Your Two Weeks of Fame and your Grandmother's

James Cook
UC Berkeley [*]
jcook@cs.berkeley.edu

Atish Das Sarma
Google Research
atish.dassarma@gmail.com

Alex Fabrikant
Google Research
fabrikant@google.com

Andrew Tomkins
Google Research
atomkins@gmail.com

ABSTRACT

Did celebrity last longer in 1929, 1992 or 2009? We investigate the phenomenon of fame by mining a collection of news articles that spans the twentieth century, and also perform a side study on a collection of blog posts from the last 10 years. By analyzing mentions of personal names, we measure each person's time in the spotlight, and watch the distribution change from a century ago to a year ago. We expected to find a trend of decreasing durations of fame as news cycles accelerated and attention spans became shorter. Instead, we find a remarkable consistency through most of the period we study. Through a century of rapid technological and societal change, through the appearance of Twitter, communication satellites and the Internet, we do not observe a significant change in typical duration of celebrity. We also study the most famous of the famous, and find different results depending on our method for measuring duration of fame. With a method that may be thought of as measuring a spike of attention around a single narrow news story, we see the same result as before: stories last as long now as they did in 1930. A second method, which may be thought of as measuring the duration of public interest in a person, indicates that famous people's presence in the news is becoming longer rather than shorter, an effect most likely driven by the wider distribution and higher volume of media in modern times. Similar studies have been done with much shorter timescales specifically in the context of information spreading on Twitter and similar social networking site. However, to the best of our knowledge, this is the first massive scale study of this nature that spans over a century of archived data, thereby allowing us to track changes across decades.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and behavioral sciences

Keywords

social media, time series, historical trends, fame durations, news archives

1. INTRODUCTION

Beginning in the 19th century, long-distance communication transitioned from foot to telegraph on land, and from sail to steam to cable by sea. Each new form of technology began with a limited number of dedicated routes, then expanded to reach a large fraction of the accessible audience, eventually resulting in near-complete deployment of digital electronic communication. Each transition represented an opportunity for news to travel faster, break more uniformly, and reach a broad audience closer to its time of inception.

Even today, the increasing speed of the news cycle is a common theme in discussions of the societal implications of technology. Stories break faster, are covered in less detail, and news sources quickly move on to other topics. Online and cable outlets aggressively search for novelty in order to keep eyeballs glued to screens. Popular non-fiction dedicates significant coverage to this trend, which by 2007 prompted a satirical website entitled *The Onion* to offer the following commentary on cable news provider CNN's offerings: "CNN is widely credited with initiating the acceleration of the modern news cycle with the fall 2006 debut of its spin-off channel CNN:24, which provides a breaking news story, an update on that story, and a news recap all within 24 seconds."

With this speed-up of the news cycle comes an associated concern that, whether or not causality is at play, attention spans are shorter, and consumers are able to focus for increasingly brief periods on a particular news subject. Stories that might previously have occupied several days of popular attention might emerge, run their course, and vanish in a single day. This theory is consistent with a suggestion by Herbert Simon [9] that as the world grows rich in information, the attention necessary to process that information becomes a scarce and valuable resource.

The speed of the news cycle is a difficult concept to pin down. We focus our study on the most common object of news: the individual. An individual's fame on a particular day might be thought of as the frequency with which a person reading the news at random would see their name. From this idea we develop two notions of the duration of the interval of discussion of an individual. The first is based on falloff from a peak, and intends to capture the spike around a narrow news story. The second looks for period of sustained public interest in an individual, from the time the public first notices that person's existence until the public loses interest and the name stops appearing in the news. We study the interaction of these two notions of "duration of fame" with the radical shifts in the news cycle we outline above. For this purpose, we employ Google's public news

[*] Work done while at interning at Google.

WWW 2012, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1229-5/12/04.

archive corpus, which contains over sixty million pages covering 250 years, and we perform what we believe to be the first study of the dynamics of fame over such a time period.

Data within the archive is heterogeneous in nature, ranging from directly captured digital content to optical character recognition employed against microfilm representations of old newspapers. The crawl is not complete, and we do not have full information about which items are missing. Rather than attempt topic detection and tracking in this error-prone environment, we instead directly employ a recognizer for person names to all content within the corpus; this approach is more robust, and more aligned with our goal of studying fame of individuals.

Based on these different notions of periods of reference to a particular person, we develop at each point in time a distribution over the duration of fame of different individuals.

Our expectation upon undertaking this study was that in early periods, improvements to communication would cause the distribution of duration of coverage of a particular person to shrink to the left. Through the 20th century, we hypothesized that the continued deployment of technology, and the changes to modern journalism resulting from competition to offer more news faster, would result in the duration of fame continuing to shrink over the course of the century into the present day.

Summary of findings.

We did indeed observe effects through the early 20th century in line with our hypothesis regarding communications. However, from 1920 to 1990, after newspapers had stabilized into roughly their current form, we saw a quite different picture. Over the course of a century, through a world war, a global depression, a two order of magnitude growth in volume, and a technological curve moving from horse-drawn carriages to satellite communication, we saw little change in the distribution of story durations or the distribution of continuous public attention. A side study on a corpus consisting entirely of blog posts over the last eight years from the Blogger blog hosting organization, which has a radically different focus from professional news media, shows once again the same distribution. This distribution is heavy-tailed, with power law exponent around -2.5.

We repeated our procedure after removing all but the most famous of the famous names: in one experiment, we kept names which were in the top 1000 in frequency for some year, and in a second, we kept names which were in the top 0.1% for some year. In all cases, found that the more popular names tended to have longer periods of fame. When we measured story durations, we saw the same thing as before: the distribution of durations of news stories for the most-mentioned names did not change between 1920 and 1990. However, when instead looked at the duration of public attention toward an individual, we found that the most famous of the famous have found steadily longer and longer durations in the news starting in 1920.

In the case of taking top 1000 names in each year, the increasing could be explained as follows: as the corpus increases in volume toward later years, a larger number of names appear, representing more draws from the same underlying distribution of fame durations. The quantiles of the distribution of duration for the top 1000 elements will therefore grow over time as the corpus volume increases. On the other hand, our experiment that took the top 0.1% of names still showed in increasing trend, although with a smaller rate of increase. We therefore conclude that the increasing trend is not completely caused by an increase in corpus volume.

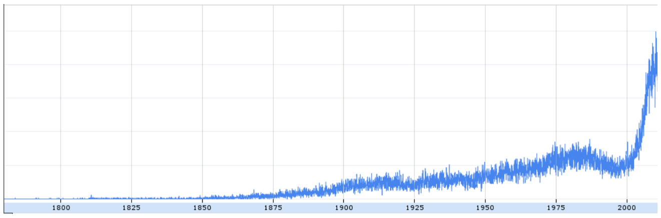

Figure 1: The volume of news articles by date.

