ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: García-Gavilanes R, Tsvetkova M, Yasseri T. 2016 Dynamics and biases of online attention: the case of aircraft crashes. *R. Soc. open sci.* **3**: 160460. http://dx.doi.org/10.1098/rsos.160460

Received: 28 June 2016
Accepted: 12 September 2016

Subject Category:
Physics

Subject Areas:
complexity/human-computer interaction/behaviour

Keywords:
collective attention, Wikipedia, attention economy, aircraft crash, media biases, media coverage

Author for correspondence:
Taha Yasseri
e-mail: taha.yasseri@oii.ox.ac.uk

THE ROYAL SOCIETY PUBLISHING

Dynamics and biases of online attention: the case of aircraft crashes

Ruth García-Gavilanes, Milena Tsvetkova and Taha Yasseri

Oxford Internet Institute, University of Oxford, Oxford, UK

TY, 0000-0002-1800-6094

The Internet not only has changed the dynamics of our collective attention but also through the transactional log of online activities, provides us with the opportunity to study attention dynamics at scale. In this paper, we particularly study attention to aircraft incidents and accidents using Wikipedia transactional data in two different language editions, English and Spanish. We study both the editorial activities on and the viewership of the articles about airline crashes. We analyse how the level of attention is influenced by different parameters such as number of deaths, airline region, and event locale and date. We find evidence that the attention given by Wikipedia editors to pre-Wikipedia aircraft incidents and accidents depends on the region of the airline for both English and Spanish editions. North American airline companies receive more prompt coverage in English Wikipedia. We also observe that the attention given by Wikipedia visitors is influenced by the airline region but only for events with a high number of deaths. Finally we show that the rate and time span of the decay of attention is independent of the number of deaths and a fast decay within about a week seems to be universal. We discuss the implications of these findings in the context of attention bias.

1. Introduction

The Internet has drastically changed the flow of information in our society. Online technologies enable us to have direct access to much of the world's established knowledge through services such as Wikipedia and to informal user-generated content through social media. There is no theoretical limit to the information bandwidth on the Internet but human attention has its own limits. Public attention to emerging topics decays over time or suffers the so-called memory buoyancy from users, which is a metaphor of information objects sinking down in the digital memory with decreasing importance and usage, increasing their distance to the user [1].

Nowadays, the online footprints of users have rendered the level of attention given to new and past events and its decay an observable phenomenon. The digital nature of Internet-based technologies enables us to analyse the variances of attention at a scale and with an accuracy that have not been feasible in relation to other communication technologies. Researchers have used logs generated by online users' activities such as tweets, search queries and web navigation paths to cover a wide range of topics on attention. For example, Lehmann *et al.* [2] characterize attention by analysing the time-series of tweets with popular tags from a dataset of 130 million tweets from 6.1 million users and found four clusters based on dynamics, semantics and information spread. Yeung *et al.* [3] focus on how events are remembered for specific years by looking at temporal expressions in the text of 2.4 million articles in English from Google news archive; they find more references to more recent events. Other studies have concentrated on attention decay. Wu & Huberman [4] discover a very short time span of collective attention with regard to news items on the digg.com linksharing website. Simkin & Roychowdhury [5] study blogs and news from more than 100 websites and find that decay in accessibility is owing to aspects of visibility such as link positioning and attractiveness. Researchers have also linked online attention to more practical matters, from predicting election outcomes [6] and detecting memory patterns in human activities [7], all the way to analysing trading behaviour in financial markets [8] or the appropriate time when to publish news to gain more attention [9].

While several aspects of online attention increase and decay have been fairly well investigated, much less is known about how geography, event impact and differences across populations with different languages affect attention. Thus, the question of whether online technologies have improved or worsen the fairness and equality with which news are released to the public, influencing their attention, is still open. The question is particularly important to investigate with regard to high impact events such as the terrorist attacks in Paris and Beirut in November 2015. It was reported [10] that only 11% of the top media outlets covered the Beirut attacks in the first 24 h in comparison with 51% for Paris. Furthermore, user attention for the Beirut bombings within the first hour was only 5% of what Paris achieved within the same time period in spite of the Paris attacks starting almost 15 h after Beirut. What determines what is covered by the media and when? What determines the level of public attention to new events? Does the decay of public attention vary depending on the event? In this paper, we answer these questions at scale by analysing editorial and traffic information on a set of articles in two different language editions of Wikipedia. We study how events are covered, what aspects determine attention to them, how attention decays, and whether there are differences between languages. Focusing on depth rather than breadth, we limit our analyses to one specific type of event—aircraft incidents and accidents—and to the two most popular Wikipedia language editions by number of active users—English and Spanish.

Wikipedia is a unique resource to study collective attention. Written and edited by volunteers from all around the world, it has become the number one source of online information in many languages, with close to 40 million articles in around 300 language editions (and counting) and with open access to logs and metadata. There is a high correlation between search volume on Google and visits to the Wikipedia articles related to the search keywords [11,12]. This indicates that Wikipedia traffic data are a reliable reflection of web users' behaviour in general. The high response rate and pace of coverage in Wikipedia in relation to breaking news [13,14] is another feature that makes Wikipedia a good research platform to address questions related to collective attention. For instance, researchers have analysed Wikipedia edit records to identify and model the most controversial topics in different languages [15,16], to study the European food culture [17] and to highlight entanglement of cultures by ranking historical figures [18]. Wikipedia traffic data have also been used to predict movie box office revenues [19], stock market moves [20], electoral popularity [21] and influenza outbreaks [22,23].

To answer our research questions, we develop an automatic system to extract editorial and traffic information on the Wikipedia articles about aircraft incidents and accidents and factual information about the events. By comparing the English and the Spanish Wikipedia, we contribute to this research field in the following ways:

— we study the coverage of the events in Wikipedia and its dynamics over time considering the airline region, the event locale and the number of deaths;
— we analyse the role of the airline region and number of deaths on the viewership data to Wikipedia articles; and
— we model attention decay over time.

We present the results from our study in the next section, after which we continue with discussion and conclude with implications. Details for our data collection and analysis strategy can be found in the last section, §4.

2. Results

Figure 1 shows a map of all the aircraft incidents and accidents from English Wikipedia coloured according to the airline region, which is where the airline company for the flight is located, and sized according to the number of deaths caused by the event. For simplicity, we divide the Americas into two regions: North America and Latin America. Latin America includes all countries or territories in the Americas where Romance languages are spoken as a first language (in this case, Spanish, Portuguese and French) and all Caribbean islands, while North America includes the rest (i.e. mostly United States and Canada). Furthermore, all headquarters in the EuroAsia region are labelled as Asia (e.g. Russia and Turkey). We observe that the locales of the events overlap most of the time with the airline regions.

