Online Social Networks and Media 13 (2019) 100047

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

Collective attention patterns under controlled conditions

Marijn ten Thij[a,*], Andreas Kaltenbrunner[b], David Laniado[c], Yana Volkovich[d]

[a] Center for Social and Biomedical Complexity, Indiana University Bloomington, 919 E. 10th Street, Bloomington, IN 47408, United States
[b] Universitat Pompeu Fabra, Carrer de Tànger, 122-140, Barcelona 08018, Spain
[c] Eurecat – Technology Centre of Catalonia, Carrer de Bilbao, 72, Barcelona 08005, Spain
[d] Xandr, 28 West 23rd Street, 4th Floor, New York, NY 10010, United States

ARTICLE INFO

Article history:
Received 2 May 2019
Revised 26 July 2019
Accepted 29 July 2019
Available online 9 September 2019

Keywords:
Wikipedia
Activity dynamics
Popularity modeling

ABSTRACT

Increasing concerns over widespread manipulation have led to an examination of how the direct promotion of content affects user behavior. Several studies have hypothesized that promotions exert their effects through a combination of social influence and attention redirection that are presently not well understood. The particulars of the promotion mechanism itself, however, might play an important role. Here we focus on a particular type of promotion mechanism: one that promotes identical content for the entire community of users. Through a large-scale analysis of Wikipedia page-views we show that fundamentally different processes underpin the longitudinal dynamics of user attention and preferences on Wikipedia compared to other promotion systems. Whereas other platforms follow history dependent processes, we found that the effects of community-wide promotions can be modeled using a history independent process. These findings allow to effectively compare view dynamics across online services.

1. Introduction

Today, hundreds of millions of users share their everyday thoughts and activities on on-line platforms such as Facebook, Instagram, Reddit and Twitter. As a result, researchers now have access to larger quantities of information for the analysis of on-line human behavior than ever before. For instance, using Twitterdata researchers have predicted political election results [31], the future activity on Twitter [9,10], stock market prices [6], and box-office revenue of a film [3]. On-line data from social platforms can provide deeper insights into the underlying processes that govern on-line human behavior; for example, the effects of session length on the quality of user comments on Reddit [27], correlations between emotion and usage of Facebook [25], the effect of displayed emotion in a users timeline on their own mood [14] and the effects of on-line activity on off-line behavior [2]. A complicating factor in this respect, however, is that on-line behavior is increasingly driven and shaped by algorithms that aim to promote or favor certain content which has an impact on the popularity of content [24]. A better understanding of how promotion drives online activity and attention is therefore important. Here, we study how on-line promotion affects the popularity of an article of an on-line platform. We define the popularity of a promoted article as the number of views this article receives during the time it is promoted [33]. Since we want to mimic only the effects of promotion, we restrict our analysis to the time when an article is promoted.

1.1. Related work

As a means of classifying popularity dynamics, Lehmann et al. [16] observe four different classes of popularity peaks in Twitterdata. Based on Digg and YouTube data, a multiplicative model is proposed to capture the dynamics of the growth of content popularity on the platforms [28,41]. Furthermore, Panzarasa and Bonaventura [21] and Szabo and Huberman [28] find that the growth of popularity in Digg, Forum posts and YouTube displays an inflection point, after which the popularity growth slows down. Thus far, investigations into underlying processes of popularity dynamics have predominantly been performed for platforms that tailor the content they show to the preferences of the user (e.g., in [28,41]), or on general websites where potentially many pages with novel content compete for audience simultaneously [15]. For these platforms, several different stochastic processes have been suggested that fit usage patterns on on-line platforms. Kumar and Tomkins [15] find that the inter-arrival times to particular websites follow a Poisson process. Other studies have proposed extensions of this model that better fit these activity patterns, i.e., an in-homogeneous Poisson process [5,18], a priority system [4,20], the Self Feeding process [32] and Rest-Sleep-Comment [8]. Another stochastic process that is also used in this context, is the

* Corresponding author
E-mail address: mtenthij@indiana.edu (M. ten Thij).

https://doi.org/10.1016/j.osnem.2019.07.003
2468-6964/

Hawkes process [11], which models and predicts popularity on social media as a self-exciting point process. Crane and Sornette [7] investigate several classes of this model to fit the popularity behavior of YouTube videos. Moreover, Zhao et al. [43] and Mishra et al. [19] use the Hawkes process to model information cascades and retweet activity on Twitter. Rizoiu et al. [23] extend the Hawkes process to include external influences by combining data from Twitter and YouTube.

1.2. Contribution

In this paper, we study the effects of promotion for the first time on Wikipedia, the fifth-most visited website as of July 2019 [1] with more than 250 million daily visits. Unlike other platforms studied previously Wikipedia provides a controlled and standardized environment for this type of studies for the following reasons: it is a platform in which the effects of optimization through algorithms are absent, it is a website of general interest for internet users and its promoted content is not exposed to internal competition. In Wikipedia, a single promoted article per day (in some language editions for even longer regular time intervals) is chosen by the editors based on the quality of the content of a page. This predefined time of exposure, together with the fact that only one article is promoted per time interval, makes the Wikipedia promotion mechanism unique in its regularity compared to the promotion mechanisms of other large and popular platforms. To study the universality of Wikipedia's promotion mechanism, we use page-view data from multiple languages in our study. We derive a simple model for page view dynamics and show that, if the daily activity cycles of users are accounted for, the activity dynamics on Wikipedia can be explained by a Poisson process. We furthermore compare our findings with the results for Twitter [23,43] and YouTube [7] and discuss the difference in the obtained processes.

2. Materials and methods

2.1. Data set

Since we measure the popularity of a promoted article as the number of page-views this article receives during the time it is displayed on the 'Main page', we retrieve the number of hourly page-views from a public data set that is provided by *WikiMedia* [34]. This data set contains the page-views of every WikiMedia project that are stored in files that contain all page-view data for a given hour. The format of how these page-views are specified is given in Table 1. There are certain intervals of missing data in the aforementioned data set. Therefore, we exclude all possible pages for which we cannot retrieve the page-view statistics during the time they are displayed on the 'Main page'.

To study the universality of the promotion mechanism in Wikipedia, we use page-view data from multiple languages. We chose languages that have a clear database, which lists previously promoted articles including the day of promotion. Therefore, we selected the languages English, Spanish, Dutch and German, as a basis to model the effect of promotion on the page-view dynamics on Wikipedia.

Table 1
A few example entries of the page-view data (from the file `pagecounts-20150601-140000.gz`). The first column specifies the project (`en` specifies the English Wikipedia), the second column specifies the title of the page, the third column reflects the number of requests, and the fourth specifies the size of the content (in bytes) that is returned.

Language	Name of page	Number of page-views	Bytes transferred
`en`	`Main_Page`	1,179,987	22,078,289,830
`en`	`Silenus`	7	298,184
`en`	`Rick_Astley`	95	3,485,487

2.2. Promoted articles

By promoted content, we mean that these articles are placed under a specific headline on the 'Main page'. For instance, promoted articles on the English Wikipedia are placed under the headline "From today's featured article" on the "Main page" [36] of the English Wikipedia. For the Spanish, Dutch, and German Wikipedias promoted articles are placed under the headlines "Artículo destacado", "Uitgelicht", and "Artikel des Tages", respectively, which are displayed on their respective 'Main pages' called "Wikipedia:Portada" [39], "Hoofdpagina" [40], and "Wikipedia:Hauptseite" [35].

Although some aspects of the promotion mechanism are different across these languages, they all use the following mechanism: after a predefined display time, the promoted content is replaced by a new article at a predefined time of the day (measured in UTC). The specific values of these characteristics of the promotion procedure for each language and the time intervals for which we retrieved the page-view data for our analysis are presented in Table 2. As an addition to showing the promoted article, both the English and the Spanish Wikipedia display the previous three promoted articles as links below the text of the article that is currently being promoted. See Fig. 1 for an example of the promoted article for the English Wikipedia.

Every Wikipedia user can nominate articles with certain characteristics (*featured articles*) to the pool of possible future promoted articles [38], although preference is given to the articles' primary editors. Only articles that meet specific criteria can be promoted [37]. A detailed analysis of the changes articles undergo before they can become featured, has been the topic of other studies [17,22].

For each language, we select all the articles promoted in the considered time-span of the corresponding Wikipedia. In the following, this set of articles is denoted by $\mathcal{S}$. The total number of articles included in $\mathcal{S}$ per language are indicated in Table 2. Recall that we restrict our modeling to the time that the promoted article is displayed on the 'Main page', thus we only study the popularity, defined as the number of page-views per hour, during these hours. Let v_t denote the number of page-views received by any article from $\mathcal{S}$ at hour t when it is displayed on the 'Main page'. Fig. 2 depicts the average number of page-views, which is denoted by $\mu(v_t)$, a promoted article in the data set attracts during its hours of exposure.

2.3. Problem: activity and circadian rhythms

In our study, we analyze the page-view activity on Wikipedia for four different languages. Fig. 2 clearly shows that the popularity patterns for the average promoted articles exhibit circadian rhythms, i.e. they display a clear day-night activity cycle.

We are interested in modeling the underlying activity pattern that is caused by promotion, thus we have to remove the circadian rhythm form the data. Since the promoted articles that we are studying are presented on the 'Main page', we analyze the circadian rhythms for both the total page-views in each language edition as a whole and the number of page-views for the 'Main page' of each language edition. Table 3 indicates the average hourly number of page-views and Fig. 3 shows the daily cycles per language as a percentage of the total activity in UTC, based on the page-views we retrieved. To facilitate the comparison, the English cycle is displayed in UTC-5 and the Spanish is displayed in UTC-6, whereas the Dutch and German cycles are shown in UTC. Fig. 3 indicates that the page-view activity follows a circadian rhythm, as

Table 2
Overall statistics of the promoted articles we obtained. The time interval for which these articles are retrieved, the number of articles received, the number of days a promoted article is displayed on the 'Main page', the number of links to previously displayed articles and the time of replacement of the article.

Language	First day	Last day	Articles	Display time	# Links	Changed at
English	2013-01-01	2015-06-30	818	One day	Three	00:00h UTC
Dutch	2015-01-01	2015-11-30	317	One day	–	23:00h UTC
German	2015-01-01	2015-11-30	323	One day	–	23:00h UTC
Spanish	2015-01-01	2015-11-30	64	Four days	Three	01:00h UTC

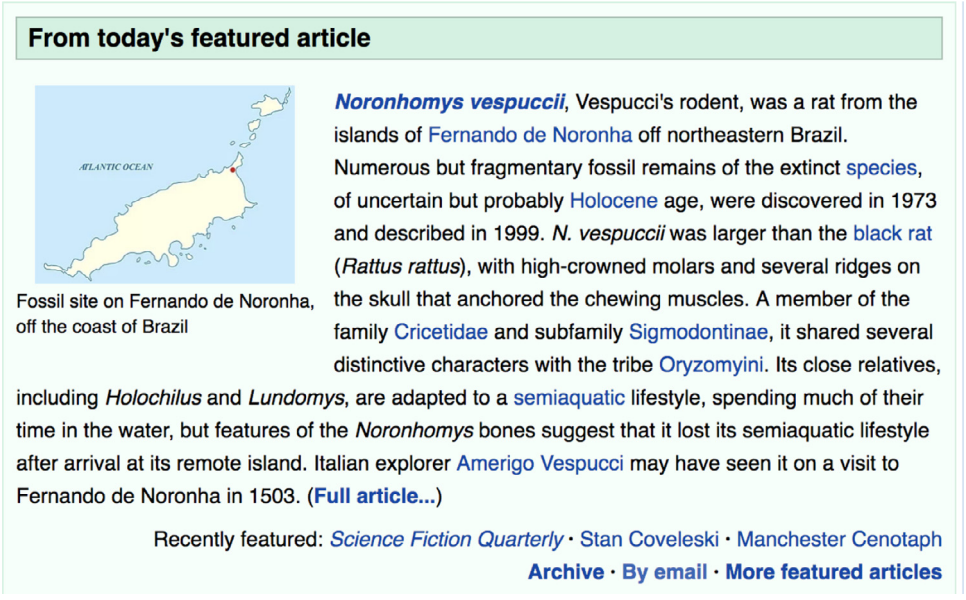


Fig. 1. Example of the promotion section "From today's featured article" of the English Wikipedia on the 'Main Page' taken on July 15, 2019. Links on the bottom right lead to the three previously promoted articles.

