# Meme-tracking and the Dynamics of the News Cycle

Jure Leskovec[*†] Lars Backstrom[*] Jon Kleinberg[*]
[*]Cornell University [†]Stanford University
jure@cs.cornell.edu lars@cs.cornell.edu kleinber@cs.cornell.edu

## ABSTRACT

Tracking new topics, ideas, and "memes" across the Web has been an issue of considerable interest. Recent work has developed methods for tracking topic shifts over long time scales, as well as abrupt spikes in the appearance of particular named entities. However, these approaches are less well suited to the identification of content that spreads widely and then fades over time scales on the order of days — the time scale at which we perceive news and events.

We develop a framework for tracking short, distinctive phrases that travel relatively intact through on-line text; developing scalable algorithms for clustering textual variants of such phrases, we identify a broad class of memes that exhibit wide spread and rich variation on a daily basis. As our principal domain of study, we show how such a meme-tracking approach can provide a coherent representation of the *news cycle* — the daily rhythms in the news media that have long been the subject of qualitative interpretation but have never been captured accurately enough to permit actual quantitative analysis. We tracked 1.6 million mainstream media sites and blogs over a period of three months with the total of 90 million articles and we find a set of novel and persistent temporal patterns in the news cycle. In particular, we observe a typical lag of 2.5 hours between the peaks of attention to a phrase in the news media and in blogs respectively, with divergent behavior around the overall peak and a "heartbeat"-like pattern in the handoff between news and blogs. We also develop and analyze a mathematical model for the kinds of temporal variation that the system exhibits.

## 1. INTRODUCTION

A growing line of research has focused on the issues raised by the diffusion and evolution of highly dynamic on-line information, particularly the problem of tracking topics, ideas, and "memes" as they evolve over time and spread across the web. Prior work has identified two main approaches to this problem, which have been successful at two correspondingly different extremes of it. Probabilistic term mixtures have been successful at identifying long-range trends in general topics over time [5, 7, 16, 17, 30, 31]. At the other extreme, identifying hyperlinks between blogs and extracting rare named entities has been used to track short information cascades through the blogosphere [3, 14, 20, 23]. However, between these two extremes lies much of the temporal and textual range over which propagation on the web and between people typically occurs, through the continuous interaction of news, blogs, and websites on a daily basis. Intuitively, short units of text, short phrases, and "memes" that act as signatures of topics and events propagate and diffuse over the web, from mainstream media to blogs, and vice versa. This is exactly the focus of our study here.

Moreover, it is at this intermediate temporal and textual granularity of memes and phrases that people experience news and current events. A succession of story lines that evolve and compete for attention within a relatively stable set of broader topics collectively produces an effect that commentators refer to as the *news cycle*. Tracking dynamic information at this temporal and topical resolution has proved difficult, since the continuous appearance, growth, and decay of new story lines takes place without significant shifts in the overall vocabulary; in general, this process can also not be closely aligned with the appearance and disappearance of specific named entities (or hyperlinks) in the text. As a result, while the dynamics of the news cycle has been a subject of intense interest to researchers in media and the political process, the focus has been mainly qualitative, with a corresponding lack of techniques for undertaking quantitative analysis of the news cycle as a whole.

**Our approach to meme-tracking, with applications to the news cycle.** Here we develop a method for tracking units of information as they spread over the web. Our approach is the first to scalably identify short distinctive phrases that travel relatively intact through on-line text as it evolves over time. Thus, for the first time at a large scale, we are able to automatically identify and actually "see" such textual elements and study them in a massive dataset providing essentially complete coverage of on-line mainstream and blog media. Working with phrases naturally interpolates between the two extremes of topic models on the one hand and named entities on the other. First, the set of distinctive phrases shows significant diversity over short periods of time, even as the broader vocabulary remains relatively stable. As a result, they can be used to dissect a general topic into a large collection of threads or memes that vary from day to day. Second, such distinctive phrases are abundant, and therefore are rich enough to act as "tracers" for a large collection of memes; we therefore do not have to restrict attention to the much smaller collection of memes that happen to be associated with the appearance and disappearance of a single named entity.

From an algorithmic point of view, we consider these distinctive phrases to act as the analogue of "genetic signatures" for different

This research was supported in part by the MacArthur Foundation, a Google Research Grant, a Yahoo! Research Alliance Grant, and NSF grants CCF-0325453, CNS-0403340, BCS-0537606, and IIS-0705774.


*KDD* '09 Paris, France

memes. And like genetic signatures, we find that while they remain recognizable as they appear in text over time, they also undergo significant mutation. As a result, a central computational challenge in this approach is to find robust ways of extracting and identifying all the mutational variants of each of these distinctive phrases, and to group them together. We develop scalable algorithms for this problem, so that memes end up corresponding to clusters containing all the mutational variants of a single phrase.

As an application of our technique, we use it to produce some of the first quantitative analysis of the global news cycle. To do this, we work with a massive set of 90 million news and blog articles that we collected over the final three months of the 2008 U.S. Presidential Election (starting August 1).[1] In this context, the collection of distinctive phrases that will act as tracers for memes are the set of quoted phrases and sentences that we find in articles — that is, quotations attributed to individuals. This is natural for the domain of news: quotes are an integral part of journalistic practice, and even if a news story is not specifically about a particular quote, quotes are deployed in essentially all articles, and they tend to follow iterations of a story as it evolves [28]. However, each individual quote tends to exhibit very high levels of variation across its occurrence in many articles, and so the aspects of our approach based on clustering mutational variants will be crucial.

Thus, our analysis of the news cycle will consist of studying the most significant groups of mutational variants as they evolve over time. We perform this analysis both at a global level — understanding the temporal variation as a whole — and at a local level — identifying recurring patterns in the growth and decay of a meme around its period of peak intensity. At a global level, we find a structure in which individual memes compete with another over short time periods, producing daily and weekly patterns of variation. We also show how the temporal patterns we observe arise naturally from a simple mathematical model in which news sources imitate each other's decisions about what to cover, but subject to recency effects penalizing older content. This combination of imitation and recency can produce synthetic temporal patterns resembling the real data; neither ingredient alone is able to do this.

At a local level, we identify some of the fine-grained dynamics governing how the intensity of a meme behaves. We find a characteristic spike around the point of peak intensity; in both directions away from the peak the volume decreases exponentially with time, but in an 8-hour window of time around the median, we find that volume $y$ as a function of time $t$ behaves like $y(t) \approx a \log(t)$. This function diverges at $t = 0$ — indicating an explosive amount of activity right at the peak period. Further interesting dynamics emerge when one separates the websites under consideration into two distinct categories — news media and blogs. We find that the peak of news-media attention of a phrase typically comes 2.5 hours earlier than the peak attention of the blogosphere. Moreover, if we look at the proportion of phrase mentions in blogs in a few-hour window around the peak, it displays a characteristic "heartbeat"-type shape as the meme bounces between mainstream media and blogs. We further break down the analysis to the level of individual blogs and news sources, characterizing the typical amount by which each source leads or lags the overall peak. Among the "fastest" sources we find a number of popular political blogs; this measure thus suggests a way of identifying sites that are regularly far ahead of the bulk of media attention to a topic.