Our results are divided in three sections: the first part deals with the editorial coverage of the events, the second with the immediate collective attention quantified by viewership statistics and the third with the modelling of attention decay.

2.1. Editorial coverage

Table 1 compares the number of aircraft accidents and incidents covered in English and Spanish Wikipedias with cases reported by the Aviation Safety Network (ASN)[1] in different continents. While ASN provides data from 1945, excluding military accidents, corporate jets and hijackings, our dataset includes these cases and dates back to the year 1897. There are 1081 articles in English Wikipedia that do not have a Spanish equivalent and most of them are about events that happened in North America (265), Asia (261) and Europe (252). On the other hand, there are 71 articles in Spanish Wikipedia with no English equivalent and most of them are about events that happened in Latin America (39).

With regard to the number of deaths, the lowest average numbers correspond to Australia, North America and Europe, respectively, for English Wikipedia, whereas Latin America and North America have the lowest average number of deaths for Spanish Wikipedia. This is because some low impact events (many with 0 deaths) that occurred in Australia, North America and Europe are only included in English Wikipedia and some low impact events in Latin America are only considered notable in Spanish Wikipedia. With regard to the articles in English that do not have a Spanish equivalent, the average number of deaths is 39 and for those that do not have an English equivalent the average is 12. These numbers indicate that the articles in Spanish without an English equivalent are low impact events concentrated in Latin America.

We also investigate the time lag between the occurrence of the event and the creation of the corresponding Wikipedia article. Our dataset contains articles about events that happened before and after Wikipedia was launched (see figure 7 in appendix A). Post-Wikipedia events (399 for English and 224 for Spanish) are shown on the upper row panels of figure 2, where the horizontal and vertical axes show the time of the occurrence of the event and the creation of the corresponding Wikipedia page, respectively. The convergence of the data points towards the diagonal line indicates that the community of Wikipedia editors reacts increasingly fast to this kind of events. English Wikipedia has been faster at covering events as the diagonal trend starts earlier. A possible explanation is the larger number of users in English Wikipedia compared with the Spanish version.

Figure 2(*c*,*d*) shows the coverage of the pre-Wikipedia events. The colour of the curve corresponds to the airline's region and the *x*-axis shows the year of the Wikipedia page creation. For English Wikipedia (1078 cases) a quicker coverage of North American events is evident. African, Australian and South American events exhibit sharp increases as the addition of these articles was concentrated in specific periods. On the other hand, Spanish Wikipedia (264 cases) shows a slightly faster coverage for events related to European companies with sharp jumps for African and Australian companies (there are only 34 and five cases, respectively). Most importantly, however, not only did English Wikipedia cover more pre-Wikipedia events, but it also did it faster. Again, this can be explained considering the larger size of the editorial community of English Wikipedia.

[1] http://aviation-safety.net/statistics/geographical/continents.php.

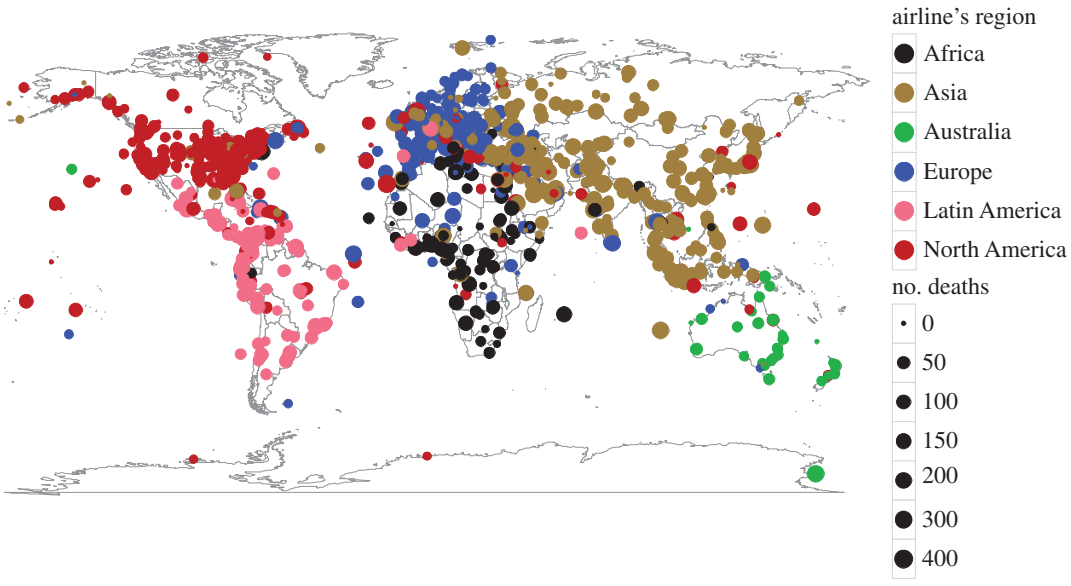

Figure 1. 1496 Geolocated incidents and accidents since 1897 reported in English Wikipedia. Each dot represents an event. The size of the dots is proportional to the number of reported deaths and the colour codes the location of the operating company.

Table 1. Breakdown by region of the number of aircraft incidents and accidents covered in Wikipedia compared with the data available at The Aviation Safety Network (ASN) website. (The column 'events' is the ratio with regard to the row 'total'.)

	Wikipedia						ASN		
	English			Spanish					
		deaths			deaths			deaths	
continent	events	avg	total	events	avg	total	events	avg	total
Africa	0.08	49	5967	0.07	58	1981	0.10	20	8108
Asia	0.24	50	17 987	0.22	61	6618	0.17	27	19 351
Australia	0.03	21	873	0.01	52	260	0.03	12	1448
Europe	0.22	36	11 818	0.17	59	4963	0.24	23	23 423
L. America	0.08	47	5789	0.24	40	4695	0.19	16	12 942
N. America	0.23	27	9052	0.16	45	3517	0.23	13	12 958
others	0.12	45	8353	0.13	80	4941	0.02	32	2712
total	1496	40	59 839	488	55	26 975	4223	19	80 942

2.2. Immediate attention

Now we turn to the viewership data. To capture the immediate attention to an event right after its occurrence, we choose the articles that were created up to 3 days after the event and extract the maximum number of views within 7 days after the page was created (see figure 4 for an example). We discuss the choice of 7 days in §2.3.

A baseline hypothesis would be that the larger the number of deaths the event caused, the more attention it attracts. However, this is not always the case; attention is driven by other factors such as media coverage, location, people involved, etc. This is reflected in figure 3. The plot shows the normalized maximum daily views versus the number of deaths in log scale for the English and Spanish Wikipedias.