Table 3
Average hourly page-view statistics of both the 'Main page' and all pages together for the languages we studied.

Language	'Main page'-views	Total page-views
English	527,583	10,221,711
Dutch	3307	177,916
German	33,956	1,035,199
Spanish	15,277	1,036,259

was found earlier for Wikipedia editing behavior [42] and for many other platforms, e.g., Digg [28], Instagram [26], Slashdot [13], and Twitter [30]. Moreover, the amplitude of the English cycle is small compared to the other languages. This can be explained due to the global nature of the use of the English language.

Since the 'Main page' links to the promoted article, an obvious choice would be to use the 'Main page'-view cycles in our analysis as a proxy for the circadian rhythms. However, due to inconsistencies in the 'Main page'-views we chose the total page-views as a proxy for the circadian rhythm. Further details on the choice of proxy in our analysis are given in Appendix A.

2.4. Solution: time redistribution

For the removal of these circadian rhythms, we rescale the page-view data by measuring it in the number of page-views rather than in minutes. This approach to remove the circadian rhythm has been previously used in [28] for the analysis of the popularity of Digg stories and in [12] for the analysis of phone-call activity patterns. We refer to this rescaled time as *redistributed time* in the rest of this paper. This procedure of removing the circadian rhythms is discussed in more detail in Appendix B.

Using the redistributed time we eliminate the effect of the circadian rhythm of Wikipedia page-view behavior on the effect of promotion, exposing the underlying effect that is caused by the promotion. Fig. 4 depicts the number of page-views for the average promoted article together with the redistributed time version of the same page-view data.

It is clear from Fig. 4 that the redistribution has diminished the circadian rhythm of the data, however it has not removed it completely. Therefore, we extend the redistribution method by removing a fraction $\eta \in [0, 1]$ of the minimum activity in the circadian rhythm. See Fig. B.2 for a visualization of the effects of this removal on the redistribution method. We formally define the extended redistribution process as follows. Let $h(t)$ denote the average number of total page-views for a given hour $t = 1, \ldots, 24$. We define the redistribution parameter T^*_η as

$$T^*_\eta = \sum_{t=1}^{24} \left[h(t) - \eta \cdot \min_t h(t) \right], \tag{1}$$

and we use T^*_η to calculate the values of v_{t^*}, the number of page-views of an average promoted page at time t^*, the new redistributed time defined by T^*_η. A new hour t^* is therefore the time interval during which $h(t) - \eta \cdot \min h(t)$ accumulates from $\frac{t^*-1}{24} T^*_\eta$ to $\frac{t^*}{24} T^*_\eta$ page-views.

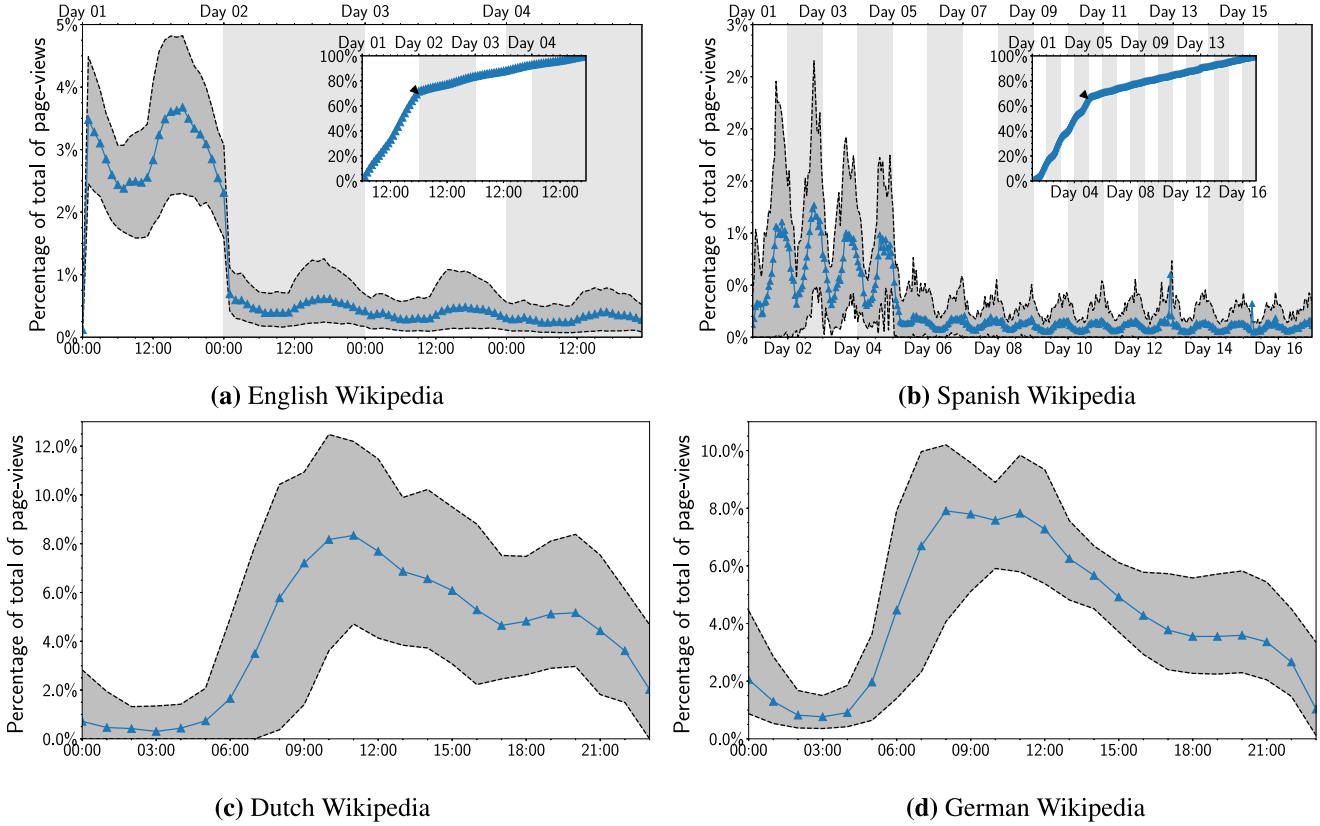


Fig. 2. The dynamics of promoted articles' page-views per language, normalized using the total page-views per article. The blue line represents the average of all articles ($\mu(v_t)$) and the black dashed lines indicate the 5- and 95-percentiles of the hourly page-view data. (a) and (b) display the cumulative page-views during the promotion. The arrows indicate the inflection point [21,28] of the growth of popularity.

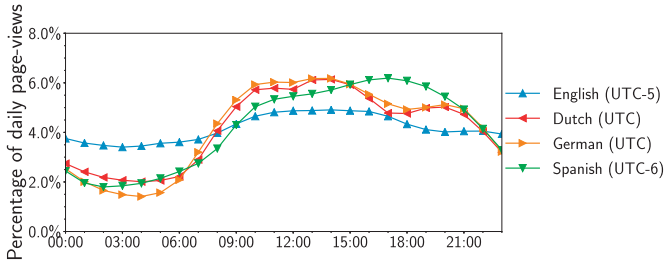


Fig. 3. The average daily percentage page-view activity cycles of the English (orange line, in UTC-5), Spanish (green line, in UTC-6), Dutch (red line, in UTC) and German (blue line, in UTC) Wikipedia's measured in terms of the total number of page-views.

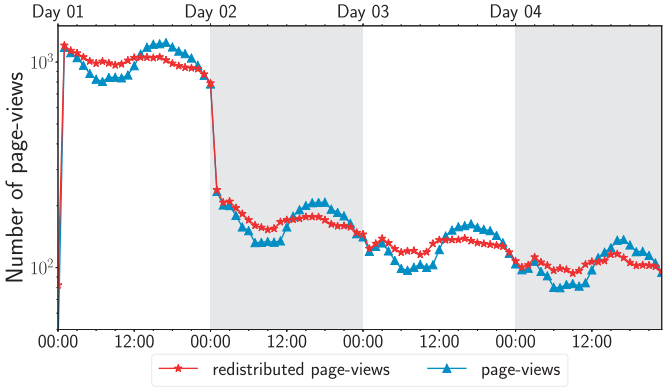


Fig. 4. Example of the "detrending" on the English Wikipedia: the redistributed page-views (v_{t^*}, red line) and the number of page-views for an average promoted article (v_t, blue line).

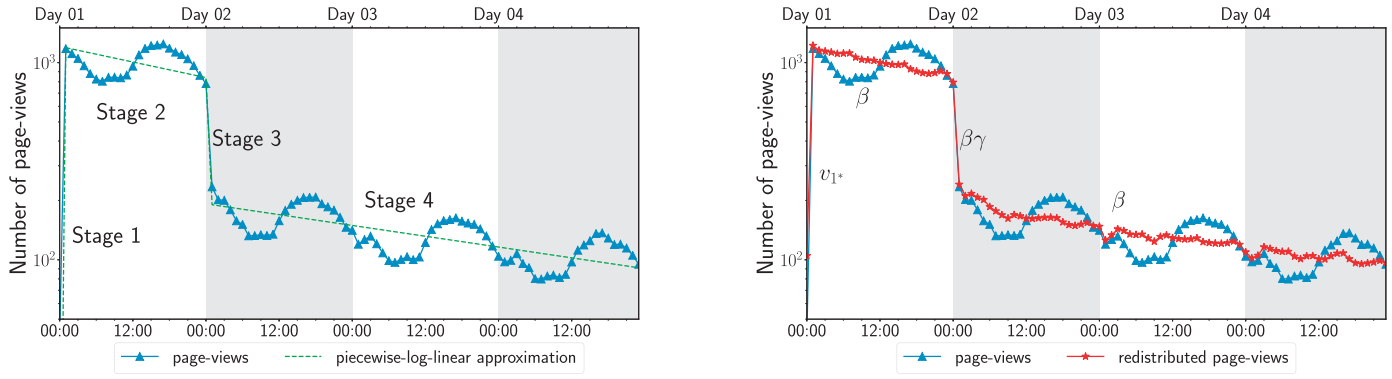
3. Results

3.1. Stage representation of page-view behavior

Fig. 4 displays the average page-view pattern of the English Wikipedia and its detrended version that is calculated using the redistribution method. Based on this detrended version, we divide the promotion pattern into four stages. The first stage is the increase in page-views during the first hour of the article's promotion on the 'Main page'. The second stage contains the remaining time of the promotion on the 'Main page'. The third stage is characterized by the sharp decline in page-views that occurs when the article gets replaced by a new one. Finally, the last stage contains the page-view dynamics during the days of being promoted only as a "Recently featured" link below the main promoted article of that day. Note that the effects of the first and third stage are instantaneous. Using this stage representation, we construct a piecewise-log-linear approximation of the number of page-views for the English. This approximation is referred to as $g(t)$ and it is depicted in Fig. 5(a) by a green dash-dotted line. Similarly, a piecewise-log-linear approximation can also be constructed for the Spanish Wikipedia. Since the Dutch and German Wikipedia's do not display links to previous promoted articles, these languages only display the first two stages during their time on the 'Main page'.

