SOCIAL SCIENCES

The memory remains: Understanding collective memory in the digital age

Ruth García-Gavilanes,[1]* Anders Mollgaard,[2]* Milena Tsvetkova,[1] Taha Yasseri[1,3]†

Recently developed information communication technologies, particularly the Internet, have affected how we, both as individuals and as a society, create, store, and recall information. The Internet also provides us with a great opportunity to study memory using transactional large-scale data in a quantitative framework similar to the practice in natural sciences. We make use of online data by analyzing viewership statistics of Wikipedia articles on aircraft crashes. We study the relation between recent events and past events and particularly focus on understanding memory-triggering patterns. We devise a quantitative model that explains the flow of viewership from a current event to past events based on similarity in time, geography, topic, and the hyperlink structure of Wikipedia articles. We show that, on average, the secondary flow of attention to past events generated by these remembering processes is larger than the primary attention flow to the current event. We report these previously unknown cascading effects.

INTRODUCTION

The way individuals collectively remember, forget, and recall events, people, places, etc., has been a prominent topic of research on collective memory (*1*). However, the notion of collective memory as a socially generated common perception of an event itself has been introduced and studied only recently (*2*), about the time when our society started to become highly connected through new channels of communication. Maurice Halbwachs is generally recognized as the father of collective memory research. Halbwachs developed the concept of collective memory, arguing that individual memories are only understood within the context of a group, unifying the nation or community through time and space (*2*). After Halbwachs, different scholars from various academic disciplines have used the concept of collective memory as an interdisciplinary concept. Research on collective memory is often based on theoretical concepts, the study of historical and archival sources, oral histories, case studies, interviews, surveys, and discourse analysis (*3*). For example, one group of researchers carried out several interviews to investigate the possible narrative template of younger and older American adults for three wars, namely, the Civil War, World War II, and the Iraq War. Although Americans of different ages recalled similar events, the interpretation of some events changed over the generations: Both younger and older adults recalled the bombings of Hiroshima and Nagasaki; however, they differed in how they rated the bombings (*4*).

More recently, memory study scholars tend to stress the significance of the media in shaping collective memories: "Culture and individual memory are constantly produced through, and mediated by, the technologies of memory. The question of mediation is thus central to the way in which memory is conceived in the fields of study of visual culture, cultural studies and media studies." (*5*). Under this perspective, offline research methodologies often involve the hiring of coders for content analysis of news and the use of surveys or interviews for analyzing the public memory agenda. For example, a group of researchers (*6*) compared "media memory agenda" and "public memory agenda" to understand the influence of the media on the shaping of collective perceptions of the past by asking coders to analyze the content of the news and request the public to fill in surveys. Alternatively, scholars have studied the role of journalists as collective memory agents by manually analyzing the stories journalists tell as professionals and the stories they tell about their profession (*7*). In all cases, most of the research methodologies applied on memory studies rely on long and costly procedures.

However, developments in digital technologies in recent years have significantly influenced how we keep track of events both as individuals and as a collective. Digital technologies have also provided us with huge amounts of data, which researchers are already using to study different aspects of our social behavior utilizing automatic procedures on much larger samples of data.

"The Internet doesn't forget." On the one hand, the Internet has had strong impacts on memory and the processes of remembering and forgetting, and on the other hand, it has converted collective memory into an observable phenomenon that can be tracked and measured online at scale. Analyzing different Web documents, researchers have shown that more recent past events are remembered more vividly in the present. For example, previous studies (*8*, *9*) investigated news corpora and concluded that most of the temporal expressions are from the near past. Campos *et al.* (*10*) analyzed 63,000 Web query logs and found that 10% had temporal references, mostly to the near past or future. Further, Jatowt *et al.* (*11*) studied how microbloggers collectively refer to time and found that although several posts are about past events, the "here and now" is what they mostly refer to and care about.

Aiming to enhance our knowledge of online collective memory, we use pageview logs of articles on Wikipedia, the largest online encyclopedia. These data provide remarkable granularity and accuracy to study online memory. There is a high correlation between search volume on Google and visits to Wikipedia articles related to the search keywords (*12*, *13*). This indicates that Wikipedia traffic data reliably reflect the Internet users' behavior in general. The high response rate and pace of coverage in Wikipedia in relation to breaking news (*14*, *15*) are features that make Wikipedia a good research platform to address questions related to collective memory.

Other researchers have previously used Wikipedia to study collective memory. In particular, Ferron *et al.* (*16*–*18*) thoroughly studied editors' behavior to confirm the interpretation of Wikipedia as a global memory place. They explored edit activity patterns with regard to commemoration processes, the sentiments of edits in old and recent traumatic and non-traumatic events, and the evolution of emotions in talk pages. However, these studies focused only on editorial activities in Wikipedia; only a few studies address collective memory considering Wikipedia visitors and

[1]Oxford Internet Institute, University of Oxford, Oxford, U.K. [2]Niels Bohr Institute, Copenhagen, Denmark. [3]Alan Turing Institute, London, U.K.
*These authors contributed equally to this work.
†Corresponding author. Email: taha.yasseri@oii.ox.ac.uk

their patterns of attention. For example, Yucesoy and Barabási (*19*) used Wikipedia viewership data to study the popularity and fame of current and retired elite athletes and found that performance dictates visibility and memory. More specifically, Kanhabua *et al.* (*20*) tackle remembering signals using pageviews in Wikipedia to identify factors for memory triggering. They calculate a remembering score made up of different combinations of time series analysis techniques and study how the score varies with regard to time and location. However, this work is limited to empirical observations and fails to give any general understanding of the phenomenon.

