

# MEASURING AGENDA SETTING WITH ONLINE SEARCH TRAFFIC: INFLUENCES OF ONLINE AND TRADITIONAL MEDIA

Laura A. Granka  
Stanford University  
Google, Inc.

## ABSTRACT

This paper addresses the patterns of influence between the news media and the public, by specifically targeting breaking stories, or shocks, to a news system. Specifically, we assess media agenda setting and selective exposure by looking at the relative public attention spans to hard and soft news (as measured by query volume), in comparison with the volume of news coverage (in print, broadcast, and Web content) for these selected news events. We measure the dynamic distribution of issue coverage in the news media, and how this volume of coverage ultimately influences online search traffic. In order to assess sustained interest in a given topic, distributions of query volume and news coverage were fit with Gamma distributions of appropriate parameters. Findings indicate that there are significant differences in the public attention spans for hard and soft news issues, particularly relative to what news coverage might predict. Soft news events produced a slower rate of decline in query volume, matching the slow tapering off of issue coverage found in Web content. Conversely, for hard, substantive news issues, query volume drops off quite quickly, more closely paralleling the distribution of coverage in broadcast news.

Prepared for delivery at the 2010 Annual Meeting of the  
American Political Science Association, September 2-5, 2010.  
Copyright by the American Political Science Association.

## INTRODUCTION

The increased choice, availability, and access to information afforded by digital technology may require a re-conceptualization of media effects. Of particular concern is that digital media – namely blogs, online news, and search engines – do little to encourage public affairs and political news consumption amongst the electorate. In fact, one working hypothesis is the increased choice online may simply overwhelm potential news consumers, making individuals more prone to engage in selective exposure and polarization (Sunstein, 2007). There is more opportunity for readers to seek out content that is already ideologically aligned with their views, or just mere entertainment (Bennett & Iyengar, 2008; Baum, 2002; Prior, 2005). While digital media is not responsible for discouraging existing newsreaders from partaking in their routine news consumption, the abundance of choice online may simply make it easier for the disinterested to tune out altogether.

Most of the existing research regarding selective exposure has focused on differences in political knowledge based on an individual's self-reported media preferences for soft news, hard news, and pure entertainment, by assessing the type of programming attended to (talk show, punditry, local news, evening news). *Within* a given program, however, individuals are likely to be more interested or attentive to certain issues that are reported – for example, paying more attention to the drama unfolding about Tiger Woods' affairs instead of the health care bill. By comparing public attention spans to given news items, we may find a stronger indication of selective exposure. This present research attempts to determine how long public attention is sustained in a given news event, depending on how substantive or sensational it is. The likelihood of sustained public interest will be compared with the corresponding amount of attention (i.e., coverage) given to the issue by the news media. Ultimately, we hope to answer: is

the public attention span skewed towards sensational and entertaining content, and is this significantly different than that of news media organizations?

This paper will address the patterns of influence between news media and the public, by specifically targeting breaking stories, or shocks, to a news system. Specifically, we assess the dynamic between issue coverage in offline media (network news transcripts and newspapers) and online media (popular political blogs), similar to the cue-taking paradigm analyzed by Shaw and Sparrow (1999). This study selects 22 major news events and tracks news coverage of each issue for a 50-70 day period. This volume of news coverage is then used to assess agenda setting effects; the unit of measurement for agenda setting is online search queries. As will be discussed further, search queries are one way to assess explicit audience interest in a topic: a searcher first needs to hear something about the event, and second, have enough interest in the topic to conduct a search to find out more. We pinpointed the instances when a story peaked in three forms of news media (newspapers, broadcast, web content), as well as when the topic peaked in online search traffic, and analyzed the corresponding decay curves, to determine the relative “legs” of the topic in each form of media. This enables us to assess the moment when interest in a given topic is at a high, per media, and how long interest in the given story is sustained.

We expect to affirm agenda-setting, in that query volume will follow news coverage, though in accordance with online selective exposure, we will expect that public interest in a topic, as measured by search queries, will be sustained longer for sensational topics than for substantive, hard news, topics. Overall, we hypothesize a strong level of convergence between search queries and news volume. The more interesting insights in our analysis will be the deviant cases – instances where the search query volume for a topic or issue exceeds what might be expected by its respective news coverage.

## DYNAMICS OF ONLINE NEWS MEDIA

An increasing number of individuals are getting their information via online or digital sources, as opposed to traditional television broadcasts and newspapers (Fox, 2008). Three-quarters of American adults are online, and over half of these individuals depend on the Internet for information. Several distinct changes are already apparent with the growing reliance on information technology for news information. The most notable change is of course the increasing amount of individual choice and control that individuals are able to exercise with respect to the media content that is consumed.