3.2. Using the extended redistribution method

To remove the circadian rhythm as much as possible, we use our previously described extended redistribution method, for



(a) Stage approximation $g(t)$ **(b)** Time redistribution and model parameters

Fig. 5. Visualization of the model definition using the average promoted article of the English Wikipedia. Displayed are the redistributed page-views (v_{t^*}, red line), the number of page-views (v_t, blue line) and the stage representation ($g(t)$, dashed green line). The black annotations on the left indicate the four stages of attention during an article's promotion and the black annotations on the right refer to the parameters that model these four stages.

Table 4
Deleted fractions of the lower bound of page-views per language.

Language	English	Spanish	Dutch	German
Value of η	0.4199	0.0000	1.0000	0.8016

which we set

$$\eta = \arg\min \left[\sum_{t^*} \left(\log\left(v_{t^*}\right) - g(t^*) \right)^2 \right]. \tag{2}$$

In doing so, we are removing constant views from the cycle, such that the squared difference between a linear fit of the page-view behavior and the redistributed page-views is minimized. Effectively, by choosing this particular value for η, we fit the circadian rhythm of the total page-views to the circadian rhythm that is present in the average page-views per promoted article. Thus, a low value of η indicates that the amplitude of the circadian rhythm of the cycle is almost equal to the data, whereas a large value of η indicates a much larger amplitude in the activity data than the amplitude of the cycle.

Fig. 5 (b) displays the effect of this optimized redistribution process to the popularity pattern of the English Wikipedia. The optimal values for η per language, indicated in Table 4, give a good insight into the circadian rhythms in our data. For instance, the fact that $\eta = 0$ for the Spanish Wikipedia indicates that the amplitude of the data is equal to or less than the cycle we use. For the Dutch Wikipedia, the difference between the amplitudes of both circadian rhythms is so large, that we need to compensate for it by setting $\eta = 1$. For the English and the German cycles, we find exact matches for η that correct the difference in amplitude of the cycles.

3.3. Page-view behavior model

Using the detrended average page-view behavior (depicted as a red line in Fig. 5(b)), we propose a model which describes the page-view dynamics of a promoted article during its promotion time. This model is defined by a single parameter: a constant interest-decrease factor for all days of the promotion. If the Wikipedia also displays previously promoted articles, a second parameter is used to capture the effects of the decline of the popularity after the exposure on the 'Main page'. The number of page-views a selected article receives during the first hour of the promotion (v_1) is used as the only input value of the model.

The definition of the model is inspired by the shape of the page-view behavior pattern, or more exactly by the normalized number of page-views per discrete time unit $w_{t^*} = v_{t^*}/v_{1^*}$ in rescaled time (with the circadian rhythm removed). We use v_t or w_t when referring to observed data and $\hat{v}_t$ or $\hat{w}_t$ for the model curves. Recall that the sub-index t stands for the real time and t^* for the redistributed time. Based on the log-linear fit of w_{t^*} and using $w_{1^*} = 1$ we define the model as:

$$\begin{aligned} \hat{w}_{1^*} &= 1 \\ \hat{w}_{t^*} &= f_{t^*} \cdot \hat{w}_{t^*-1} \end{aligned} \tag{3}$$

Based on the four stages of attention to a promoted article in Wikipedia, we need two options for f_{t^*}: $f_{t^*} = \beta$ to model the decrease of the number of page-views in a typical hour of the promotion time and $f_{t^*} = \beta\gamma$ as an expression for the sharp decline in the number of page-views after the promoted article gets moved to "Recently featured" position. Thus, we model the shape of the article popularity by stage: the first stage of the promoted article is characterized by $\hat{w}_{1^*}$, the second by the interest-decrease factor β, the third by $\beta\gamma$, and the fourth again by the same factor β as in the second phase. Thus, our model assigns parameters to the four stages of user attention during the promotion of an article. Fig. 5(b) shows the average article popularity both in regular and redistributed time scales, with the addition of the model parameters that capture each stage in the promotion of an article.

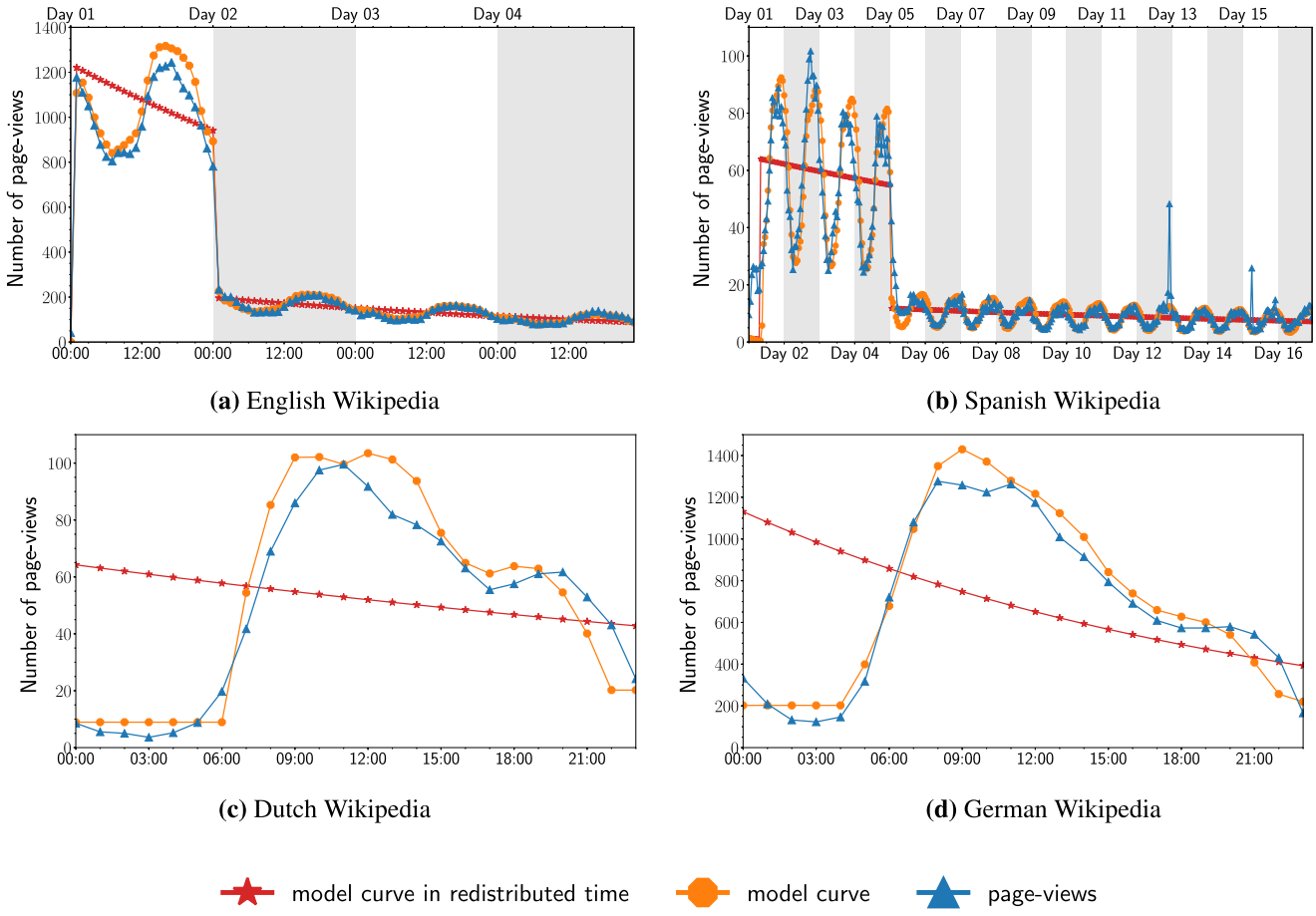
This modeling leads to the following equations for $\hat{w}_{t^*}$ for the considered languages.

$$\begin{aligned} \text{English: } \hat{w}_{t^*} &= \begin{cases} \beta^{t^*-1} & \text{for} \quad 1 \le t^* \le 24 \quad \text{(day 1)}, \\ \gamma\beta^{t^*-1} & \text{for} \quad 25 \le t^* \le 95 \quad \text{(days 2-4)}. \end{cases} \\ \text{Spanish: } \hat{w}_{t^*} &= \begin{cases} \beta^{t^*-1} & \text{for} \quad 1 \le t^* \le 96 \quad \text{(days 1-4)}, \\ \gamma\beta^{t^*-1} & \text{for} \quad 97 \le t^* \le 384 \quad \text{(days 5-16)}. \end{cases} \\ \text{Dutch: } \hat{w}_{t^*} &= \beta^{t^*-1} \quad \text{for} \quad 1 \le t^* \le 24 \quad \text{(day 1)}. \\ \text{German: } \hat{w}_{t^*} &= \beta^{t^*-1} \quad \text{for} \quad 1 \le t^* \le 24 \quad \text{(day 1)}. \end{aligned} \tag{4}$$

Next, we apply the time-redistribution to find $\hat{w}_t$, based on $\hat{w}_{t^*}$ as is discussed in Appendix B. With this, we calculate the number of page-views of the promoted article during the tth hour as

$$\hat{v}_t = \frac{v_1}{\hat{w}_1} \cdot \hat{w}_t, \tag{5}$$

where v_1 is the number of page-views of the promoted article after the first hour of exposure in the original time scale.



(a) English Wikipedia

(b) Spanish Wikipedia

(c) Dutch Wikipedia

(d) German Wikipedia

Fig. 6. The average page-views per promoted article per language and their corresponding model curves. The blue line displays the original page-view data, $\mu(v_t)$, the red line displays the model in redistributed time, $\mu(\hat{v}_{t^*})$, and the orange line represents the model in normal time, $\mu(\hat{v}_t)$.

Table 5
The hour used as a start time for the model in the fit of the page-view behavior data.

	English	Spanish	Dutch	German
Start	1	8	0	0

Finally, we have to account for the fact that the overall circadian rhythms of the considered languages are centered around different times, as is demonstrated in Fig. 3. To account for this, we added the notion of a start time in our fitting process for the page-view dynamics. This start parameters captures the offset from UTC in the page-view behavior process. To determine the appropriate start time for a particular language, we ran model fits using multiple start times and chose the start time that minimized the squared difference between the actual and projected page-views as our final choice for the start parameter. Table 5 indicates the start times used in the results that are displayed in the article. Before these times, we set the page-views of our model to 0, as can be seen in Fig. 6(a) and (b).

3.4. Parameter estimation

Recall that $\mu(v_t^*)$ denotes the hourly page-views that an average promoted article receives. We estimate the model parameters β and γ by calculating

$$\{\beta, \gamma\} = \arg\min \left(\sum_{t^*} \left(\log\left[\hat{v}_{t^*}(\beta, \gamma)\right] - \log\left[\mu(v_{t^*})\right] \right)^2 \right). \quad (6)$$

Table 6
95% confidence intervals for all parameters after executing 10,000 estimates, which are based on 100 sampled articles from $\mathcal{S}$ for each estimate and the parameter estimate using Eq. (6).

Language	Parameter	CI	Estimate
English	β	[0.9868,0.9910]	0.9889
	γ	[0.1802,0.2373]	0.2079
Spanish	β	[0.9976,0.9987]	0.9982
	γ	[0.1719,0.2392]	0.2139
Dutch	β	[0.9690,0.9962]	0.9825
German	β	[0.9517,0.9584]	0.9550

Thus, we minimize the squared difference between the actual and projected page-views for the average promoted article. Using Eq. (6), we obtain parameter estimates for all four languages, which are presented in Table 6.