To summarize, we find that the most popular figures in today's news stay in the limelight for longer than their counterparts did in the past. At the same time, however, the average person remains in the limelight for essentially the same amount of time today as in the past, and the length of individual news stories has not changed, either for the average person or for the most popular.

2. WORKING WITH THE NEWS CORPUS

We perform our main study on a collection of the more than 60 million news articles in the Google archive that are both (1) in English, and (2) searchable and readable by Google News users at no cost. In Section 6, we cross-validate our observations against the corpus of public blog posts on Blogger, which is described there.

The articles of the news corpus span a wide range of time, with the relative daily volume of articles over the range of the corpus shown in Figure 1. There are a handful of articles from the late 18th century onward, and the article coverage grows rapidly over the course of the 19th century. From the last decade of the 19th century through the end of the corpus (March 2011), there is consistently a very substantial volume of articles per day, as well as a wide diversity of publications. For the sake of statistical significance, our study focuses on the years 1895–2011.

The news corpus contains a mix of modern articles obtained from the publisher in the original digital form, as well as historical articles scanned from archival microform and OCRed, both by Google and by third parties. For scanned articles, per-article metadata such as titles, issue dates, and boundaries between articles are also derived algorithmically from the OCRed data, rather than manually curated.

Our study design was driven by several features that we discovered in this massive corpus. We list them here to explain our study design. Also, data mining for high-level behavioral patterns in a diachronous, heterogeneous, partially-OCRed corpus of this scale is quite new, precedented on this scale perhaps only by [8] (which brands this new area as "culturomics"). But, with the rapid digitization of historical data, we expect such work to boom in the near future. We thus hope that the lessons we have learned about this corpus will also be of independent interest to others examining this corpus and other similar archive corpuses.

2.1 Corpus features, misfeatures, and missteps

2.1.1 News mentions as a unit of attention

Our 116-year study of the news corpus aims to extend the rich literature studying topic attention in online social media like Twitter, typically over the span of the last 3–5 years. Needless to say, 100-year-old printed newspapers are an imperfect proxy for the attention of individuals, which has only recently become directly observable via online behavior. Implicit in the heart of our study is the assumption that news articles are published to serve an audience, and the media makes an effort, even if imperfect, to cater to the audience's information appetites. We coarsely approximate a unit of attention as one occurrence in a Google News archive article, and we leave open a number of natural extensions to this work, such as weighting articles by historical publication subscriber counts, or by size and position on the printed page.

Due to the automated OCR process, not every "item" in the corpus can be reasonably declared a news article. For example, a single photo caption might be extracted as an independent article, or a sequence of articles on the same page might be misinterpreted as a single article. Rather than weighting each of these corpus items equally when measuring the attention paid to a name, we elected to count multiple mentions of a name within an item separately, so that articles will tend to count more than captions, and there is no harm in mistakenly grouping multiple articles as one.

We manually examined (A) a uniform sample of 50 articles from the whole corpus (which, per Fig. 1, contains overwhelmingly articles from the last decade), and (B) a uniform sample of 50 articles from 1900–1925. We classified each sample into:

- News articles: timely content, formatted as a standalone "item", published without external sponsorship, for the benefit of part of the publication's audience,
- News-like items: non-article text chunks where a name mention can qualify as that person being "in the news" — e.g. photo captions or inset quotes,
- Non-news: ads and paid content, sports scores, recipes, news website comments miscategorized as news, etc.

The number of items of each type in the two samples are given in the following table.

	full corpus sample	1900–1925 sample
news articles	31	28
news-like items	3	2
non-news items	16	20

We expect that the similarity in these distributions should result in minimal noise in the cross-temporal comparisons, and leave to future work the task of automatically distinguishing real news stories from non-news.

2.1.2 Compensating for coverage

Even once we discard the more sparsely covered 18th and 19th centuries, there is still more than an order of magnitude difference between article volume in 1895 and 2011. We address these coverage differences by downsampling the data down to the same number of articles for each month in this range. We address the nuanced effects of this downsampling on our methodology in Section 3.3.

2.1.3 Evolution of discourse and media — why names?

We set out originally to understand changes in the public's attention as measured by news story topics. There are a myriad heuristics to define a computationally feasible model of a "single topic" that can be thought to receive and lose the public's attention. But over the course of a century, the changes in society, media formatting, subjects of public discourse, writing styles, and even language itself are substantial enough that neither sophisticated statistical models trained on plentiful, well-curated training data from modern media nor simple generic approaches like word co-occurrence in titles are guaranteed to work well. Very few patterns connect articles from 1910 newspapers' "social" sections (now all but forgotten) about tea at Mrs. Smith's, to 1930 articles about the arrival of a trans-oceanic liner, to 2009 articles about a viral Youtube video.

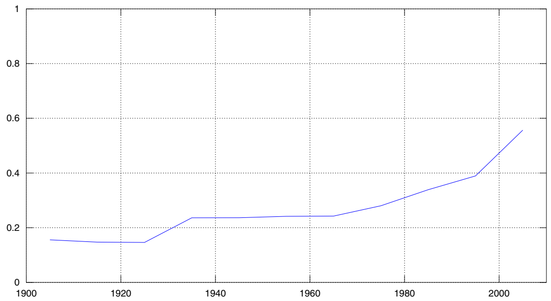

Figure 2: Articles with recognized personal names per decade

After trying out general proper noun phrases produced inconclusively noisy results, we decided to focus on occurrences of personal names, detected in the text by a proprietary state-of-the-art statistical recognizer. Personal names have a relatively stable presence in the media: even with high OCR error rates in old microform, over 1/7th of the articles even in the earliest decades since 1900 contain recognized personal names (see Figure 2).

But personal names are not without historical caveats, either. A woman appearing in 2005 stories as "Jane Smith" would be much more likely to be exclusively referenced as "Mrs. Smith", or even "Mrs. John Smith", in 1915. Also, the English-speaking world was much more Anglo-centric in 1900 than now, with much less diversity of names. An informal sample suggests that most names with non-trivial news presence 100 years ago referred overwhelmingly to a single bearer of that name for the duration of a particular news topic, but many names are not unique when taken across the duration of the whole corpus — for instance, "John Jacob Astor", appearing in the news heavily over several decades (Fig. 3), in reference to a number of distinct relatives. On account of both of these phenomena, among others, we aim to focus on name appearance patterns that are most likely to represent a single news story or contiguous span of public attention involving that person, rather than trying to model the full media "lifetime" of individuals, as we had considered doing at the start of this project.

