. Predicting consumer behavior with Web search. *Proceedings of the National academy of sciences* 107(41): 17486–17490.

Hickmann, K. S.; Fairchild, G.; Priedhorsky, R.; Generous, N.; Hyman, J. M.; Deshpande, A.; and Del Valle, S. Y. 2015. Forecasting the 2013–2014 influenza season using Wikipedia. *PLoS Comput Biol* 11(5): e1004239.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Hofman, J. M.; Sharma, A.; and Watts, D. J. 2017. Prediction and explanation in social systems. *Science* 355(6324): 486–488.

Janosov, M.; Battiston, F.; and Sinatra, R. 2019. Success and luck in creative careers. *arXiv preprint arXiv:1909.07956* .

Keegan, B.; Gergle, D.; and Contractor, N. 2011. Hot off the wiki: dynamics, practices, and structures in Wikipedia's coverage of the Tōhoku catastrophes. In *Proceedings of the 7th international symposium on Wikis and open collaboration*, 105–113.

Keegan, B.; Gergle, D.; and Contractor, N. 2012. Do editors or articles drive collaboration? Multilevel statistical network analysis of Wikipedia coauthorship. In *CSCW*, 427–436.

Kobayashi, R.; and Lambiotte, R. 2016. TiDeH: Time-Dependent Hawkes Process for Predicting Retweet Dynamics. In *ICWSM*, 191–200.

Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabási, A.-L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; et al. 2009. Computational social science. *Science* 323(5915): 721–723.

Lehmann, J.; Gonçalves, B.; Ramasco, J. J.; and Cattuto, C. 2012. Dynamical classes of collective attention in twitter. In *WWW*, 251–260.

Lin, Y.-R.; Margolin, D.; Keegan, B.; Baronchelli, A.; and Lazer, D. 2013. # Bigbirds never die: Understanding social dynamics of emergent hashtags. In *ICWSM*.

Matsubara, Y.; Sakurai, Y.; Prakash, B. A.; Li, L.; and Faloutsos, C. 2012. Rise and fall patterns of information diffusion: model and implications. In *KDD*, 6–14.

Medvedev, A. N.; Lambiotte, R.; and Delvenne, J.-C. 2017. The anatomy of Reddit: An overview of academic research. In *Dynamics on and of Complex Networks*, 183–204. Springer.

Mestyán, M.; Yasseri, T.; and Kertész, J. 2013. Early prediction of movie box office success based on Wikipedia activity big data. *PloS one* 8(8): e71226.

Mishra, S.; Rizoiu, M.-A.; and Xie, L. 2018. Modeling popularity in asynchronous social media streams with recurrent neural networks. In *ICWSM*, 201–210.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Petrović, S.; Osborne, M.; and Lavrenko, V. 2010. Streaming first story detection with application to twitter. In *NAACL*, 181–189.

Proskurnia, J.; Grabowicz, P.; Kobayashi, R.; Castillo, C.; Cudré-Mauroux, P.; and Aberer, K. 2017. Predicting the Success of Online Petitions Leveraging Multidimensional Time-Series. In *WWW*, 755–764.

Ratkiewicz, J.; Flammini, A.; and Menczer, F. 2010. Traffic in social media I: paths through information networks. In *SOCIALCOM'10*, 452–458.

Reynolds, D. A. 2009. Gaussian Mixture Models. *Encyclopedia of biometrics* 741.

Rich, E. 2008. Management fads and information delays: An exploratory simulation study. *Journal of Business Research* 61(11): 1143–1151.

Rizoiu, M.-A.; Xie, L.; Sanner, S.; Cebrian, M.; Yu, H.; and Van Hentenryck, P. 2017. Expecting to be HIP: Hawkes intensity processes for social media popularity. In *WWW*, 735–744.

Rogers, E. M. 2010. *Diffusion of innovations*. Simon and Schuster.

Ruchansky, N.; Seo, S.; and Liu, Y. 2017. Csi: A hybrid deep model for fake news detection. In *CIKM*, 797–806.

Salganik, M. J.; Dodds, P. S.; and Watts, D. J. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762): 854–856.

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics* 6(2): 461–464.

Singer, P.; Lemmerich, F.; West, R.; Zia, L.; Wulczyn, E.; Strohmaier, M.; and Leskovec, J. 2017. Why we read wikipedia. In *WWW*, 1591–1600.

Strang, D.; and Macy, M. W. 2001. In search of excellence: Fads, success stories, and adaptive emulation. *American journal of sociology* 107(1): 147–182.

Szabo, G.; and Huberman, B. A. 2010. Predicting the popularity of online content. *Communications of the ACM* 53(8): 80–88.

Tatar, A.; De Amorim, M. D.; Fdida, S.; and Antoniadis, P. 2014. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications* 5(1): 8.

Tsytsarau, M.; Palpanas, T.; and Castellanos, M. 2014. Dynamics of news events and social media reaction. In *KDD*, 901–910.

Valera, I.; and Gomez-Rodriguez, M. 2015. Modeling Adoption and Usage of Competing Products. In *ICDM*, 409–418.

Vinh, N. X.; Epps, J.; and Bailey, J. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research* 11: 2837–2854.

Weng, L.; Flammini, A.; Vespignani, A.; and Menczer, F. 2012. Competition among memes in a world with limited attention. *Scientific reports* 2: 335.

West, R.; Weber, I.; and Castillo, C. 2012. Drawing a data-driven portrait of Wikipedia editors. In *WikiSym*, 3.

Wikimedia. 2020. Wikimedia Downloads. URL https://dumps.wikimedia.org/. Online; accessed 14-January-2020.

Wu, S.; Rizoiu, M.-A.; and Xie, L. 2019. Estimating attention flow in online video networks. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): 1–25.

Yang, J.; and Leskovec, J. 2011. Patterns of temporal variation in online media. In *WSDM*, 177–186.

Yasseri, T.; and Bright, J. 2016. Wikipedia traffic data and electoral prediction: towards theoretically informed models. *EPJ Data Science* 5(1): 22.

Yasseri, T.; Sumi, R.; Rung, A.; Kornai, A.; and Kertész, J. 2012. Dynamics of conflicts in Wikipedia. *PloS one* 7(6): e38869.

Yoshida, M.; Arase, Y.; Tsunoda, T.; and Yamamoto, M. 2015. Wikipedia page view reflects web search trend. In *Web Sci*, 65: 1–2.

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *KDD*, 1513–1522.