We test the performance of our parameter estimation by choosing 100 articles at random from $\mathcal{S}$ and applying the same parameter estimation using Eq. (6) on the average of these 100 articles. After we repeat this procedure 10,000 times, we use the results to determine 95% confidence intervals (CI) for the parameters β and γ, which are indicated in Table 6. Since all estimated parameter values are inside these confidence intervals, our estimation method of the parameters is consistent.

Using the values obtained by using Eq. (6), we construct the model curves for the number of page-views for every article. Fig. 6 shows the average page-views per language, including the model curves for the redistributed time ($\hat{v}_{t^*}$) and normal time ($\hat{v}_t$).

Table 7
Mean absolute error (MAE) and mean absolute percentual error (MAPE), displayed in parentheses, for the model based on all articles in $\mathcal{S}$ per language.

Language	Baseline	General: Eq. (6)	Per article: Eq. (7)
English	205.03 (194.24%)	99.71 (205.36%)	78.01 (145.96%)
Dutch	33.87 (433.69%)	31.84 (131.62%)	17.74 (99.31%)
German	384.97 (335.82%)	239.65 (215.48%)	197.76 (189.13%)
Spanish	15.83 (294.74%)	18.57 (213.34%)	13.40 (89.68%)

These curves indicate that our model is able to visually capture the behavior of the corresponding page-view curve (v_t).

Another way of using the model is to estimate the parameters for each single article directly based on the page-views of that particular article. Using this approach, we minimize the squared difference between the actual and projected page-views for the promoted article that we analyze to obtain β and γ, i.e., we calculate

$$\{\beta, \gamma\} = \arg\min \left(\sum_{t^*} \left(\log(\hat{v}_{t^*}(\beta, \gamma)) - \log(v_{t^*}) \right)^2 \right). \tag{7}$$

Examples of the model fit using this approach are shown in Appendix C for a random selection of articles.

3.5. Model evaluation

In the previous section, we addressed two ways in which the model can be used. Here, we will compare the performance of both approaches with a baseline. The comparison is performed by using both the mean absolute error (MAE) and the mean absolute percentual error (MAPE) and the results are displayed in Table 7. The first column displays the errors for a baseline where we use the average as an estimate. The second column displays the errors for the model fits that are obtained estimating β and γ using Eq. (6), which we will refer to as the general parameter estimate. The third column shows the fit when the parameter estimates are obtained for each article individually, using Eq. (7). From Table 7, we see that the general estimation performs better than the baseline. Moreover, the fit improves when the parameters β and γ are estimated per article. However, the precise improvement of the per article approach varies greatly per language considered. To further investigate these differences, we analyzed the distribution of the MAE and MAPE for all languages.

The MAE and the distribution of the absolute errors for all articles considered from the English Wikipedia are shown in Fig. 7. First, Fig. 7(a) displays the MAE for both approaches and the baseline. Here, we clearly see that the general parameter estimates provides a better fit in the first day of the promotion than the per article estimates and both approaches outperform the baseline. However, the per article estimate has lower average errors during the days in which the article is displayed as a link. When we compare the mean error from Fig. 7(a) with the median error, displayed in Fig. 7(c) and (e) we find large differences between the two, which indicates that the error distributions are very skewed. The MAPE, displayed in Fig. 7(b), (d), and (f), shows similar results as the MAE. Looking at the median of these distributions, we see that the error is at varies around 20% throughout the exposure duration of the articles for the approach that uses Eq. (7). The approach that uses Eq. (6) has a slight increase in the median error after the first day, but it stays between 30% and 50% per hour, averaging at 40%. The median normalized error for the approach that uses Eq. (7) ranges between 50% and 100% throughout the exposure duration of the article for the other languages we considered. Again, using Eq. (6) increases the errors as was the case for the English Wikipedia.

The estimation approach that uses Eq. (7) clearly outperforms the general estimation using Eq. (6). Given the volatile nature of the individual page-view time series for the promoted articles, this is not surprising. We have not investigated the addition of different features, such as the article text or topic, to improve fit of the model. We did check whether the model performance improved if two different parameters are used for the second and fourth stage of the promotion. Since using separate parameters for these stages did not improve the results, we omit these results for brevity.

3.6. Model interpretation

There are two ways in which our model can be interpreted, by using two different perspectives on the activity dynamics. The first addresses the activity dynamics from the perspective of the promoted page. The second addresses the dynamics from the perspective of an individual visitor to the promoted page.

First, we consider the activity dynamics from the perspective of the promoted page. Thus far, several studies that use this same perspective have found the underlying process that govern the activity dynamics on different platforms to be a Hawkes process [11] (e.g., Twitter [19,43] and YouTube [7]). The arrival rate at time t of such a Hawkes process is of the following form;

$$\varphi(t) + \sum_{i:\tau_i<t} \psi(t - \tau_i) \tag{8}$$

where τ_i denotes the arrival time of event i that occurred before t. The Hawkes process consists of two terms, one of which is a time-dependent arrival rate and the other that captures the effects of previous arrivals.

Now, let us consider the model equations as shown in Eq. (4). This equation only consists of a single term, which is time-dependent. Thus, when we compare Eq. (4) to Eq. (8), we find that the activity dynamics on Wikipedia are equal to a Hawkes process for which $\psi(t) = 0$ and therefore, it reduces to an in-homogeneous Poisson process.

The second interpretation of our results can be formulated using the perspective of a visitor to the promoted page. Following Kumar and Tomkins [15], the inter-arrival times to particular websites can be modeled as a homogeneous Poisson process. We assume that this process has rate $\lambda = -\ln(\beta)$. Consider T_{1^*}[1] the time of the first arrival of a visitor to the main page after a new article is promoted page expressed in redistributed time. By using the definition of the Poisson process, we find that

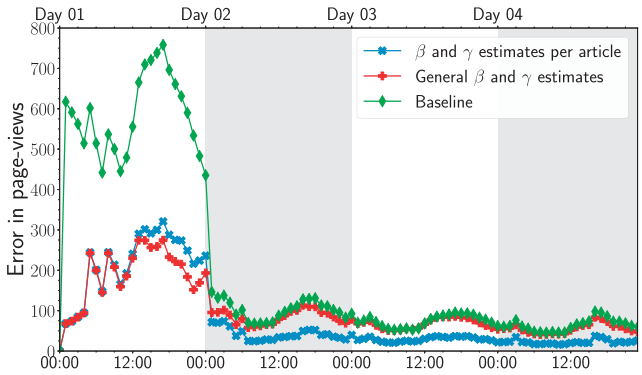
$$\mathbb{P}(T_{1^*} > x) = e^{-\kappa x} = \beta^x, \tag{9}$$

thus the probability that the first arrival occurs in the interval $(x-1, x]$ can be expressed as follows

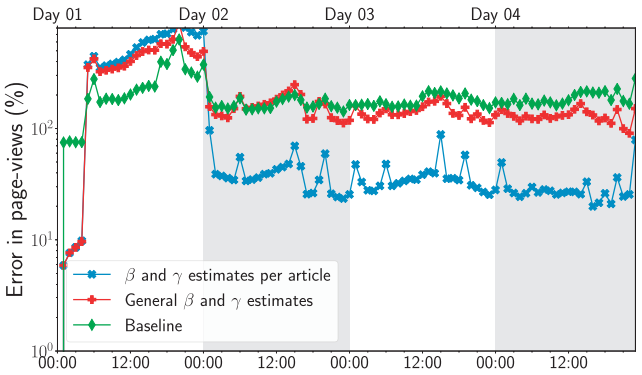
$$\mathbb{P}(x-1 < T_{1^*} \leq x) = \mathbb{P}(T_{1^*} > x-1) - \mathbb{P}(T_{1^*} > x) = \beta^{x-1}(1-\beta). \tag{10}$$

This expression corresponds to the likelihood that a user sees the link to a new promoted article for the first time during the xth hour of exposure on the *Main page*. It is, apart from the constant factor $(1-\beta)$, identical to the decrease factor β^{x-1} of our model. The number of users that actually click the link and visit the promoted article depends on its attractiveness and implies a constant re-scaling of Eq. (10). This is represented in our model by v_1. The other parameter of the model (γ) corresponds to the decrease in the likelihood of visiting an article after it has passed to the "Recently featured" section.

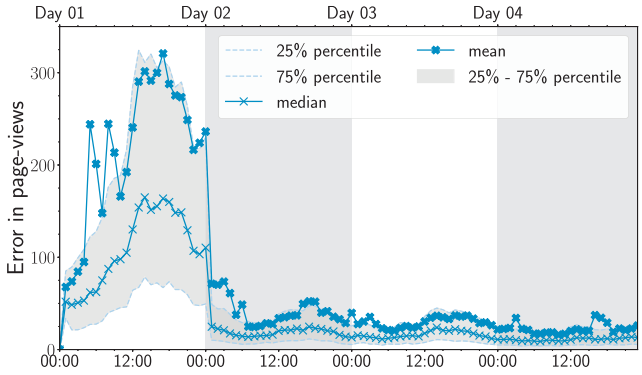
[1] Note that we use redistributed time here.



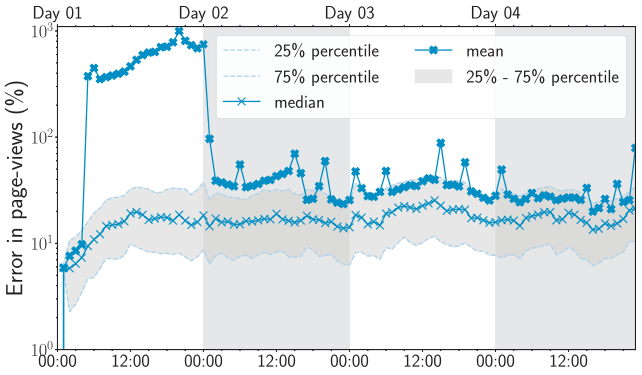
(a) MAE for both approaches and the baseline



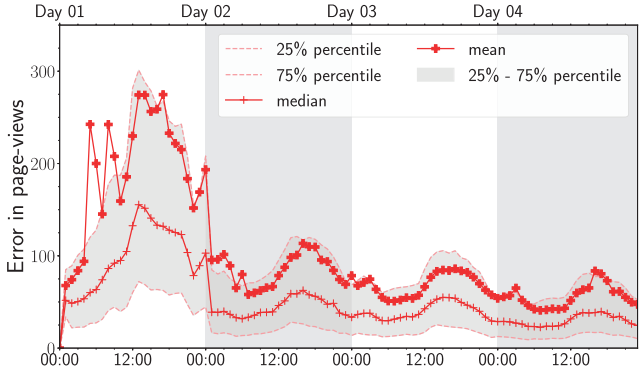
(b) MAPE for both approaches and the baseline



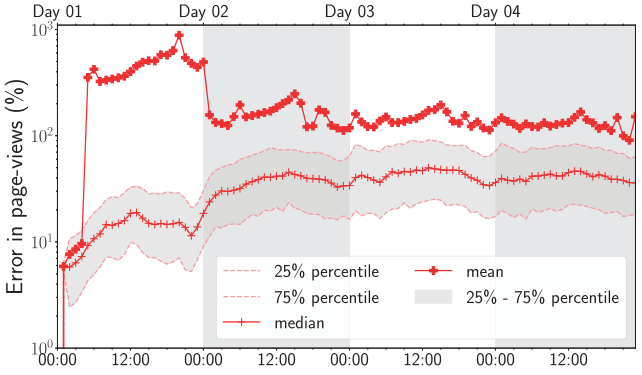
(c) AE distribution for per article estimates



(d) APE distribution for per article estimates



(e) AE distribution for general estimates



(f) APE distribution for general estimates

Fig. 7. MAE, MAPE and error distributions for the English Wikipedia.