[1]This is of course a period when news coverage was particularly high-intensity, but it gives us a chance to study the news cycle over precisely the kind of interval in which people's general intuitions about it are formed — and in which it is enormously consequential. In the latter regard, studying the effect of communication technology on elections is a research tradition that goes back at least to the work of Lazarsfeld, Berelson, and Gaudet in the 1940s [22].

**Further related work.** In addition to the range of different approaches for tracking topics, ideas, and memes discussed above, there has been considerable work in computer science focused on news data in particular. Two dominant themes in this work to date have been the use of algorithmic tools for organizing and filtering news; and the role of blogging and the production of news by individuals rather than professional media organizations. Some of the key research issues here have been the identification of topics over time [5, 11, 16], the evolving practices of bloggers [25, 26], the cascading adoption of stories [3, 14, 20, 23], and the ideological divisions in the blogosphere [2, 12, 13]. This has led to development of a number of interesting tools to help people better understand the news (e.g. [5, 11, 12, 13, 16]).

Outside of computer science, the interplay between technology, the news media, and the political process has been a focus of considerable research interest for much of the past century [6, 22]. This research tradition has included work by sociologists, communication scholars, and media theorists, usually at qualitative level exploring the political and economic contexts in which news is produced [19], its effect on public opinion , and its ability to facilitate either polarization or consensus [15].

An important recent theme within this literature has been the increasing intensity of the news cycle, and the increasing role it plays in the political process. In their influential book *Warp Speed: America in the Age of the Mixed Media*, Kovach and Rosenstiel discuss how the excesses of the news cycle have become intertwined with the fragmentation of the news audience, writing, "The classic function of journalism to sort out a true and reliable account of the day's events is being undermined. It is being displaced by the continuous news cycle, the growing power of sources over reporters, varying standards of journalism, and a fascination with inexpensive, polarizing argument. The press is also increasingly fixated on finding the 'big story' that will temporarily reassemble the now-fragmented mass audience" [19]. In addition to illuminating their effect on the producers and consumers of news, researchers have also investigated the role these issues play in policy-making by government. As Jayson Harsin observes, over time the news cycle has grown from being a dominant aspect of election campaign season to a constant feature of the political landscape more generally; "not only," he writes, "are campaign tactics normalized for governing but the communication tactics are themselves institutionally influenced by the twenty-four hour cable and internet news cycle" [15].

Moving beyond qualitative analysis has proven difficult here, and the intriguing assertions in the social-science work on this topic form a significant part of the motivation for our current approach. Specifically, the discussions in this area had largely left open the question of whether the "news cycle" is primarily a metaphorical construct that describes our perceptions of the news, or whether it is something that one could actually observe and measure. We show that by tracking essentially *all* news stories at the right level of granularity, it is indeed possible to build structures that closely match our intuitive picture of the news cycle, making it possible to begin a more formal and quantitative study of its basic properties.

## 2. ALGORITHMS FOR CLUSTERING MUTATIONAL VARIANTS OF PHRASES

We now discuss our algorithms for identifying and clustering textual variants of quotes, capable of scaling to our corpus of roughly a hundred million articles over a three-month period. These clusters will then form the basic objects in our subsequent analysis.

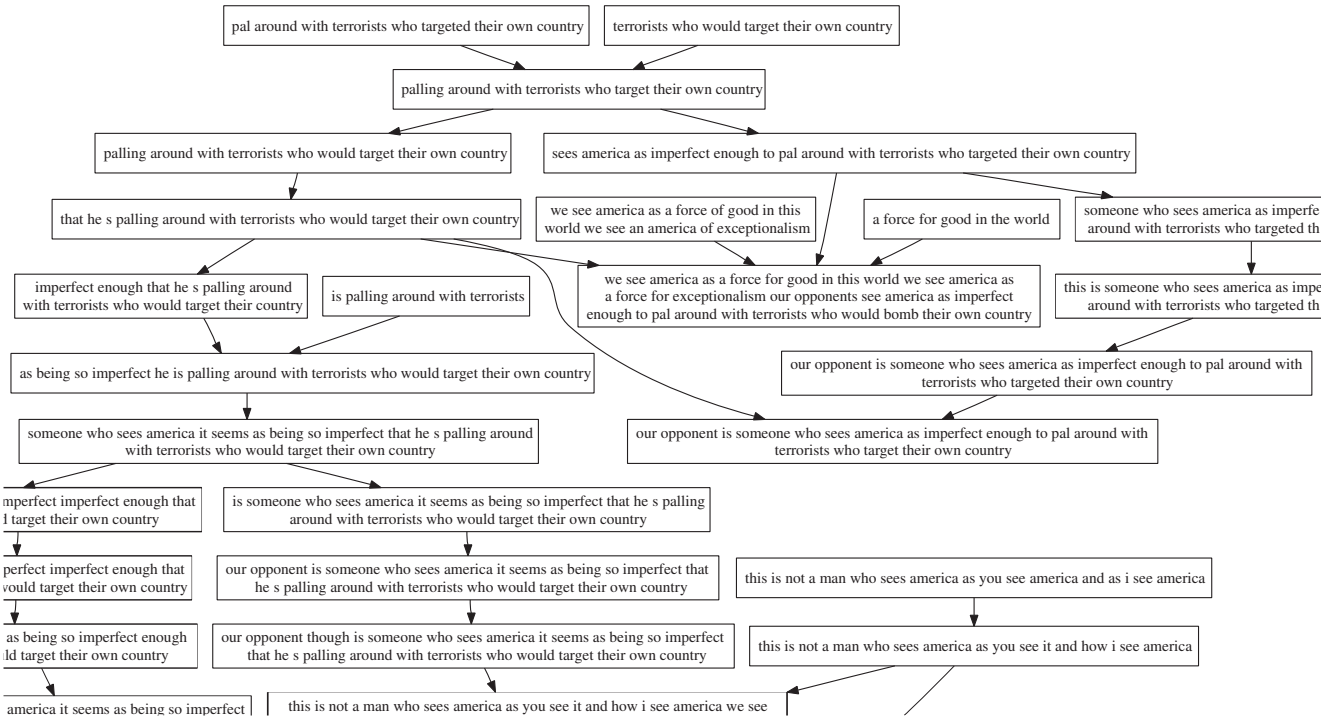


**Figure 1: A small portion of the full set of variants of Sarah Palin's quote, "Our opponent is someone who sees America, it seems, as being so imperfect, imperfect enough that he's palling around with terrorists who would target their own country." The arrows indicate the (approximate) inclusion of one variant in another, as part of the methodology developed in Section 2.**

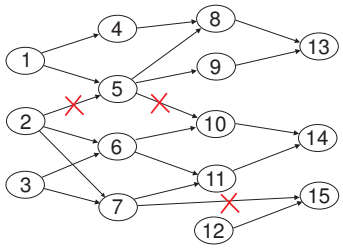


**Figure 2: Phrase graph. Each phrase is a node and we want to delete the least edges so that each resulting connected component has a single root node/phase, a node with zero out-edges. By deleting the indicated edges we obtain the optimal solution.**

To begin, we define some terminology. We will refer to each news article or blog post as an *item*, and refer to a quoted string that occurs in one or more items as a *phrase*. Our goal is to produce *phrase clusters*, which are collections of phrases deemed to be close textual variants of one another. We will do this by building a *phrase graph* where each phrase is represented by a node and directed edges connect related phrases. Then we partition this graph in such a way that its components will be the phrase clusters.