In English Wikipedia, we have identified two regimes: low-impact events (less than 40 deaths), where there is no correlation between impact and attention, and high-impact events (greater than or equal to 40 deaths), where the maximum number of daily page views increases proportionally to the event impact with $r = 0.71$, $p < 0.001$. To separate these two regimes, we used visual inspection to accommodate the

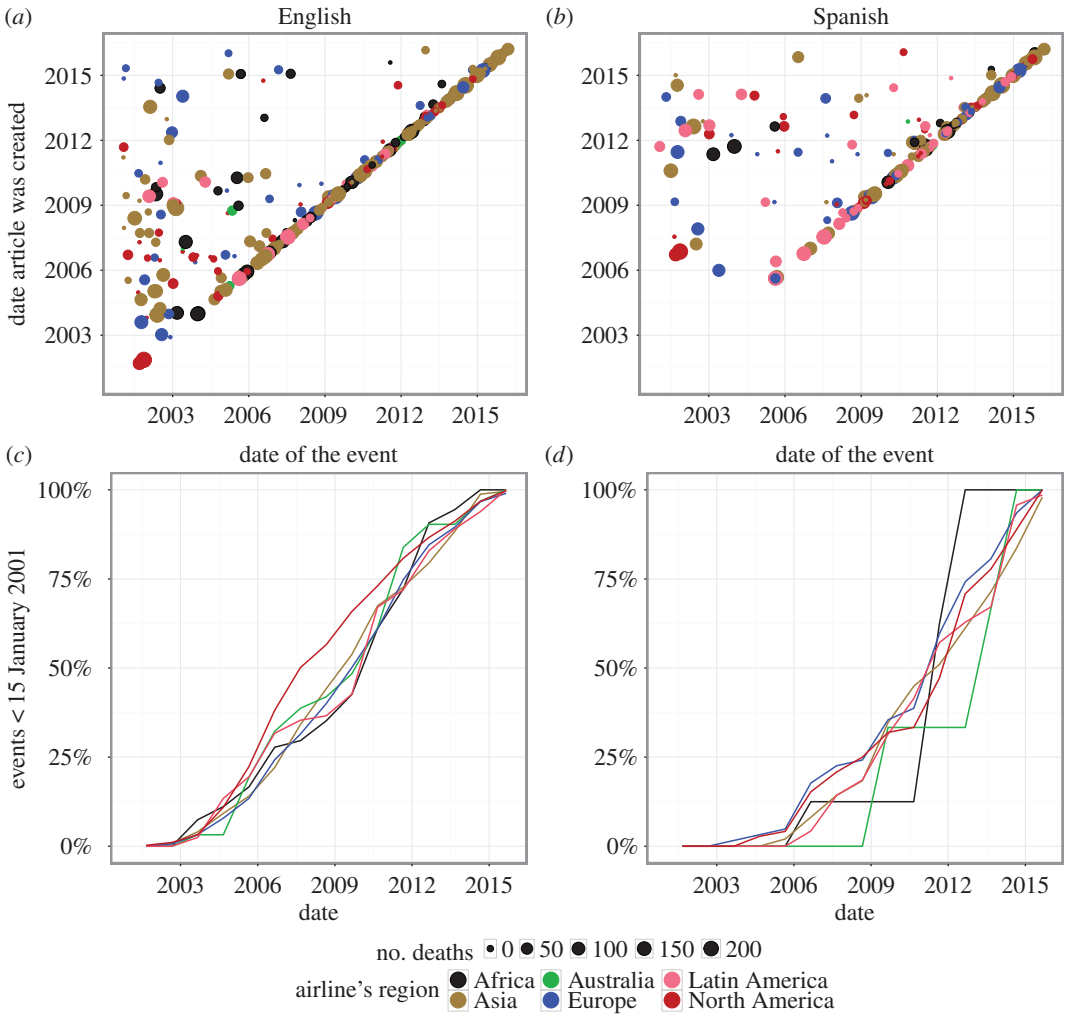

Figure 2. Coverage of aircraft incidents and accidents in the English and Spanish Wikipedia: (*a*,*b*) the lag between the occurrence of the event and the creation of the corresponding article in Wikipedia for post-Wikipedia events; and (*c*,*d*) the corresponding percentage of covered pre-Wikipedia events in time.

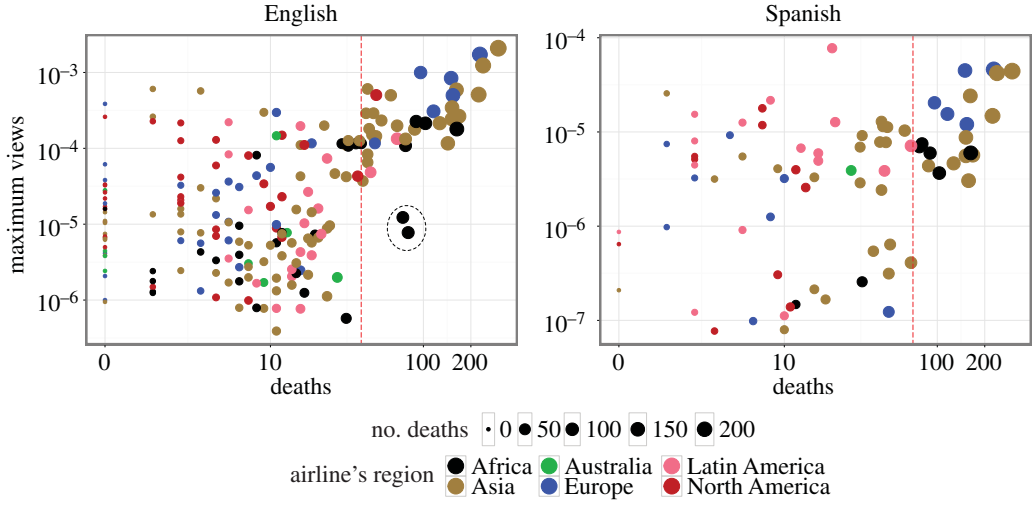

Figure 3. Normalized maximum number of page views versus the number of deaths of each event, both on log-scale for (*a*) English (En) and (*b*) Spanish (Sp) Wikipedia. The two outliers in (*a*) are removed from the analysis.