*Increased choice.* Information on the Internet more closely represents a pull rather than a push paradigm, meaning that individuals must specifically seek out and identify individual sources and stories that they want to consume. Digital media is inherently interactive, forcing a user to click, type, and search for specific sources and bits of information. As such, digital media has a unique ability to engage and create an active audience and active consumer of news content. News content is not “pushed” onto individuals as it may be in a 20 to 30 minute news broadcast, with editors serving as the gatekeepers by screening and selecting content (Shoemaker, 1991).

Instead, the increased degree of choice implies that individuals are in some ways their own gatekeepers of online news content, exercising some selectivity in their exposure to media messages. Individuals have the ability to immediately access breaking news, share this information with anyone of their choosing, and also publicly post their own opinions and analyses. Studies have also shown that the abundance of online information and online sources can result in increased political polarization (Iyengar et al, 2008; Adamanic, & Glance, 2005; Iyengar & Hahn, 2007). Further, individuals have more opportunity to ignore current affairs

messages and click on only entertainment content. We speculate that search queries can be used as a measure of the types of news that the public is tuning into.

*Offline-Online Dynamic.* Another change with online media is that it opens new avenues for cue-taking between news outlets. In other words, what is the dynamic between online and offline forms of news media: do some news media first rely on others' reporting before inserting a given story in their own repertoire? Early on, Bruns essentially classified bloggers as "gatewatchers," meaning that they rarely make their own news, but rather accumulate and repack information from traditional media elites (Bruns, 2003).

More recently, with a corpus of over 90 million news stories, Leskovic and colleagues conducted a large-scale analysis of the basic properties of news cycles, and found that traditional media (network broadcasts and print newspapers) are still extremely influential in promoting news topics and issues, generally peaking coverage of a story about 2.5 hours before it reaches the blogosphere (Leskovic, et al, 2009). Further, they found that the rates of story dissemination and decay were faster and higher in offline media, whereas blog volume picked up, and decayed, at slower rates. This indicates that blogs generally take cues from mainstream media, and (potentially due to the minimal production costs of uploading web content) have the luxury to keep it in the blogosphere for a longer time period. Because their analysis is aggregated across all news events and sources, it depicts overall trends, and cannot pinpoint whether specific types of news events caused unique distributions. The research presented here will attempt to classify these distributions based on news type.

*Agenda-Setting.* It is well documented that news coverage is heavily influential in defining the issues people care most about. Agenda setting reasons that the news media shape public perceptions of issue importance simply through the sheer volume of issue-specific coverage

(McCombs & Shaw, 1972; Iyenger & Kinder, 1984; Behr & Iyengar, 1985). In other words, the issues receiving the greatest news coverage are judged to be the most important issues of the day. Traditional measures of agenda setting depend heavily on surveys and experimental research to measure attitude changes and public opinion, and this research seeks to affirm whether online search traffic can be equally strong a predictor.

#### SEARCH AS A MEASURE OF INTEREST AND PUBLIC ATTENTION

While surveys and experiments allow for greater control over individual-level variables (e.g., education, age, gender, political affiliation), self-report measures are also costly to instrument and susceptible to social-desirability bias, particularly when questions assess political interest and awareness (for more on this, see Prior, 2009). Further, self-reported methods do nothing to capitalize on the structural and sampling advantages of digital media. Online digital media affords researchers with a naturally robust platform on which to track actual user behaviors in an immediate and inexpensive manner. To alleviate some the aforementioned challenges of self-report, we propose a new measure of agenda setting— aggregate online search queries.

Online search is an active medium, meaning that a user has to explicitly make the effort to acquire information about a given topic by manually typing in a query. Because of this, online search queries may be a strong behavioral indicator of what issues and topics are the most compelling, interesting, or important. This, coupled with the lack of self-report bias makes search queries an attractive way to implicitly measure fluctuations and changes in political issue interest over time. Individuals are increasingly using online search engines to gather information about news and current events. In a recent survey by the author, only 8 out of 1,000 Americans

sampled reported *not* using an online search engine, and 54.4% of those individuals report using a search engine on a daily basis (unpublished research). It is important to understand how news-related search behavior fits in with models of existing media use and news consumption.

Online search queries are publicly available for download (e.g., Google Insights for Search, <http://google.com/search/insights/>). Existing research using search queries has shown how search volume is both indicative and predictive of external events, from predicting flu outbreaks (Ginsberg, 2008) to consumer behavior (Choi & Varian, 2008, Goel et al, 2010). Choi & Varian developed auto-regressive models predicting several types of consumer behavior, including automotive sales, home sales, and travel. By adding aggregate search queries to these models, the authors were able to outperform the standard models, reducing the error of the estimate, improving model accuracy. Goel and colleagues conducted similar analyses, predicting the relative success of consumer products, including movies, video games, and music.