Several other prediction tasks have been done using Wikipedia data and metadata. For example, researchers have used Wikipedia viewership data to predict movie box office revenues (*21*), stock market moves (*22*), electoral popularity (*23*, *24*), and influenza outbreaks (*25*, *26*). Further, researchers have predicted the click-through rate between Wikipedia pages, which enables determination of which existing and potential Wikipedia links are useful. They performed this analysis using Web server logs (*27*) and navigational paths (*28*). Researchers have also used pageview counts to predict the dynamics of Wikipedia pages. For example, Thij *et al.* (*29*) predicted that the attention to promoted content on Wikipedia decays exponentially over time.

Using Wikipedia viewership data, we study how new events trigger a flow of attention to past events, which is how we operationalize collective memory. We limit our focus to aircraft incidents and accidents as reported in English Wikipedia, which is the largest language edition of the online encyclopaedia. We quantify and model the attention that flows from articles about recent accidents to articles about past accidents and study the effect of different dimensions of the event on the distribution of attention flow.

RESULTS

To calculate the effect of a new event on the attention to a past event, we pair the pageview time series of the corresponding Wikipedia articles. Here, we focus on all aircraft incidents or accidents reported in English Wikipedia. We call the events that occurred within the period 2008–2016 as "source events" and their Wikipedia articles as "source articles." We pair the source events with older aircraft incidents or accidents, called "target events," and their Wikipedia articles, called "target articles" (see Materials and Methods).

View flow

As an example, Fig. 1A shows the flow of attention from the Germanwings Flight 9525 accident to the American Airlines Flight 587 accident represented by the viewership time series of their corresponding Wikipedia articles. The Germanwings accident occurred on 24 March 2015, when the copilot deliberately crashed the plane into a mountain in the Alps, thereby killing 150 people. The American Airlines accident took place in November 2001 and was caused by a pilot error, which resulted to the plane crashing into the Bell Harbor neighborhood outside New York, thereby killing 265 people. We see an increase in the views to the American Airlines Flight 587 article on the day of the Germanwings crash and this lasted for several days. Note that there was no Wikipedia hyperlink between the two articles during this period. The area of the shaded region measures the increase of the views to the target article relative to the average daily views of the previous year (Fig. 1A, dashed line), called "prior activity." We refer to this area as the view flow, and it will be the central variable of interest in our study. The view flow is calculated over the week after the first edit of the source article. In particular, we focus on the first week where the attention is expected to be maximal (*30*). Note that any area below the dashed line will count negatively, so the view flow can theoretically be negative as well.

Our data set includes 84,761 pairs of source and target events (see Materials and Methods). In Fig. 1B, we show the view flow from the 98 source events (vertical axis) to all 123 target events from 2000 to 2007 (horizontal axis). We notice that some source events trigger a strong view flow on many target events, whereas others have triggering effect on only few or no target events. In the following section we analyze the influence of a range of factors on the view flow between pairs.

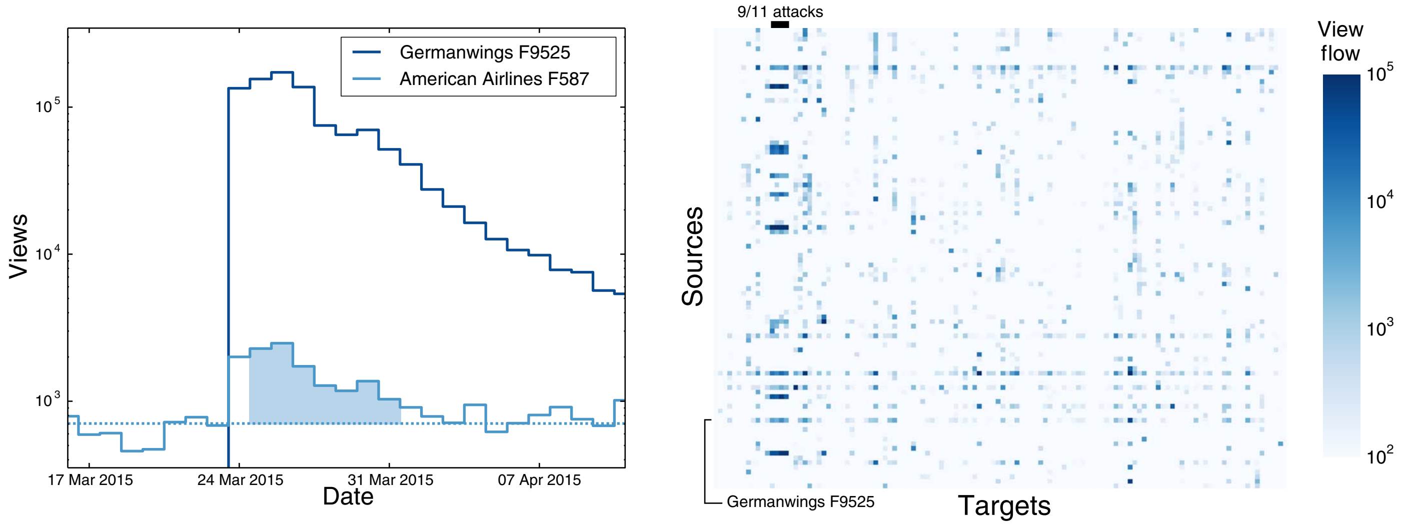

Fig. 1. View flow. Left: Daily Wikipedia article view count on a logarithmic scale for the Wikipedia articles representing Germanwings Flight 9525 (source) and American Airlines Flight 587 (target). The colored area measures the increase in views relative to the daily average of the previous year (dashed line). **Right:** View flow from 98 sources (2008–2016) to all 123 target events from the period 2000–2007. The color of the pixels shows the strength of the view flow on a logarithmic scale. Both axes are sorted according to the date of the accident such that going down or going right brings the reader to more recent events. Some source events, like Germanwings Flight 9525 (see pointer), trigger a lot of target events. We also point to the articles for the 9/11 crashes, which are triggered often and always in unity.