2.1.4 OCR errors in data and metadata

We empirically discovered another downfall of studying long-term "media lifetimes" of individuals. In an early experiment, we measured, for each personal name, the 10th and 90th percentiles of the dates of that name's occurrence in the news. We then looked at the time interval between 10th and 90th percentiles, postulating that a large enough fraction of names are unique among newsworthy individuals that the distribution of these *inter-quantile gaps* could be a robust measure of media lifetime. After noticing a solid fraction of the dataset showing inter-quantile gaps on the scale of 10-30 years, we examined a heat map of gap durations, and discovered a regular pattern of gap durations at exact-integer year offsets, which, other than for Santa Claus, Guy Fawkes, and a few other clear exceptions, seemed an improbable phenomenon.

This turned out to be an artifact of OCRed metadata. In particular, the culprit was single-digit OCR errors in the *scanned article year*. While year errors are relatively rare, every long-tail name that occurred in fewer than 10 articles (often within a day or two of each other), and had a mis-OCRed error for one of those occurrences contributed probability mass to integral-number-of-years media lifetimes. As extra evidence, the heat map had a distinct outlier segment of high probability mass for inter-quantile range of exactly 20 years, for end dates ranging from 1980 to 1989 — the digits 6 and 8 being particularly easy to mistake on blurry microfilm. Note that short-term phenomena are relatively safe from OCR date errors, thanks to the common English convention of written-out month names, and to the low impact of OCR errors in the day number.

OCR errors in the article text itself are ubiquitous. Conveniently, the edit distance between two recognizable personal names is rarely very short, so by agreeing to discard any name that occurs only once in the corpus, we are likely to discard virtually all OCR errors as well, with no impact on data on substantially newsworthy people. We should note that OCR errors are noticeably more frequent on older microfilm, but the reasonable availability of recognizable personal names even in 100-year-old articles, per Fig. 2, suggests that this problem is not dire. A manually-coded sample of 50 articles with recognized names from the first decade of the 1900s showed only 8 out of 50 articles having incorrectly recognized names (including both OCR errors and non-names mis-tagged as names).

2.1.5 Simultaneity and publishing cycles

There are also pitfalls with examining short timelines. In the earliest decades we examine, telegraph was widely available to news publishers, but not fully ubiquitous, with rural papers often reporting news "from the wire" several days after the event. An informal sample seems to suggest that most news by 1900 propagated across the world on the scale of a few days. Also, many publications in the corpus until the last 20 years or so were either published exclusively weekly or, in the case of Sunday newspaper issues, had substantially higher volume once a week, resulting in many otherwise obscure names having multiple news mentions separated by one week — a rather different phenomenon than a person remaining in the daily news for a full week. On account of both of these, we generally disregard news patterns that are shorter than a few days in our study design.

3. MEASURING FAME

We begin by producing a list of names for each article. To do this, we extract short capitalized phrases from the body text of each article, and keep phrases recognized by an algorithm to be personal names.

For every name that appears in the input, we consider that name's *timeline*, which is the multiset of dates at which that name appears, including multiple occurrences within an article. We intend the timeline to approximate the frequency with which a person browsing the news at random on a given day would encounter that name. The accuracy of this approximation will depend on the volume of news articles available. In order to avoid the possibility that any trends we detect are caused by variations in this accuracy caused by variations in the volume of the corpus, we randomly choose an approximately equal number of articles to work with from each month. We describe and analyze this process in Section 3.3.

In general, our method can be applied to any collection of timelines. In Section 6, we apply it to names extracted from blog posts.

3.1 Finding Periods of Fame

Once we have computed a timeline for each name that appears in the corpus, we select a time during which we consider that name to have had its period of fame, using one of the two methods described below. In order to compare the phenomenon of fame at different points in time, we consider the joint distribution of two variables over the set of names: the *peak date* and the *duration* of the name's period of fame. We try the following two methods to compute a peak date and duration for each timeline.

- **Spike method**. This method intends to capture the spike in public attention surrounding a particular news story. We divide time into one-week intervals and consider the name's rate of occurrence in each interval. The week with the highest rate is considered to be the peak date, and the period extends backward and forward in time as long as the rate does not drop below one tenth its maximum rate. Yang and Leskovec [12] used a similar method in their study of digital media, using a time scale of hours where we use weeks.

- **Continuity method**. This method intends to measure the duration of public interest in a person. We define a name's period of popularity to be the longest span of time within which there is no seven-day period during which it is not mentioned. The peak date falls halfway between the beginning and the end of the period. (In Section 5, we will find that the durations are short compared to the time span of the study, so using any choice of peak date between the beginning and end will produce a similar distribution.)

To demonstrate the distinction between these two methods, Figure 3 shows the occurrence timeline for Marilyn Monroe. The "continuity method" picks out the bulk of her fame — 1952-02-13 ("A") through 1961-11-15 ("D"), by which point her appearance in the news was reduced to a fairly low background level. The "spike method" picks out the intense spike in interest surrounding her death, yielding the range 1962-7-18 ("E") – 1962-8-29 ("H").

Very often these two methods identify short moments of fame within a much longer context. For example, in Figure 3, we see the timeline for the name "John Jacob Astor", normalized by article counts. The spike method identifies as the peak the death of John Jacob Astor III of the wealthy Astor family, with a duration of 38 days (March 8 to February 15, 1890). The continuity method identifies instead the death of his nephew John Jacob Astor IV, who died on the Titanic, with a period of five months [11]. The period begins

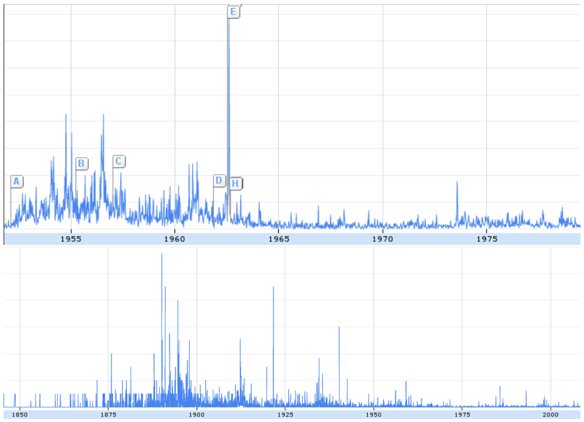

Figure 3: Timelines for "Marilyn Monroe" (top) and "John Jacob Astor" (bot).

on March 23, 1912, three weeks before the Titanic sank, and ends August 31. Many of the later occurrences of the name are historical mentions of the sinking of the Titanic.

3.2 Choosing the Set of Names

Basic filtering.

In all our experiments, to reduce noise, we discard the names which occurred less than ten times, or whose fame durations are less than two days. (In both methods, a name whose fame begins Monday and ends Wednesday is considered to have a duration of two days.) We also remove peaks that end in 2011 or later, since these peaks might extend further if our news corpus extended further in the future.

Top 1000 by year.