Table 2. Results from regression analyses with logarithm of the maximum number of page views as a dependent variable. (The column for $\beta 1$ corresponds to a model that only considers the number of deaths (log-transformed) as the independent variable, whereas $\beta 2$ reports a model which considers log(deaths) and the airline region as independent variables. Significance codes: *** <0.001, ** <0.01, * <0.05.)

	all events							
	English ($n = 204$)				Spanish ($n = 80$)			
	β_1		β_2		β_1		β_2	
intercept	−12.18	***	−13.19	***	−13.89	***	−15.24	***
deaths	0.61	***	0.69	***	0.41	**	0.53	***
Asia			0.79	*			0.7	
Australia			0.22				0.99	
Europe			1.42	**			1.21	
Latin America			0.23				1.68	*
North America			1.67	***			0.96	
adj. R^2	0.22	***	0.28	***	0.11	**	0.12	*
	low-impact							
	English ($n = 166$)				Spanish ($n = 60$)			
	β_1		β_2		β_1		β_2	
intercept	−11.44	***	−12.27	***	−13.3	***	−15.87	***
deaths	0.04		0.14		0.1		0.14	
Asia			0.47				2.42	
Australia			0.07				2.95	
Europe			0.99	*			2.1	
Latin America			0.46				3.18	*
North America			1.39	**			2.3	
adj. R^2	−0.01		0.04		−0.01		0.02	
	high-impact							
	English ($n = 38$)				Spanish ($n = 20$)			
	β_1		β_2		β_1		β_2	
intercept	−12.61	***	−12.95	***	−18.03	***	−18.73	***
deaths	0.97	***	0.92	***	1.33	**	1.45	**
Asia			0.49				−0.22	
Australia								
Europe			1.01	*			0.88	
Latin America			−0.21					
North America			1.72	*				
adj. R^2	0.38	***	0.48	***	0.28	**	0.50	**

largest empty square on the lower-right region of the diagram. Regardless of the high correlation of this region, impact does not always reflect attention: the plot shows two African outliers with less attention than expected from the overall trend. In Spanish Wikipedia, the separation of the two phases at around 70 deaths is less evident but still exists. The correlation in the high impact regime is $r = 0.67$, $p < 0.005$. Also note that in the high impact regime, the level of attention increases almost quadratically with the number of deaths. However, we hesitate fitting a function here owing to the small number of data points.

To analyse the importance of the airline's region and number of deaths on level of attention, we use linear regression models. We have removed the two outlier events from the English sample shown in figure 3. We then model all the data points using a simple linear model considering the number of deaths as the only parameter (table 2). In the English case, the number of deaths alone can only explain around

Table 3. Death equivalence ratios based on the viewership data from English and Spanish Wikipedias. (The matrix is calculated according to the coefficients reported on the upper part of table 2. For six different airline continents, the matrix shows the ratio of triggered attention, controlling for the number of deaths. For example, the attention given to events caused by a North American Airline in English Wikipedia is on average 2 and 47 times larger than to the events caused by European and African companies, respectively. In Spanish Wikipedia, the level of attention given to events related to Latin America is three times larger than the European events, five times larger than North American and 10 times larger than Asian events.)

	English Wikipedia					
	Africa	Australia	Latin America	Asia	Europe	North America
Africa	1	2	2	6	26	47
Australia		1	1	4	16	28
Latin America			1	4	16	28
Asia				1	4	8
Europe					1	2
North America						1
	Spanish Wikipedia					
	Africa	Asia	North America	Australia	Europe	Latin America
Africa	1	5	10	10	16	48
Asia		1	2	2	3	10
North America			1	1	2	5
Australia				1	2	5
Europe					1	3
Latin America						1

22% of the variation in the level of the immediate attention. If we add the airline region as a categorical variable using Africa as the reference category, we increase the explanatory power to 28%. Here, we observe that events related to North American companies attract more views than companies from other regions ($\beta_1 = 1.67$). On the other hand, Latin American companies play the same role in Spanish Wikipedia ($\beta_1 = 1.68$).

If we split the data points into high- and low-impact events and recalculate the linear model separately for each regime, we see that the addition of the airline region in cases with a high number of deaths increases the explanatory power of the regression. In both language editions, the proportion variance explained increases considerably. The explanatory power we obtain for the low-impact events, however, is negligibly small.

Based on the results of the categorical regression analysis including the location of the operating companies, one can estimate the relative level of attention paid to pairs of events from different regions on average. These ratios are reported in table 3. For instance, controlling for the number of deaths, a North American event triggers about 50 times more attention among English Wikipedia readers compared with an African event. This ratio for North American versus European is about 2. In Spanish Wikipedia, however, a Latin American event triggers about 50 times more attention than an African and five times more than a North American event.

2.3. Modelling attention decay

Now we focus on attention decay by analysing the viewership time-series after the event. After the initial boost in viewership, which in 73% of the cases happens in less than 5 days after the date of the page creation, an exponential decay follows (see figure 4 for an example). This phenomenon also occurs both because of the decay of novelty [4] as well as limitations in human capacity to pay attention to older items in competition with newer ones [24].

To model the attention decay, we use a segmented regression model with two break points to fit the normalized daily page-view counts in logarithmic scale (see §4 for details). Figure 4 shows a typical example of the time series of the viewership of an article and the fit of the segmented regression model.

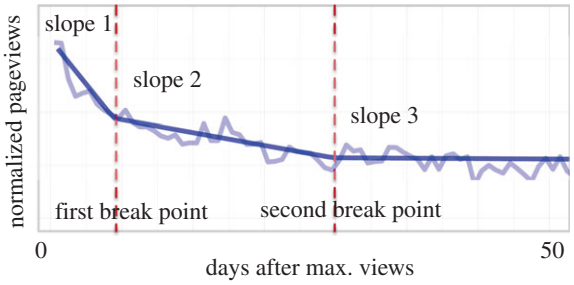

Figure 4. Typical example of the viewership time-series of a Wikipedia article related to an airplane crash fitted with segmented regression with two break points. The *y*-axis is in logarithmic scale.

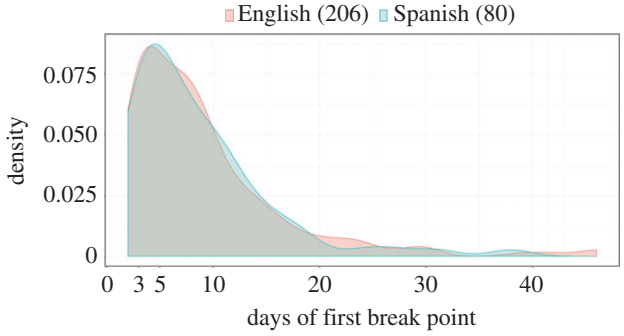

Figure 5. Distribution of the position of the first break point in number of days for a set of articles in English and Spanish Wikipedia.

Table 4. The distribution of normalized maximum daily views of each article and adj. R^2 of the segmented regressions as well as the distribution of the model parameters, calculated half-life (reverse of the absolute value of the slope) and the number of deaths for each event. (All distributions are based on 206 and 80 observations for English (En) and Spanish (Sp) Wikipedias.)