Triggering factors

Here, we limit the analysis to the 11 largest sources (9823 source-target pairs) because the view flow of smaller sources is dominated by the natural noise of the targets (see Materials and Methods). All error bars presented in this section reflect SDs due to sampling error, which are computed using bootstrapping. Presented *P* values test the hypothesis that the mean of population 2 is larger than the mean of population 1. These are likewise computed using bootstrapping (10,000 samplings). Mann-Whitney *U* tests have been performed on all presented population pairs and generally yield *P* values below 10^{-9}.

Figure 2A shows the average view flow for different groups of source-target article pairs. As expected, we find that target articles about recent events are triggered much more often than those about older events ($P = 0.000$). We find that the number of deaths in the target event has an impact: events with more casualties are more likely to be triggered ($P = 0.000$). We also find that the previous viewership of the target articles has a very large impact on the flow of views ($P = 0.000$).

We find very little impact from the location of the operating company of the target flight, namely, whether it is western (North American) ($P = 0.225$) or whether it shares the continent with the source flight ($P = 0.282$). We further check the effect of having the target and source articles appear in a common Wikipedia category, as an indication of similarity (see Materials and Methods). We find that a shared category has a very large impact on the view flow ($P = 0.000$). Finally, we check whether there has been a link from the source article to the target article during any of the 7 days under study. We observe that a direct hyperlink has a huge impact on the viewership flow ($P = 0.000$). However, by removing all linked pairs (75 pairs) and performing the same analysis, we get the same qualitative findings (see Fig. 2B), except that western companies now are triggered significantly more often ($P = 0.020$). The average view flow only drops by 32%, thereby showing that links are not the main driving force responsible for view flow.

Up to this point, we analyzed the view flow considering all variables as binary; but we can get even better resolution with analysis of numbers of years of separating events, numbers of deaths, and prior activity. In Fig. 3A, we show the view flow as a function of years of separation between the source and target events. Although the error bars are rather large, it is clear that there is a strong drop in view flow over the first 45 years. In Fig. 3B, we show the view flow as a function of numbers of deaths involved in the target event. As expected, there is increased view flow with large numbers of deaths, but surprisingly, there is greater view flow to target articles about events with no deaths compared to those with small numbers (~20) of deaths. The average view flow drops from 1112 ± 242 for events with zero deaths to 159 ± 37 for target events of ~20 deaths. One possible explanation is that events with zero deaths are reported in Wikipedia because they are remarkable in some other way. Hijackings are a major contributor, but there are other examples, such as the 1940 Brocklesby mid-air collision, where two planes collided mid-air and got locked together but still managed to land safely. In Fig. 3C, we present (in log-log scale) the view flow as a function of the prior activity of the target article, again measured 1 year before the source event. The trend nicely follows the fitted power law $Ce^{\alpha x}$, with $C = 2.19 \pm 0.24$ and $\alpha = 1.23 \pm 0.03$. The goodness of fit is $R^2 = 0.999$, whereas a linear fit only yields $R^2 = 0.737$.

Although the source articles, combined, received 7.4 million views during their respective first weeks, we estimate the combined view flow to all the target articles to be 10.5 million. The ratio between the two is 1.42 ± 0.26, thereby indicating that the flow of attention is, on average, greater than the attention received by the main event itself. If we remove all linked pairs, then we are still left with a ratio of 0.96 ± 0.24. These results tell us that view flow is not a minor player in attention dynamics, but rather a driving force, at least in Wikipedia, even if we cannot generalize this finding to the whole Internet.

Modeling remembering

In the previous section, we showed that online views of two different articles can be strongly coupled. Therefore, one cannot describe the attention to a topic as an isolated phenomenon. We will now model the