We first discuss how to construct the graph, and then how we partition it. The dominant way in which one finds textual variants in our quote data is excerpting — when phrase $p$ is a contiguous subsequence of the words in phrase $q$. Thus, we build the *phrase graph* to capture these kinds of inclusion relations, relaxing the notion of inclusion to allow for very small mismatches between phrases.

**The phrase graph.** First, to avoid spurious phrases, we set a lower bound $L$ on the word-length of phrases we consider, and a lower bound $M$ on their frequency — the number of occurrences in the full corpus. We also eliminate phrases for which at least an $\varepsilon$ fraction occur on a single domain — inspection reveals that frequent phrases with this property are exclusively produced by spammers. (We use $\varepsilon = .25$, $L = 4$, and $M = 10$ in our implementation.)

After this pre-processing, we build a graph $G$ on the set of quoted phrases. The phrases constitute the nodes; and we include an edge $(p, q)$ for every pair of phrases $p$ and $q$ such that $p$ is strictly shorter than $q$, and $p$ has directed edit distance to $q$ — treating words as tokens — that is less than a small threshold $\delta$ ($\delta = 1$ in our implementation) or there is at least a $k$-word consecutive overlap between the phrases we use $k = 10$). Since all edges $(p, q)$ point from shorter phrases to longer phrases, we have a directed acyclic graph (DAG) $G$ at this point. In general, one could use more complicated natural language processing techniques, or external data to create the edges in the phrase graph. We experimented with various other techniques and found the current approach robust and scalable.

Thus, $G$ encodes an approximate inclusion relationship or long consecutive overlap among all the quoted phrases in the data, allowing for small amounts of textual mutation. Figure 1 depicts a very small portion of the phrase DAG for our data, zoomed in on a few of the variants of a quote by Sarah Palin. Only edges with endpoints not connected by some other path in the DAG are shown.

We now add weights $w_{pq}$ to the edges $(p, q)$ of $G$, reflecting the importance of each edge. The weight is defined so that it decreases in the directed edit distance from $p$ to $q$, and increases in the frequency of $q$ in the corpus. This latter dependence is important, since we particularly wish to preserve edges $(p, q)$ when the inclusion of $p$ in $q$ is supported by many occurrences of $q$.

**Partitioning the phrase graph.** How should we recognize a good phrase cluster, given the structure of $G$? The central idea is that we are looking for a collection of phrases related closely enough that they can all be explained as "belonging" either to a single long phrase $q$, or to a single collection of phrases. The outgoing paths from all phrases in the cluster should flow into a single root node $q$, where we define a *root* in $G$ to be a node with no outgoing edges

(*e.g.*, nodes 13, 14, 15 in Fig. 2). So, the phrase cluster should be a subgraph for which all paths terminate in a single root node.

Thus, informally, to identify phrase clusters, we would like delete edges of small total weight from the phrase graph so it falls apart into disjoint pieces, with the property that each piece "feeds into" a single root phrase that can serve as the exemplar for the phrase cluster. More precisely, we define a directed acyclic graph to be *single-rooted* if it contains exactly one root node. (Note that every DAG has at least one root.) We now define the following *DAG partitioning problem*:

> DAG Partitioning: Given a directed acyclic graph with edge weights, delete a set of edges of minimum total weight so that each of the resulting components is single-rooted.

For example, Figure 2 shows a DAG with all edge weights equal to 1; deleting indicated edges forms the unique optimal solution.

We now show that DAG Partitioning is computationally intractable to solve optimally. We then discuss the heuristic we use for the problem on our data, which we find to work well in practice.

PROPOSITION 1. *DAG Partitioning is NP-hard.*

*Proof Sketch.* We show that deciding whether an instance of DAG Partitioning has a solution of total edge weight at most $W$ is NP-complete, using a reduction from an NP-complete problem in discrete optimization known as the *Multiway Cut problem* [9, 10]. In an instance of Multiway Cut, we are given a weighted undirected graph $H$ in which a subset $T$ of the nodes has been designated as the set of *terminals*. The goal is to decide whether we can delete a set of edges of total weight at most $W$ so that each terminal $T$ belongs to a distinct component. Due to space constraints we give the details of this construction at the supporting website [1]. ■

**An alternate heuristic.** Given the intractability of DAG Partitioning, we develop a class of heuristics for it that we find to scale well and to produce good phrase clusters in practice.

To motivate the heuristics, note that in any optimal solution to DAG Partitioning, there is at least one outgoing edge from each non-root node that has not been deleted. (For if a non-root node had all its outgoing edges deleted, then we could put one back in and still preserve the validity of the solution.) Second, a subgraph of the DAG where each non-root node has only a single out-edge must necessarily have single-rooted components, since the edge sets of the components will all be in-branching trees. Finally, if — as a thought experiment — for each node $v$ we happened to know just a single edge $e$ that was not deleted in the optimal solution, then the subgraph consisting of all these edges $e$ would have the same components (when viewed as node sets) as the components in the optimal solution of DAG Partitioning. In other words, it is enough to find a single edge out of each node that is included in the optimal solution to identify the optimal components.

With this in mind, our heuristics proceed by choosing for each non-root node a single outgoing edge. Thus each of the components will be single-rooted, as noted above, and we take these as the components of our solution. We evaluate the heuristics with respect to the total amount of edge weight kept in the clusters if a random edge out of each phrase is kept. We found that keeping an edge to the shortest phrase gives 9% improvement over the baseline, while keeping an edge to the most frequent phrase gives 12% improvement. Proceeding from the roots down the DAG and greedily assigning each node to the cluster to which it has the most edges gives 13% improvement over the baseline. We also experimented with simulated annealing but that did not improve the solution, suggesting further evidence for the effectiveness of our heuristics.

