Lost in Propagation? Unfolding News Cycles from the Source

Chenhao Tan
Department of Computer Science
Cornell University
Ithaca, NY
chenhao@cs.cornell.edu

Adrien Friggeri
Facebook
1 Facebook way
Menlo Park, CA
friggeri@fb.com

Lada A. Adamic
Facebook
1 Facebook way
Menlo Park, CA
ladamic@fb.com

Abstract

The news media play an important role in informing the public on current events. Yet it has been difficult to understand the comprehensiveness of news media coverage on an event and how the reactions that the coverage evokes may diverge, because this requires identifying the origin of an event and tracing the information all the way to individuals who consume the news. In this work, we pinpoint the information source of an event in the form of a press release and investigate how its news cycle unfolds. We follow the news through three layers of propagation: the news articles covering the press release, shares of those articles in social media, and comments on the shares. We find that a news cycle typically lasts two days. Although news media in aggregate cover the information contained in the source, a single news article will typically only provide partial coverage. Sentiment, while dampened in news coverage relative to the source, again rises in social media shares and comments. As the information propagates through the layers, it tends to diverge from the source: while some ideas emphasized in the source fade, others emerge or gain in importance. We also discover how far the news article is from the information source in terms of sentiment or language does not help predict its popularity.

Introduction

News reaches us via different ways. For instance, a person may learn about a presidential speech by listening to the original speech, or from a friend who comments and links on social media to a news report on the speech. In the latter case, the information has propagated through several channels, which can affect how it is perceived. While information diffusion is an active research area (Bakshy et al. 2011; Gruhl et al. 2004; Lerman and Ghosh 2010; Liben-Nowell and Kleinberg 2008; Simmons, Adamic, and Adar 2011) that ranges from news dynamics (Leskovec, Backstrom, and Kleinberg 2009) to the role of networks (Bakshy et al. 2012), it remains unknown how news may differ from *the source* as a result of different propagation processes.

The news media traditionally act as the first channel between information sources and individuals in the propagation process of current events. They are expected to and strive to provide comprehensive and unbiased coverage. However, there have been concerns that the news media sometimes introduce biases, e.g., through selective coverage. In politics, Puglisi, Snyder, and James showed that democratic-leaning newspapers provide relatively more coverage of scandals involving republican candidates than scandals involving democratic politicians and vice-versa. Similarly, science coverage may be distorted. For example, Saguy and Almeling (2008) found that news media overdramatize studies of obesity and are more likely than the original scientific articles to highlight individual blame for weight.[1] Overall, there is a lack of quantitative understanding of how the news media may report the same event differently and how the difference affects individuals' perceptions. This is partly due to the difficulty of identifying sources of information in general, especially for complex and dynamic events.

Fortunately, presidential speeches, formal statements and press releases from organizations can serve as observable information sources. Although press releases may contain certain biases (e.g., a university press release may overstate the significance of the results in a scientific study, while a government press release may emphasize benefits while omitting less desirable outcomes), they present the most accurate version of the information that news articles cover, from the perspective of the source. In addition, it is now possible to capture individuals' reactions via social media as news production and consumption is happening online. These two data sources offer unique opportunities to trace how information from the source propagates.

We draw inspirations from mass communication models (Lazarsfeld, Berelson, and Gaudet 1968; Katz and Lazarsfeld 1955; Merton 1957) and employ a layered model. The first layer is the initial press release, which we also refer to as the *information source*. Relevant *news articles* on an information source constitute the second layer. The final part of propagation captures the reactions of individuals. Online social media such as Facebook present at least two layers: individuals can share news articles and add text with their *shares*; and then further down the propagation, individuals

[1] Anecdotally, a recent intentionally faulty study was not only published by nominally peer-reviewed scientific journals but also succeeded in spreading the erroneous message that chocolate helps weight loss to millions after news media reported on the story (Bohannon 2015).

Figure 1: An example of word clouds generated from the information source, news articles, shares, comments on President Obama's speech about the deaths of Warren Weinstein and Giovanni Lo Porto (The White House 2015d). Green words are positive, red words are negative according to the LIWC dictionary (Pennebaker, Francis, and Booth 2007). The size of a word represents word frequency. Word clouds were generated using (Mueller 2015), and stopwords were filtered out.

who see these shares can respond with *comments*. These four layers represent different stages of information dissemination from the source to individuals.

This setup resembles a "telephone game" where the four layers act as players. As the information propagates, these layers can present different or even conflicting pictures of the same event. Figure 1 illustrates an example from President Obama's speech on the deaths of Warren Weinstein and Giovanni Lo Porto in a U.S. counterterrorism operation (The White House 2015d). In the original speech, the information source in this case, Obama placed emphasis on "families" and "people," maybe to arouse empathy from the audience, and on "al qaeda" as a common enemy. He avoided using "killed" and called the deaths a "loss." In propagation, news articles brought up the term "drone" and started to use the verb, "killed." When individuals shared news articles about this speech, "hostage" gained more prominence and "war" started to get attention. Finally, in the comments on the shares, "war" and "drone(s)" dominated the conversation. Motivated by such different pictures, we aim to understand how information may diverge from the source at different stages of propagation.

Organization and contributions. In this paper, we present the first large-scale study on the entire propagation process from the source to individuals, to other individuals, through news media and social media. We build a dataset that leverages press releases from various organizations spanning politics, science, technology and finance. After identifying relevant news articles, we analyze, in aggregate, de-identified shares and comments of those articles on Facebook.

With this novel dataset, our first contribution is to uncover how the news cycle develops for an information source by examining the volume of content in the layers. We find that a news cycle usually lasts two days. There are two temporal peaks in news media coverage, the second being aligned with the largest volume of reactions from individuals. We also discover that individuals rarely share the original links of information sources. This confirms the essential role that the news media play in how individuals access information.