In order to be useful, query volume should be used comparatively, such as comparing volume within a selected time period, or against different issues or topics. Overall changes in search query volume, or sudden spikes in query volume, are two potential ways to assess how prominent or "important" an issue may be at any given moment. This research will analyze the changes in search query volume as a measure of interest in a given issue. This rise and fall in query volume will be compared with the fluctuations of news coverage by traditional and online news sources. We will determine whether the public and news media are equally likely to sustain interest in a given news event, dependent upon its being a substantive or sensational event.

As we are considering how long coverage or attention to a given issue is sustained, it should be noted that Web content and newspapers are at the distinct advantage: these mediums,

particularly Web, are much less constrained in terms of space and time. There are fewer adverse consequences when considering whether to include an update about an “old” news event in a back page of a newspaper, whereas such coverage in a 30-minute news broadcast is much more difficult and obtrusive. Thus, of the news mediums, we might expect that broadcast networks have the quickest rates of decay, and newspapers and Web content to have the slowest. Depending on the news event, we might expect queries to more closely align with print and Web sources, as a searcher has unlimited time to seek out the additional details of a news story. Conversely, one might expect search queries and broadcast news to be more closely aligned, assuming the public attention span is continually refreshed with a more recent story.

## METHODS

As per agenda setting, it is expected that media coverage of news events will drive public awareness and interest in these issue. Thus, the first step is to determine whether the behavioral data obtained via search query volume is also consistent with the conclusions of agenda setting research: how consistently do real world events and news coverage motivate search traffic for political and current events? If there is in fact a relationship between searches and news, we will continue by comparing the distributions of news media coverage and search query volume to determine how quickly interest in each topic dissipates. It is likely that there are some topics and issues that sustain the news media interest more than the public interest, and alternatively, that spikes in query volume for some issues are sustained longer than the news coverage.

### DATA COLLECTION

Twenty-two news events were selected, all of which have a defined start date, and varied in degrees of their newsworthiness and sensationalism. Specifically, the events selected were



meant to represent three categories: 1) substantive news, including both international events (e.g., the Iran election or Israel- Gaza war) and domestic events (e.g., the Arizona immigration bill or the Sotomayor nomination); 2) sensational news, including political sensational events (e.g., Joe the Plumber’s mention in the 2008 presidential election, or Joe Wilson shouting “you lie” to Obama), as well as celebrity events (e.g., Michael Jackson’s death and Tiger Woods’ extramarital affairs); and 3) disaster events (e.g., the 2004 tsunami, or 2010 Haiti earthquake).

Query and news volume were measured at the daily level, for a two-three month time period, beginning with the month of the event. United States news coverage was collected for four media types: broadcast (including both television and radio), newspapers, and Web content (including blogs and Web-only publications). The daily counts of news stories were obtained through the Lexis Nexis database archives, which maintains transcripts and print archives of news publications through the Vanderbilt transcript archive. Keywords were used to search for these events in Lexis Nexis, and the resulting daily counts were used to represent the volume of coverage on each day during the time interval. Appendix A lists the news stories analyzed in this paper, along with the corresponding query terms used to obtain the data in Google Insights for Search and the Lexis Nexis database.

Only non-wire US-based news outlets were included in the newspaper and broadcast sample; the Web archive of course included worldwide sources. Among the sample of news sources used were: 691 total US newspapers (including major newspapers, i.e. The NY Times or Washington Post, as well as local and regional dailies; excluding wire services), broadcast transcripts (from ABC News, American Public Media, CBS News, CNBC News, CNN, Fox News Network, MSNBC, National Public Radio), and 4,959 Web-based sources (including 4,561 blogs and 298 Web-based publications) (see Lexis, 2010, for full list).

Search query volume was collected from Google Insights for Search. The search volume is pre-normalized according to a scale with a max of 100 (equaling the highest volume of coverage in the given time period), and to simplify analyses and graphical comparisons across media types and news topics, the news coverage was normalized in the same manner. Further, as another measure of data preparation, in order to obtain the Maximum Likelihood Estimation parameter estimates of the Gamma functions fitted to the data, direct optimization of the log-likelihood is performed using `optim` (an R function in the Stats package). The optimization calculation leads to values of infinity when faced with zeros in the gamma distribution, so for all dates where there was no query or news volume, the zeros were smoothed by being replaced with 0.0001. This change ultimately reflects a limitation of the form of data collection, in that there was likely to be a single search query issued, or news story reported that was simply not covered in the collected sample.