4. Discussion and conclusion

In this paper, we formulated a model that captures the page-view pattern of a promoted article in Wikipedia during its exposure duration using two parameters (β and γ) and one input parameter (v_1). The model shows that the page-views of an article decrease exponentially with a constant rate if the circadian rhythm in Wikipedias activity is removed. The input parameter v_1, i.e. the number of views an article obtains during the first hour of promotion, can be seen as a measurement of the interest for the article on a global scale. The γ parameter captures the additional decline in popularity that is observed when an article is moved to one of the previously promoted pages. Therefore, this parameter is only included in the model for languages which have this previously promoted feature in their promotion scheme. The parameter β can be seen as a popularity-decay factor. Therefore, β should be a useful parameter to account for (and compare) the half-live of a piece of content on a given site.

Even though users do not necessarily have to land on the promoted pages through the 'Main page', we found that our model fits

best when we use the 'Main page'-views as a proxy to account for the circadian rhythm in the data. Moreover, we extended a previously used redistribution method [12,28] to optimize the removal of a circadian rhythm from the page-view data.

By comparing our model to results of similar analyses on other platforms, we find clear differences in how the dynamics on the platform can be modeled. Based on our model, we can represent the activity dynamics on Wikipedia by a Poisson process. This contrasts results of previous studies into the activity dynamics of other platforms (e.g., Twitter [19,43] and YouTube [7]), where the underlying process was shown to be a Hawkes process [11]. There are two major differences between the platform we investigate here and those other platforms. First, the content in Wikipedia is not personalized for each user whereas it is personalized on the other platforms. Second, there is no competition for attention between promoted articles in Wikipedia, whereas there is a continuous competition for attention between all the promoted items on the other platforms. Therefore, our results imply that the presence of these mechanisms triggers the process, governing the human behavior on the platform to become history dependent.

A possible extension of our analysis is to include a maximum likelihood estimation of the parameters of the assumed stochastic process to quantify the difference in performance between the aforementioned processes. To perform such an analysis, one needs the exact times at which the page-views occurs. The page-view data-set only provides hourly totals and therefore cannot be used in such an analysis. That is why we leave this analysis for future work.

The presented model and its interpretation provide additional insight into how human temporal behavior can be modeled using a stochastic process, e.g a Poisson process [18], a priority system [4,20] or a Hawkes process [7,19,43] on a given platform. Identifying the correct model for human temporal behavior on a specific platform allows to effectively compare and interpret view dynamics on other websites or parts of websites with similar update strategies, e.g. online newspapers which are updated on a daily basis, or a list of todays recommended items (mobile apps, products, etc.). The findings might also be useful to select the right forecasting method to predict the success rate of new online advertisements or sponsored content in general by factoring in the appropriate model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Oliver Keyes and the WikiMedia Foundation for providing the page-view data that is used to select the proxy for the circadian rhythm in the data and Johan Bollen for his insightful comments during the preparation of this manuscript.

Appendix A. Circadian rhythms

Since we are interested in modeling the underlying activity pattern that is caused by promotion, we have to remove the circadian rhythm form the data. While comparing the daily cycle in total page-views between languages, shown in Fig. 3, we see that the cycles are very similar. The only difference is that the number of page-views of the English Wikipedia is more stable over time, which is an artifact of the global use of the English language.

Intuitively, the 'Main page'-views can be a good proxy for the circadian rhythms in the page-views of promoted articles, as visitors reach these promoted pages through the 'Main page'. This choice of proxy is supported by the results of our earlier work [29]. For German, Dutch, and Spanish, we see very similar cycles for the 'Main page'-views as for the total page-views (see Fig. A.1). However, for the English Wikipedia, Fig. A.1(a) clearly indicates that the 'Main page'-cycle is different compared to the total page-views cycle with a second peak in activity during the morning hours.

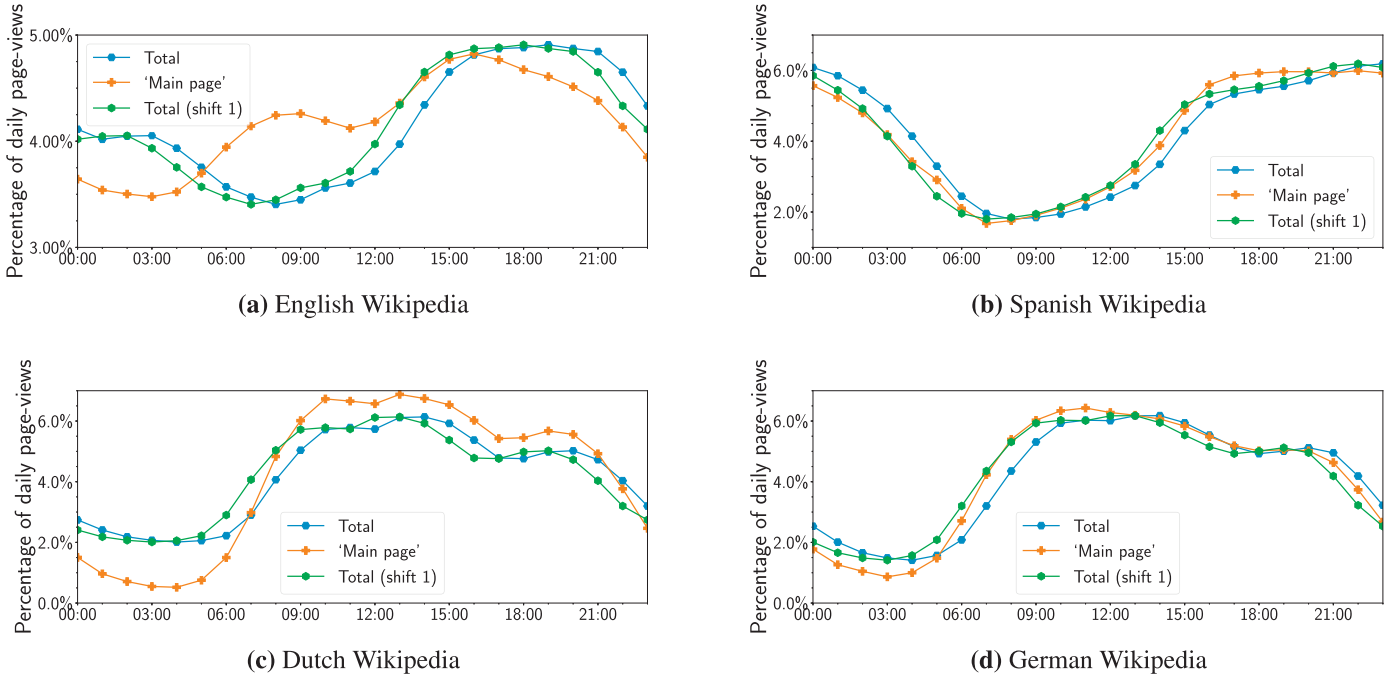
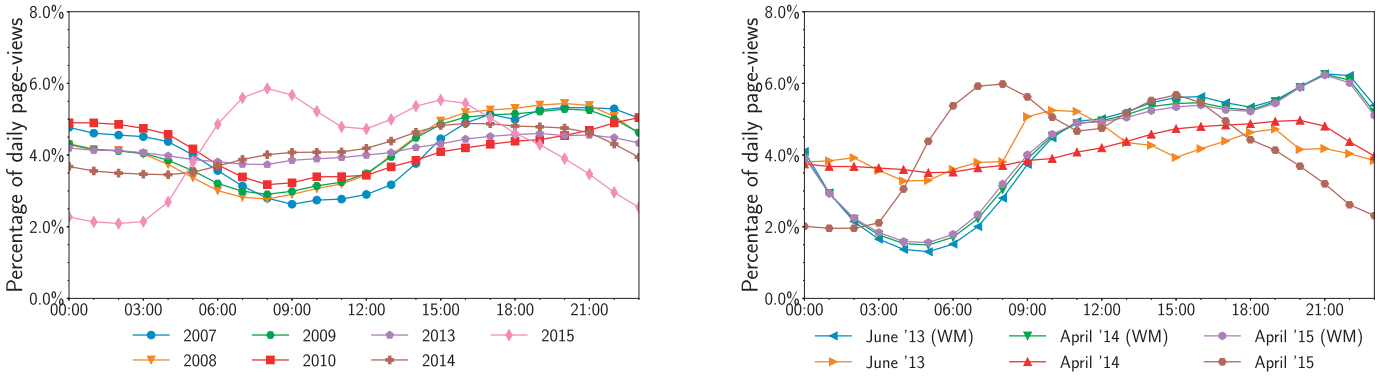


Fig. A.1. Daily cycles of the 'Main-page' page-views and the total page-views shown in UTC. Two versions of the total page-view cycle are displayed, i.e., the original cycle and a one hour shifted version.



(a) Yearly 'Main page'-views

(b) WikiMedia vs page-view dump

Fig. A.2. Normalized cycles of page-view behavior for 'Main page' of the English Wikipedia. (a) shows the year-to-year development of this page-view behavior and (b) shows the comparison between our retrieved page-view cycles and the internally retrieved page-view cycles by the WikiMedia Foundation, for three separate months.

In an effort to explain these differences in the 'Main page'-views and the total page-views for the English Wikipedia, we combine the page-view data from this study with the page-view data from our previous analysis into the effect of promotion in the English Wikipedia [29]. We analyze the year-to-year development of the page-view cycle for the English 'Main page', which is shown in Fig. A.2(a). Throughout the years, there is a clear increase of the number of page-views in the morning hours. Since this effect is only present in the circadian rhythms of the 'Main page'-views and not in the total page-views, we hypothesized that this difference might be caused by the method that is user to log the page-views in the page-view database [34]. We confirmed this hypothesis by comparing page-view cycles based on a month of data from our own data set with page-view cycles of the same month that were supplied by the WikiMedia Foundation (WMF). These WMF page-view cycles are acquired using an internal logging of page-views that included IP-address and time logs per view. The WMF cycles are calculated by inferring a location from these IP-addresses and subsequently adjusting the time of the page-view to UTC. Both our monthly cycles and the WMF monthly cycles are displayed in Fig. A.2(b). Since the change in the circadian rhythm is only present in our data, we chose to not use the circadian rhythm in the 'Main page'-views as a proxy in our analysis.

Finally, a comparison of the circadian rhythms in the 'Main page'-views and the total page-views, as depicted by the orange and blue lines in Fig. A.1, shows that these rhythms do not overlap. We found that the circadian rhythm of the 'Main page'-views does overlap with a one hour shifted total page-view cycle for the Spanish, Dutch and German Wikipedias. These shifted total page-view cycles are also displayed in Fig. A.1 by green lines. Based on this observation and the fact that the 'Main page'-views are the intuitive proxy, we use this shifted total page-view cycle in our analysis as a proxy for the circadian rhythm in the page-view data.

Appendix B. Circadian rhythms correction

In this section, we present a detailed description of the process used to correct for the circadian rhythms in our data set. For the removal of these circadian rhythms, we use a redistributed time where we measure by the number of page-views rather than by minutes. Thus, an hour in redistributed time is the time interval during which the considered circadian rhythm accumulates from a twenty-fourth of the total page-views. Effectively, this redistributed time is an interpolation of the cumulative page-views.