For each peak type, we repeat our experiment with the set of names restricted in the following way. We counted the total number of times each name appeared in each year (counting repeats within an article). For each year, we produced the set of the 1000 most frequently mentioned names in that year. We took the union of these sets over all years, and ran our experiments using only the names in this set. Note that a name's peak of popularity need not be the same year in which that name was in the top 1000: so if a name is included in the top-1000 set because it was popular in a certain year, we may yet consider that name's peak date to be a different year.

Top 0.1% by year.

We consider that filtering to the top 1000 names in each year might introduce the following undesirable bias. Suppose names are assigned peak durations according to some universal distribution, and later years have more names, perhaps because of the increasing volume of news. If a name's frequency of occurrence is proportional to its duration, then selecting the top 1000 names in each year will tend to produce names with longer durations of fame in years with a greater number of names. With this in mind, we considered one more restriction on the set of names. In each year y, we considered the total number of distinct names n_y mentioned in that year. We then collected the top $n_y/1000$ names in each year y. We ran our experiments using only the names in the union of those sets. As with the top-1000 filtering, a name's peak date will not necessarily be the same year for which it was in the top 0.1% of names.

3.3 Sampling for Uniform Coverage

The spike and continuity methods for identifying periods of fame may be affected by the volume of articles available in our corpus. For example, suppose a name's timeline is generated stochastically, with every article between February 1 and March 31 containing the name with a 1% probability. If the corpus contains 10000 articles in every week, then both the spike and continuity methods will probably decide that the article's duration is two months. However, if the corpus contains less than 100 articles in each week, then the durations will tend to be short, since there will be many weeks during which the name is not mentioned.

We propose a model for this effect. Each name ν has a "true" timeline which assigns to each day t a probability $f_\nu(t) \in [0, 1]$ that an article on that day will mention ν.[1] For each day, there is a total number of articles n_t; we have no knowledge of the relation between n_t and ν, except that there is some lower bound $n_t > n_{\min}$ for all t within some reasonable range of time. Then we suppose the timeline for name ν is a sequence of independent random variables $X_{\nu,t} \sim \text{Binom}(f_\nu(t), n_t)$. Our goal is to ensure that any measurements we take are independent of the values n_t.

To accomplish this independence of news volume, we randomly sampled news articles so that the expected number in each month was $n_{\min}$. Let $X'_{\nu,t}$ be the number of sampled articles containing name ν. If we were to randomly sample $n_{\min}$ articles without replacement, then we would have $X'_{\nu,t} \sim \text{Binom}(f_\nu(t), n_{\min})$. Notice that the joint distribution of the random variables $X'_{\nu,t}$ is unaffected by the article volumes n_t. Any further measurement based on the variables $X'_{\nu,t}$ will therefore also be unrelated to the sequence n_t. In practice, instead of sampling exactly $n_{\min}$ articles without replacement, we flipped a biased coin for each of the n_t articles at time t, including each article with probability $n_{\min}/n_t$. For a large enough volume of articles, the resulting measurements will be the same.

We removed all articles published before 1895, since the months before 1895 had less than our target number $n_{\min}$ of articles. We also removed articles published after the end of the year 2010, to avoid having a month with news articles at the beginning but not the end of the month, but with the same number of sampled articles.

As an example of the effect of downsampling, the blue dotted lines in Figure 9 show the 50th, 90th and 99th percentiles of the distribution of fame durations using the continuity method. We see that they increase suddenly in the last ten years, when our coverage of articles surges with the digital age. The red lines show the same measurement after downsampling: the surge no longer appears.

3.4 Graphing the Distributions

We graph the joint distribution of peak dates and durations in two different ways. We consider the set of names which peak in successive five-year periods. Among each set of names, we graph the 50th, 90th and 99th percentile durations of fame. These appear as darker lines in the graphs;

[1] In fact, articles could mention the name multiple times, but in the limit of a large number of articles, this will not affect our analysis.

for example, the top of Fig. 6 shows the distribution for the spike method. The lighter solid red lines show the same three quantiles for shorter three-month periods. For comparison, the dashed light blue lines show the same results if the article sampling described in Sec. 3.3 is not performed (and articles before 1895 and after 2010 are not removed). Fig. 9 shows the same set of lines using the continuity method. All the later figures are produced in the same way, except they do not include the non-sampled full distributions.

The second type of graph focuses on one five-year period at a time. The bottom of Fig. 6 shows a cumulative plot showing the number of names with duration greater than that shown on the x-axis. This is plotted for many five-year periods. The graphs of measurements using the spike method look more like step functions because that method measures durations in seven-day increments, whereas the longest-stretch method can yield any number of days. (Recall that peaks that last less than two days are removed.)

3.5 Estimating Power Law Exponents

We test the hypothesis that the tail of the distribution of fame durations follows a power law. For a given five-year period, we collect all names which peak in that period, and consider 20% of the names with the longest fame durations – that is, we set $d_{\min}$ to be the 80th percentile of durations, and consider durations $d > d_{\min}$. Among those 20%, we compute a maximum likelihood estimate of the power law exponent $\hat{\alpha}$, predicting that the probability of a duration $d > d_{\min}$ is $p(d) \propto d^{\hat{\alpha}}$. Clauset et al [2] show that the maximum likelihood estimate $\hat{\alpha}$ is given by $\hat{\alpha} = 1+(\sum_{i=1}^{n} \ln(d_i/d_{\min}))$. We include a line on each plot of cumulative distributions of fame durations, of slope $\hat{\alpha}+1$ on the log-log graph because we plot cumulative distributions rather than density functions. The $\hat{\alpha}$ values we measure are discussed in the following sections, and summarized in Figure 4 for the news corpus and Figure 5 for the blog corpus.

3.6 Statistical Measurements

We used bootstrapping to estimate the uncertainty in the four statistics we measured: the 50th, 90th and 99th percentile durations and of the best-fit power law exponents. For selected five-year periods, we sampled $|S|$ names with replacement from the set S of names that peaked in that period of time. For each statistic, we repeated this process 25000 times, and reported the range of numbers within which 99% of our samples fell. The results are presented in Figures 4 (for the news corpus) and 5 (for the blog corpus).

4. RESULTS: SPIKE METHOD

We measure periods of popularity using the spike method described in Section 3, and plot the distribution of durations as it changes over time. We find that for most of the period of our study (1920-1990), the distribution of durations changes little. When we restrict to more popular names as described in Section 3.2, we observe longer durations across time, but we still see the same flatness in the range 1920-1990. This indicates that the length of time stories stay in the news has not changed much in the past several decades.

4.1 Basic filtering

When we use the basic filtering described in Section 3.2, most of the names have the shortest possible duration of one week[2] (Figure 6). The 90th and 99th percentiles show a slight decreasing trend before 1920 and a slight increasing trend after 1990 but change little between the years 1920-90.