## ANALYSIS

Data from all sources were aligned at the event start date (further referenced as  $T_0$ ). This frequently also happened to be the max volume for the given time period. As noted above, gamma frequency distributions were fit against the volume of news coverage and search query volume, from  $T_0$  to the last day of data collection. The gamma distribution was selected as the ideal distribution with which to fit the data, as it is a continuous probability distribution, most frequently used to model waiting times. In the case of this research, there is a definitive start time,  $T_0$  – the time of the event – followed by the continuous flow of news coverage until it dissipates and falls out of the media or public zeitgeist. While an artificial end point was in fact instated for the data, most of the volume had shrunk to 1-4% of the maximum height, which for

the purposes of this study is a reasonable approximation of “end.” The gamma distribution has two parameters – shape and scale – which describe both the onset and rate of decay.

The fit of the gamma models were compared to those fit by the Weibull distribution. The Weibull is also frequently used in survival analysis, modeling time until “failure;” failure in the case of this research would be the ceasing of interest in a given topic. The Weibull has similarities to the gamma function, having both a shape and a scale function. Both the gamma and the Weibull distributions exhibit distinct properties when the shape parameter moves from less than, equal to, or greater than 1. These differences in shape can indicate an early failure onset (less than 1), a fairly constant rate of failure over time (equal to 1), and a failure onset that occurs later in time (greater than one). The difference in curvature from varying the shape parameter can be clearly seen in the graphic examples throughout this paper.

There were no major differences between the gamma distribution and the Weibull distribution, so all results reported in the paper are using the gamma distribution, as it is a slightly more common and generalizable model. Appendix B shows plots of the raw news and query data for two specific events, with the fitted Weibull and Gamma distributions.

## RESULTS

### RELATIONSHIP BETWEEN SEARCH QUERIES AND NEWS MEDIA

The first phase was simply meant to validate and assess the extent to which agenda setting effects can be measured in online search traffic, as result of issue coverage in online and print sources. Plots and simple correlations (Spearman rho, due to the non-linearity and non-normality of the data) were conducted to obtain a general sense of the relationships between queries and news for the different types of news events. These correlations are reported in Table 1. As predicted and consistent with agenda setting, search query volume is typically highly

correlated with the fluctuations in media coverage – overall, it appears that indeed, search volume of real-world events does closely correlate with fluctuations in news media volume.

However, from the correlations it is clear that there is much between-event variation, particularly with respect to the given form of media that is most closely correlated with the search volume. There is no single consistent relationship across all news events; in other words, the volume of coverage from one given media (whether it be Web content or broadcast coverage) does not consistently match with search volume. By fitting gamma distributions to search volume, and each form of news, we will be able to better specify what type of news media search volume is likely to follow, and whether this differs between substantive and sensational events.

**Table 1: Spearman Correlations of Search Queries and News Media**

<b>Event Name</b>	<b>Query-News Correlation</b>	<b>Event Name</b>	<b>Query-News Correlation</b>
Iran Election		Iraq Election	
<i>Web</i>	0.834	<i>Web</i>	0.430
<i>Newspapers</i>	0.776	<i>Newspapers</i>	0.512
<i>Broadcast</i>	0.844	<i>Broadcast</i>	0.471
David Cameron		Gaza War	
<i>Web</i>	0.538	<i>Web</i>	0.934
<i>Newspapers</i>	0.541	<i>Newspapers</i>	0.749
<i>Transcripts</i>	0.420	<i>Broadcast</i>	0.891
Arizona immigration		Sotomayor nomination	
<i>Web</i>	0.780	<i>Web</i>	0.849
<i>Newspapers</i>	0.606	<i>Newspapers</i>	0.861
<i>Broadcast</i>	0.775	<i>Broadcast</i>	0.865
Health care bill		Cash for Clunkers	
<i>Web</i>	0.890	<i>Web</i>	0.318
<i>Newspapers</i>	0.889	<i>Newspapers</i>	0.278
<i>Broadcast</i>	0.779	<i>Broadcast</i>	0.450
Joe the Plumber		Joe Wilson	
<i>Web</i>	0.867	<i>Web</i>	0.730
<i>Newspapers</i>	0.889	<i>Newspapers</i>	0.815
<i>Broadcast</i>	0.835	<i>Broadcast</i>	0.697
Sniper Fire		Wikileaks	
<i>Web</i>	0.714	<i>Web</i>	0.600
<i>Newspapers</i>	0.762	<i>Newspapers</i>	0.611
<i>Broadcast</i>	0.724	<i>Broadcast</i>	0.410
Climategate		Hudson plane crash	
<i>Web</i>	0.901	<i>Web</i>	0.573
<i>Newspapers</i>	0.604	<i>Newspapers</i>	0.604