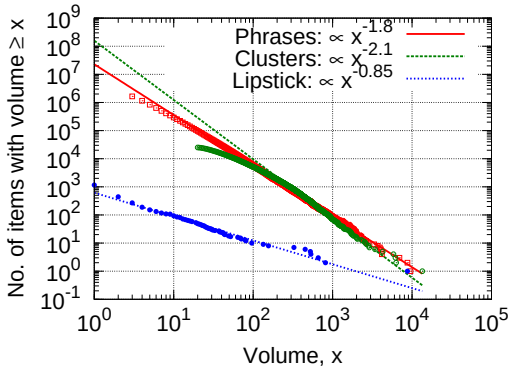


**Figure 3: Phrase volume distribution. We consider the volume of individual phrases, phrase-clusters, and the phrases that compose the "Lipstick on a pig" cluster. Notice phrases and phrase-clusters have similar power-law distribution while the "Lipstick on a pig" cluster has much fatter tail, which means that popular phrases are have unexpectedly high popularity.**

**Dataset description.** Our dataset covers three months of online mainstream and social media activity from August 1 to October 31 2008 with about 1 million documents per day. In total it consist of 90 million documents (blog posts and news articles) from 1.65 million different sites that we obtained through the Spinn3r API [27]. The total dataset size is 390GB and essentially includes complete online media coverage: we have all mainstream media sites that are part of Google News (20,000 different sites) plus 1.6 million blogs, forums and other media sites. From the dataset we extracted the total 112 million quotes and discarded those with $L < 4$, $M < 10$, and those that fail our single-domain test with $\varepsilon = .25$. This left us with 47 million phrases out of which 22 million were distinct. Clustering the phrases took 9 hours and produced a DAG with 35,800 non-trivial components (clusters with at least two phrases) that together included 94,700 nodes (phrases).

Figure 3 shows the complementary cumulative distribution function (CCDF) of the phrase volume. For each volume $x$, we plot the number of phrases with volume $\geq x$. If the quantity of interest is power-law distributed with exponent $\gamma$, $p(x) \propto x^{-\gamma}$, then when plotted on log-log axes the CCDF will be a straight line with slope $-(\gamma + 1)$. In Figure 3 we superimpose three quantities of interest: the volume of individual phrases, phrase clusters (volume of all phrases in the cluster), and the individual phrases from the largest phrase-cluster in our dataset (the "lipstick on a pig" cluster). Notice all quantities are power-law distributed. Moreover, the volume of individual phrases decays as $x^{-2.8}$, and of phrase-clusters as $x^{-3.1}$, which means that the tails are not very heavy as for $\gamma > 3$ power-law distributions start to have finite variances. However, notice that volume of the "lipstick on a pig" cluster decays as $x^{-1.85}$ in which case the tail is much heavier. In fact, for $\gamma < 2$ power-laws have infinite expectations. This means that variants of popular phrases, like "lipstick on a pig," are much more "stickier" than what would be expected from overall phrase volume distribution. Popular phrases have many variants and each of them appears more frequently than an "average" phrase.

# 3. GLOBAL ANALYSIS: TEMPORAL VARIATION AND A PROBABILISTIC MODEL

Having produced phrase clusters, we now construct the individual elements of the news cycle. We define a *thread* associated with a given phrase cluster to be the set of all items (news articles or blog posts) containing some phrase from the cluster, and we then

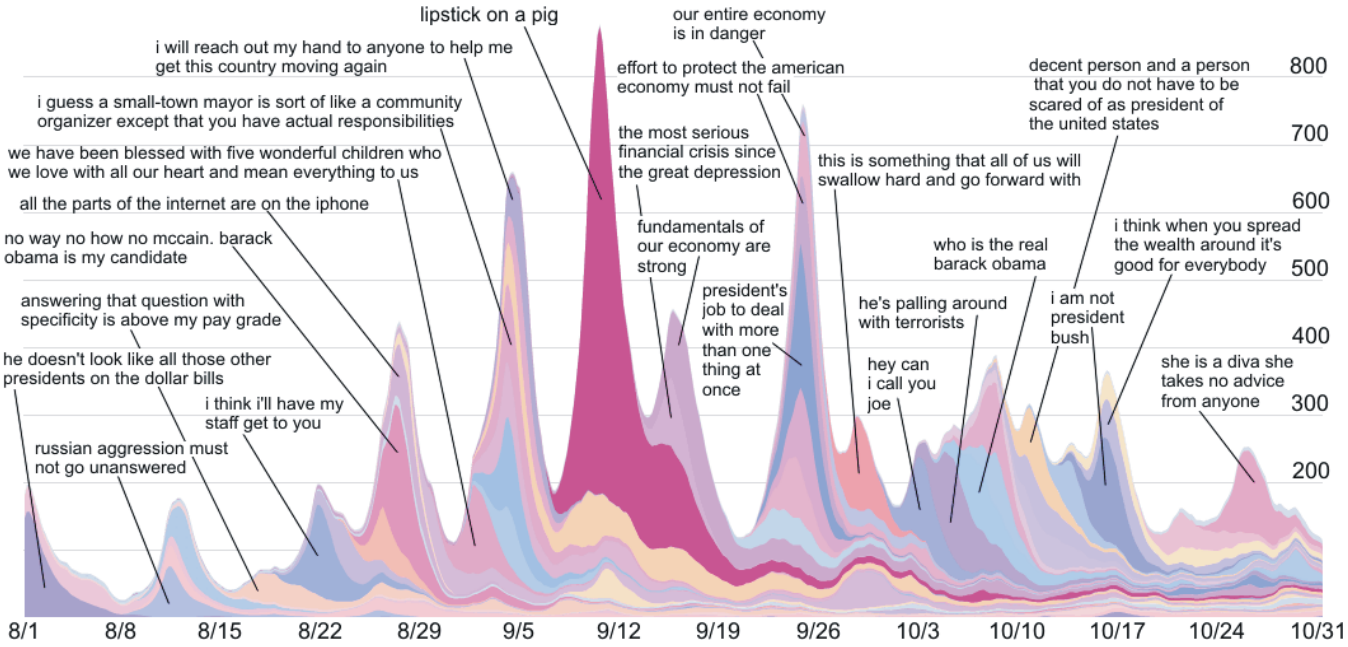


**Figure 4: Top 50 threads in the news cycle with highest volume for the period Aug. 1 – Oct. 31, 2008. Each thread consists of all news articles and blog posts containing a textual variant of a particular quoted phrases. (Phrase variants for the two largest threads in each week are shown as labels pointing to the corresponding thread.) The data is drawn as a stacked plot in which the thickness of the strand corresponding to each thread indicates its volume over time. Interactive visualization is available at http://memetracker.org.**

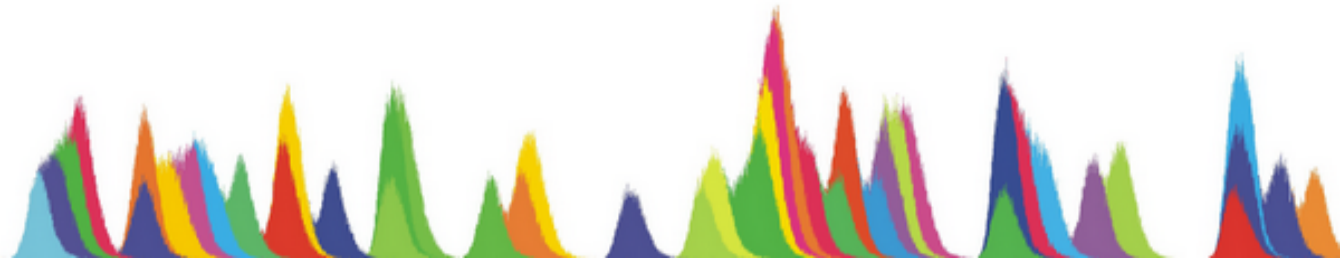

**Figure 5: Temporal dynamics of top threads as generated by our model. Only two ingredients, namely imitation and a preference to recent threads, are enough to qualitatively reproduce the observed dynamics of the news cycle.**

track all threads over time, considering both their individual temporal dynamics as well as their interactions with one another.