	English			Spanish		
	min	distribution	max	min	distribution	max
adj. R^2	0		1	0		1
max.views	3.9×10^{-7}		2.1×10^{-3}	7.7×10^{-8}		7.7×10^{-5}
slopes						
slope 1	−4.2		0.08	−4.2		0.0
slope 2	−2.6		1.9	−3.4		3.9
slope 3	−0.7		0.6	−1.0		1.1
half-life (days)	0 .3		24	0.3		30
break points position						
1st b.p (days)	2		46	2		38
2nd b.p (days)	4		49	4		49
number of deaths						
deaths	0		298	0		298

The distributions of fit parameters are reported in table 4. These distributions confirm the assumptions that we make in developing our segmented regression model with two break points as well as similarities between the two language editions that we study. For instance, in both cases, the half-life of the attention in the first phase and the detected position of the first break point show similar patterns.

In figure 5, we show the distribution of the location of the first break point in larger scale. This parameter indicates the time span of the initial attention paid to the event. The first break point is localized approximately 3–10 days for both English and Spanish Wikipedia.

In figure 6, we consider other parameters that the best fit of the model assigns to each event. We observe that there is no significant correlation between the position and the value of attention at the first break point and the number of deaths, meaning that the rate of decay in attention and the first

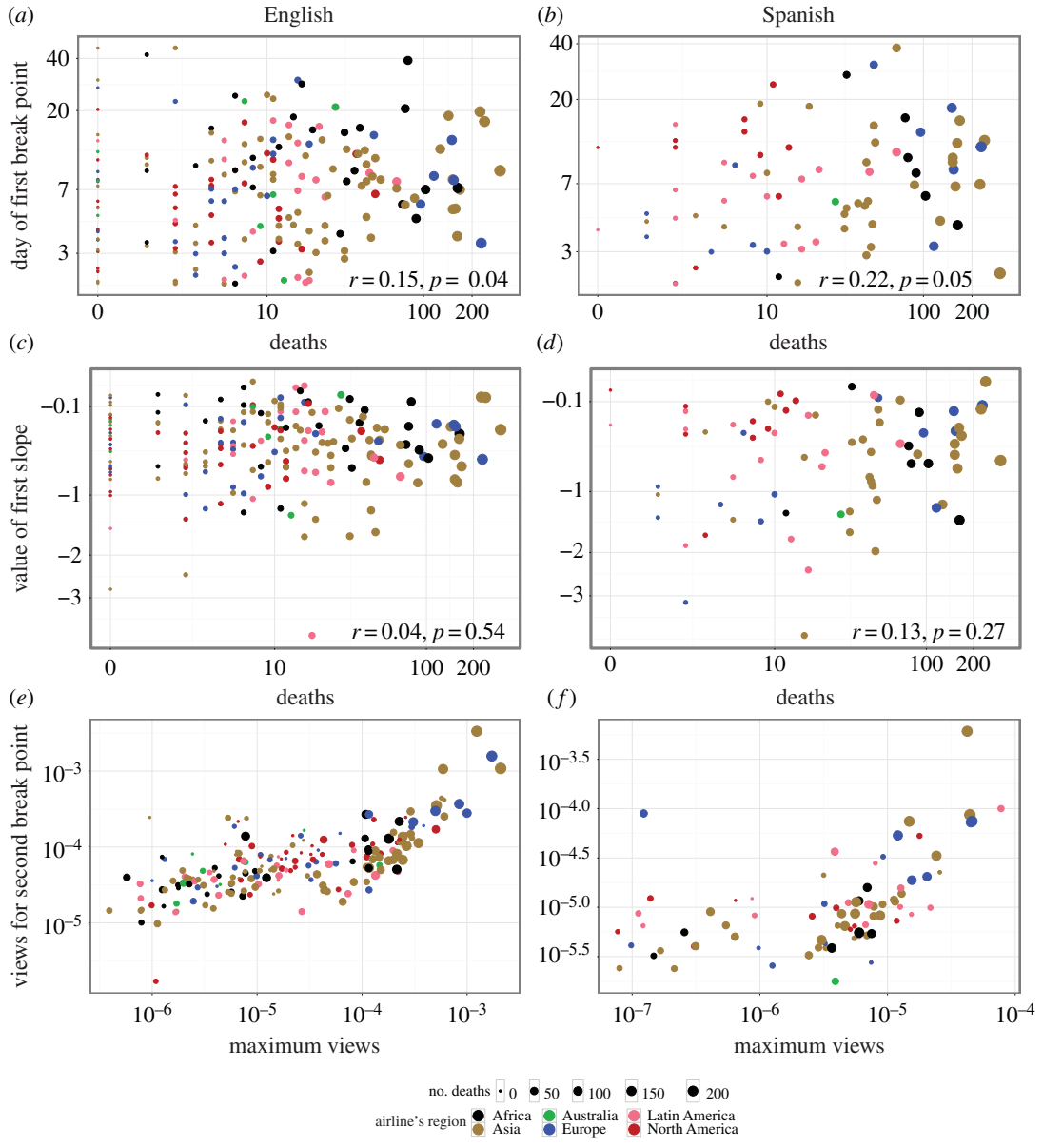

Figure 6. Log–log scatter plots of model parameters against the number of deaths of each event: (*a*,*b*) the location of the first break point (days) versus the number of deaths, (*c*,*d*) the slope of the first segment versus the number of deaths, and (*e*,*f*) the intercept of the last segment versus the maximum daily page views. The four first plots report the Spearman's rank correlation coefficient and the corresponding *p*-value between the *x*- and *y*-axes.

attention phase time span are independent of the impact of the event (upper and middle rows). However, in the lower row of the same figure we show that the relationship between the level of attention at the second break point, which can be interpreted as the level of the long-lasting attention, and the immediate attention in the initial phase, is similar to what is observed in figure 3; i.e. for low impact events, the long lasting attention is independent of the initial attention, whereas for high impact events, the initial attention is a good predictor of the long-term attention to the event.

3. Discussion and conclusion

We studied online attention to aircraft incidents and accidents using editorial and viewership data for the English and Spanish editions of Wikipedia. Overall, we found certain universal patterns.

We found some differences in event coverage between the two languages but often, they can be attributed to the same underlying biases. For example, attention on English Wikipedia is more focused on

events concerning North American and European airlines while attention on Spanish Wikipedia gives priority to Latin American airlines. English Wikipedia tends to cover more events in North America, while Spanish Wikipedia tends to cover more events in Latin America.

Our findings suggest that crashes of flights operated by North American companies, which mostly happened also in North America, receive higher publishing priority in English Wikipedia regardless of the impact, while accidents from other locales, especially older accidents, are published later and have to be more impactful to receive the same level of editorial attention. Similar editorial biases in different contexts have been studied and reported before [25,26]. Although one can argue that English Wikipedia is mostly edited and used by North American users, previous research has shown that only about half of the editorial activity on English Wikipedia originates from North America [27] and English should be considered as the *lingua franca* of Wikipedia [28]. Also note that the difference that we see within each Wikipedia language edition is consistent regardless of the language of the study and hence the origin of viewers.