	<i>Broadcast</i>	0.564		<i>Transcripts</i>	0.564
Roman Polanski	<i>Web</i>	0.897	Tiger Woods	<i>Web</i>	0.837
	<i>Newspapers</i>	0.817		<i>Newspapers</i>	0.765
	<i>Transcripts</i>	0.737		<i>Transcripts</i>	0.835
Michael Jackson	<i>Web</i>	0.500	Sandra Bullock	<i>Web</i>	0.370
	<i>Newspapers</i>	0.753		<i>Newspapers</i>	0.428
	<i>Transcripts</i>	0.622		<i>Transcripts</i>	0.263
Haiti earthquake	<i>Web</i>	0.723	Italy earthquake	<i>Web</i>	0.609
	<i>Newspapers</i>	0.827		<i>Newspapers</i>	0.569
	<i>Broadcast</i>	0.858		<i>Broadcast</i>	0.577
Tsunami	<i>Web</i>	0.649	Hurricane Katrina	<i>Web</i>	0.664
	<i>Newspapers</i>	0.627		<i>Newspapers</i>	0.548
	<i>Broadcast</i>	0.930		<i>Broadcast</i>	0.617

---

## EVENT DURATION

In order to assess sustained interest in a given topic, distributions of query and news volume were fit with Gamma distributions of appropriate parameters. Each gamma parameter (shape and rate) was calculated with Maximum Likelihood Estimation (MLE) using the `fitdistr` function in the MASS R library. The shape and scale parameters of the gamma distributions were then plotted and compared across all news types and search queries. To test whether the decays in news and query volume come from similar distributions (e.g., whether search query volume declines in a rate equivalent to either of the three news sources), the Kolmogorov-Smirnov (KS) fit test was calculated, a non-parametric test for distributions. In this analysis, the primary interest is whether the distribution of query volume significantly matches one of newspapers, broadcast, and Web news content, so comparisons of distributions within news media are not reported in this paper. Further, as the query distribution for each event is compared against three media types (broadcast, newspaper, Web content), all p-values were Bonferroni adjusted (this the most conservative multiple testing approach, requiring p-values to

be multiplied by the number of comparisons; in this case, 3). The results and data reported in this paper related to the Gamma distribution reflect these adjusted p-values.

## RESULTS

### HARD NEWS

Distributions of interest for both the international and domestic hard news topics showed a distinct pattern and relationship across search queries and news media. Four international news issues were compared in this analysis: the Dec 2008-Jan 2009 Israel-Gaza war; David Cameron (the leader of the 2010 UK election); the June 2009 Iran Election, and the March 2009 Iraq election. As exhibited in Figure 1, public attention, as measured by search volume, most closely models the rise and fall of broadcasts, whereas newspapers and Web content follow a significantly different trajectory, one of slower decay. This pattern is consistent across all international news events. Further tests of the distributions with the KS test indicate that search queries have a significantly different distribution than all news types, with the exception of the war in Gaza, in which the distribution of search queries was not significantly different than broadcast coverage. The D values (which reflect distributional similarity) from the KS test comparing search queries and broadcast news were notably smaller than the other news types, yet only reached the level of insignificance in the case of the Gaza war. The gamma shape parameter of both newspaper and Web coverage for international hard news is consistently greater than 1, indicating that newspapers and Web take a longer time before declining coverage on these topics (Table 2 for all parameters). Conversely, the shape parameter of the search query and broadcast transcripts is consistently less than one, indicating a much sooner onset of decay.

The national news events offered similar results (also in Table 2). The domestic event sample included the May 2010 Arizona immigration bill, the Supreme Court nomination of

Sonia Sotomayor, the March 2010 passage of the health care bill, and the July-September 2009 “cash for clunkers” program. It was anticipated that the domestic news events would indicate a later onset of decay in query volume and broadcast coverage, due to the direct impact these domestic events might have on the public. The cash for clunkers program in particular was expected to sustain more interest in query volume, due to it being directly actionable. However, when fitted, these domestic events generally paralleled the distribution of international hard news events, with some exceptions: the query distribution for the cash for clunkers program and the Arizona immigration bill were only significantly different from newspaper coverage (Clunkers:  $D=0.3913$ ,  $p\text{-value}<0.001$ ; AZ immigration:  $D=0.3247$ ,  $p\text{-value}=0.001$ ). Search volume was not significantly different than broadcast or Web content for these events. Overall, there is a clear indication that broadcast networks are the first to drop a hard news story, sometimes even before the public interest falls. Newspapers, as expected, continue their coverage of hard news topics the longest.