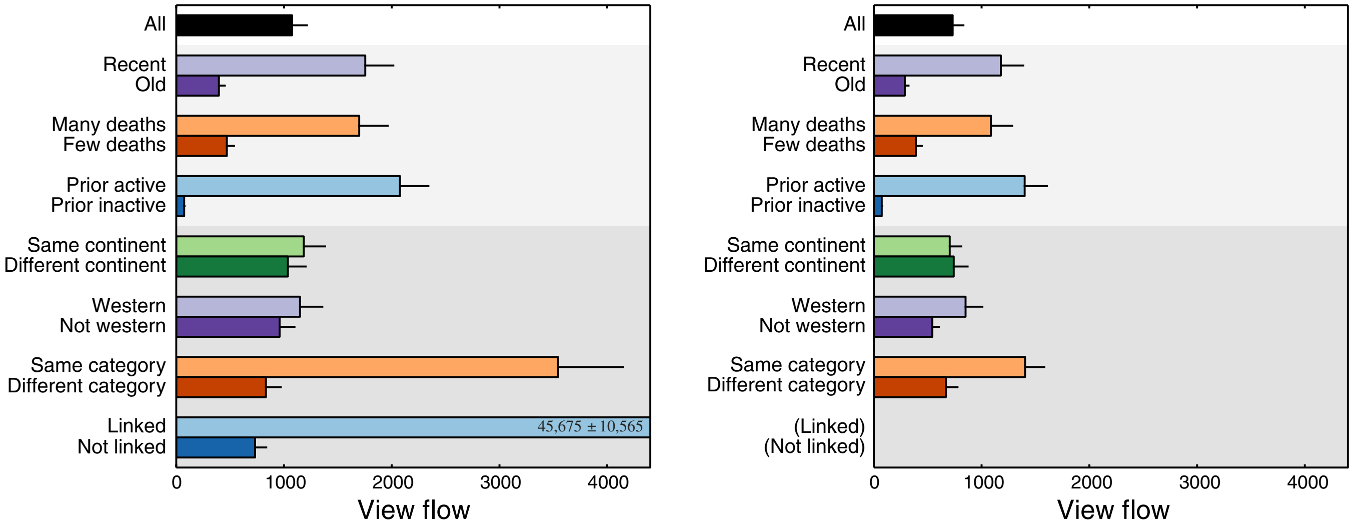

Fig. 2. Triggering factors for view flow. Left: Average view flow among pairs belonging to different groups according to different factors. The black bar labeled "All" includes all pairs, the bars labeled "Recent" and "Old" split the source-target pairs into those that are separated by more or less than 29 years (the median separation between pairs). The bars "Many deaths" and "Few deaths" split the pairs according to the number of deaths of the target event (at the median value of 22 deaths). The next two bars split the pairs according to the prior activity of the target article. The bins in dark gray area are based on whether the source and target flights were operated by companies located on the same continent, whether the operating company is located in Europe, Australia, or North America (Western), whether the source and target articles belong to the same article categories, and whether there is a direct hyperlink from the source article to the target article. **Right:** The same as in the left panel, but pairs with a hyperlink from source to target have been removed from the sample.

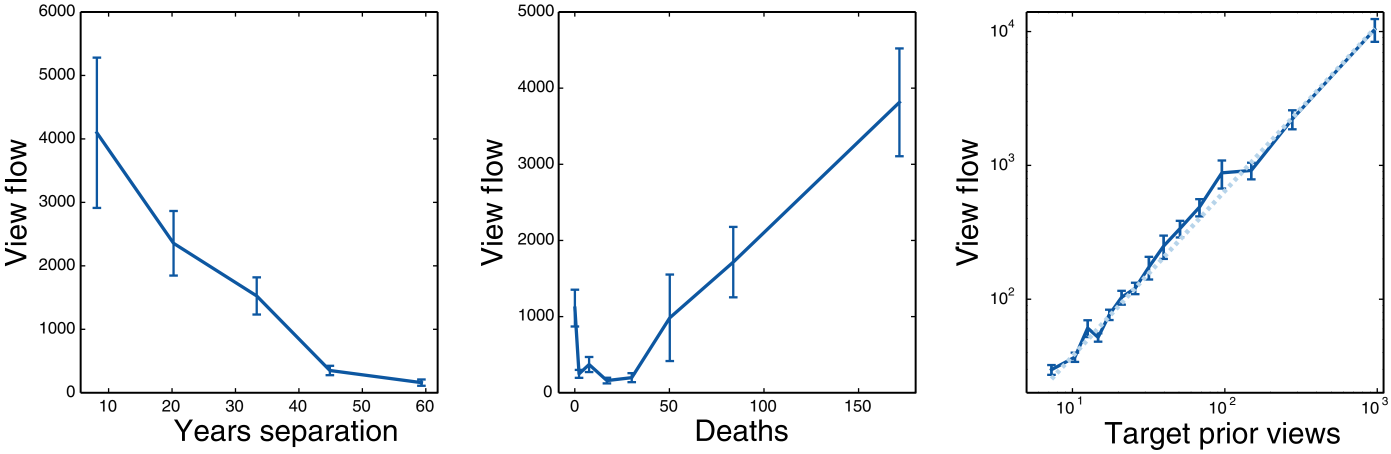

Fig. 3. Detailed analysis of triggering factors for view flow. Left: Average view flow against the separation in years between source and target event. **Center:** Average view flow against the number of deaths involved in the target event. **Right:** Average daily views of the target article during the year before the source event. A power law fit with an exponent of 1.23 is also shown.

coupling between the source and target and thereby show that a big fraction of the target views may be explained from the source views alone. More formally, we aim to predict the views of a target article y based on the views of a source article x and a number of factors that couple the two. The goal is to maximize the coefficient of determination in predicting y. We introduce a model with three terms

$$y = y_{\text{offset}} + y_{\text{link}} + y_{\text{triggered}} \tag{1}$$

The first term, y_{offset}, comes from the fact that some target articles receive more attention than others on average. We model this as $y_{\text{offset}} = a_{\text{history}} \cdot y_{\text{history}}$, where y_{history} is the average weekly views for the previous year. With this term alone, we are able to explain 24 ± 9% of the variance among the target views. The estimate is based on a fivefold cross-validation, and the error bars are given by the spread in the results of the five samplings. We then include view flow mediated by links (y_{link}) in the model. To do this, we estimate the number of views to the source article that is exposed to a link to the target article and call this variable "exposure to target link" represented as x_{link} (see Materials and Methods). We then model link flow as $y_{\text{link}} = a_{\text{link}} \cdot x_{\text{link}}$, which in combination with y_{offset} allows us to explain 30 ± 8% of the variance in the views among the target articles.