These biases in Wikipedia can be driven by the biases in mainstream media [29]. Previous research has shown that a considerable dominance of references to Western media exists in Wikipedia [30], and therefore, events of less importance for the Western media are more sparsely covered in Wikipedia. In the case of aircraft crashes, for example, in 1981, 10 people died in the controversial flight *FAB 001* belonging to the Ecuadorian Air Force. It is a controversial flight because the former president of Ecuador Jaime Roldós was among the victims and the cause of the crash is still a mystery. Although there are articles in several languages in Wikipedia covering the biography of Jaime Roldós and the type of airplane used in the crash, there is no article equivalent to the specific flight that caused his death and thus this case is missing in our dataset. The same happens for the flight that killed the former president of the Philipines Ramón Magsaysay or the Iraqi former president Abdul Salam Arif, among others.

In both languages, we observed two attention regimes for events—low-impact regime, where the level of maximum attention is independent of the number of deaths and high-impact regime, where the airline region and the impact of the event significantly influence attention. In addition, focusing on the immediate attention to the event, we found that the time span and rate of the exponential decay (the slope of the fit to the first segment exemplified in the semi-log diagram of figure 4) is independent of the impact of the event and the language of the article. The short span of attention that we observed (on the order of a few days) is in accordance with previous findings by other researchers [4,31,32].

Our study needs further generalization to include other type of events, such as natural disasters, political and cultural events. Moreover, our analysis has been limited to the English and Spanish editions of Wikipedia. Although these two are among the largest Wikipedia language editions, we might see variations in results studying attention patterns in different language editions.

4. Material and methods

4.1. Data collection

We collect data from Wikipedia using two main sources: the MediaWiki API and Wikidata. Wikidata[2] is a Wikipedia partner project that aims to extract facts included in Wikipedia articles and fix inconsistencies across different editions [33]. Although content in Wikidata is still somewhat limited, the availability of such structured information makes it easier for researchers to obtain data from a set of Wikipedia articles in a systematic way.

To complete the data missing from Wikidata, we automatically crawl Wikipedia infoboxes[3] and collect features of events (see below).

We first focus on a set of articles classified as aircraft accidents or incidents in English Wikipedia, belonging to the categories *Aviation accidents and incidents by country* and *Aviation accidents and incidents by year*, and their subcategories, which cover all airline accidents and incidents in different countries and throughout history available in Wikipedia. In total, we obtain 1606 articles from which 1496 are specifically about aircraft crashes or incidents (we discard articles of biographies, airport attacks, etc.). From the 1496 articles, we obtain the following: date of the event, number of deaths, coordinates of the event and airline region.

We extract all editorial information for the articles in the sample using the MediaWiki API. We extract the date when the article was created and alternative names for the article. We use the latter to merge all

[2]Using https://cran.r-project.org/web/packages/WikidataR/index.html.

[3]Using https://cran.r-project.org/web/packages/WikipediR/index.html.

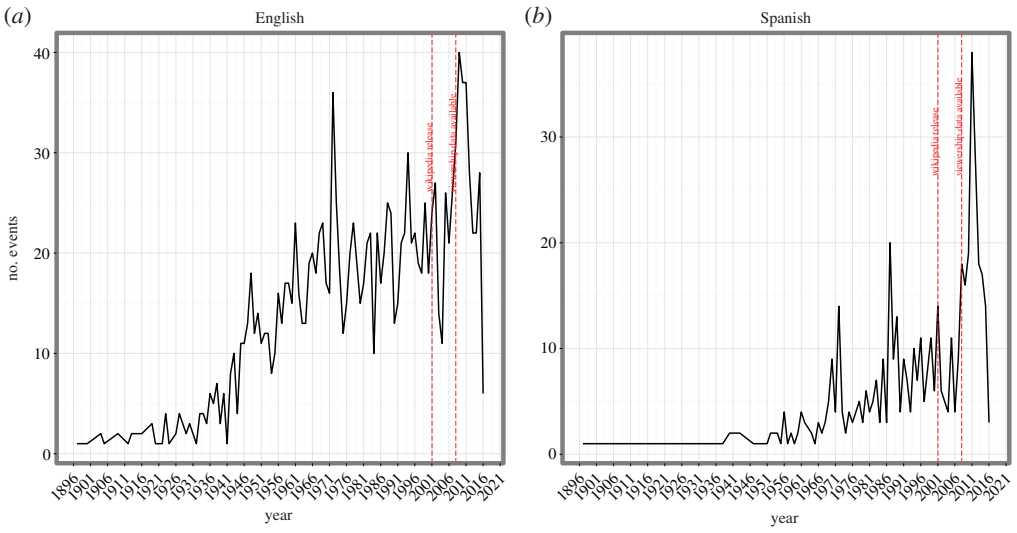

Figure 7. The number of aircraft incidents and accidents per year reported in (*a*) English and (*b*) Spanish Wikipedia.

traffic statistics to the main title. Next, we extract all available articles in the same categories considered in English Wikipedia from Spanish and follow the same procedure to extract the features of the articles in the Spanish edition. In total, we obtain 525 articles in Spanish Wikipedia from which 488 are about aircraft incidents or accidents.

Finally, we extract the daily traffic to the articles in English and Spanish from the Wikipedia pageview dumps[4] through a third party interface.[5]

4.2. Data analysis

To control for the changes in the overall popularity of Wikipedia, we normalize the viewership counts by the overall monthly traffic to Wikipedia.[6] To numerically model attention dynamics, we apply segmented regression analysis to viewership data during 50 days after the first pick due to the occurrence of the event. We use segmented regression as implemented in the R package 'segmented'.[7] Segmented regression models are models where the relationship between the response and one or more explanatory variables are piecewise linear, represented by two or more straight lines connected at values called break points [34]. To find those break points, the algorithm first fits a generic linear model then fits the piecewise regression through an iterative procedure that uses starting break point values given by us at the beginning. In our specific case, three piecewise regressions are fitted in each iteration and the two break point values are updated accordingly as to minimize the gap γ between the segments. The model converges when the gap between the segments is minimized. We refer the reader to the paper by Muggeo [34] for a detailed explanation. Additionally, the package description explains that bootstrap restarting is used to make the algorithm less sensitive to starting values.

Although alternative approaches could be undertaken to model nonlinear relationships, for instance via splines, the main appeal of the segmented model lies in its simplicity and the interpretability of the parameters.

We have chosen two break points (three segments) for the analysis but our main results are robust against changing this number (see figure 9 in appendix A). This choice is informed by previous research that identifies three phases in the evolution of collective reactions to events: communicative interaction, floating gap and cultural memory (stabilization phase) [35].