We formally define the redistribution process as it is used in [12,28]. Let $h(t)$ denote the average number of total page-views for a given hour $t = 1, \ldots, 24$. We define the redistribution parameter T^* as follows:

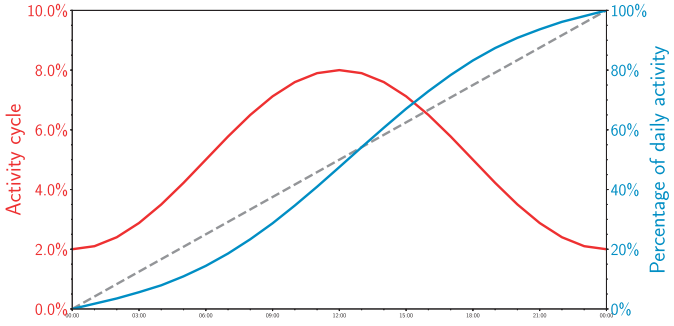
$$T^* = \sum_{t=1}^{24} h(t),$$

and use T^* to calculate the values of v_{t^*}, the number of page-views of an average promoted page at time t^*, the new redistributed time defined by T^*. A new hour t^* is therefore the time interval during which the analyzed cycle accumulates from $\frac{t^*-1}{24}T^*$ to $\frac{t^*}{24}T^*$ page-views.

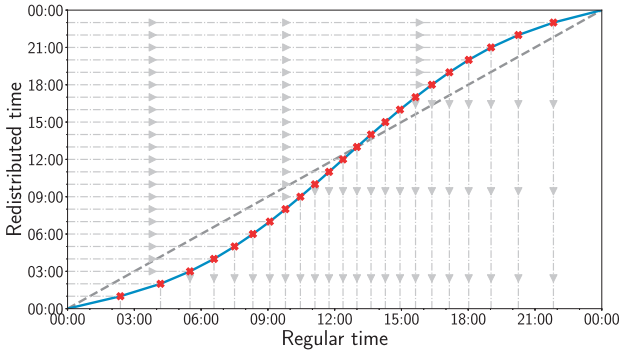
Fig. B.1 illustrates this approach using a sinusoidal activity cycle, which is depicted as a solid red line in Fig. B.1(a). The first step into redistribute the cycle is to find the times at which the cumulative page-views, blue line in Fig. B.1(a), reach a next twenty-fourth part of its total. Thus, we divide the rescaled time axis, shown on the y-axis in Fig. B.1(b), in 24 evenly sized intervals. These intervals are indicated by the horizontal gray lines in Fig. B.1(b). Next, we find the times at which the cumulative page-views reaches these values (shown as red markers in Fig. B.1(b)). Using these points, we find the corresponding times (in minutes) on the x-axis. This conversion is visualized through the dashed gray lines in Fig. B.1(b). The gray arrows on the gray lines are displayed to indicate the direction of the mapping. The times that correspond to these points on the x-axis will be the new 'hours'. Fig. B.1(c) shows the new hours both in the original time (measured in minutes) and in redistributed time (measured in page-views). Note that all areas under the page-view curves are of equal size.

In a similar manner, this process can also be used to convert redistributed time in regular time. By applying the same routine to the redistributed time scale t^*, we can reintroduce the circadian rhythm in the data. This is visualized by dividing the regular time, shown on the x-axis in Fig. B.1(d), in 24 evenly sized intervals. Equal to the visualization in Fig. B.1(b), we display the mapping with gray lines and red markers. By doing so, we have made sure that the area under the redistributed curve of these times corresponds to the area under the curve of the original hours in the circadian rhythm, as is illustrated in Fig. B.1(e).

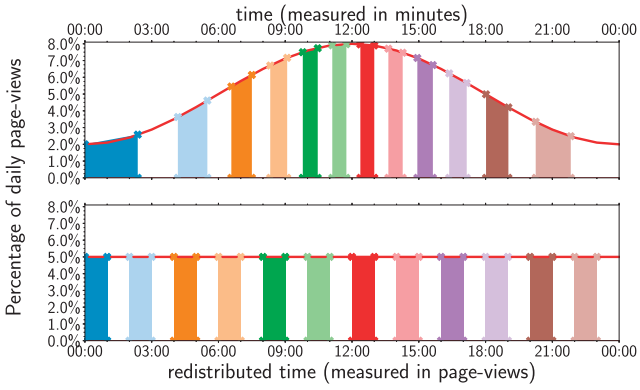
As an optimization of the redistribution process, we extend the previously described method by removing a fraction $\eta \in [0, 1]$ of the minimum activity in the circadian rhythm (i.e. $\min h(t)$), which is demonstrated in Fig. B.2(a). After rescaling the new cycle, we obtain a cycle with a larger amplitude as is shown in Fig. B.2(b). Thus, the removal of a fraction of the minimum activity



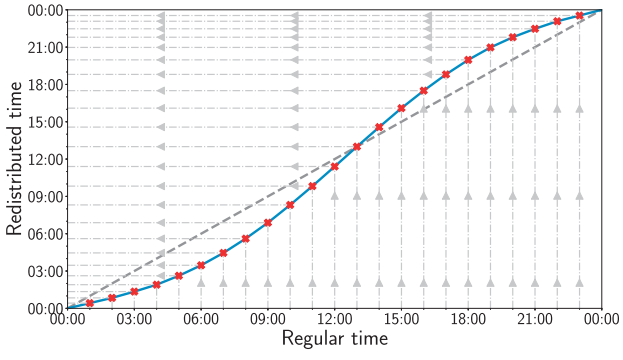
(a) The cycle used as an illustration of the redistribution. The red line represent the cycle in hourly form and the blue line represents its cumulative form.



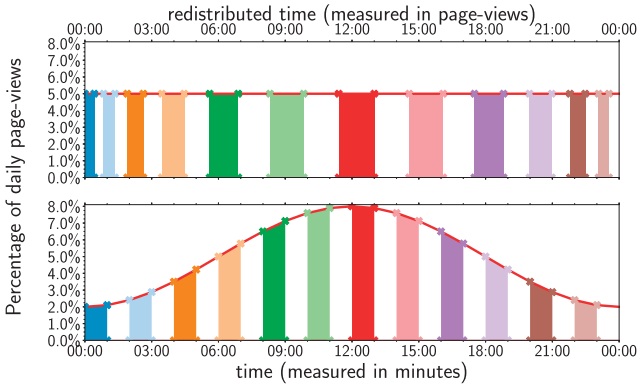
(b) Mapping time in minutes (t) to time in page-views (t^*). The gray arrows indicate the direction of the mapping.



(c) Visualization of the effects of time redistribution.

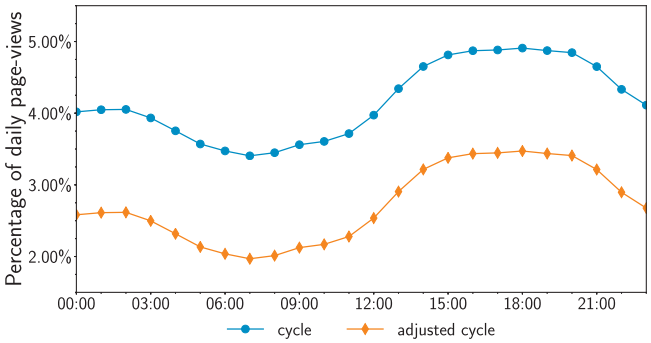


(d) Mapping time in page-views (t^*) to time in minutes (t). The gray arrows indicate the direction of the mapping.

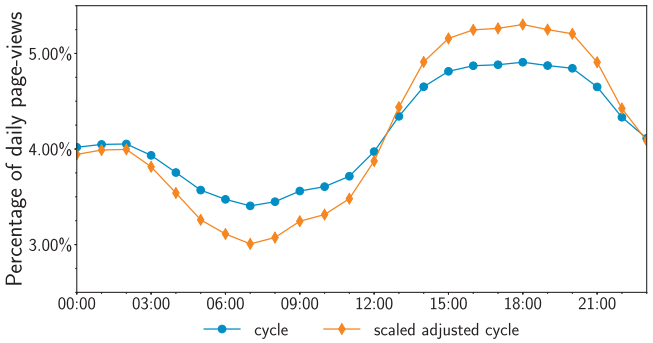


(e) Visualization of the effects of the reversion of the time redistribution.

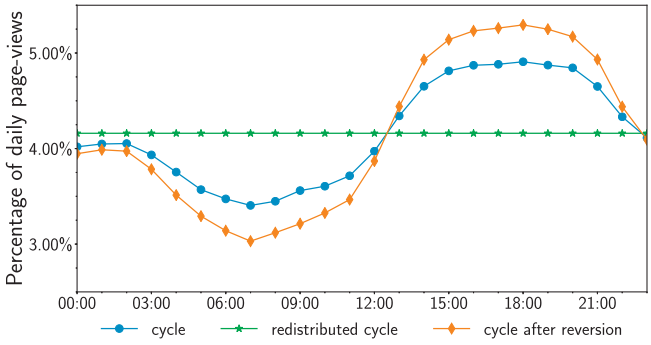
Fig. B.1. An example of the redistribution of time, based on a simple sinusoidal activity cycle.



(a) Removing a constant fraction



(b) Rescaling the adjusted cycle



(c) Effect of optimizing the redistribution

Fig. B.2. Impact of removing a constant fraction of the page-view cycle. (a) indicates the actual and adjusted version of the cycle and (b) shows that removing a constant part triggers a larger amplitude after normalizing the sum of the newly acquired cycle. (c) depicts the English total page-view cycle, which has been redistributed using the intervals calculated by using the full cycle, then the trend has been reapplied by using the intervals calculated by the optimized scheme that used the adjusted cycle.

and subsequent rescaling of the cycle emphasizes the circadian rhythm during the redistribution calculation. An example of how this rescaling influences the redistribution process, is shown in Fig. B.2(c). In this figure, the data is redistributed using the regular cycle and subsequently the circadian rhythm is reintroduced to the data using the adjusted cycle, which leads to a larger amplitude.

Appendix C. Examples

To illustrate the fit of our final model for individual articles, we show the obtained model curves for several articles from each language in Figs. C.1–C.4 in which we use Eq. (6) to estimate β and Eq. (7) to estimate γ.