Our main contribution is to provide an understanding of how information in the source propagates through the layers. We begin by investigating how closely content from news media mirrors the information sources they cover. We show that quotation is less prevalent in finance and technology compared to politics and science. We further demonstrate that on average a single news article cannot cover the information source although all news articles combined cover all the words that occur in the information source.

We then study the differences in sentiment between layers. Perhaps as expected, news articles tend to have fewer subjective words than information sources do, while shares and comments use subjective words more. We further propose two hypotheses to explain the increasing subjectivity in the content by individuals compared to news articles, and show that the main reason is adding "novel" content instead of magnifying existing subjective parts from news articles. As for positivity, the balance of positive to negative sentiment words decreases with each layer of propagation.

We also study how language differs between layers by focusing on the most frequent words, and how they. We demonstrate that language gets more and more different from the information source in propagation. The increasing distance is related to a concentration of usage on certain words. We further examine how specific words fare by comparing the rank in word frequency across layers. We observe interesting patterns in how some words that information sources emphasize fail to propagate.

Given that news articles usually provide partial coverage of the source and are slightly more negative, our final contribution is to examine whether these factors are correlated with the popularity of a news article. We find that distances from the source in terms of sentiment and language do not improve prediction performance in popularity, showing that articles staying true to the source enjoy no advantage. Although the prediction problem is quite difficult with a low accuracy of 55% if we focus on comparable news domains, an important strategy to gain popularity is to publish the news article early after the source.

Related work

Information diffusion. Among the studies mentioned in the introduction, most relevant to this work are studies of

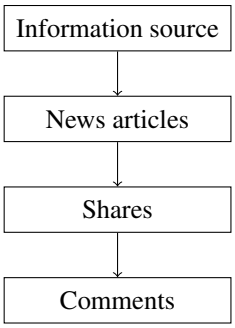


Figure 2: A four-layer structure.

information source	news article
https://www.whitehouse.gov/the-press-office/2015/05/15/remarks-president-national-peace-officers-memorial	http://time.com/3860376/barack-obama-peace-officers-memorial-service/
http://news.berkeley.edu/2015/06/01/alzheimers-protein-memory-loss/	http://www.thehindu.com/todays-paper/tp-in-school/sleep-well-to-avoid-memory-loss/article7275836.ece
http://googleblog.blogspot.com/2015/04/android-wear-wear-what-you-want-get.html	http://www.theverge.com/2015/4/20/8447971/android-wear-update-wifi-support-emoji-smartwatch

Table 1: Samples of information sources and relevant news articles.

how memes evolve as they propagate in mainstream and social media. Leskovec, Backstrom, and Kleinberg proposed a method for tracking temporal dynamics in the volume of quotes in news media and in blogs, and Simmons, Adamic, and Adar subsequently found that the text of quotes was indeed evolving. In comparison with these studies, we focus on an event, identify the information source and investigate how information evolves at different stages of propagation starting from the source.

Media coverage and bias. Media coverage has been studied in politics (Prat and Strömberg 2013) and in science communication (Winsten 1985). Media bias has also been the subject of substantial prior work (Brechman, Lee, and Cappella 2009; Baum and Groeling 2008; Groseclose and Milyo 2005; Lin, Bagrow, and Lazer 2011). The most relevant work is a recent study that examined how different news media quote presidential speeches and demonstrated the existence of systematic biases in quotations (Niculae et al. 2015). In our work, instead of addressing biases of individual news outlets, we demonstrate properties of the entire propagation process of news starting from a press release as the information source.

Role of news media. Understanding how news spreads to the public is a central question in communication (Lasswell 1948). There are two major theories in explaining the role of news media: the "hypodermic needle" model posits that mass media exert direct and relatively strong effects on public opinion; the "two-step flow" model argues that mass media influence the public only indirectly through opinion leaders (Lazarsfeld, Berelson, and Gaudet 1968; Katz and Lazarsfeld 1955; Merton 1957). Recent technology developments enable studies to validate these theories. Wu et al.(2011) provided evidence of the two-step flow model in the context of Twitter.

Data collection

A four-layer structure. In order to understand the propagation from the information source all the way to the individuals who eventually consume and react to the news, we need to study the composition of information at several propagation layers. As discussed in the introduction, we collect data from information sources, news media, and social media.

We start from *press releases* that are issued by organizations that are often subsequently covered by news media. These statements serve as relatively static information sources.[2] To make our findings robust, we extract press releases from March to June in 2015 spanning four topics:

- Politics. The White House publishes speeches and remarks from the President and the Vice President of the US (The White House 2015c). We focus on presidential speeches in this work. This period includes Obama's speeches in Selma in memory of the American Civil Rights Movement, a eulogy in honor of Beau Biden, on Memorial day, etc.
- Science. A common way that news media learn about new scientific discovery is from university press releases. We consider MIT, Stanford and UC Berkeley. For example, our final dataset includes a study on mass extinction and a study on poor sleep and Alzheimer's protein.
- Technology. We collect press releases from big technology companies such as Google (Official Blog), Facebook, and Microsoft, which mostly announce new products and product features.
- Finance. We use statements from the Federal Open Market Committee, which holds regular meetings to discuss monetary policies and releases a press statement after each meeting. As the actions undertaken are consequential, news media report heavily on these statements. There were three statements issued during the time span.

The second layer in our data are relevant news articles for each information source. Through a process described below, we identify relevant news articles that were first shared on Facebook within 7 days of the issuing date of the corresponding information source. As we will show later, 7 days is sufficient for including relevant news articles.