The final term in the model, $y_{\text{triggered}}$, represents triggering of memory. Three conditions must be met for a source event to trigger the memory of a target event. First, one must hear about the source event, and second, one must already have the target event stored in his or her long-term memory. Finally, the coupling between the two events needs to be sufficiently strong to trigger the memory. We expect the number of people who hear about the source event to be proportional to the number of views to the source article, which we name x. Likewise, we expect the number of people who have the target event stored in their memory to be proportional to the previous average views of the corresponding target article, y_{history}. Finally, there is the coupling between the two events α, which is the probability that hearing about the source event will trigger the memory of the target event. The first-order approximation of the triggered views can then be written as

$$y_{\text{triggered}} = x \times y_{\text{history}} \times \alpha \tag{2}$$

For simplicity, we model the coupling using a linear combination of the remaining variables (indexed as z_i)

$$\alpha = \sum_i a_i z_i + a_0 \tag{3}$$

Here, we have not included geographical variables, which proved to be negligible in the above analysis. Instead, we have used information regarding years of separation, numbers of deaths in the target event, shared Wikipedia category (0 or 1), and target article link (0 or 1). By including the triggering term in the model, we increase the explained variance from 30 ± 8% to 35 ± 11% (see Materials and Methods for parameter values).

DISCUSSION

We introduced "view flow," or the attention to an old topic induced by a new topic, as a quantitative measure of remembering. We then used this measure to study the factors of remembering for the case of aircraft accidents and incidents, using data from Wikipedia. In particular, we studied how time, similarity, geography, previous attention, and links affect the view flow from a source event to a target event. We found that the memory of an aircraft incident effectively lasts around 45 years. This 45-year limit might reflect the fact that people who were adults at the time of the accident might not use Wikipedia, or may have died in the interim or simply may have forgotten about the accident during that time interval, such that only written records are left in the end. Incidents with either many (50+) or no deaths are remembered the most on Wikipedia. The latter result may be explained by a bias in Wikipedia, which tends to keep records of "no death" incidents only if they are remarkable in some other way. If we reinsert the data points into the fitted model (Eq. 1), then we reproduce the U shape, despite the fact that the model is linear in the death variable. This indicates that the U shape is not inherent to the deaths but is rather contained in the other variables.

Generally, we do not find that geographical similarity has any significant impact on remembering aircraft incidents, even though the level of attention paid to individual incidents is considerably driven by location (30). Links were found to greatly increase the view flow between source

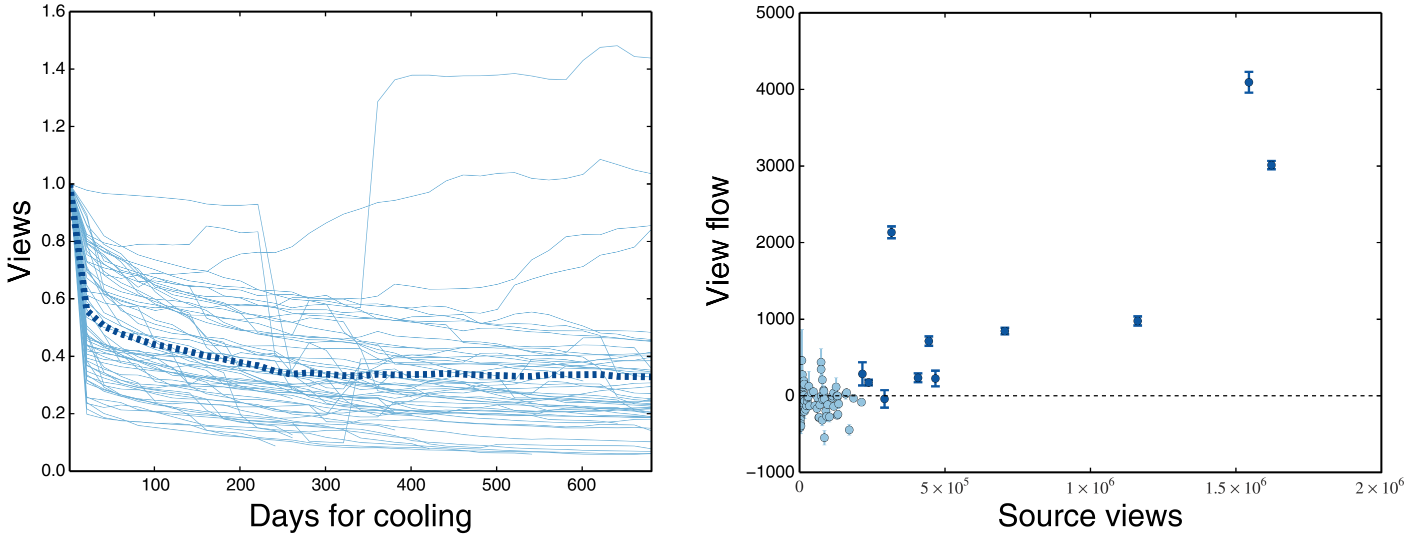

Fig. 4. Filtering. Left: View curves of all articles from 2008 to 2016. The vertical axis measures the average views during the 1-year sampling period. The horizontal axis determines the period used for sampling. Specifically, it determines the days of separation between the incident and the beginning of the sampling period. Note that the computation relaxes after approximately 1 year of cooling. We therefore require that sources and targets are separated by at least 2 years because this ensures that the target has relaxed before sampling its average view rate. **Right:** We show the average view flow from each source against the views of that source during the same period. The dark dots represent the 11 largest source events, which were used in Triggering factors: Malaysia Airlines flight 370, Malaysia Airlines flight 17, Air France flight 447, Germanwings flight 9525, 2010 Polish Air Force Tu-154 crash, Indonesia AirAsia flight 8501, Asiana Airlines flight 214, 2011 Lokomotiv Yaroslavl air disaster, Metrojet flight 9268, and Colgan Air Flight 3407. The smaller source events (light blue) have not been included because noise dominates in this region.