Using our approach we completely automatically created and also automatically labeled the plot in Figure 4, which depicts the 50 largest threads for the three-month period Aug. 1 – Oct. 31. It is drawn as a stacked plot, a style of visualization (see e.g. [16]) in which the thickness of each strand corresponds to the volume of the corresponding thread over time, with the total area equal to the total volume. We see that the rising and falling pattern does in fact tell us about the patterns by which blogs and the media successively focus and defocus on common story lines.

An important point to note at the outset is that the total number of articles and posts, as well as the total number of quotes, is approximately constant over all weekdays in our dataset. (Refer to [1] for the plots.) As a result, the temporal variation exhibited in Figure 4 is not the result of variations in the overall amount of global news and blogging activity from one day to the next. Rather, the periods when the upper envelope of the curve are high correspond to times when there is a greater degree of convergence on key stories, while the low periods indicate that attention is more diffuse, spread out over many stories. There is a clear weekly pattern in this (again, despite the relatively constant overall volume), with the five large peaks between late August and late September corresponding, respectively, to the Democratic and Republican National Conventions, the overwhelming volume of the "lipstick on a pig" thread, the beginning of peak public attention to the financial crisis, and the negotiations over the financial bailout plan. Notice how the plot captures the dynamics of the presidential campaign coverage at a very fine resolution. Spikes and the phrases pinpoint the exact events and moments that triggered large amounts of attention.

Moreover, we have evaluated competing baselines in which we produce topic clusters using standard methods based on probabilistic term mixtures (e.g. [7, 8]).[2] The clusters produced for this time period correspond to much coarser divisions of the content (politics, technology, movies, and a number of essentially unrecognizable clusters). This is consistent with our initial observation in Section 1 that topical clusters are working at a level of granularity different from what is needed to talk about the news cycle. Similarly, producing clusters from the most linked-to documents [23] in the dataset produces a much finer granularity, at the level of individual articles. For reasons of space, we refer the reader to the supporting website [1] for the full results of these baseline approaches.

**Global models for temporal variation.** From a modeling per-

[2] As these do not scale to the size of the data we have here, we could only use a subset of 10,000 most highly linked-to articles.

spective, it is interesting to ask for a minimal set of dynamic behaviors that will produce this type of sustained temporal variation over time. Rather than trying to fit the curve in Figure 4 exactly, the question here is to find basic ingredients that can produce synthetic dynamics of a broadly similar structure.

To begin with, there are interesting potential analogies to natural systems that contain dynamics similar to what one sees in the news cycle. For example, one could imagine the news cycle as a kind of species interaction within an ecosystem [18], where threads play the role of species competing for resources (in this case media attention, which is constant over time), and selectively reproducing (by occupying future articles and posts). Similarly, one can see analogies to certain kinds of biological regulation mechanisms such as follicular development [21], in which threads play the role of cells in an environment with feedback where at most one or a few cells tend to be dominant at any point in time. However, the news cycle is distinct in that there is a constant influx of new threads on a time scale that is comparable to the rate at which competition and selective reproduction is taking place.[3] A model for the dynamics of the news cycle must take this into account, as we now discuss.

We argue that in formulating a model for the news cycle, there are two minimal ingredients that should be taken into account. The first is that different sources *imitate* one another, so that once a thread experiences significant volume, it is likely to persist and grow through adoption by others. The second, counteracting the first, is that threads are governed by strong *recency* effects, in which new threads are favored to older ones. (There are other effects that can be included as well, including the fact that threads differ in their initial *attractiveness* to media sources, with some threads having inherently more likelihood to succeed. However, we omit this and other features from the present discussion, which focuses on identifying a minimal set of ingredients necessary for producing the patterns we observe.)

We seek to capture the two components of imitation and recency in a stylized fashion using the following model, whose dynamics we can then study. The model can be viewed as incorporating a type of preferential attachment [4], but combined with factors related to the effects of novelty and attention [32]. Time runs in discrete periods $t = 1, 2, 3, \ldots, T$, and there is a collection of $N$ media sources, each of which reports on a single thread in one time period. Simply for the sake of initialization, we will assume that each source is reporting on a distinct thread at time 0. In each time step, a new thread $j$ is produced.

Also in each time step $t$, each source must choose which thread to report on. A given source chooses thread $j$ with probability proportional to the product $f(n_j)\delta(t-t_j)$, where $n_j$ denotes the number of stories previously written about thread $j$, time $t$ is the current time, and time $t_j$ is the time when $j$ was first produced. The function $\delta(\cdot)$ is monotonically decreasing in $t - t_j$. One could take this decrease to be exponential in some polynomial function of $t - t_j$ based on research on novelty and attention [32]; or, following research on human response-time dynamics [24, 29], one could take it to be a heavy-tailed functional form. The function $f(\cdot)$ is monotonically increasing in $n_j$, with $f(0) > 0$ since otherwise no source would ever be the first to report on a thread. Based on considerations of preferential attachment [4], it is natural to consider functional forms for $f(\cdot)$ such as $f(n_j) = (a + bn_j)$ or more generally $f(n_j) = (a+bn_j)^\gamma$. Again, we note that while the imitative effect created by $f(\cdot)$ causes large threads to appear, they cannot persist for very long due to the recency effects imposed by $\delta(\cdot)$.

[3]We thank Steve Strogatz for pointing out the analogies and contrasts with these models to us.

**Analysis and simulation results.** We find through simulation that this model produces fluctuations that are similar to what is observed in real news-cycle data. Figure 5 shows the results of a simulation of the model with the function $f$ taking a power-law functional form, and with a exponentially decaying form for the recency function $\delta$. (The threads of highest volume are depicted by analogy with Figure 4.) We see that although the model introduces no exogenous sources of variability as time runs forward, the distribution of popular threads and their co-occurrence in time can be highly non-uniform, with periods lacking in high-volume threads punctuated by the appearance of popular threads close together in time.

In Figure 6, we illustrate the basic reasons why one cannot produce these effects with only one of the two ingredients. When there is only a recency effect but no imitation (so the probability of choosing thread $j$ is proportional only to $\delta(t - t_j)$ for some function $\delta$), we see that no thread ever achieves significant volume, since each is crowded out by newer ones. When there is only imitation but no recency effect, (so the probability of choosing thread $j$ is proportional only to $f(n_j)$ for some function $f$), then a single thread becomes dominant essentially forever: there are no recency effects to drive it away, although its dominance shrinks over time simply because the total number of competing threads is increasing.