The third layer is shares of relevant news articles on Facebook. When individuals share news articles on Facebook, they can add their own caption, either highlighting a portion of the article by quoting it or expressing their own view or commentary.[3] The final layer is individuals' comments

[2]In our data, the average distance of news articles from the source does not change in the first two days, which supports this assumption. Plots are omitted for space reasons.

[3]We remove sharing text that matches the title of the corre-

posted in reply to these shares of relevant news articles. For shares and comments, we consider data in English within 14 days of the issuing date of the information source so that the last article gets a week to accumulate reactions. All data used were de-identified and analyzed in aggregate.

In summary, we employ a four-layer structure as in Figure 2. Lower layers get further away from information sources. For each information source, we refer to its entire propagation process, including news articles, shares, and comments, as its *news cycle*. We call the propagation between two neighboring layers a *transmission step*. In this paper, each news cycle is treated as a sample. For all text-related computation, stopwords are filtered out.

Identifying relevant news articles. Developing a general system that identifies relevant news articles for all possible information sources is a research problem of its own (Allan 2012). Since this is not the focus of this study, we only keep information sources for which we can identify relevant news articles with high recall and high precision. Our approach consists of two stages: 1) a rough pass through all news articles to collect a set of candidate relevant articles that may oversample; 2) a "semi-supervised" method that combines manual labeling and information retrieval techniques to identify relevant articles.

In the rough pass, we first curate a list of more than 1,800 news media domains worldwide, and only consider news articles within these domains. We deduplicate multiple forms of the same url by matching titles and descriptions within the same domain.[4] For each information source, we manually specify keywords to narrow down an initial set of candidate news articles.[5] For example, the keywords used for the presidential speech in Figure 1 are "deaths, warren, weinstein, giovanni, drone." Information sources that are too general, e.g. Obama's speech on middle class economics (The White House 2015b), were removed because it was difficult to identify unambiguous keywords for them. This procedure ensures that we achieve a *high recall* for selected information sources.

In the second stage, we manually label the 20 most shared news articles for each information source as relevant or not. This manual step offers a *seed set* for filtering irrelevant articles under the assumption that relevant news articles should be more similar to the seed set than irrelevant news articles. We remove information sources that had fewer than 200 shares for their 20 most shared relevant news articles in the manual step, so that we can focus on information sources that got a reasonable number of shares and comments to study the propagation process. We filter the rest of the news articles based on their minimal distance to the *seed set* using tf-idf vectors. We found that using a threshold with limited manual filtering around the boundary is sufficient for identifying relevant news articles with high precision.

Table 1 gives a sample of pairs found with our approach. We manually labeled a random sample of 100 information source and news article pairs and derived a precision of 93%. Note that our approach ensures that the precision is even better for highly shared news articles. Table 2 gives a summary of the final dataset.

Table 2: Dataset summary. "sources" gives the number of press releases from information sources in a topic, while the last three columns give the average number of contributions in the other three layers. For instance, on average, a presidential speech was covered by 185 news articles, which were then shared 46,761 times with 51,142 comments.

domain	sources	articles	shares	comments
all	85	184	22,242	19,971
politics	22	185	46,761	51,142
science	5	87	50,587	33,285
tech	55	195	10,968	7,349
finance	3	126	1,877	598

News cycle of a story

We begin by studying basic properties of the news cycle for an information source. We find that a news cycle typically lasts two days and that individuals learn about the information source primarily *indirectly* from the news media instead of *directly* accessing the information source.

Volume over time. To capture the rate at which news media and individuals react, we compute the fraction of content produced in each hour for news articles, shares, and comments respectively. We do not always have a reliable timestamp of an information source as most press releases only record the dates, but rather use the first share of a relevant news article to Facebook as a proxy. Similarly, we approximate the time when a news article is posted by its first share on Facebook. Since many media outlets share articles to Facebook via their Facebook pages, we expect the share timestamp to be close to the article's.

Figure 3 shows that the news cycle of a press release lasts roughly two days: only a tiny fraction of news articles, shares, and comments are produced thereafter. The two peaks in news articles suggest that some news media may have anticipated or had access to the information source ahead of it becoming publicly available, while other sources may react with delay. Furthermore, news articles, shares, and comments align well in the second peak. This may be related to the "two-step flow" model (Katz and Lazarsfeld 1955) wherein news media and individuals react to news being picked up by opinion leaders. It is worth noting that the trend of volume over time varies between different topics, e.g. volume fluctuates less for university press releases.

Sharing of news media coverage vs. the original source. The news media traditionally play an important role in the public's access to information. Since press releases can be shared as a link in the age of social media, information

sponding news article exactly because this text is likely to be generated by automatic tools for sharing.

[4]We do not deduplicate news articles that were in different domains but originated from the same news wire service such as Associated Press, as this is part of the process that affects individuals' downstream perception.

[5]We define separate sets of keywords for title and description and for the main content in a webpage to achieve better precision. We extract the main content of webpages using (Cuthbertson 2015) to avoid noise such as sidebars and comment sections.

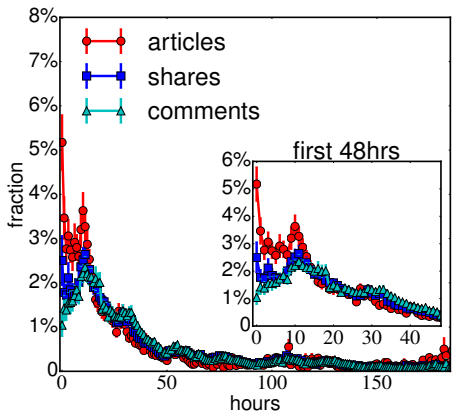


Figure 3: The fraction of total content contributions by hour, for news articles, shares, and comments, as a function of the number of hours elapsed since any news article relating to an information source was shared. In all figures, error bars represent standard errors.