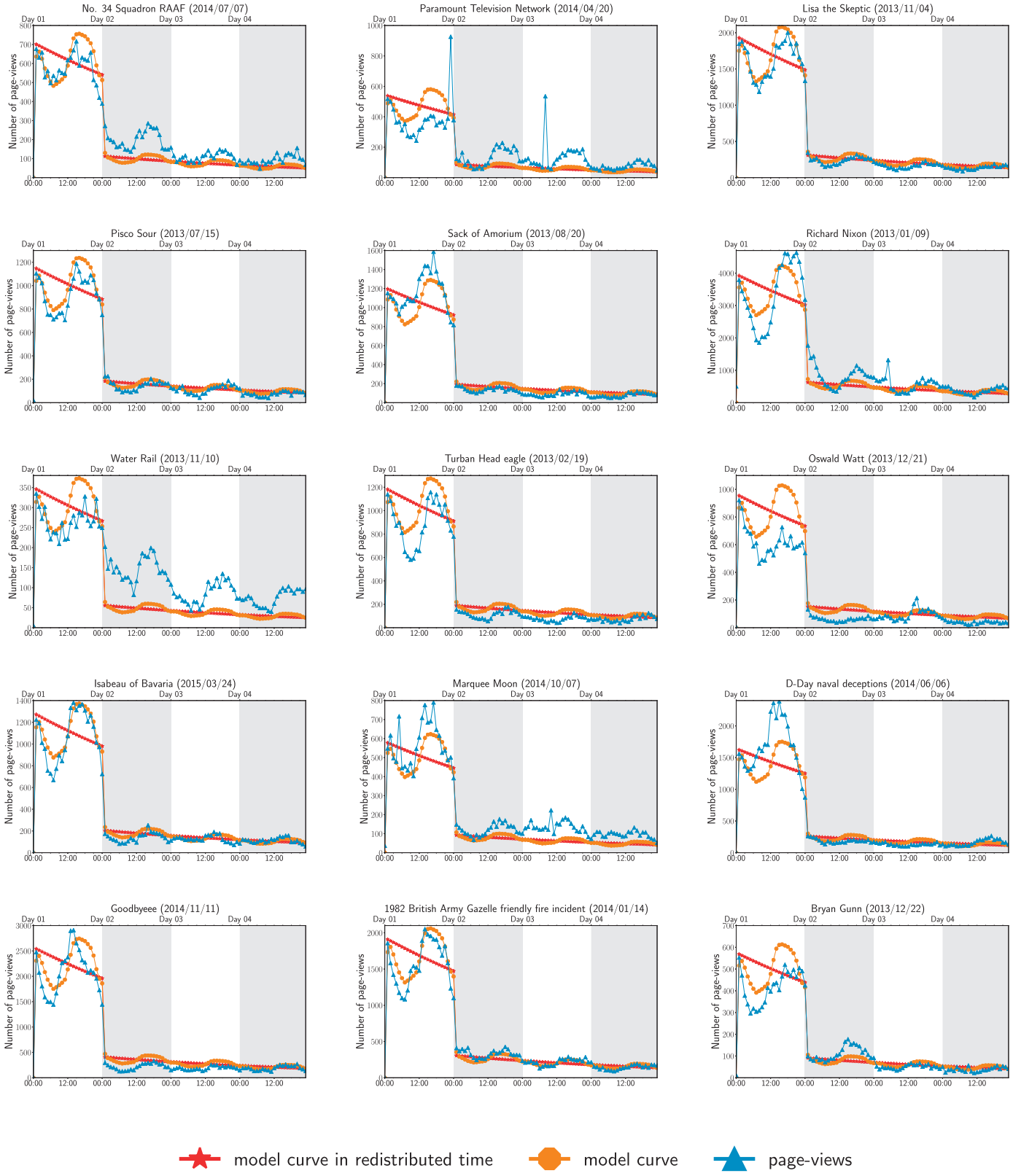


Fig. C.1. Example of the model fit on a sample of English promoted articles. For each example, the title and the promotion date are shown in the title of the figure. The blue line displays the original page-view data (v_t), the red line displays the model in redistributed time ($\hat{v}_{t^*}$) and the orange line represents the model in normal time ($\hat{v}_t$).

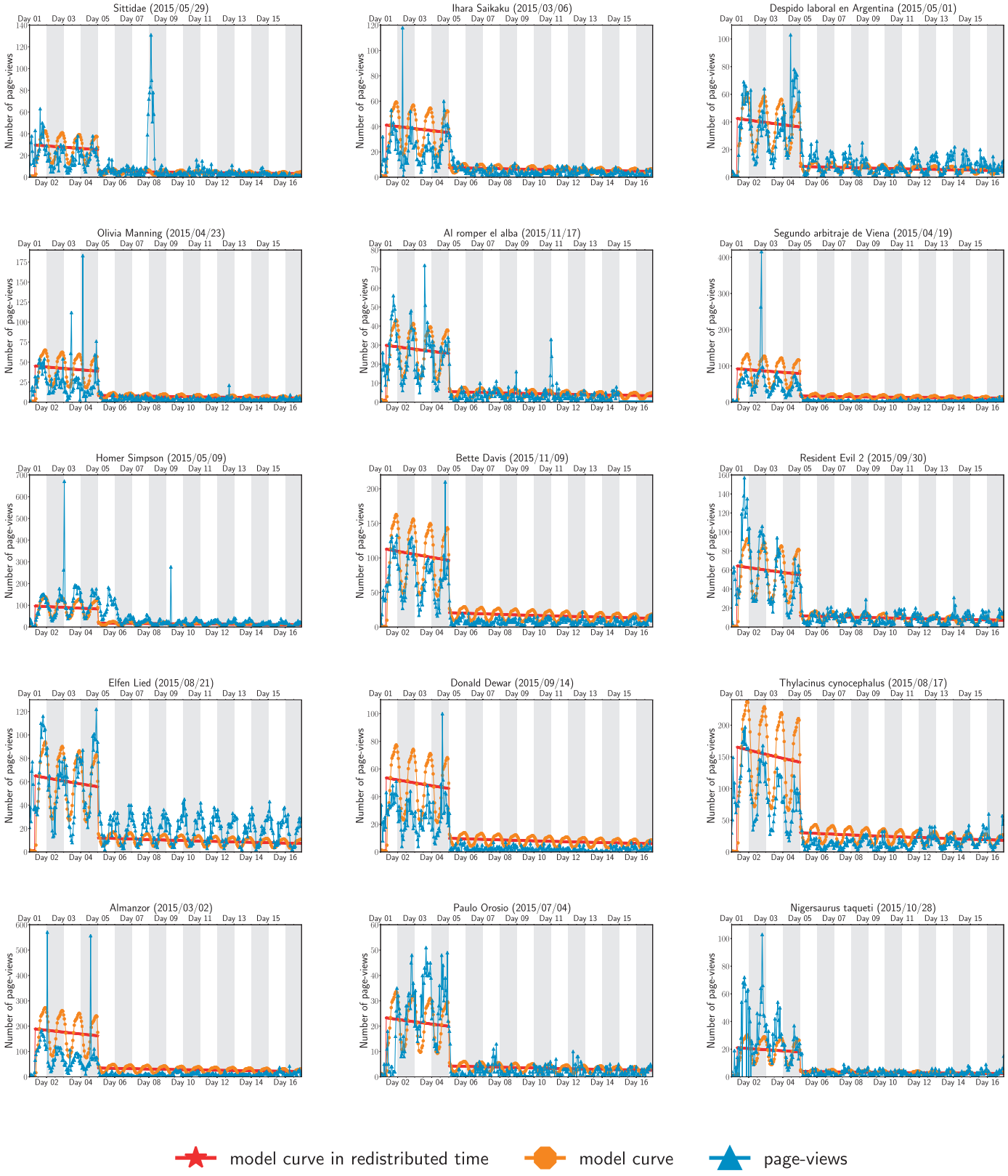


Fig. C.2. Example of the model fit on a sample of Spanish promoted articles. For each example, the title and the promotion date are shown in the title of the figure. The blue line displays the original page-view data (v_t), the red line displays the model in redistributed time ($\hat{v}_{t^*}$) and the orange line represents the model in normal time ($\hat{v}_t$).

Fig. C.3. Example of the model fit on a sample of Dutch promoted articles. For each example, the title and the promotion date are shown in the title of the figure. The blue line displays the original page-view data (v_t), the red line displays the model in redistributed time ($\hat{v}_{t^*}$) and the orange line represents the model in normal time ($\hat{v}_t$).

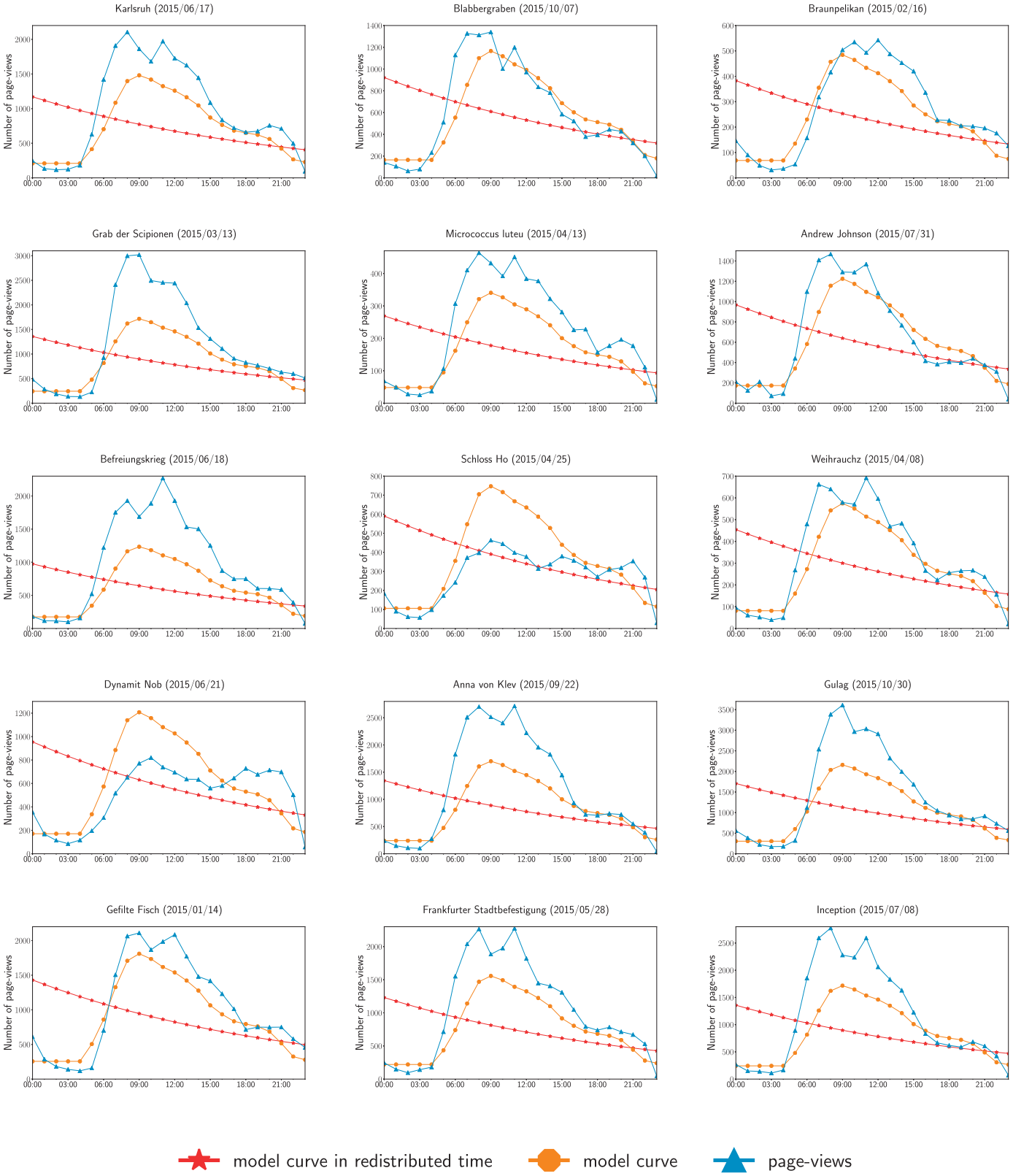


Fig. C.4. Example of the model fit on a sample of German promoted articles. For each example, the title and the promotion date are shown in the title of the figure. The blue line displays the original page-view data (v_t), the red line displays the model in redistributed time ($\hat{v}_{t^*}$) and the orange line represents the model in normal time ($\hat{v}_t$).

References

[1] Alexa, Wikipedia.org traffic, demographics and competitiors, 2019, Retrieved on 23 July 2019, www.alexa.com/siteinfo/wikipedia.org.

[2] T. Althoff, P. Jindal, J. Leskovec, Online actions with offline impact: how online social networks influence online and offline user behavior, in: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, in: WSDM, ACM, New York, NY, USA, 2017, pp. 537–546, doi:10.1145/3018661.3018672.

[3] S. Asur, B.A. Huberman, Predicting the future with social media, in: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT, 1, IEEE, 2010, pp. 492–499.

[4] A.-l. Barabási, The origin of bursts and heavy tails in human dynamics, Nature 435 (7039) (2005) 207–211.

[5] C. Bauckhage, C. Ojeda, R. Sifa, Circadian cycles and work under pressure: a stochastic process model for e-learning population dynamics, in: Data Science – Analytics and Applications, iDSC, Springer Fachmedien Wiesbaden, Wiesbaden, 2017, pp. 13–18.