and target, but because they are only present for a small fraction of source-target pairs, they cannot explain most of the observed view flow. It is worth mentioning that the flow between pairs without a direct Wikipedia link remains as an open question because our data do not provide any explanation of the underlying mechanisms. The reported flow could be mediated by the external channels online or offline. Of more general importance is the previous attention to the target article, which has a superlinear effect on the view flow. This shows that regardless of the strength of the coupling between events, some past events are consistently more memorable. The view flow is especially strong when the source and target are similar in some way, as measured by a shared Wikipedia category.

Overall, we find that a source event induces a combined view flow, which on average is ~142% of the views given to the source event itself. This tells us that view flow is a major force that should not be ignored. Interaction between signals has previously been studied in economics (*31*) however, signal interactions are not included in current models of social spreading. The typical approach in previous studies is to make predictions for the popularity of a topic based on the recent history of that topic alone (*32*–*35*). Future models should not consider spreading phenomena as stand-alone objects but should also account for cross-correlations. Concepts, ideas, videos, and so on are not stand-alone objects but instead form a large network with attention flowing from one to another.

We made a first attempt to model remembering. We proposed to model remembering with a product between the current attention to the source event, the previous attention to the target event, and a coupling between the two. The rationale behind this model is threefold. To trigger the memory of the target event, one must hear about the source event, the target event must already be stored in long-term memory, and the coupling between the two events must be large enough to trigger the memory. Our model allowed us to explain 35% of the variance in views among the Wikipedia articles about target events. Note that no information regarding the internal dynamics of the target article views was used to produce this result. A big limitation of our model is the linear expression for coupling between articles, which, if improved, might allow much more variance to be explained. Furthermore, we do not account for any spreading processes induced by the triggering of memory. These processes might be responsible for the superlinear relationship observed in Fig. 3C.

In summary, we argue that the flow of attention between different events and concepts is mediated by memory or, more generally, associativity. We find that source events generate a flow of attention to previous events, which is even greater than the attention given to the source itself. A first model to explain remembering in the case of airline crashes has been provided. The theoretical framework and the mathematical formulation in Eq. 2 can be easily generalized to explain collective online memory in a broader context, whereas the coupling α must be modeled to fit the particular setting.

MATERIALS AND METHODS

Data collection

We collected data from Wikipedia using two main sources: MediaWiki API and Wikidata, using https://cran.r-project.org/web/packages/WikidataR/index.html. The MediaWiki API is a Web service that provides access to wiki features, data, and metadata of articles such as links and categories. On the other hand, Wikidata is a Wikipedia partner project that aims to store structured data from other Wikimedia projects, including Wikipedia, and fix inconsistencies across different editions (*36*). Examples of such structured data include the date or geographical coordinates of an event.

We focused on a set of articles in English Wikipedia in the categories "aviation accidents and incidents by country" and "aviation accidents and incidents by year" and their subcategories. These categories cover all airline accidents and incidents in different countries and throughout history that are available in English Wikipedia. Using the MediaWiki API, we obtained 1606 articles

Table 1. Model parameters. Least square fit of the parameters in the model to the data. Error bars are estimated using bootstrapping.

a_{history}	a_{link}	a_{deaths}	a_{years}
0.83 ± 0.04	0.05 ± 0.02	$3.3 \times 10^{-9} \pm 2.2 \times 10^{-9}$	$-2.0 \times 10^{-8} \pm -1.5 \times 10^{-8}$
a_{category}	a_{linked}	a_0	
$0.0 \times 10^{-7} \pm 8.7 \times 10$	$8.2 \times 10^{-6} \pm 3.9 \times 10^{-6}$	$1.5 \times 10^{-6} \pm 0.8 \times 10^{-6}$	

from which 1496 are specifically about aircraft crashes or incidents (we discarded articles of biographies, airport attacks, etc.). Furthermore, we extracted editorial information for the articles in the data set: the date when the article was created, the alternative names for the article through time, and the article links and categories. We collected the links from the page history for the 7 days after the first edit of the article and for each link, and we calculated the fraction of the day that it remained in the article. For the 1496 articles, we systematically collected structured data from Wikidata: the date of the event, geographical coordinates of where the event occurred, number of deaths, and the continent of the aircraft company. Unfortunately, Wikidata did not have complete information for all articles. To remedy this deficiency, we obtained the missing data by automatically crawling Wikipedia infoboxes, using https://cran.r-project.org/web/packages/WikipediR/index.html, or manually checking the information in the articles.

Finally, we extracted the daily traffic to the articles between 01 January 2008 and 10 April 2016 from the Wikipedia pageview dumps available at https://dumps.wikimedia.org/other/pagecounts-raw/ through a third-party interface, http://stats.grok.se. There are no data available before this interval. We used the alternative title names of each article to merge all traffic statistics to the current title. Article views have been normalized according to the global traffic of English Wikipedia, which is available at https://tools.wmflabs.org/, such that all views correspond to the January 2015 values.