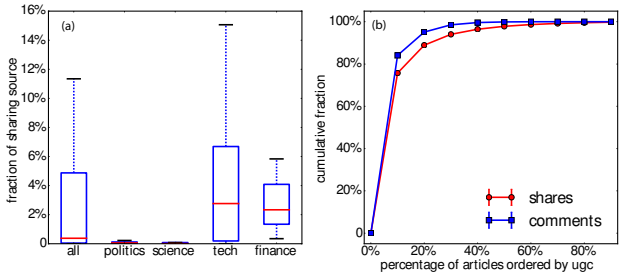


Figure 4: (a) fraction of shares directly pointing to the information source. (b) cumulative fraction of shares and comments of the most shared/commented news articles. Since the number of news articles varies among information sources, the x-axis shows the percentage of news articles.

sources are now able to talk to the public directly. Our dataset offers an opportunity to quantify the importance of information sources vs. the news media. We compute the fraction of shares of an information source over all shares of relevant news articles about the source and the source itself. As shown in Figure 4(a), while tech press releases are shared directly in a few cases, overall the original source is shared rarely, and almost never in politics and science. This suggests that the news media still play an essential role for the public to access information.

Furthermore, not all news articles are equal in getting the public's attention. To understand how concentrated the public's attention is, we compute the cumulative distribution of comments and shares on news articles. Figure 4(b) shows that 10% of news articles get 80% of shares/comments. We also find that it is not the same set of news media outlets that always dominate. We omit the plot for space reasons.

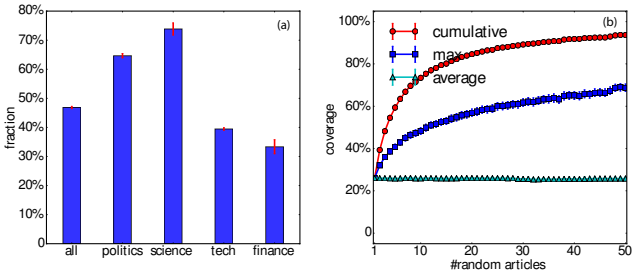


Figure 5: (a) fraction of news articles containing at least one quote of the information source; (b) "average" and "max" give the average and the maximal coverage respectively for a random sample of a given size; "cumulative" shows the cumulative coverage of the combination of the random sample.

News propagation through the layers

In this section, we delve further in a news cycle and investigate how information may diverge from the source in propagation. Given that the news media are the main channel from which individuals learn about information sources, we first explore how completely news media cover information sources and then study how sentiment and language differ from the source in propagation.

News media coverage of the source

We quantify coverage of information sources by news articles from two perspectives: direct quotation and general word reuse in a bag-of-words approach.

Quoting information sources. Since the use of a direct quote indicates an exact replication of a speaker's words (Gibson and Zillmann 1993), quoting can partly reflect how well the news media cover the source. Our first metric is thus the fraction of relevant news articles quoting the information source.[6]

According to Figure 5(a), quoting frequency differs significantly across topics. Press releases from universities are the most highly quoted in the four topics, which echoes the finding that news articles on scientific journal articles cite press releases (De Semir, Ribas, and Revuelta 1998). University press releases also frequently contain quotes by the researchers, which can then be copied. Similarly, presidential speeches are heavily quoted in the news media, as any coverage may be expected to include segments of the speech. In contrast, quoting the source is less common in the coverage of technology press releases and of FOMC statements. This may be explained by either the format of the information sources themselves, or the journalists' or financial analysts' comfort level in writing interpretations without quoting the information source in these two topics.

[6]We consider the text between quotation marks as *quotes* from the information source if over 80% of 4-grams found between the quotation marks are present in the information source.

Bag-of-words coverage. News articles need not quote directly to cover the information conveyed in the source. A common way to represent textual data is to ignore word ordering and view texts as bags of words. We define *coverage* as the fraction of words from the information source that occur in relevant news articles. Two questions arise naturally: first, how does a single news article cover the information source on average (*average coverage*)? Second, how do news articles collectively cover the information source (*cumulative coverage*)? Here we show results for unigrams, while similar findings hold for bigrams and trigrams.

On average, a news article covers 20% of the words in the information source. For each information source and a size n, we randomly sample n articles 100 times and compute the mean value of the corresponding metrics.[7] Figure 5(b) shows that although on average a news article only covers 20% of the source, cumulative coverage of a random sample of articles grows quickly and exceeds the maximal coverage of any single article. This suggests that several articles combined can cover the information source reasonably well and it is not due to a single article citing the source verbatim.

Given the limited circulation via shares of the information source itself, the above analysis shows that it is unlikely that people can be exposed to complete coverage of the content in the information source via the individual articles they read. It is interesting that piecing together different articles nearly recovers the full content of the source. This demonstrates that relevant news articles do not highlight the same 20% and that if coverage is the goal, reading multiple articles can provide it. However, it is important to note that complete coverage may not be the goal of any news article, and furthermore, that providing additional information beyond what is contained in the source may be what is valued and expected by the public.

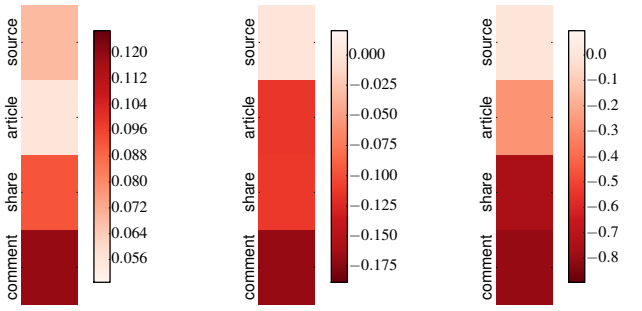
Sentiment in propagation

We now move beyond the first transmission step and explore the entire propagation process across four layers. We track sentiment in terms of subjectivity and positivity as information from the source propagates through the layers. We find that news articles contain fewer sentiment-laden words than information sources, but that individuals' reactions in the form of share text and comments use sentiment-laden words more often. Positivity declines in each transmission step. We further investigate reasons that may explain the changes of sentiment in propagation.