We find that most of the events are fitted well, with high adjusted R^2 (average 0.84 for English and 0.80 for Spanish). However, in some cases, this model is not able to capture the overall dynamics, mostly

[4]https://dumps.wikimedia.org/other/pagecounts-raw/.

[5]http://stats.grok.se.

[6]The data are obtained from https://stats.wikimedia.org/EN/Tablespage-viewsMonthlyCombined.htm.

[7]We use the R package *segmented*: https://cran.r-project.org/web/packages/segmented/.

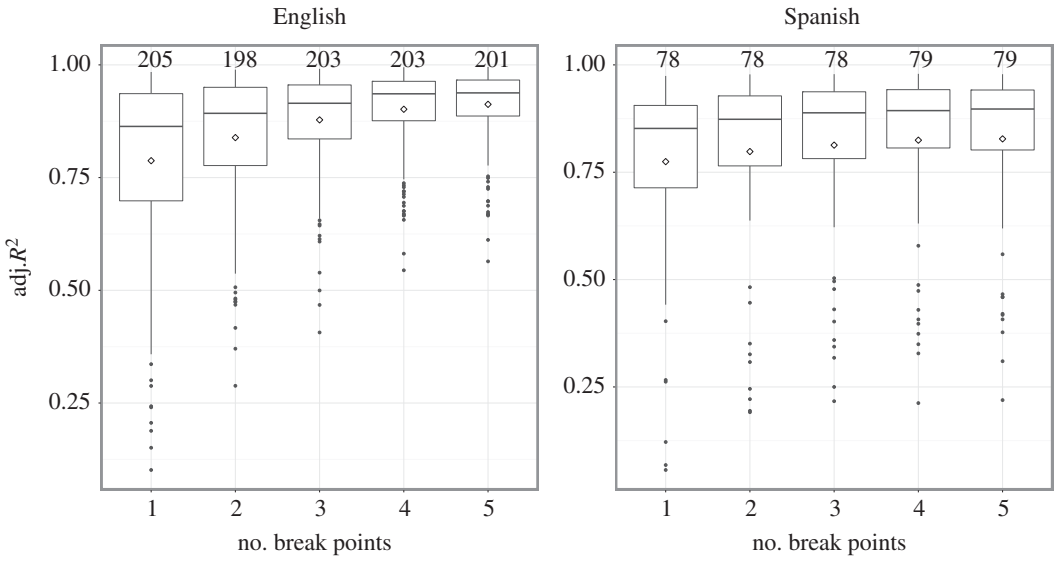

Figure 8. Boxplot of the variance explained (adj.R^2) of the viewership time series (up to 50 days after the event) of Wikipedia articles for different numbers of break points. The numbers at the top represent the total count of data points for each model.

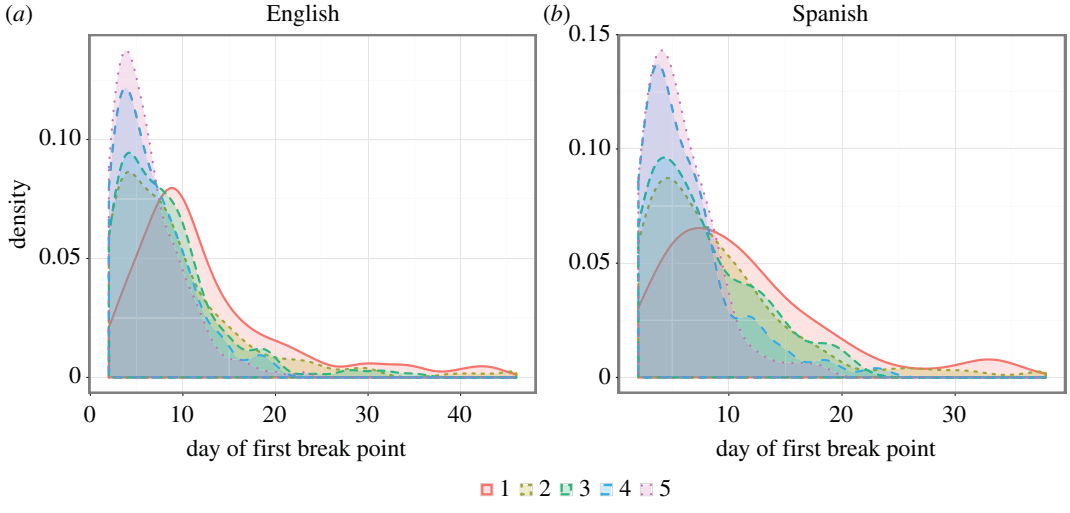

Figure 9. Distribution of the location of the first break point (days) for segmented regressions with different numbers of break points.

owing to secondary shocks driven by new triggering factors that are too close to the event, e.g. the discovery of the corresponding airplane black box or other related newsworthy events.

Data accessibility. The datasets supporting this article have been uploaded to Dryad and is available via https://dx.doi.org/10.5061/dryad.34mn3 [36].

Authors' contributions. R.G.-G. collected and analysed the data, participated in the design of the study and drafted the manuscript; M.T. participated in the design of the study and helped draft the manuscript; T.Y. conceived, designed and coordinated the study, and helped draft the manuscript. All authors gave final approval for publication.

Competing interests. The authors declare no competing interests.

Funding. This research is part of the project *Collective Memory in the Digital Age: Understanding Forgetting on the Internet* funded by Google.

Appendix A

See figures 7–9.