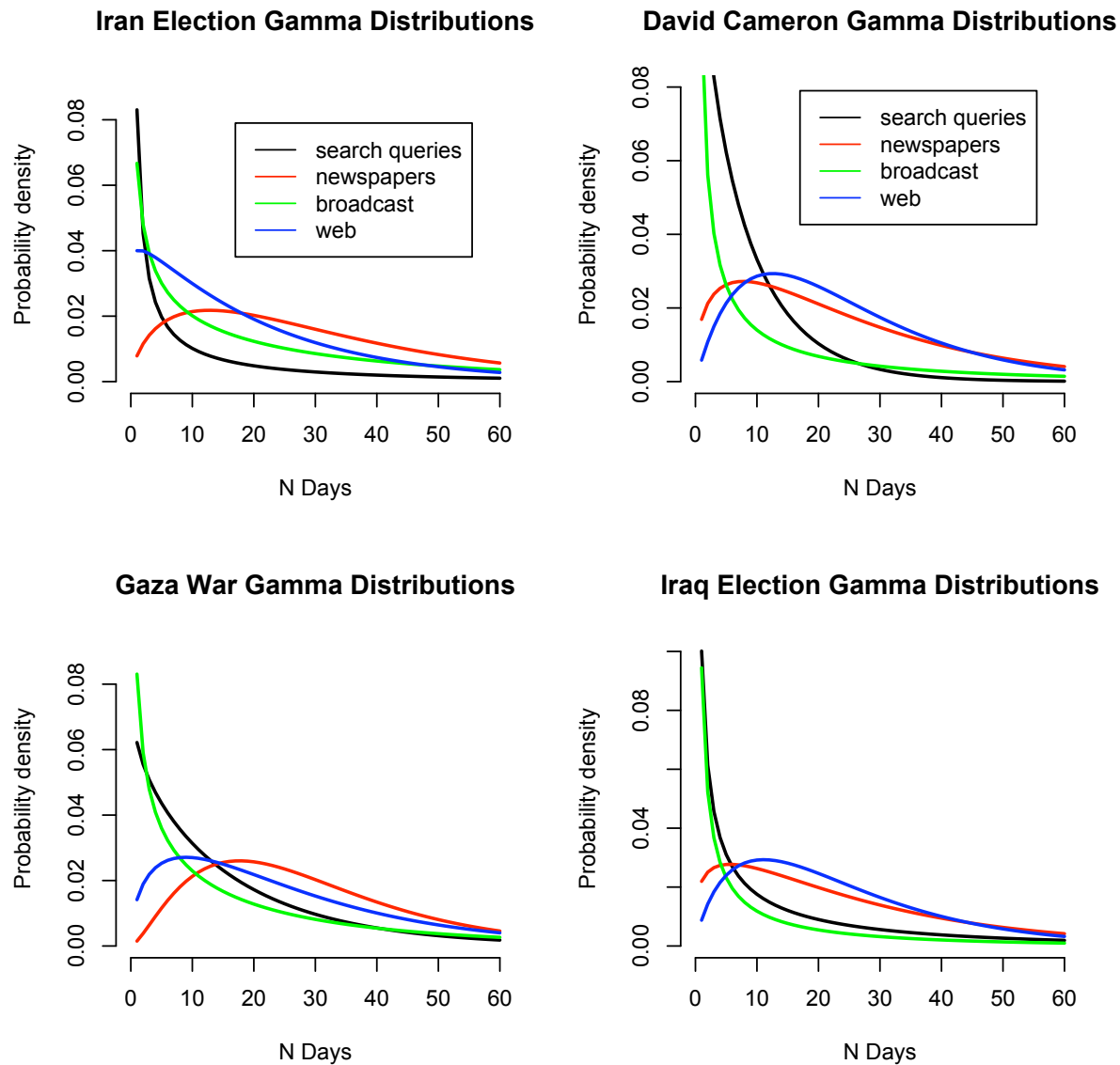
**Table 2: Gamma Parameters for International and Domestic News**

Event	Queries	Broadcast	Newspapers	Web
Iran	Shape	0.1452 (0.0210)	0.5595 (0.5595)	1.6275 (0.2866)
	Rate	0.0150 (0.0057)	0.0182 (0.0044)	0.0500 (0.0103)
	KS D-stat		0.537**	0.537**
			0.6852**	
Iraq	Shape	0.3224 (0.0377)	0.1785 (0.0200)	1.2589 (0.1666)
	Rate	0.0204 (0.0044)	0.0205 (0.0056)	0.0730 (0.0114)
	KS D-stat		0.5761**	0.4565**
			0.4239**	
Gaza	Shape	0.9184 (0.1407)	0.5719 (0.0838)	2.4852 (0.4099)
	Rate	0.0542 (0.0109)	0.0274 (0.0060)	0.0566 (0.0107)
	KS D-stat		0.2462	0.4**
			0.5538**	