Sentiment difference across layers. Subjectivity. We evaluate the subjectivity of language in each layer in aggregate. Following (Godbole, Srinivasaiah, and Skiena 2007), we define the subjectivity score for a layer L as the fraction of sentiment-laden words based on the LIWC dictionary:

$$\text{subjectivity}(L) = \frac{\sum_{w \in \text{POSEMO} \cup \text{NEGEMO}} C_L(w)}{\sum_w C_L(w)},$$

[7] In addition to random sampling, we used the most shared n articles and found no difference in coverage between random samples and the most shared ones.



(a) LIWC subjective (b) Relative LIWC positivity (c) Relative Vader positivity

Figure 6: (a) average LIWC subjectivity scores in different layers. Darker color indicates *higher level of subjectivity*. (b) relative LIWC positivity scores decline in each layer except between news articles and shares. (c) relative Vader positivity scores decline in each layer. All differences within each figure are significant according to a paired t-test with p-value < 0.001 after Bonferroni correction except the difference between news articles and shares in Figure 6b.

where POSEMO and NEGEMO refer to sets of positive words and negative words in LIWC respectively.[8]

As shown in Figure 6a, news articles use fewer subjective terms than information sources, which is congruent with journalism's aim of objectivity. In contrast, shares and comments use more subjective terms as they capture individuals' opinions and reactions.

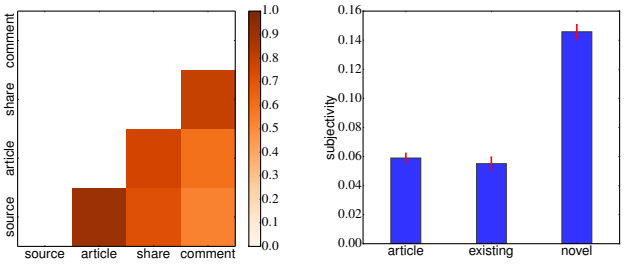
Positivity. Another dimension of sentiment is the level of positivity. As the language in the four layers may differ in nature, we employ two approaches to measuring positivity to increase the robustness of our results: 1) we define the positivity score in layer L as the fraction of positive words among sentiment words based on the LIWC dictionary:

$$\text{LIWCPositive}(L) = \frac{\sum_{w \in \text{POSEMO}} C_L(w)}{\sum_{w \in \text{POSEMO} \cup \text{NEGEMO}} C_L(w)};$$

2) we use the Vader score to better capture the subtlety in informal social media content (Hutto and Gilbert 2014), and take the difference between the positivity of a layer and the corresponding information source to derive a "relative positivity."

According to both metrics (Figure 6b and 6c), positivity generally declines in propagation, except the transmission step from news articles to shares in LIWCPositive, where the difference is not statistically significant. Note that the decrease in positivity in the text of shares and comments may be complementing the perceived positive endorsements of share and like actions, both of which are light-weight, typically positive expressions. Our findings, therefore, reflect only the sentiment *in text* and not the overall sentiment of all interactions on Facebook.

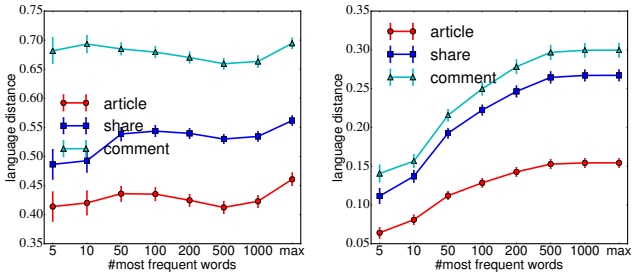
[8] We remove a small set of frequent words in our dataset that are considered emotional in LIWC but are not in our setting, such as "friend", "interest", and "share."



(a) Correlation in subjectivity between layers.

(b) "existing" vs. "novel" in shares.

Figure 7: (a) correlation in subjectivity scores. All correlations are significant ($p < 0.001$) after Bonferroni correction. Similar observations hold for positivity. (b) the subjectivity of words in the article, words in share text that also occur in the article, and words that are original to the share text.



(a) Truncated-vocab distance from the source

(b) Projected-source distance from the source

Figure 8: Contrast of language distance when considering language models based on varying numbers of words in (a) both the layer and in the source, and (b) only projections of the words in the source.

Correlation between layers. Does sentiment in shares and comments get more negative because of a certain level of persistent negativity in online discussions regardless of topic or any specific article (Wolchover 2015)? Or is sentiment related to information sources in some way? To explore this, we compute Pearson correlation for sentiment between different layers. In Figure 7a, we present results just for subjectivity scores, but similar results hold for positivity.

Correlations between all layers are significantly positive. Even the minimal correlation, which is between information sources and comments, is 0.46. This suggests that sentiment is actually maintained in the propagation process. However, we indeed observe that the correlation with earlier layers decreases as propagation happens.

Magnifier or creator? Although the change in sentiment is intriguing, it remains a question what driving force leads to such changes. Here we study the subjectivity increase in shares compared to news articles as an example. There are two possible hypotheses if we look at shares of each news article:

- **Individuals as magnifiers.** Individuals choose to focus on sharing the subjective part within the news article, which leads to the increasing subjectivity in shares.
- **Individuals as creators.** Individuals add subjectivity in their "original content," which can be interpretations of the news article or novel points that they introduce.