Rigorous analysis of the proposed model appears to be quite complex. However, one can give an argument for the characteristic shape of thread volume over time in Figure 5 through an approximation using differential equations. If we focus on a single thread $j$ in isolation, and view all the competing threads as a single aggregate, then the volume $X(t)$ of $j$ at time $t$ can be approximated by $X(t+1) = cf(X(t))\delta(t)$, where for notational simplicity we translate time so $t_j = 0$, and we use $c$ to denote a normalizing constant for the full distribution. Subtracting $X(t)$ from both sides to view it as a difference equation, we can in turn approximate this using the differential equation $dx/dt = cf(x)\delta(t) - x$.

For certain choices of $f(\cdot)$ and $\delta(\cdot)$ we can solve this analytically in closed-form, obtaining an expression for the volume $x(t)$ as a function of time. For example, suppose we have $f(x) = qx$ and $\delta(t) = t^{-1}$; then

$$\frac{dx}{dt} = cqxt^{-1} - x = x(cqt^{-1} - 1).$$

Dividing through by $x$ we get

$$\int \frac{1}{x}\frac{dx}{dt}dt = \int (cqt^{-1} - 1)dt$$

and hence $x = At^{cq}e^{-t}$. This function has the type of "saw-tooth" shape of increase followed by exponential decrease as in Figure 5.

Again, this functional form arises from a particular choice of $f(\cdot)$ and $\delta(\cdot)$ and is intended to give a sense for the behavior of volume over time. At a more general level, we feel that any model of the news cycle will need to incorporate, at least implicitly, some version of these two ingredients; we view a more general and more exact analysis of such models as an interesting open question.

# 4. LOCAL ANALYSIS: PEAK INTENSITY AND NEWS/BLOG INTERACTIONS

So far we have examined the dynamics of the news cycle at a global level, and proposed a simple model incorporating imitation and recency. We now analyze the process at a more fine-grained level, focusing on the temporal dynamics around the peak intensity of a typical thread, as well as the interplay between the news media and blogs in producing the structure of this peak.

**Thread volume increase and decay.** Recall that the *volume* of a

**Figure 6: Only a single aspect of the model does not reproduce dynamic behavior. With only preference to recency (left) no thread prevails as at every time step the latest thread gets attention. With only imitation (right) a single thread gains most of the attention.**

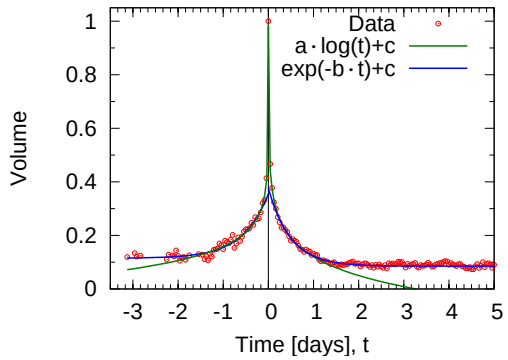


**Figure 7: Thread volume increase and decay over time. Notice the symmetry, quicker decay than buildup, and lower baseline popularity after the peak.**

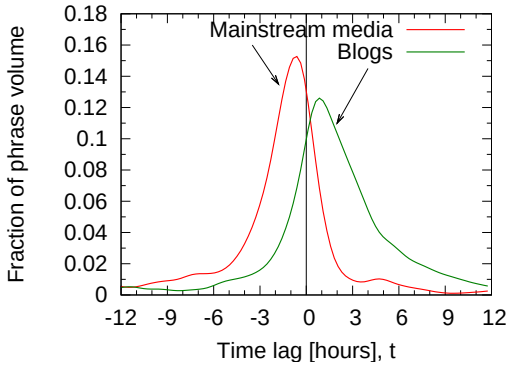


**Figure 8: Time lag for blogs and news media. Thread volume in blogs reaches its peak typically 2.5 hours after the peak thread volume in the news sources. Thread volume in news sources increases slowly but decrease quickly, while in blogs the increase is rapid and decrease much slower.**

thread at a time $t$ is simply the number of items it contains with timestamp $t$. First we examine how the volume of a thread changes over time. A natural conjecture here would be to assume an exponential form for the change in the popularity of a phrase over time. However, somewhat surprisingly we show next that the exponential function does *not* increase fast enough to model the behavior.

Given a thread $p$, we define its *peak time* $t_p$ as the median of the times at which it occurred in the dataset. We find that threads tend to have particularly high volume right around this median time, and hence the value of $t_p$ is quite stable under the addition or deletion of moderate numbers of items to $p$. We focus on the 1,000 threads with the largest total volumes (i.e. the largest numbers of items). For each thread, we determine its volume as a function of time; we then normalize and aligne these curves so that $t_p = 0$ for each, and so that the volume of each at time 0 was equal to 1. Finally, for each time $t$ we plot the median volume at $t$ over all 1,000 phrase-clusters. This is depicted in Figure 7.

In general, one would expect the overall volume of a thread to be very low initially; then as the mass media begins joining in the volume would rise; and then as it percolates to blogs and other media it would slowly decay. However, it seems that the behavior tends to be quite different from this. First, notice that in Figure 7 the rise and drop in volume is surprisingly symmetric around the peak, which suggests little or no evidence for a quick build-up followed by a slow decay. We find that no one simple justifiable function fits the data well. Rather, it appears that there are two distinct types of behavior: the volume outside an 8-hour window centered at the peak can be well modeled by an exponential function, $e^{-bx}$, while the 8-hour time window around the peak is best modeled by a logarithmic function, $a|\log(|x|)|$. The exponential function is increasing too slowly to be able to fit the peak, while the logarithm has a pole at $x = 0$ ($|\log(|x|)| \to \infty$ as $x \to 0$). This is surprising as it suggests that the peak is a point of "singularity" where the number of mentions effectively diverges. Another way to view this is as a form of Zeno's paradox: as we approach time 0 from either side, the volume increases by a fixed increment each time we shrink our distance to time 0 by a constant factor.

Fitting the function $a\log(t)$ to the spike we find that from the left $t \to 0^-$ we have $a = 0.076$, while as $t \to 0^+$ we have $a = 0.092$. This suggests that the peak builds up more slowly and declines faster. A similar contrast holds for the exponential decay parameter $b$. We fit $e^{bt}$ and notice that from the left $b = 1.77$, while after the peak $b = 2.15$, which similarly suggests that the popularity slowly builds up, peaks and then decays somewhat more quickly. Finally, we also note that the background frequency before the peak is around 0.12, while after the peak it drops to around 0.09, which further suggests that threads are more popular before they reach their peak, and afterwards they decay very quickly.

**Time lag between news media and blogs.** A common assertion about the news cycle is that quoted phrases first appear in the news media, and then diffuse to the blogosphere, where they dwell for some time. However, the real question is, how often does this happen? What about the propagation in the opposite direction? What is the time lag? As we show next, using our approach we can determine the lag within temporal resolution of less than an hour.