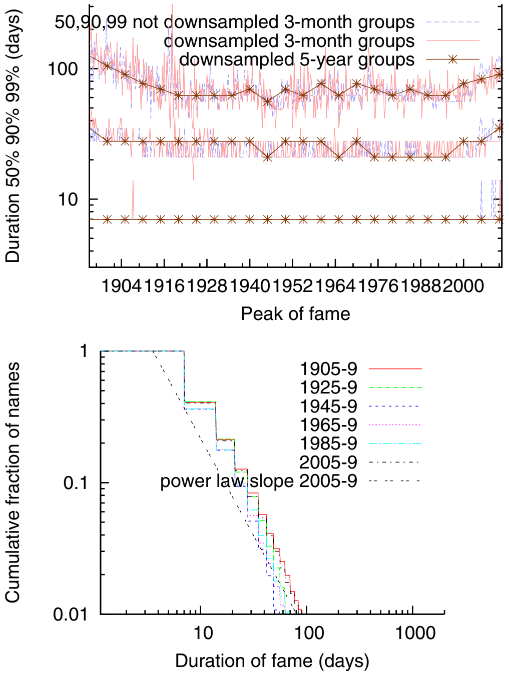

Figure 6: Fame durations measured using the spike method, plotted as the 50th, 90th and 99th percentiles over time (top) and for specific five-year periods (bottom). The bottom graph also includes a line showing the max-likelihood power law exponent for the years 2005-9. (The slope on the graph is one plus the exponent from Fig. 4, since we graph the cumulative distribution function.) To illustrate the effect of sampling for uniform article volume, the first graph includes measurements taken before sampling; see Sec. 3.3.

The upper-right corner of the table in Figure 4 shows the maximum likelihood estimates of the power law exponents for various five-year-long peak periods. The values have a fairly small range, between -2.63 and -2.32. We include a reference line of slope[3] -1.48 on the log-log graph for the period 2005-9, which has an estimated exponent of -2.48.

4.2 Top 1000 by year

When we restrict to names that were in the 1000 most popular for some year (Section 3.2), we see a median duration of three to four weeks (Figure 7). This is longer than the median of one week we see with basic filtering, indicating that more popular names tend to have longer-lasting news stories. As with the basic filtering, we do not see any clear trend of change between the years 1920 and 1990, but we

[2] Note that fame periods where the two weeks adjacent to the highest-rate week are both less than 10% of the maximum rate are considered to have a duration of zero, and were therefore removed according to our rule that we ignore peaks that last for less than two days. The smallest duration that can appear is therefore one week, meaning exactly one of the two adjacent weeks was above the 10% threshold.

[3] The slope is one more than the power law exponent, since our chart is of the cumulative distribution function.

method	filtering	period	50th %ile	90th %ile	99th %ile	power law exponent
spike	all	1905-9	7 (7 .. 7)	28 (28 .. 28)	91 (78 .. 106)	-2.45 (-2.55 .. -2.21)
spike	all	1925-9	7 (7 .. 7)	28 (28 .. 28)	65 (63 .. 78)	-2.63 (-2.74 .. -2.33)
spike	all	1945-9	7 (7 .. 7)	21 (21 .. 28)	56 (49 .. 63)	-2.44 (-2.50 .. -2.38)
spike	all	1965-9	7 (7 .. 7)	21 (21 .. 28)	63 (56 .. 70)	-2.37 (-2.44 .. -2.31)
spike	all	1985-9	7 (7 .. 7)	21 (21 .. 28)	70 (63 .. 78)	-2.32 (-2.36 .. -2.27)
spike	all	2005-9	7 (7 .. 7)	28 (28 .. 28)	84 (78 .. 91)	-2.48 (-2.53 .. -2.43)
spike	top 1000	1905-9	21 (21 .. 21)	63 (56 .. 70)	155 (133 .. 192)	-2.75 (-3.15 .. -2.56)
spike	top 1000	1925-9	21 (14 .. 21)	49 (46 .. 56)	91 (78 .. 113)	-3.22 (-3.74 .. -2.99)
spike	top 1000	1945-9	21 (14 .. 21)	49 (42 .. 49)	91 (70 .. 130)	-3.33 (-3.73 .. -2.89)
spike	top 1000	1965-9	21 (21 .. 21)	56 (49 .. 63)	119 (99 .. 164)	-2.90 (-3.54 .. -2.65)
spike	top 1000	1985-9	21 (21 .. 28)	63 (56 .. 78)	161 (121 .. 366)	-2.85 (-3.19 .. -2.57)
spike	top 1000	2005-9	35 (28 .. 35)	99 (84 .. 119)	309 (224 .. 439)	-2.64 (-2.96 .. -2.44)
spike	top 0.1%	1905-9	35 (28 .. 42)	122 (91 .. 155)	289 (161 .. 381)	-2.82 (-3.96 .. -2.36)
spike	top 0.1%	1925-9	28 (21 .. 35)	63 (56 .. 82)	145 (91 .. 218)	-3.49 (-4.82 .. -2.92)
spike	top 0.1%	1945-9	21 (21 .. 28)	56 (49 .. 67)	133 (84 .. 161)	-3.35 (-4.32 .. -2.78)
spike	top 0.1%	1965-9	28 (21 .. 35)	70 (63 .. 99)	162 (119 .. 494)	-2.90 (-3.77 .. -2.47)
spike	top 0.1%	1985-9	35 (28 .. 35)	90 (70 .. 113)	327 (140 .. 443)	-2.66 (-3.13 .. -2.35)
spike	top 0.1%	2005-9	35 (35 .. 42)	119 (99 .. 140)	338 (263 .. 557)	-2.76 (-3.10 .. -2.44)
continuity	all	1905-9	7 (7 .. 7)	20 (19 .. 21)	70 (64 .. 79)	-2.67 (-2.76 .. -2.59)
continuity	all	1925-9	7 (7 .. 7)	18 (17 .. 19)	64 (56 .. 71)	-2.64 (-2.72 .. -2.53)
continuity	all	1945-9	7 (7 .. 7)	16 (15 .. 16)	53 (49 .. 58)	-2.74 (-2.82 .. -2.66)
continuity	all	1965-9	7 (7 .. 7)	17 (16 .. 18)	66 (58 .. 75)	-2.58 (-2.69 .. -2.52)
continuity	all	1985-9	7 (7 .. 7)	18 (17 .. 18)	77 (71 .. 83)	-2.48 (-2.56 .. -2.44)
continuity	all	2005-9	7 (7 .. 7)	21 (20 .. 21)	101 (96 .. 108)	-2.43 (-2.46 .. -2.40)
continuity	top 1000	1905-9	24 (23 .. 26)	69 (62 .. 76)	166 (136 .. 229)	-3.01 (-3.35 .. -2.70)
continuity	top 1000	1925-9	22 (21 .. 24)	58 (53 .. 66)	176 (131 .. 338)	-3.01 (-3.39 .. -2.67)
continuity	top 1000	1945-9	27 (25 .. 29)	66 (57 .. 80)	211 (169 .. 332)	-2.92 (-3.32 .. -2.59)
continuity	top 1000	1965-9	34 (32 .. 35)	92 (81 .. 104)	262 (203 .. 622)	-2.75 (-3.11 .. -2.48)
continuity	top 1000	1985-9	52 (49 .. 56)	135 (118 .. 147)	312 (231 .. 739)	-3.20 (-3.62 .. -2.83)
continuity	top 1000	2005-9	87 (80 .. 91)	229 (211 .. 250)	649 (532 .. 752)	-2.97 (-3.32 .. -2.75)
continuity	top 0.1%	1905-9	66 (59 .. 79)	146 (126 .. 176)	968 (209 .. 4287)	-3.29 (-5.20 .. -2.24)
continuity	top 0.1%	1925-9	53 (47 .. 61)	125 (104 .. 161)	476 (258 .. 2498)	-2.67 (-3.72 .. -2.20)
continuity	top 0.1%	1945-9	57 (52 .. 66)	150 (123 .. 194)	419 (218 .. 1089)	-3.19 (-4.26 .. -2.52)
continuity	top 0.1%	1965-9	69 (61 .. 79)	168 (143 .. 214)	713 (261 .. 874)	-3.01 (-4.01 .. -2.45)
continuity	top 0.1%	1985-9	85 (78 .. 94)	187 (158 .. 216)	732 (276 .. 892)	-3.40 (-4.30 .. -2.80)
continuity	top 0.1%	2005-9	113 (107 .. 119)	271 (246 .. 306)	681 (614 .. 874)	-3.16 (-3.59 .. -2.85)