Cameron				
Shape	0.8710 (0.1371)	0.2256 (0.0315)	1.4055 (0.2304)	2.0043 (0.3370)
Rate	0.1084 (0.0226)	0.0177 (0.0054)	0.0522 (0.0103)	0.0799 (0.0153)
KS D-stat		0.3115*	0.6885**	0.7705**
Arizona immigration				
Shape	1.2256 (0.1769)	0.7830 (0.1086)	1.8700 (0.2785)	1.4468 (0.2116)
Rate	0.0506 (0.0090)	0.0306 (0.0058)	0.0626 (0.0107)	0.0604 (0.0105)
KS D-stat		0.1558	0.3247*	0.1818
Sotomayor				
Shape	0.5716 (0.0813)	0.6323 (0.0908)	1.9504 (0.3078)	0.8704 (0.1288)
Rate	0.0829 (0.0177)	0.0253 (0.0053)	0.1035 (0.0186)	0.0692 (0.0136)
KS D-stat		0.5652**	0.6812**	0.4058**
Health care bill				
Shape	0.3614 (0.0464)	0.5694 (0.0761)	1.2664 (0.1821)	0.8486 (0.1179)
Rate	0.0681 (0.0155)	0.0270 (0.0054)	0.0659 (0.0116)	0.0631 (0.0117)
KS D-stat		0.6667**	0.7051**	0.5769**
Cash for Clunkers				
Shape	0.7404 (0.0935)	0.2682 (0.0309)	1.1417 (0.1498)	0.6747 (0.0844)
Rate	0.0438 (0.0077)	0.0166 (0.0038)	0.0374 (0.0061)	0.0409 (0.0073)
KS D-stat		0.2065	0.3913**	0.1848

**Table 2:** Standard errors are reported in parentheses below each parameter estimate. The K-S test D values (distributional differences between search and news media) follow. P-values less than 0.001 are indicated with \*\*, and p-values less than 0.05 are indicated with \*.





**Figure 1: Fitted gamma distributions for international hard news events**

Query distributions were significantly different from all media, with the exception of broadcast coverage of the Gaza War. Note the similarities between search queries and broadcast coverage.

## SOFT NEWS

Sensational news stories appear to be covered very differently by the media, and similarly search volume has its own unique distribution (Figure 2 and Table 3). Likely because these events are less nuanced than hard, international news, they drop off more quickly across all

forms of media and in search volume. However, it is clear that for some stories, certain forms of media dropped the story more quickly than others, and search volume did not consistently follow broadcast coverage, as it did with hard international news. Specifically for the cases of Joe the Plumber, Sniper Fire, Wikileaks, and Climategate, search volume was significantly different than the interest reflected in broadcast networks, and *not* different than the rate of coverage on the Web and in newspapers. There was a significantly slower rate of decline in search queries than broadcast coverage for the topics of Joe the Plumber ( $D=.333$ ,  $p\text{-value}=0.028$ ), Sniper Fire ( $D=0.413$ ,  $p\text{-value}=0.023$ ), Climategate ( $D=0.3778$ ,  $p\text{-value}=0.009$ ), and Wikileaks ( $D=0.771$ ,  $p\text{-value}<.001$ ); queries for Wikileaks were also significantly different from newspaper coverage:  $D=.637$ ,  $p\text{-value}<.001$ ). While these topics shot up quite quickly across the broadcast airwaves, they also dissipated quite quickly, and only remained a hot topic for broadcasters for a short time. It is likely that broadcaster news dropped these stories fairly quickly due to the space constraints of the media, yet it is interesting to note that query volume did not follow broadcast coverage as it did for international events. There were no significant differences in the rates of coverage between query volume and Web coverage.