Sampling

The source articles were created in Wikipedia within the period 01 January 2008 and 10 April 2016. These articles have viewership data available from the moment they were created to the last day of the period. To capture the immediate attention to a source event right after its occurrence, we chose the corresponding source articles created up to 1 day after the source event. Furthermore, we removed small source events that are proximate to large source events. This was done to avoid false positives, that is, small sources that are credited with the view flow triggered by large source events. We defined proximity by a 10-day range because the main attention of an aircraft event has been shown to decay over this time scale (*30*). The process of removing false positives was performed as follows: (i) we sorted all the source articles by their total number of views during the first week in Wikipedia; (ii) starting from the article with the most views, we removed all source articles that were created within a 10-day range; and (iii) we continued with the next article with the highest views and repeated (ii) and so on. In the end, this procedure left us with 98 source events.

We then paired each one of the 98 source events with target events from our entire data set such that each of the target events occurred at least 2 years before the source event. This assured that the views of the target article had at least 1 year to stabilize such that the calculation of the average views before the source event is representative. In the left panel of Fig. 4, we justified the 1-year stabilization period by showing that the view average stabilizes after approximately 300 days. The 2-year separation criterion reduced the number of source-target pairs from 189,430 to 144,773.

In Triggering factors, we restricted our study to the 11 largest sources with the argument that the noise of the view flow is comparable to the signal for the smaller sources. We illustrate this in Fig. 4, which shows the average view flow from any target to its sources. The error bars were computed as the spread in the target views during the year before the source event. We removed all sources below the 11th largest source because the natural noise of the target views dominates in this region, as illustrated in Fig. 4. Our analysis has also been applied to the complete set of sources, which yielded the following changes to the results: The average view flow is a full scale smaller, and the noise is a bit more dominant. The effects of categories and geography are enhanced, whereas the superlinear effect of the previous views on the view flow almost disappears.

Category similarity

Categories in Wikipedia form a pseudohierarchical structure, and their function is to group other regular Wikipedia articles to a common subject (*37*). In general, categories are socially annotated, and editors can classify an article into a category simply by appending one to it. The categories appended to a Wikipedia article are generally found at the bottom of it. Here, we considered the common categories among target and source Wikipedia articles as a similarity feature.

Hyperlinks

In the context of this project, hyperlinks are internal links in Wikipedia linking a page to another page within English Wikipedia. The blue hyperlinks are an essential feature in Wikipedia because an article can often only be understood in the context of related articles, and internal links make it easy to explore this context (*28*). Here, we predicted the views of the source article flowing to the target article due to an internal link in the source article. To do this, we used "exposure to target link" (x_{link}) as an independent input variable for predicting the views of the target article. The variable was calculated using the revision histories of the source articles, which allowed us to track the fraction of a given day with an internal link to the target article. We then constructed x_{link} by multiplying this fraction with the number of views of the source article in that day. In the prediction model, we added the resulting number of views for all the days considered in the prediction, which, in this case, is 7 days after the source article was created.

Parameter values

In Table 1, we show the fitted parameter values with error bars estimated from 10,000 bootstrapping samples. The a_{linked} parameter is part of the coupling constant and should not be confused with a_{link}, which is in the y_{link} term.