Figure 4: Percentiles and best-fit power-law exponents for five-year periods of the news corpus. Each entry is of the form x (a .. b), where x is an estimate based on all articles in the period, and 99% of bootstrap estimates fell within the range a .. b. See Section 3.6 and Sections 4 and 5.

see an increasing trend after 1990, and a decreasing trend before 1920 in the 90th and 99th percentiles.

The power law exponents for this set of names are shown in the second block of Figure 4. They are significantly greater in magnitude than the exponents measured using basic filtering, and also show more variation.

4.3 Top 0.1% by year

When we restrict our experiment to names which were in the top thousandth of popularity in at least one year, we see a median duration of about one month (Fig. 8), again longer than the median from basic filtering. As before, we do not see a clear trend of change between 1920-90. The 99th percentile shows more fluctuation than with basic or top-1000 filtering; this may happen because this experiment involves a smaller total number of names.

The power law exponents for this set of names are shown in the third block of Fig. 4. Again, they are greater in magnitude than the exponents measured with basic filtering. They show more variation than the basic-filtering exponents, and the variation parallels the variation with top-1000 filtering.

5. RESULTS: CONTINUITY METHOD

We measure periods of popularity using the continuity method described in Sec. 3, and plot the distribution of durations as it changes over time. In contrast to our measurements using the spike method (Sec. 4) we find that the more popular names (top-1000 and top-0.1% as described in Sec. 3.2) show progressively longer durations of fame over the past 90 years, indicating that while the typical duration of a story in the news has stayed the same, the typical duration of public attention to a person is growing longer.

5.1 Basic Filtering

Using basic filtering (Sec. 3.2), we find a distribution of durations similar to that which me measured using the spike method (Fig. 9). Most of the names have a duration of one week,[4] and the distribution changes little between the years 1920-90. The fourth block of the table in Fig. 4 shows maximum likelihood estimates of the power law exponents. The exponents range between -1.77 and -1.45, slightly greater in magnitude than exponents measured using the spike method.

5.2 Top 1000 by year

We restrict to names which were in the top 1000 for some year (Section 3.2). Here we see a new trend which did not

[4] Unlike the spike method, the continuity method can produce durations of less than one week. Recall that the basic filtering excludes names whose periods of fame are shorter than two days.

method	filtering	period	50th %ile	90th %ile	99th %ile	power law exponent
spike	all	2000-4	7 (7 .. 7)	35 (28 .. 35)	123 (84 .. 189)	-2.37 (-2.52 .. -2.23)
spike	all	2005-9	7 (7 .. 7)	28 (21 .. 28)	75 (63 .. 84)	-2.34 (-2.76 .. -2.27)
spike	top 1000	2000-4	21 (14 .. 21)	56 (49 .. 63)	265 (148 .. 479)	-2.51 (-2.83 .. -2.18)
spike	top 1000	2005-9	14 (14 .. 21)	49 (42 .. 54)	109 (91 .. 151)	-2.74 (-3.03 .. -2.41)
spike	top 0.1%	2000-4	39 (28 .. 56)	189 (106 .. 305)	717 (286 .. 840)	-2.26 (-3.05 .. -1.85)
spike	top 0.1%	2005-9	28 (25 .. 35)	88 (74 .. 102)	213 (113 .. 1674)	-3.29 (-5.40 .. -2.23)
continuity	all	2000-4	7 (7 .. 7)	22 (20 .. 23)	114 (95 .. 160)	-2.38 (-2.49 .. -2.28)
continuity	all	2005-9	6 (6 .. 7)	18 (17 .. 19)	80 (66 .. 93)	-2.62 (-2.72 .. -2.53)
continuity	top 1000	2000-4	20 (18 .. 21)	71 (59 .. 83)	387 (237 .. 819)	-2.32 (-2.54 .. -2.12)
continuity	top 1000	2005-9	21 (20 .. 22)	59 (53 .. 73)	408 (211 .. 1057)	-2.37 (-2.62 .. -2.18)
continuity	top 0.1%	2000-4	102 (89 .. 123)	372 (236 .. 768)	2010 (768 .. 2238)	-2.24 (-3.15 .. -1.86)
continuity	top 0.1%	2005-9	83 (70 .. 93)	302 (193 .. 617)	2083 (954 .. 2991)	-2.12 (-2.75 .. -1.79)

Figure 5: Percentiles and best-fit power-law exponents for five-year periods of the blog corpus. Each entry is of the form x (a .. b), where x is an estimate based on all articles in the period, and 99% of bootstrap estimates fell within the range a .. b. See Sections 3.6 and 6.