We labeled each of our 1.6 million sites as *news media* or *blogs*. To assign the label we used the following rule: if a site appears on Google News then we label it as news media, and otherwise we label it as a blog. Although this rule is not perfect we found it to work quite well in practice. There are 20,000 different news sites in Google News, which a tiny number when compared to 1.65 million sites that we track. However, these news media sites generate about 30% of the total number of documents in our dataset. Moreover, if we only consider documents that contain frequent phrases then the share of news media documents further rises to 44%.

By analogy with the previous experiment we take the top 1000 highest volume threads, align them so that each has an overall peak at $t_p = 0$, but now create two separate volume curves for each thread: one consisting of the subsequence of its blog items, and the other consisting of the subsequence of its news media items. We will refer to the sizes of these as the *blog volume* and *news volume* of the thread. Figure 8 plots the median news and blog volumes and reveals that this time our intuition was right. First, notice that

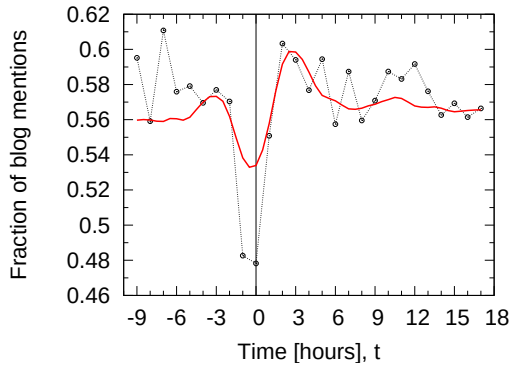


**Figure 9: Phrase handoff from news to blogs. Notice a heartbeat like pulse when news media quickly takes over a phrase.**

the median for a thread in the news media typically occurs first, and then a median of *2.5 hours* later the median for the thread among blogs occurs. Moreover, news volume both increases faster, and higher, but also decreases quicker than blog volume. For news volume we make an observation similar to what we saw in Figure 7: the volume increases slower than it decays. However, in blogs this is exactly the opposite. Here the number of mentions first quickly increases, reaches its peak 2.5 hours after the news media peak, but then decays more slowly.

One interpretation is that a quoted phrase first becomes high-volume among news sources, and is then "handed off" to blogs. The news media are slower to heavily adopt a quoted phrase and subsequently quick in dropping it, as they move on to new content. On the other hand, bloggers rather quickly adopt phrases from the news media, with a 2.5-hour lag, and then discuss them for much longer. Thus we see a pattern in which a spike and then rapid drop in news volume feeds a later and more persistent increase in blog volume for the same thread.

**Handoff of phrases from news media to blogs.** To further investigate the dynamics and transitions of phrases from the news media to the blogosphere we perform the following experiment: we take the top 1000 threads, align them so that they all peak at time $t_p = 0$, but now calculate the ratio of blog volume to total volume for each thread as a function of time.

Figure 9 shows a "heartbeat"-like like dynamics where the phrase "oscillates" between blogs and mainstream media. The fraction of blog volume is initially constant, but it turns upward about three hours before the peak as early bloggers mention the phrase. Once the news media joins in, around $t = -1$, the fraction of blog volume drops sharply; but it then jumps up after $t = 0$ once the news media begins dropping the thread and blogs continue adopting it. The fraction of blog mention peaks around $t = 2.5$, and after 6-9 hours the hand-off is over and the fractions stabilize. It is interesting that the constant fraction before the peak ($t \leq -6$) is $56\%$, while after the peak ($t \geq 9$) is actually *higher*, which suggests a persistent effect in the blogosphere after the news media has moved on. This provides a picture of the very fine-scale temporal dynamics of the handoff of news from mainstream media to blogs, aggregated at the very large scale of 90 million news articles.

**Lag of individual sites on mentioning a phrase.** We also investigate how quickly different media sites mention a phrase. Thus, we define the *lag* of a site with respect to a given thread to be the time at which the site first mentions the associated quoted phrase, minus the phrase peak time. (Negative lags indicate that the site mentioned the quoted phrase before peak attention.) This measure gives us a sense for how early or late the site takes part in the thread, relative to the bulk of the coverage. Table 1 gives a list of sites with the minimum (i.e. most negative) lags. Notice that early mentioners are blogs and independent media sites; behind them, but still well ahead of the crowd, are large media organizations.

| Rank | Lag [h] | Reported | Site |
|---|---|---|---|
| 1 | -26.5 | 42 | hotair.com |
| 2 | -23 | 33 | talkingpointsmemo.com |
| 4 | -19.5 | 56 | politicalticker.blogs.cnn.com |
| 5 | -18 | 73 | huffingtonpost.com |
| 6 | -17 | 49 | digg.com |
| 7 | -16 | 89 | breitbart.com |
| 8 | -15 | 31 | thepoliticalcarnival.blogspot.com |
| 9 | -15 | 32 | talkleft.com |
| 10 | -14.5 | 34 | dailykos.com |
| 16 | -14 | 54 | blogs.abcnews.com |
| 30 | -11 | 32 | uk.reuters.com |
| 34 | -11 | 72 | cnn.com |
| 40 | -10.5 | 78 | washingtonpost.com |
| 48 | -10 | 53 | online.wsj.com |
| 49 | -10 | 54 | ap.org |

**Table 1: How quickly different media sites report a phrase. Lag: median time between the first mention of a phrase on a site and the time when its mentions peaked. Reported: percentage of top 100 phrases that the site mentioned.**

**Quotes migrating from blogs to news media.** The majority of phrases first appear in news media and then diffuses to blogs where it is then discussed for longer time. However, there are also phrases that propagate in the opposite way, percolating in the blogosphere until they are picked up the news media. Such cases are very important as they show the importance of independent media. While there has been anecdotal evidence of this phenomenon, our approach and the comprehensiveness and the scale of our dataset makes it possible to automatically find instances of it.

To extract phrases that acquired non-trivial volume earlier in the blogosphere, we use the following simple heuristic. Let $t_m$ denote the median time of news volume for a thread. Then let $f_b$ be the fraction of the total thread volume consisting of blog items dated at least a week before $t_m$. We look for threads for which $0.15 < f_b < 0.5$. Here the threshold of $0.15$ ensures that the phrase was sufficiently mentioned on the blogosphere well before the news media peak, and $0.5$ selects only phrases that also had a significant presence in the news media.

Table 2 lists the highest-volume thread as automatically returned by our rule. Manual inspection indicates that almost all correspond to intuitively natural cases of stories that were first "discovered" by bloggers. Moreover, out of 16,000 frequent phrases we considered in this experiment 760 passed the above filter. Interpreting this ratio in light of our heuristic, it suggests that about 3.5% of quoted phrases tend to percolate from blogs to news media, while diffusion in the other direction is much more common.