To distinguish these two roles, we categorize words in the shares of a news article into two groups: "existing," which includes words that occurred in the news article and "novel," which consists of words that never occurred in the news article. We examine which group has a larger subjectivity score.

Figure 7b presents the results for articles that were shared more than 100 times. We see that there is a striking difference in subjectivity scores between "existing" words and "novel" words in shares. While individuals do not magnify existing subjectivity in the corresponding news article at all, novel words that individuals introduce in shares are twice as subjective as the corresponding news article.

Language in propagation

In addition to sentiment, we try to understand how language may diverge in propagation. We demonstrate that the language gets more different from the source in more distant layers. We further investigate the "life" of specific words, i.e., how words surface up or fade away in propagation.

Language distance across layers. Measuring language distance. In general, it is challenging to measure the distance between languages from different contexts. For instance, given that comments are conversational and press releases are formal, the difference between these two layers may not be meaningful if we consider all words. However, as the four layers are about the same event, intuitively, we would expect the most frequent words to be similar across different layers. After all, if the source is about climate change, "climate" would be expected to be a frequent word in all layers. It would be surprising if "climate" does not occur frequently in comments. Thus, we propose a distance metric that is based on the most frequent words in each layer (*truncated-vocab distance*) and another metric that is based on the most frequent words in the source (*projected-source distance*) to track the change on the words from the source in propagation.

Formally, we define the most frequent n words in a layer L as $Freq(L, n)$ and denote the number of occurrences of word w in L as $C_L(w)$. For a vocabulary size n and a layer L, we compute a truncated unigram language model:

$$P_{L,n}(w) = \begin{cases} \frac{C_L(w)}{\sum_{w \in Freq(L,n)} C_L(w)}, & \text{if } w \in Freq(L,n) \\ 0, & \text{otherwise.} \end{cases}$$

Truncated-vocab distance between layer L_1 and layer L_2 for n is defined by Jensen-Shannon divergence (Lin 1991) between truncated language models in the two layers:[9]

$$TruncatedVocabDist(L_1, L_2, n) = JSD(P_{L_1,n} || P_{L_2,n}),$$

where Jensen-Shannon distance is defined as $JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$, $M = \frac{1}{2}(P + Q)$, $D(P||M) =$

[9]We find similar results using different metrics, including cosine distance between Tf-Idf vectors and Jaccard distance.

$\sum_w P(w) log_2 \frac{P(w)}{M(w)}$. This distance metric is between 0 and 1. A larger value indicates a farther distance between language models.

In order to track the words originating from the source, we define projected unigram language models by only considering the most frequent n words in the information source:

$$ProjP_{L,n}(w) = \begin{cases} \frac{C_L(w)}{\sum_{w \in Freq(S,n)} C_L(w)}, & \text{if } w \in Freq(S,n) \\ 0, & \text{otherwise,} \end{cases}$$

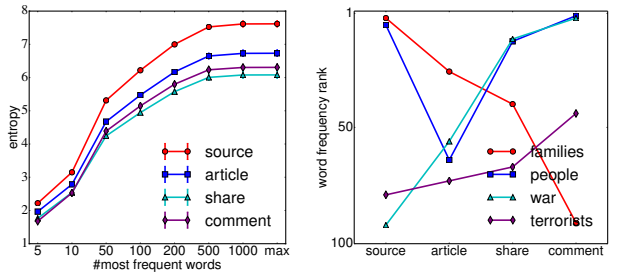
where S represents the corresponding source. Similarly, we define *projected-source distance* based on Jensen-Shannon divergence using projected unigram language models.

Language distance from the source increases in propagation. Both in truncated-vocab distance (Figure 8a) and in projected-source distance (Figure 8b) for all n, there is a *consistent* ordering in the distances from the source: articles $<$ shares $<$ comments. This suggests that as information propagates from the source to the news media, and then to social media, the words used are increasingly different from the source. It is intriguing how the distances vary with n: in Figure 8a, truncated-vocab distance is stable with n, which indicates that even the top 5 words already differ across the layers perhaps due to different nature of language. In contrast, in projected-source distance, a metric that focuses on words that are in the source (Figure 8b), the distance grows as n increases. This implies that the usage of the most frequent words from the information source is better preserved than less frequent ones in propagation.

Usage of words from the source becomes more concentrated in propagation. It is surprising that even in projected-source distance, language gets more different from the source in propagation. We now explore two possible reasons: 1) words from the source get more evenly used in later layers; 2) usage of words from the source becomes more concentrated. To capture the degree of concentration, we compute the entropy for the projected unigram language models in the four layers. Figure 9a shows that the increasing projected-source distance is connected to an increasing degree of concentration instead of a more even distribution.

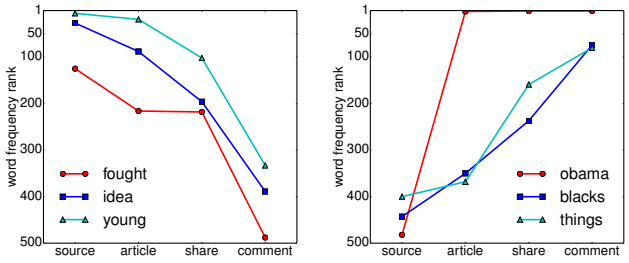
Life of words. A further question is how a specific word from the source fares in propagation. For example, as we discussed in the introduction, although President Obama tried to emphasize "families" in the speech, it disappeared in later word clouds. We attempt to unfold the "life" of words in propagation by looking at how the frequency rank of a word changes. Figure 9b shows how the rank of words changes in the example of Figure 1. In fact, "families" did not fall to the bottom immediately, instead, the rank drops slowly in each transmission step. Another word that Obama emphasized, "people", was not particularly popular among news articles, but became a core theme in the public's discussion, though our bag-of-words analysis does not reveal whether it is used in the same way. In contrast, although Obama did not stress "war" or "terrorists", these words became prominent as the story propagated through the layers.