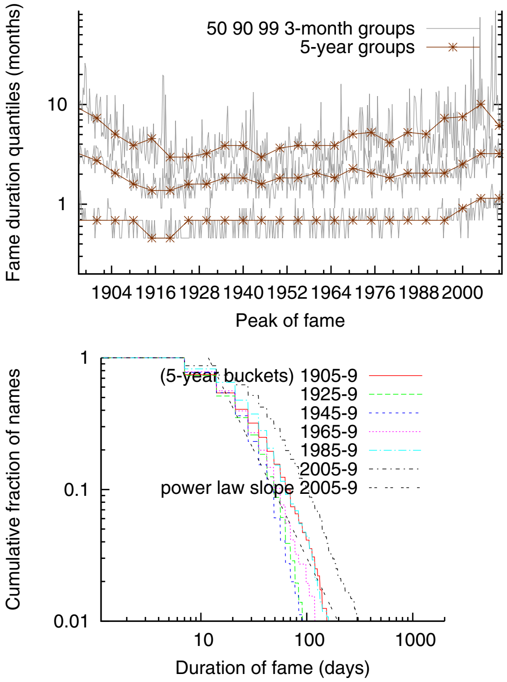

Figure 7: Fame durations, restricting to the union of the 1000 most-mentioned names in every year, using the spike method to identify periods of fame.

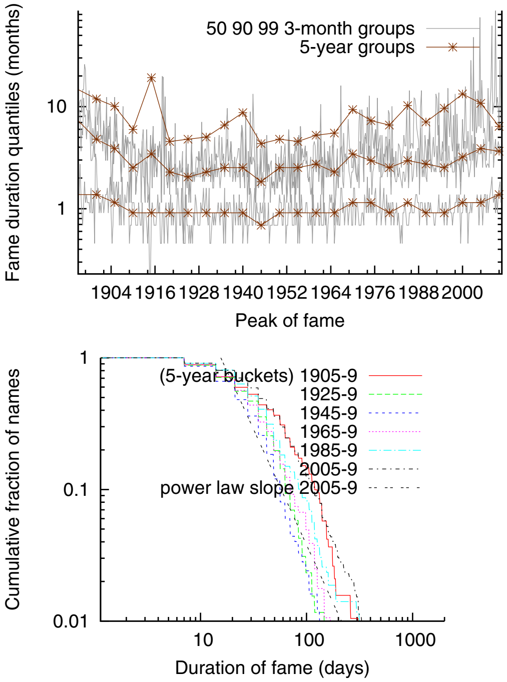

Figure 8: Fame durations, restricting to the union of the 0.1% most-mentioned names in every year, measured using the spike method.

appear when we used the spike method to determine durations of fame. The fame durations increase steadily from 1920 to 2005, the median growing from half a month to two months, and the 90th percentile growing the same way.

Power law exponents appear in the fifth block of Fig. 4.

5.3 Top 0.1% by year

Restricting to names in the top 0.1% of some year, we again see an increasing trend, though not as pronounced. The median duration in 1920 is almost two months, and in 2005 is almost three months.

Since the increasing trend is stronger with top-1000 filtering than with top-0.1% filtering, we hypothesize that the top-1000 trend is partly caused by a larger total population of names available in later years; we describe this hypothetical effect with top-0.1% filtering in Sec. 3.2. Since a trend still appears when filtering to a top 0.1% of names, we believe that the trend is not entirely explained by that effect.

The power law exponents appear at the bottom of Fig. 4.

6. BLOG POSTS

We ran our experiments on a second set of data consisting of public English-language blog posts from the Blogger service. We began by sampling so that the number of blog posts in each month in our data set was equal to the number of news articles we sampled in each month (as described in Section 3.3). The cumulative graphs of fame duration from six experiments are shown in Fig. 12. We combine the two methods for identifying periods of fame with three sets of names described in Section 3.2. The respective distributions from the news corpus are superimposed for comparison.

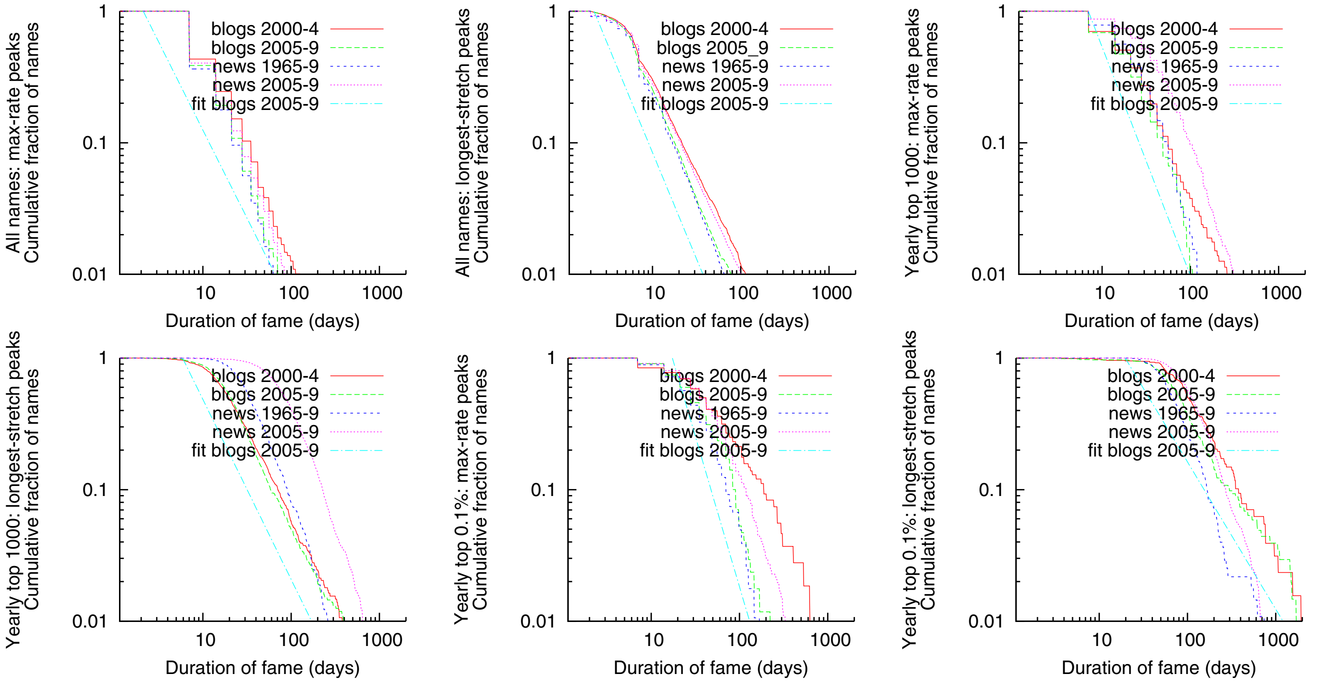

Figure 12: Cumulative duration-of-fame graphs for the blog corpus. Left-to-right top-to-bottom: all names with the spike and continuity methods, then union of 1000 top names in each year with the spike and continuity methods, then union of 0.1% of top names in each year with the spike and continuity methods.