REFERENCES AND NOTES

1. H. Ebbinghaus, *Über das Gedächtnis Untertitel: Untersuchungen zur experimentellen Psychologie* (Duncker & Humblot, 1885).
2. M. Halbwachs, L. A. Coser, *On Collective Memory* (University of Chicago Press, 1992).
3. T. E. Bosch, "Memory studies," Working paper, Media, Conflict and Democratisation (MeCoDEM, Leeds, U.K.), 2016.
4. F. Zaromb, A. C. Butler, P. K. Agarwal, H. L. Roediger III, Collective memories of three wars in united states history in younger and older adults. *Mem. Cognit.* **42**, 383–399 (2014).
5. M. Sturken, Memory, consumerism and media: Reflections on the emergence of the field. *Mem. Stud.* **1**, 73–78 (2008).
6. N. Kligler-Vilenchik, Y. Tsfati, O. Meyers, Setting the collective memory agenda: Examining mainstream media influence on individuals' perceptions of the past. *Mem. Stud.* **7**, 484–499 (2014).
7. O. Meyers, Memory in journalism and the memory of journalism: Israeli journalists and the constructed legacy of *haolam hazeh*. *J. Commun.* **57**, 719–738 (2007).
8. C.-m. Au Yeung, A. Jatowt, Studying how the past is remembered: Towards computational history through large scale text mining, in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11)*, pp. 1231–1240.
9. A. Jatowt, C.-m. Au Yeung, K. Tanaka, Estimating document focus time, in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM'13)*, pp. 2273–2278.
10. R. Campos, L. I. Porto; Center for Human Language, What is the temporal value of web snippets, *The 1st International Temporal Web Analytics Workshop of the 20th International World Wide Web Conference (TWAW'11)*, pp. 9–16.
11. A. Jatowt, E. Antoine, Y. Kawai, T. Akiyama, Mapping temporal horizons: Analysis of collective future and past related attention in Twitter, in *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*, pp. 484–494
12. J. Ratkiewicz, A. Flammini, F. Menczer, Traffic in social media I: Paths through information networks, in *Proceedings of the 2010 IEEE Second International Conference on Social Computing (SOCIALCOM'10)*, pp. 452–458.
13. M. Yoshida, Y. Arase, T. Tsunoda, M. Yamamoto, Wikipedia page view reflects web search trend, in *Proceedings of the 2015 ACM Web Science Conference (WebSci'15)*, pp. 65:1–65:2.
14. T. Althoff, D. Borth, J. Hees, A. Dengel, Analysis and forecasting of trending topics in online media streams, in *Proceedings of the 21st ACM International Conference on Multimedia (MM'13)*, pp. 907–916
15. B. Keegan, D. Gergle, N. Contractor, Hot off the wiki: Dynamics, practices, and structures in Wikipedia's coverage of the Tōhoku catastrophes, in *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym'11)*, pp. 105–113.
16. M. Ferron, P. Massa, Collective memory building in Wikipedia: The case of North African uprisings, in *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym'11)*, pp. 114–123.
17. M. Ferron, P. Massa, Studying collective memories in Wikipedia. *J. Soc. Theory* **3**, 449–466 (2011).
18. M. Ferron, P. Massa, Psychological processes underlying Wikipedia representations of natural and manmade disaster, in *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration (WikiSym'12)*, pp. 2:1–2:10.
19. B. Yucesoy, A.-L. Barabási, Untangling performance from success. *EPJ Data Sci.* **5**, 17 (2016).
20. N. Kanhabua, T. N. Nguyen, C. Niederée, What triggers human remembering of events?: A large-scale analysis of catalysts for collective memory in Wikipedia, in *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL'14), pp. 341–350.
21. M. Mestyán, T. Yasseri, J. Kertész, Early prediction of movie box office success based on wikipedia activity big data. *PLOS ONE* **8**, e71226 (2013).
22. H. S. Moat, C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, T. Preis, Quantifying Wikipedia usage patterns before stock market moves. *Sci. Rep.* **3**, 1801 (2013).
23. T. Yasseri, J. Bright, Can electoral popularity be predicted using socially generated big data? *Info. Tech.* **56**, 246–253 (2014).
24. T. Yasseri, J. Bright, Wikipedia traffic data and electoral prediction: Towards theoretically informed models. *EPJ Data Sci.* **5**, 22 (2016).
25. D. J. McIver, J. S. Brownstein, Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLOS Comput. Biol.* **10**, e1003581 (2014).
26. K. S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J. M. Hyman, A. Deshpande, S. Y. Del Valle, Forecasting the 2013–2014 influenza season using Wikipedia. *PLOS ONE* **11**, e1004239 (2015).
27. A. Paranjape, R. West, L. Zia, J. Leskovec, Improving website hyperlink structure using server logs, in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM'16)*, pp. 615–624.
28. R. West, A. Paranjape, J. Leskovec, Mining missing hyperlinks from human navigation traces: A case study of Wikipedia, in *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*, pp. 1242–1252.
29. M. t. Thij, Y. Volkovich, D. Laniado, A. Kaltenbrunner, http://arXiv:1212.5943 (2012).
30. R. García-Gavilanes, M. Tsvetkova, T. Yasseri, Dynamics and biases of online attention: The case of aircraft crashes. *R. Soc. Open Sci.* **3**, 160460 (2016).
31. L. Kristoufek, What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis. *PLOS ONE* **10**, e0123923 (2015).
32. H. Pinto, J. M. Almeida, M. A. Gonçalves, Using early view patterns to predict the popularity of youtube videos, in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM'13)*, pp. 365–374.
33. O. Tsur, A. Rappoport, What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities, in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM'12)*, pp. 643–652.
34. A. Mollgaard, J. Mathiesen, Emergent user behavior on Twitter modelled by a stochastic differential equation. *PLOS ONE* **10**, e0123876 (2015).
35. Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, C. Faloutsos, Rise and fall patterns of information diffusion: Model and implications, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*, pp. 6–14.
36. C. Müller-Birn, B. Karran, J. Lehmann, M. Luczak-Rösch, Peer-production system or collaborative ontology engineering effort: What is Wikidata?, in *Proceedings of the 11th International Symposium on Open Collaboration (OpenSym'15)*, pp. 20:1–20:10.
37. A. Kittur, E. H. Chi, B. Suh, What's in Wikipedia?: Mapping topics and conflict using socially annotated category structure, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*, pp. 1509–1512.

Acknowledgments: We thank the Wikimedia Foundation Inc. for making data dumps of Wikipedia freely available and for building an application program interface that provides convenient access to wiki features, data, and metadata. **Funding:** This research is part of the project Collective Memory in the Digital Age: Understanding Forgetting on the Internet funded by Google. T.Y. has been supported by the Alan Turing Institute under the Engineering and Physical Sciences Research Council grant EP/N510129/1. **Author contributions:** R.G.-G. collected and analyzed the data, participated in the design of the study, and drafted the manuscript; A.M. analyzed 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. designed and coordinated the study and helped draft the manuscript. All authors gave final approval for publication. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper. Additional data related to this paper may be requested from the authors.

Submitted 26 September 2016
Accepted 10 February 2017
Published 5 April 2017
10.1126/sciadv.1602368

Citation: R. García-Gavilanes, A. Mollgaard, M. Tsvetkova, T. Yasseri, The memory remains: Understanding collective memory in the digital age. *Sci. Adv.* **3**, e1602368 (2017).

Science Advances

The memory remains: Understanding collective memory in the digital age

Ruth García-Gavilanes, Anders Mollgaard, Milena Tsvetkova and Taha Yasseri

Sci Adv **3** (4), e1602368.
DOI: 10.1126/sciadv.1602368

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.