In another example, Figure 10 shows words that had the largest rank decrease on average ("faded away most") and words that achieved the largest rank increase on average



(a) Source fixation (b) Example of "life" of words.

Figure 9: (a) The entropy of the projected unigram language models (defined as $-\sum_w p(w) \log_2 p(w)$) as a function of the number of words included. A smaller entropy indicates more concentration. (b) Frequency rank for individual words in different layers of the example in Figure 1. A larger rank indicates fewer occurrences.



(a) Faded away most (b) Surfaced up most

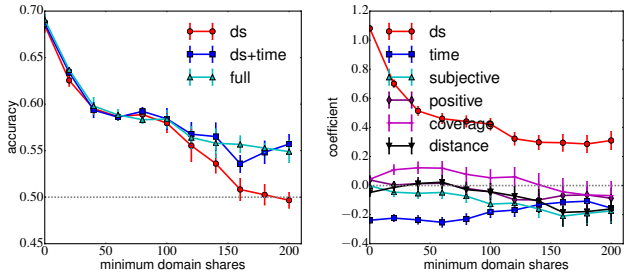
Figure 10: Life of words propagating through the layers of a news cycle from Obama's speech in Selma.

("surfaced up most") among the most frequent 500 words for Obama's speech in Selma at the 50th Anniversary of Selma to Montgomery Marches, an important moment of the American Civil Rights Movement (The White House 2015a). Obama emphasized "young" and "fought" in his speech, but the two words were not prominent in the public's discussion. On the other hand, the word "blacks", not frequent in Obama's speech in Selma, gained prominence in propagation.

Predicting news article shares

We have demonstrated that news articles usually cover the information source partially and are less subjective and less positive than the information source. But are these factors related to the popularity of a news article? In this section, we explore this question by predicting the number of shares.

Predicting shares. As the popularity of information sources varies, predicting the actual number of shares will be biased towards properties of sources that are widely covered and shared. In contrast, the volume of shares is comparable among relevant news articles of the same information source. Thus, to alleviate the above concern, we set up a balanced task that predicts whether a news article will get more shares than half of the news articles covering the same



(a) Prediction performance. (b) Coefficients in logistic prediction.

Figure 11: (a) prediction accuracy in 5-fold cross validation as a function of the minimum shares required to include a domain. As this threshold is increased, the prediction task becomes a competition between more popular domains. (b) coefficients of different features in logistic regression with full data. Error bars represent standard errors.

information source.

This binary treatment may simplify the task too much because news domains also differ in popularity. For instance, New York Times articles are usually shared more than news articles from a local newspaper. To capture the general popularity of news domains, we compute average domain shares for a news domain from all the news articles from March 2015 to June 2015 in that news domain. We then apply a minimum threshold in average domain shares to control the popularity of news domains so that we focus on a collection of news articles from *comparable* news domains. Note that the prediction task is still balanced because we predict whether a news article outperforms the median within the set of news articles given a minimum average domain shares.

We consider the following features: average domain shares, number of seconds since the first news article of the corresponding source came out ("time"), relative subjectivity from the information source ("subjectivity"), relative positivity from the information source ("positivity"), unigram coverage of the source ("coverage") and language distance from the source ("distance"). We also add subjectivity and positivity of the information source as control variables. In addition to using all features ("full"), we consider two baselines: one uses only average domain shares ("ds"); the other uses average domain shares and time ("ds+time").

In order to assess the prediction performance, we measure accuracy in 5-fold cross validation using logistic regression.[10] All features are standardized before regression.

Prediction results. Figure 11a shows that it is relatively easy to tell which news article will outperform the median if we do not control the popularity of news domains. In fact, using only average domain shares as features, we can already achieve 70%, and adding other features does not improve accuracy. However, as minimum average domain shares increase, the predictive power of average domain shares declines significantly. For news domains with more than 160 average domain shares, using average domain shares is equivalent to random guessing. Adding time since the first news article, the performance can grow slightly to 55% among popular domains. The low accuracy shows that it is difficult to predict which news article will be shared most if we look at comparable news domains. Interestingly, comparisons with the information source in language and sentiment do not improve performance over using only average domain shares and time, which suggests that news articles that stay true to the source enjoy no advantage.

As for coefficients (Figure 11b), average domain shares is consistently the dominant feature. Another consistently statistically significant feature is time since the first news article, with articles published earlier being shared more. Although comparisons with the information source do not improve prediction performance, less subjectivity and smaller distance from the information source are associated with more shares as the minimum average domain shares increases. In contrast, the coefficients of relative positivity and coverage of the information source are not statistically significant.

[10]Results are consistent with different classification models.

Concluding discussion

Summary. In this paper, we present the first large-scale study on how the news cycle of an event unfolds on social media, and on how information diverges from the information source at different stages of propagation. We find that the news media indeed mediate the public's consumption of information from sources. Shares and comments, capturing people's opinions and reactions, tend to be more subjective than both the source and the news coverage. This increasing subjectivity is mainly due to additional content that people add in the shares instead of words that are present in the news article. Language diverges from the content of the information sources in propagation with a higher degree of concentration. Furthermore, news articles that stay close to the source are not shared more or less.

Limitations, implications and further directions. Our findings are constrained by the focus on identifiable information sources across four domains, yielding a non-random sample of all possible current events. Were it possible to identify information sources for a wider variety of events, especially for dynamic and complex events such as presidential campaigns and ongoing political debates, it would be interesting to investigate how the results generalize.