## 5. CONCLUSION

We have developed a framework for tracking short, distinctive phrases that travel relatively intact through on-line text and presented scalable algorithms for identifying and clustering textual variants of such phrases that scale to a collection of 90 million articles, which makes the present study one of the largest analyses of on-line news in terms of data scale. Our work offers some of the first quantitative analyses of the global news cycle and the dynamics of information propagation between mainstream and social media. In particular, we observed a typical lag of 2.5 hours between the peaks of attention to a phrase in the news media and in blogs,

| $M$ | $f_b$ | Phrase |
|---|---|---|
| 2,141 | .30 | Well uh you know I think that whether you're looking at it from a theological perspective or uh a scientific perspective uh answering that question with specificity uh you know is uh above my pay grade. |
| 826 | .18 | A changing environment will affect Alaska more than any other state because of our location I'm not one though who would attribute it to being man-made. |
| 763 | .18 | It was Ronald Reagan who said that freedom is always just one generation away from extinction we don't pass it to our children in the bloodstream we have to fight for it and protect it and then hand it to them so that they shall do the same or we're going to find ourselves spending our sunset years telling our children and our children's children about a time in America back in the day when men and women were free. |
| 745 | .18 | After trying to make experience the issue of this campaign John McCain celebrated his $72^{nd}$ birthday by appointing a former small town mayor and brand new governor as his vice presidential nominee is this really who the republican party wants to be one heartbeat away from the presidency given Sarah Palin's lack of experience on every front and on nearly every issue this vice presidential pick doesn't show judgement it shows political panic. |
| 670 | .38 | Clarion fund recently financed the distribution of some 28 million DVDs containing the film obsession radical islam's war against the west in what many political analysts describe as swing states in the upcoming presidential elections. |

**Table 2: Phrases first discovered by blogs and only later adopted by the news media. $M$: total phrase volume, $f_b$: fraction of blog mentions before 1 week of the news media peak.**

with a "heartbeat"-like shape of the handoff between news and blogs. We also developed a mathematical model for the kinds of temporal variation that the system exhibits. As information mostly propagates from news to blogs, we also found that in only 3.5% of the cases stories first appear dominantly in the blogosphere and subsequently percolate into the mainstream media.

Our approach to meme-tracking opens an opportunity to pursue long-standing questions that before were effectively impossible to tackle. For example, how can we characterize the dynamics of mutation within phrases? How does information change as it propagates? Over long enough time periods, it may be possible to model the way in which the essential "core" of a widespread quoted phrase emerges and enters popular discourse more generally. One could combine the approaches here with information about the political orientations of the different news media and blog sources [2, 12, 13], to see how particular threads move within and between opposed groups. Introducing such types of orientation is challenging, however, since it requires reliable methods of labeling significant fractions of sources at this scale of data. Finally, a deeper understanding of simple mathematical models for the dynamics of the news cycle would be useful for media analysts; temporal relationships such as we find in Figure 8 suggest the possibility of employing a type of two-species predator-prey model [18] with blogs and the news media as the two interacting participants. More generally, it will be useful to further understand the roles different participants play in the process, as their collective behavior leads directly to the ways in which all of us experience news and its consequences.

**Acknowledgements.** We thank David Strang and Steve Strogatz for valuable conversations and the creators of Flare and Spinn3r for resources that facilitated the research.

# 6. REFERENCES

[1] Supporting website: `http://memetracker.org/supp`
[2] L. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election. *Workshop on Link Discovery*, 2005.
[3] E. Adar, L. Zhang, L. Adamic, R. Lukose. Implicit structure and dynamics of blogspace. *Wks. Weblogging Ecosystem'04*.
[4] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. of Modern Phys.*, 74:47–97, 2002.
[5] J. Allan (ed). *Topic Detection and Tracking.* Kluwer, 2002.
[6] L. Bennett. *News: The Politics of Illusion*. A. B. Longman (Classics in Political Science), seventh edition, 2006.
[7] D. Blei, J. Lafferty. Dynamic topic models. *ICML*, 2006.
[8] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, pages 3:993–1022, 2003.
[9] G. Calinescu, H. Karloff, Y. Rabani. An improved approximation algorithm for multiway cut. *JCSS* 60(2000).
[10] E. Dahlhaus, D. S. Johnson, C. H. Papadimitriou, P. D. Seymour, and M. Yannakakis. The complexity of multiterminal cuts. *SIAM J. Comput.*, 23(4):864–894, 1994.
[11] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *WWW '04*, 2004.
[12] M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. C. Kanig. Blews: Using blogs to provide context for news articles. In *ICWSM '08*, 2008.
[13] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In *ICWSM '07*, 2007.
[14] D. Gruhl, D. Liben-Nowell, R. V. Guha, and A. Tomkins. Information diffusion through blogspace. In *WWW '04*, 2004.
[15] J. Harsin. The rumour bomb: Theorising the convergence of new and old trends in mediated U.S. politics. *Southern Review: Communication, Politics and Culture*, 39(2006).
[16] S. Havre, B. Hetzler, L. Nowell. ThemeRiver: Visualizing theme changes over time. *IEEE Symp. Info. Vis.* 2000.
[17] J. Kleinberg. Bursty and hierarchical structure in streams. In *KDD '02*, pages 91–101, 2002.
[18] M. Kot. *Elements of Mathematical Ecology*. Cambridge University Press, 2001.
[19] B. Kovach and T. Rosenstiel. *Warp Speed: America in the Age of Mixed Media*. Century Foundation Press, 1999.
[20] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *CACM*, 47(12):35–39, 2004.
[21] M. Lacker and C. Peskin. Control of ovulation number in a model of ovarian follicular maturation. In *AMS Symposium on Mathematical Biology*, pages 21–32, 1981.
[22] P.F. Lazarsfeld, B. Berelson, and H. Gaudet. *The People's Choice.* Duell, Sloan, and Pearce, 1944.
[23] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, M. Hurst. Cascading behavior in large blog graphs. *SDM'07*.
[24] R. D. Malmgren, D. B. Stouffer, A. Motter, and L. A. N. Amaral. A poissonian explanation for heavy tails in e-mail communication. *PNAS*, to appear, 2008.
[25] J. Schmidt. Blogging practices: An analytical framework. *Journal of Computer-Mediated Communication*, 12(4), 2007.
[26] J. Singer. The political j-blogger. *Journalism*, 6(2005).
[27] Spinn3r API. `http://www.spinn3r.com`. 2008.
[28] M. L. Stein, S. Paterno, and R. C. Burnett. *Newswriter's Handbook: An Introduction to Journalism*. Blackwell, 2006.
[29] A. Vazquez, J. G. Oliveira, Z. Deszo, K.-I. Goh, I. Kondor, and A.-L. Barabasi. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(036127), 2006.
[30] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. *Proc. KDD*, 2006.
[31] X. Wang, C. Zhai, X. Hu, R. Sproat. Mining correlated bursty topic patterns from coordinated text streams. *KDD*, 2007.
[32] F. Wu and B. Huberman. Novelty and collective attention. *Proc. Natl. Acad. Sci. USA*, 104, 2007.