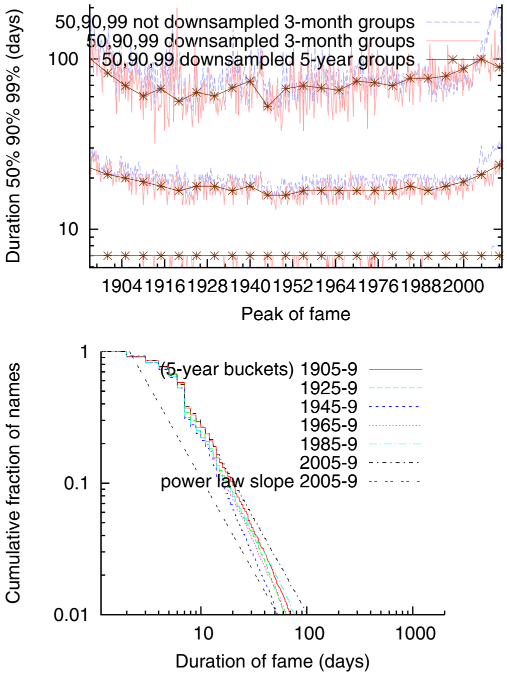

Figure 9: Fame durations measured using the continuity method, plotted as the 50th, 90th and 99th percentiles over time (top), and for specific five-year periods (bottom). To illustrate the effect of sampling, the first graph includes measurements taken before sampling; see Section 3.3.

The graphs of fame duration measured using the continuity method are much smoother for the blog corpus than for the news corpus. This happens because whereas we only know which day each news article was written, we know the time of day each blog entry was posted.

The second graph (top-center in Figure 12) has a distinctive rounded cap which surprised us at first. We believe it is caused by the following effect. Peaks with only two mentions in them are fairly common, and have a simple distinctive distribution that is the difference between two sample dates conditioned on being less than a week apart. Since two dates that are longer than one week apart cannot constitute a longest-stretch peak, the portion of the graph with durations longer than one week does not include any names from this two-sample distribution, and so it looks different. Our estimates of power-law exponents only consider the longest 20% of durations, so they ignore this part of the graph.

The estimates we computed for the power-law exponents of the duration distributions for blog data are shown in Figure 5, and can be compared to the figures for news articles in Figure 4. The estimates for the all-names distributions are fairly close to the corresponding ones for news articles, supporting a view that many different media of communication show the same patterns of fame. When we restrict to popular names, the estimated exponents are often distorted less than for the corresponding dates in the news corpus.

7. RELATED WORK

Michel *et al.* [8] study a massive corpus of digitized content in an attempt to study cultural trends. The corpus they study is even larger than ours in terms of both volume and temporal extension.

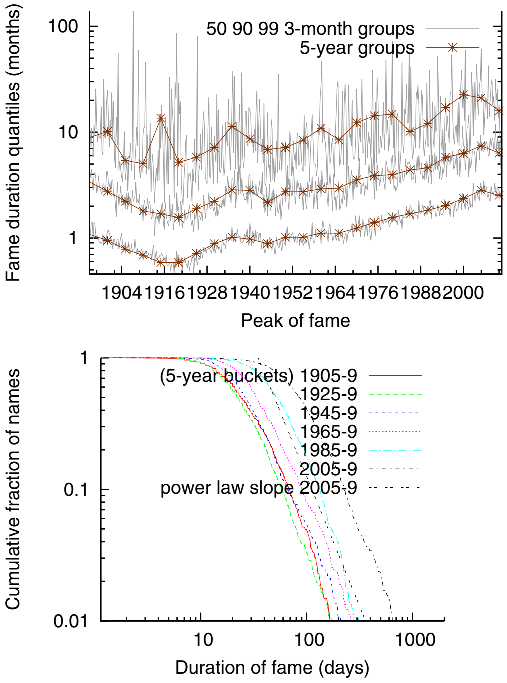

Figure 10: Fame durations, restricting to the union of the 1000 most-mentioned names in every year, measured using the continuity method.

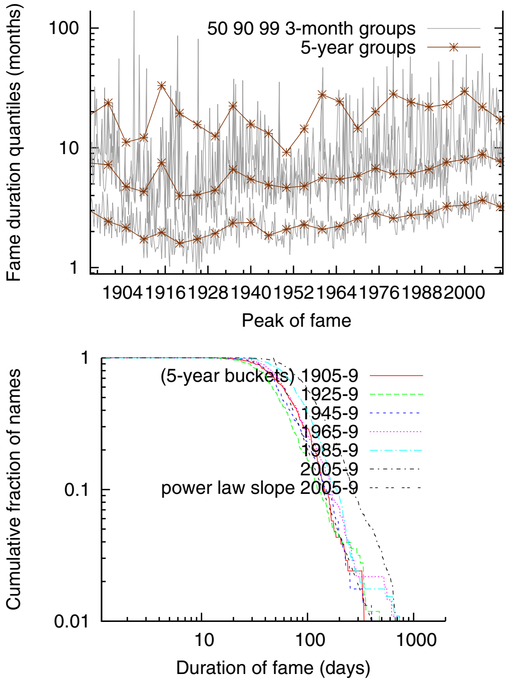

Figure 11: Fame durations, restricting to the union of the 0.1% most-mentioned names in every year, measured using the continuity method.

Leetaru [6] presents evidence that sentiment analysis of news articles from the past decade could have been used to predict the revolutions in Tunisia, Egypt and Libya.

Our spike method for identifying periods of fame is motivated in part by the work of Yang and Lescovec [12] on identifying patterns of temporal variation on the web. Szabo and Huberman [10] also consider temporal patterns, in their case regarding consumption of particular content items. Kleinberg studies other approaches to identification of bursts [5].

Numerous works have studied the propagation of topics through online media. Leskovec *et al.* [7] develop techniques for tracking short "memes" as they propagate through online media, as a means to understanding the news cycle. Adar and Adamic [1], and Gruhl *et al.* [4] consider propagation of information across blogs.

Finally, a range of tools and systems provide access to personalized news information; see Gabrilovich et al [3] and the references therein for pointers.

8. ACKNOWLEDGEMENTS

The authors would like to thank Zoran Dimitrijevic and the Google News Archive team for their help with the data; Danny Wyatt, Ed Chi, and Rachel Schutt for statistical advice; and the anonymous reviewers for helpful suggestions.

9. REFERENCES

[1] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. WI 2005, pages 207–214.

[2] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[3] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. WWW 2004, pages 482–490.

[4] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. WWW 2004, pages 491–501.

[5] J. Kleinberg. Bursty and hierarchical structure in streams. KDD 2002, pages 91–101.

[6] K. Leetaru. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9-5), 2011.

[7] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. KDD 2009, pages 497–506.

[8] J-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[9] H. A. Simon. Designing organizations for an information-rich world. 1971.

[10] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53:80–88, August 2010.

[11] Wikipedia. Astor family — Wikipedia, the free encyclopedia, 2011. [Online; accessed 10-August-2011].

[12] J. Yang and J. Leskovec. Patterns of temporal variation in online media. WSDM 2011, pages 177–186.