Furthermore, although Facebook is an important platform for communication, and is becoming a primary source of news for many individuals (Duggan and Smith 2013), there are specific mechanisms of sharing and feedback in Facebook that may make interaction with news different from interactions on other social media platforms. Specifically, the finding that text in shares and comments tends to be more negative than in either the source or news articles could be a reflection of individuals having other ways of expressing positive opinions, e.g. sharing the article without text, or liking a friend's share of an article. Future work could incorporate non-textual reactions and reactions on other social media.

An interesting avenue for future work that arises from our findings on the life of words is to predict which words from

the information source will fade away or surface up in propagation. Such understandings can enrich our knowledge of not only how information reaches us, but also how it might differ from its source. Finally, the insights presented in this paper could inform the development of tools to track coverage and reactions to coverage within news cycles.

Acknowledgments. We thank Mia Cha, Amaç Herdağdelen, Jon Kleinberg, Lillian Lee, Sendhil Mullainathan, Christy Sauper, Ves Stoyanov, Karthik Subbian, and Shaomei Wu for helpful comments and discussions. This work was supported in part by a Facebook fellowship.

References

Allan, J. 2012. *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media.

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of WSDM*.

Bakshy, E.; Rosenn, I.; Marlow, C.; and Adamic, L. 2012. The Role of Social Networks in Information Diffusion. In *Proceedings of WWW*.

Baum, M. A., and Groeling, T. 2008. New Media and the Polarization of American Political Discourse. *Political Communication* 25(4):345–365.

Bohannon, J. 2015. I fooled millions into thinking chocolate helps weight loss. here's how. `http://io9.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800`.

Brechman, J.; Lee, C.-j.; and Cappella, J. N. 2009. Lost in Translation? A Comparison of Cancer-Genetics Reporting in the Press Release and Its Subsequent Coverage in the Press. *Science Communication*.

Cuthbertson, T. 2015. python-readability. `https://github.com/gfxmonk/python-readability`.

De Semir, V.; Ribas, C.; and Revuelta, G. 1998. Press releases of science journal articles and subsequent newspaper stories on the same topic. *Jama* 280(3):294–295.

Duggan, M., and Smith, A. 2013. Social media update 2013. *Pew Research Center*.

Gibson, R., and Zillmann, D. 1993. The Impact of Quotation in News Reports on Issue Perception. *Journalism & Mass Communication Quarterly* 70(4):793–800.

Godbole, N.; Srinivasaiah, M.; and Skiena, S. 2007. Large-Scale Sentiment Analysis for News and Blogs. In *Proceedings of ICWSM*.

Groseclose, T., and Milyo, J. 2005. A measure of media bias. *The Quarterly Journal of Economics* 1191–1237.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information Diffusion Through Blogspace. In *Proceedings of WWW*.

Hutto, C. J., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Katz, E., and Lazarsfeld, P. F. 1955. *Personal Influence, The part played by people in the flow of mass communications*. Transaction Publishers.

Lasswell, H. D. 1948. The structure and function of communication in society. *The communication of ideas* 37:215–228.

Lazarsfeld, P. F.; Berelson, B.; and Gaudet, H. 1968. The peoples choice: how the voter makes up his mind in a presidential campaign.

Lerman, K., and Ghosh, R. 2010. Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. In *Proceedings of ICWSM*.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of KDD*.

Liben-Nowell, D., and Kleinberg, J. 2008. Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences* 105(12):4633–4638.

Lin, Y.-R.; Bagrow, J. P.; and Lazer, D. 2011. More Voices Than Ever? Quantifying Media Bias in Networks. In *Proceedings of ICWSM*.

Lin, J. 1991. Divergence measures based on the Shannon entropy. *Information Theory, IEEE Transactions on* 37(1):145–151.

Merton, R. K. 1957. Patterns of influence: Local and cosmopolitan influentials. *Social theory and social structure* 2:387–420.

Mueller, A. 2015. word_cloud. `https://github.com/amueller/word_cloud`.

Niculae, V.; Suen, C.; Zhang, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of WWW*.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2007. Linguistic inquiry and word count: Liwc 2007. Technical report.

Prat, A., and Strömberg, D. 2013. The political economy of mass media. In *Advances in Economics and Econometrics: Tenth World Congress*. Cambridge University Press.

Puglisi, R.; Snyder, J. M.; and James, J. 2011. Newspaper Coverage of Political Scandals. *The Journal of Politics* 73(03):931–950.

Saguy, A. C., and Almeling, R. 2008. Fat in the Fire? Science, the News Media, and the "Obesity Epidemic". *Sociological Forum* 23(1):53–83.

Simmons, M. P.; Adamic, L. A.; and Adar, E. 2011. Memes online: Extracted, subtracted, injected, and recollected. In *Proceedings of ICWSM*.

The White House. 2015a. Remarks by the president at the 50th anniversary of the selma to montgomery marches. `https://goo.gl/Df90Zy`.

The White House. 2015b. Remarks by the president on middle class economics. `https://goo.gl/RirWRx`.

The White House. 2015c. Speeches and remarks. `https://www.whitehouse.gov/briefing-room/speeches-and-remarks`.

The White House. 2015d. Statement by the president on the deaths of warren weinstein and giovanni lo porto. `https://goo.gl/uauWj0`.

Winsten, J. A. 1985. Science and the media: the boundaries of truth. *Health Affairs* 4(1):5–23.

Wolchover, N. 2015. Why is everyone on the internet so angry? `http://goo.gl/qY752f`.

Wu, S.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Who Says What to Whom on Twitter. In *Proceedings of WWW*.