References

1. Kanhabua N, Niederée C, Siberski W. 2013 Towards concise preservation by managed forgetting: research issues and case study. In *Proc. of the 10th Int. Conf. on Preservation of Digital Objects* (*iPres*), pp. 3–8. Lisbon, Portugal: Biblioteca Nacional de Portugal.
2. Lehmann J, Gonçalves B, Ramasco JJ, Cattuto C. 2012 Dynamical classes of collective attention in twitter. In *Proc. of the 21st Int. Conf. on World Wide Web*, pp. 251–260. New York, NY: ACM Press.
3. Au Yeung Cm, Jatowt A. 2011 Studying how the past is remembered: towards computational history through large scale text mining. In *Proc. of the 20th ACM Int. Conf. on Information and Knowledge Management*, pp. 1231–1240. New York, NY: ACM Press.
4. Wu F, Huberman BA. 2007 Novelty and collective attention. *Proc. Natl Acad. Sci. USA* **104**, 17 599–17 601. (doi:10.1073/pnas.0704916104)
5. Simkin MV, Roychowdhury VP. 2015 Why does attention to web articles fall with time? *J. Assoc. Inf. Sci. Technol.* **66**, 1847–1856. (doi:10.1002/asi.23289)
6. Yasseri T, Bright J. 2014 Can electoral popularity be predicted using socially generated big data? *Inf. Technol.* **56**, 246–253. (doi:10.1515/itit-2014-1046)
7. Singer P, Helic D, Taraghi B, Strohmaier M. 2014 Detecting memory and structure in human navigation patterns using Markov chain models of varying order. *PLoS ONE* **9**, e102070. (doi:10.1371/journal.pone.0102070)
8. Preis T, Moat HS, Stanley HE. 2013 Quantifying trading behavior in financial markets using Google trends. *Sci. Rep.* **3**, 1684. (doi:10.1038/srep01684)
9. Subašić I, Castillo C. 2013 Investigating query bursts in a web search engine. *Web Intell. Agent Syst.* **11**, 107–124. (doi:10.3233/WIA-130265)
10. Roy SD. 2015 Paris and Beirut: data suggest how social media shapes the coverage. Blog post online. https://goo.gl/M8Xi4J (accessed 29 January 2016).
11. Ratkiewicz J, Flammini A, Menczer F. 2010 Traffic in social media I: paths through information networks. In *Proc. IEEE Second Int. Conf. on Social Computing*, pp. 452–458. Washington, DC: IEEE Computer Society.
12. Yoshida M, Arase Y, Tsunoda T, Yamamoto M. 2015 Wikipedia page view reflects web search trend. In *Proc. ACM Web Sci. Conf.* (*poster*), pp. 65:1–65:2. New York, NY: ACM Press.
13. Althoff T, Borth D, Hees J, Dengel A. 2013 Analysis and forecasting of trending topics in online media streams. In *Proc. of the 21st ACM Int. Conf. on Multimedia*, New York, NY, USA, pp. 907–916.
14. Keegan B, Gergle D, Contractor N. 2011 Hot off the Wiki: dynamics, practices, and structures in Wikipedia's coverage of the TōHoku catastrophes. In *Proc. of the 7th Int. Symp. on Wikis and Open Collaboration*, pp. 105–113. New York, NY: ACM Press.
15. Yasseri T, Spoerri A, Graham M, Kertész J. 2014 The most controversial topics in Wikipedia: a multilingual and geographical analysis. In *Global Wikipedia: international and cross-cultural issues in online collaboration* (eds P Fichman, N Hara), pp. 25–48. Lanham, MD: Scarecrow Press.
16. Iñiguez G, Török J, Yasseri T, Kaski K, Kertész J. 2014 Modeling social dynamics in a collaborative environment. *EPJ Data Sci.* **3**, 1–20. (doi:10.1140/epjds20)
17. Laufer P, Wagner C, Flöck F, Strohmaier M. 2015 Mining cross-cultural relations from Wikipedia: a study of 31 European food cultures. In *Proc. of the ACM Web Sci. Conf.*, pp. 3:1–3:10.
18. Young-Ho E, Dima LS. 2013 Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles. *PLoS ONE* **8**, e74554. (doi:10.1371/journal.pone.0074554)
19. Márton M, Yasseri T, Kertász J. 2013 Early prediction of movie box office success based on Wikipedia activity big data. *PLoS ONE* **8**, e71226. (doi:10.1371/journal.pone.0071226)
20. Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, Preis T. 2013 Quantifying Wikipedia usage patterns before stock market moves. *Sci. Rep.* **3**, 1801. (doi:10.1038/srep01801)
21. Yasseri T, Bright J. 2016 Wikipedia traffic data and electoral prediction: towards theoretically informed models. *EPJ Data Sci.* **5**, 22. (doi:10.1140/epjds/s13688-016-0083-3)
22. McIver DJ, Brownstein JS. 2014 Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. *PLoS Comput. Biol.* **10**, e1003581. (doi:10.1371/journal.pcbi.1003581)
23. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, Del Valle SY. 2015 Forecasting the 2013–2014 influenza season using Wikipedia. *PLoS ONE* **11**, e1004239. (doi:10.1371/journal.pcbi.1004239)
24. Parolo PDB, Pan RK, Ghosh R, Huberman BA, Kaski K, Fortunato S. 2015 Attention decay in science. *J. Informetrics* **9**, 734–745. (doi:10.1016/j.joi.2015.07.006)
25. Graham M, Hogan B, Straumann RK, Medhat A. 2014 Uneven geographies of user-generated information: patterns of increasing informational poverty. *Ann. Assoc. Am. Geogr.* **104**, 746–764. (doi:10.1080/00045608.2014.910087)
26. Samoilenko A, Yasseri T. 2014 The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. *EPJ Data Sci.* **3**, 1. (doi:10.1140/epjds20)
27. Yasseri T, Sumi R, Kertész J. 2012 Circadian patterns of Wikipedia editorial activity: a demographic analysis. *PLoS ONE* **7**, e30091. (doi:10.1371/journal.pone.0030091)
28. Kim S, Park S, Hale SA, Kim S, Byun J, Oh A. 2016 Understanding editing behaviors in multilingual Wikipedia. *PLoS ONE* **11**, e0155305. (doi:10.1371/journal.pone.0155305)
29. Adams WC. 1986 Whose lives count? TV coverage of natural disasters. *J. Commun.* **36**, 113–122. (doi:10.1111/j.1460-2466.1986.tb01429.x)
30. Ford H, Sen S, Musicant RD, Miller N. 2013 Getting to the source: where does Wikipedia get its information from? In *Proc. of the 9th Int. Symp. on Open Collaboration, WikiSym + OpenSym*, pp. 1–10. New York, NY: ACM Press.
31. Gleeson JP, Cellai D, Onnela JP, Porter MA, Reed-Tsochas F. 2014 A simple generative model of collective online behavior. *Proc. Natl Acad. Sci. USA* **111**, 10 411–10 415. (doi:10.1073/pnas.1313895111)
32. Ciampaglia GL, Flammini A, Menczer F. 2015 The production of information in the attention economy. *Sci. Rep.* **5**, 9452. (doi:10.1038/srep09452)
33. Müller-Birn C, Karran B, Lehmann J, Luczak-Rösch M. 2015 Peer-production system or collaborative ontology development effort: what is Wikidata? In *Proc. of the Int. Symp. on Open Collaboration*, pp. 20:1–20:10. New York, NY: ACM Press.
34. Muggeo VMR. 2008 Segmented: an R package to fit regression models with broken-line relationships. *R News* **8**, 20–25.
35. Pentzold C. 2009 Fixing the floating gap: the online encyclopaedia Wikipedia as a global memory place. *Mem. Stud.* **2**, 255–272. (doi:10.1177/1750698008102055)
36. García-Gavilanes R, Tsvetkova M, Yasseri T. 2016 Data from: Dynamics and biases of online attention: the case of aircraft crashes. Dryad Digital Repository. (doi:10.5061/dryad.34mn3).