[6] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.

[7] R. Crane, D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system, Proc. Natl. Acad. Sci. U.S.A. 105 (41) (2008) 15649–15653, doi:10.1073/pnas.0803685105.

[8] A. Ferraz Costa, Y. Yamaguchi, A. Juci Machado Traina, C. Traina Jr., C. Faloutsos, RSC: mining and modeling temporal activity in social media, in: Proceedings of the International Conference on Knowledge Discovery and Data Mining, KDD, ACM, New York, NY, USA, 2015, pp. 269–278, doi:10.1145/2783258.2783294.

[9] S. Gao, J. Ma, Z. Chen, Modeling and predicting retweeting dynamics on microblogging platforms, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15, ACM, New York, NY, USA, 2015, pp. 107–116, doi:10.1145/2684822.2685303.

[10] J. Harada, D. Darmon, M. Girvan, W. Rand, Forecasting high tide: predicting times of elevated activity in online social media, in: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM, ACM, New York, NY, USA, 2015, pp. 504–507, doi:10.1145/2808797.2809392.

[11] A.G. Hawkes, Spectra of some self-exciting and mutually exciting point processes, Biometrika 58 (1) (1971) 83–90.

[12] H.-H. Jo, M. Karsai, J. Kertész, K. Kaski, Circadian pattern and burstiness in mobile phone communication, New J. Phys. 14 (013055) (2012) 1–17.

[13] A. Kaltenbrunner, V. Gómez, V. López, Description and prediction of Slashdot activity, in: Proceedings of the 2007 Latin American Web Conference, LA-WEB 2007, IEEE, New York, NY, USA, 2007, pp. 57–66, doi:10.1109/LA-Web.2007.21.

[14] A.D.I. Kramer, J.E. Guillory, J.T. Hancock, Experimental evidence of massive-scale emotional contagion through social networks, Proc. Natl. Acad. Sci. U.S.A. (2014) 1–3, doi:10.1073/pnas.1320040111.

[15] R. Kumar, A. Tomkins, A characterization of online browsing behavior, in: Proceedings of the Nineteenth International Conference on World Wide Web, WWW, ACM, New York, NY, USA, 2010, pp. 561–570, doi:10.1145/1772690.1772748.

[16] J. Lehmann, B. Gonçalves, J.J. Ramasco, C. Cattuto, Dynamical classes of collective attention in twitter, in: Proceedings of the Twenty-First International Conference on World Wide Web, WWW, ACM, New York, NY, USA, 2012, pp. 251–260, doi:10.1145/2187836.2187871.

[17] X. Li, Z. Luo, K. Pang, T. Wang, A lifecycle analysis of the revision behavior of featured articles on wikipedia, in: Proceedings of the International Conference on Information Science and Cloud Computing Companion, ISCC-C, IEEE/ACM, New York, NY, USA, 2013, pp. 846–851, doi:10.1109/ISCC-C.2013.16.

[18] R.D. Malmgren, D.B. Stouffer, A.E. Motter, L.A.N. Amaral, A Poissonian explanation for heavy tails in e-mail communication, Proc. Natl. Acad. Sci. U.S.A. 105 (47) (2008) 18153–18158, doi:10.1073/pnas.0800332105.

[19] S. Mishra, M.-A. Rizoiu, L. Xie, Feature driven and point process approaches for popularity prediction, in: Proceedings of the Twenty-Fifth ACM International on Conference on Information and Knowledge Management, CIKM, ACM, New York, NY, USA, 2016, pp. 1069–1078, doi:10.1145/2983323.2983812.

[20] J.G. Oliveira, A.-L. Barabási, Human dynamics: Darwin and Einstein correspondence patterns, Nature 437 (7063) (2005) 1251.

[21] P. Panzarasa, M. Bonaventura, Emergence of long-range correlations and bursty activity patterns in online communication, Phys. Rev. E – Stat. Nonlinear Soft Matter Phys. 92 (062821) (2015) 1–13, doi:10.1103/PhysRevE.92.062821.

[22] R. Picot-Clémente, C. Bothorel, N. Jullien, Social interactions vs revisions, what is important for promotion in Wikipedia? in: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, ASONAM, IEEE/ACM, New York, NY, USA, 2015, pp. 888–893, doi:10.1145/2808797.2810063.

[23] M.-A. Rizoiu, L. Xie, S. Sanner, M. Cebrian, H. Yu, P. Van Hentenryck, Expecting to be hip: Hawkes intensity processes for social media popularity, in: Proceedings of the Twenty-Sixth International Conference on World Wide Web, WWW, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2017, pp. 735–744, doi:10.1145/3038912.3052650.

[24] M.J. Salganik, P.S. Dodds, D.J. Watts, Experimental study of inequality and unpredictability in an artificial cultural market, Science 311 (5762) (2006) 854–856, doi:10.1126/science.1121066.

[25] A.M. Shaw, K.R. Timpano, T.B. Tran, J. Joormann, Correlates of Facebook usage patterns: the relationship between passive Facebook use, social anxiety symptoms, and brooding, Comput. Hum. Behav. 48 (2015) 575–580, doi:10.1016/j.chb.2015.02.003.

[26] T.H. Silva, P.O.S.V. d. Melo, J.M. Almeida, J. Salles, A.A.F. Loureiro, A picture of Instagram is worth more than a thousand words: workload characterization and application, in: Proceedings of the International Conference on Distributed Computing in Sensor Systems, DCOSS, IEEE, New York, NY, USA, 2013, pp. 123–132, doi:10.1109/DCOSS.2013.59.

[27] P. Singer, E. Ferrara, F. Kooti, M. Strohmaier, K. Lerman, Evidence of online performance deterioration in user sessions on reddit, PLoS One 11 (8) (2016) 1–16, doi:10.1371/journal.pone.0161636.

[28] G. Szabo, B.A. Huberman, Predicting the popularity of online content, Commun. ACM 53 (8) (2010) 80–88, doi:10.1145/1787234.1787254.

[29] M. ten Thij, Y. Volkovich, D. Laniado, A. Kaltenbrunner, Modelling page-view dynamics on Wikipedia, 2013. Preprint arXiv:1212.5943.

[30] M. ten Thij, S. Bhulai, P. Kampstra, Circadian patterns in Twitter, in: Proceedings of the Third International Conference on Data Analytics, DATA ANALYTICS, IARIA, Wilmington, DE, USA, 2014, pp. 12–17.

[31] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Predicting elections with Twitter: what 140 characters reveal about political sentiment, in: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, ICWSM, AAAI, 2010, pp. 178–185.

[32] P.O.S. Vaz de Melo, C. Faloutsos, R. Assunção, A. Loureiro, The self-feeding process: a unifying model for communication dynamics in the web, in: Proceedings of the International Conference on World Wide Web, WWW, ACM, New York, NY, USA, 2013, pp. 1319–1330, doi:10.1145/2488388.2488503.

[33] C. Wang, M. Ye, B.A. Huberman, From user comments to on-line conversations, in: Proceedings of the International Conference on Knowledge Discovery and Data Mining, KDD, ACM, New York, NY, USA, 2012, pp. 244–252, doi:10.1145/2339530.2339573.

[34] Wikimedia, Page view statistics for Wikimedia projects, 2019, Retrieved on 23 July 2019, https://dumps.wikimedia.org/other/pagecounts-raw/.

[35] Wikipedia, Wikipedia, die freie enzyclopädie, 2019b, Retrieved on 23 July 2019, https://de.wikipedia.org/wiki/Wikipedia:Hauptseite.

[36] Wikipedia, Wikipedia, the free encyclopedia, 2019d, Retrieved on 23 July 2019, https://en.wikipedia.org/wiki/Main_Page.

[37] Wikipedia, Wikipedia: today's featured article, 2019e, Retrieved on 23 July 2019, https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria.

[38] Wikipedia, Wikipedia: today's featured article/requests, 2019f, Retrieved on 23 July 2019, https://en.wikipedia.org/wiki/Wikipedia:Today's_featured_article/requests.

[39] Wikipedia, Wikipedia, la enciclopedia libre, 2019c, Retrieved on 23 July 2019, https://es.wikipedia.org/wiki/Wikipedia:Portada.

[40] Wikipedia, Wikipedia, de vrije encyclopedie, 2019a, Retrieved on 23 July 2019, https://nl.wikipedia.org/wiki/Hoofdpagina.

[41] F. Wu, B.A. Huberman, Novelty and collective attention, Proc. Natl. Acad. Sci. U.S.A. 104 (45) (2007) 17599–17601, doi:10.1073/pnas.0704916104.

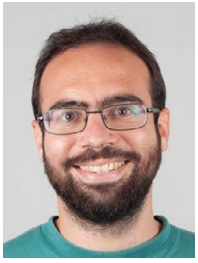
[42] T. Yasseri, R. Sumi, J. Kertész, Circadian patterns of Wikipedia editorial activity: a demographic analysis, PLoS One 7 (1) (2012) 1–8, doi:10.1371/journal.pone.0030091.

[43] Q. Zhao, M.A. Erdogdu, H.Y. He, A. Rajaraman, J. Leskovec, Seismic: a self-exciting point process model for predicting tweet popularity, in: Proceedings of the Twenty-First ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, ACM, New York, NY, USA, 2015, pp. 1513–1522, doi:10.1145/2783258.2783401.

Marijn ten Thij, PhD is a postdoctoral researcher at the Center for Social and Biomedical Complexity (CSBC) at Indiana University Bloomington. Marijn obtained a PhD in Mathematics (Business Analytics) at the Vrije Universiteit Amsterdam under supervision of Prof. dr. Sandjai Bhulai and a MSc in Applied Mathematics at the Stochastic Operations Research (SOR) group at the University of Twente. Marijn's research interests involve the usage of mathematical modelling to study and mimic human behavior through data obtained from social media. His current work is at the intersection of the fields Complex Systems, Computational Social Science, and Data Science.

Andreas Kaltenbrunner is Director of Data Analytics at NTENT, where he leads a team focusing on user behavior analysis and improvements for ranking in mobile search. Andreas is also teaching a master course on Data Driven Social Analytics at Universitat Pompeu Fabra and is involved in research activities centered on computational social science, social media and social network analysis, areas in which he has co-authored more than 70 publications.

He obtained his PhD in Computer Science and Digital Communication in 2008 from the Universitat Pompeu Fabra with a thesis about stochastic effects in human and neural communication patterns. Afterwards, he joined the technology center Barcelona Media, where he led from 2013 onwards the Social Media Research Line. Between June 2015 and August 2017 he lead the Digital Humanities Research Unit at the technology center Eurecat, before joining NTENT in September 2017.



David Laniado is Senior researcher at Eurecat, where he leads research in Computational Social Science. He obtained from Politecnico di Milano his master degree in Computer Engineering in 2007, and his PhD in Information Engineering in 2012, with a thesis on social construction of knowledge by online communities. His research focuses on the study of patterns of online conversation and discussion, participatory democracy and deliberation processes, gender patterns and socio-technical aspects of online interactions.

Dr. Yana Volkovich is a Senior Data Scientist at Xandr, AT&T's advertising and analytics company, which operates a technology platform for buying and selling digital advertising. Before joining Xandr, Yana was a Senior Research Scientist at Eurecat, one of the largest Technology Centers in Spain. In 2014 she was the visiting Marie Curie Research fellow at Cornell Tech. Yana received her PhD in Applied Probability from the University of Twente (Netherlands).

Yana's research interests include Machine/Deep Learning, Computational Social Science, Graph theory, and Heavy-tailed distributions. Dr. Volkovich has served as a program committee member in a number of top scientific conferences and as a reviewer for scientific journals in the fields of Probability, Computer Science, and Computational Social Science.