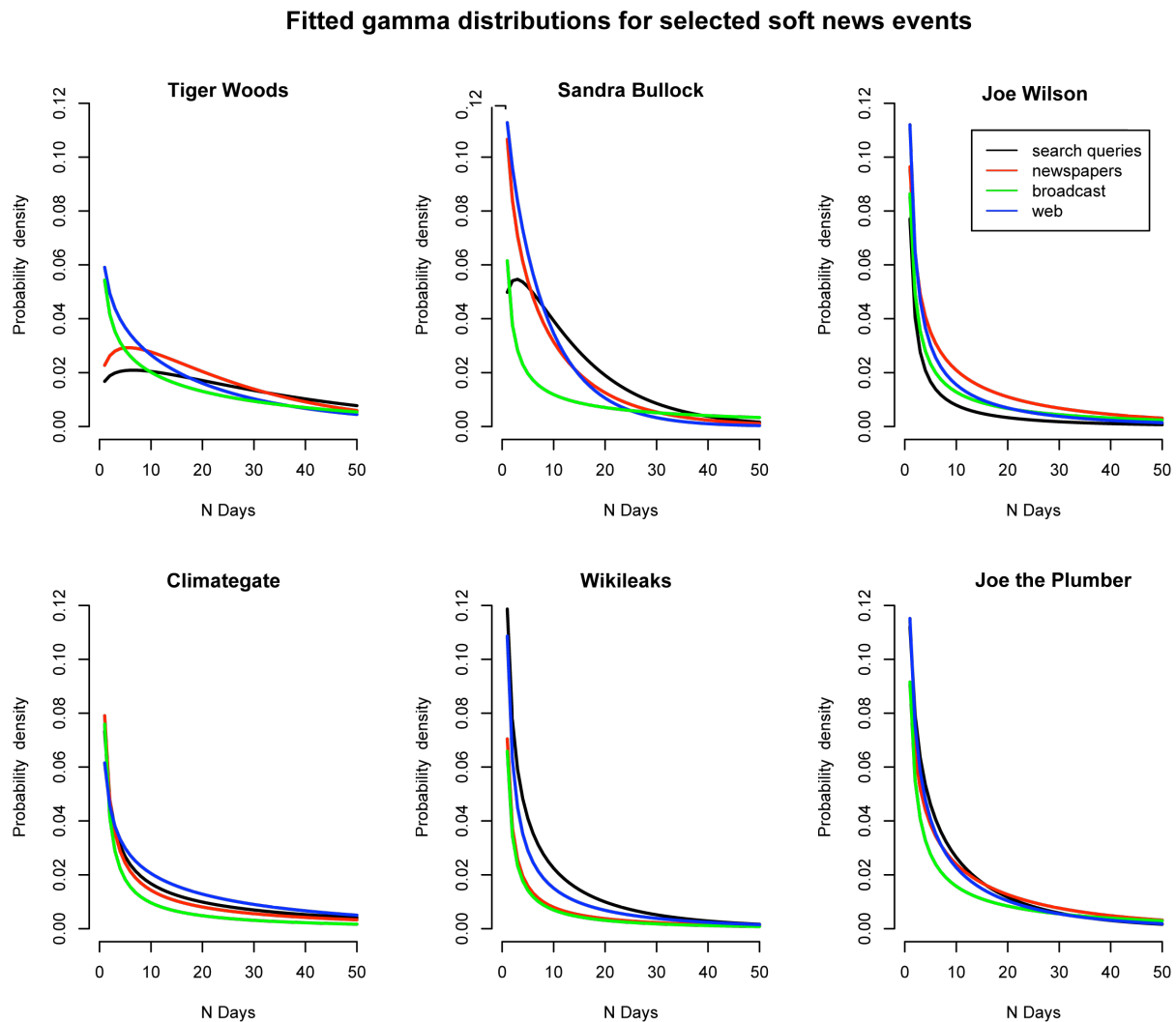
An anomaly in this set is the news story about Joe Wilson, who shouted “you lie!” to President Obama during a congressional meeting. Joe Wilson was searched for significantly less than what news coverage would have predicted, showing significant differences between the search query distribution across *all three* media forms (transcripts:  $D=0.555$ ,  $p\text{-value}<.001$ ; newspapers:  $D=0.7037$ ,  $p\text{-value}<.001$ ; Web content:  $D=.5$ ;  $p\text{-value}<.001$ ). While news coverage was discussing whether or not Joe Wilson’s apology had sufficed, the public had already moved on and was no longer interested in this topic.

The celebrity stories showed an even greater degree of interest from the searching public. There were no significant difference between search volume and newspaper coverage for the events surrounding Michael Jackson, Tiger Woods, and Sandra Bullock – indicating that search volume increased to match the coverage and attention to detail that newspapers can afford. There were also no significant differences in broadcast coverage from Michael Jackson and Tiger Woods: while broadcast networks are again the quickest to end coverage of an event, they still managed to discuss these celebrity stories for a longer period of time than they allot to domestic and international news (Figure 2). The graphs in Figure 2 reflect the similarity in attention from search queries and newspaper coverage. While the slow rate of decay is expected for newspapers (it was initially predicted that newspapers would have the slowest rates of decay), witnessing the same pattern in search queries seems indicative that the public remains interested in these celebrity dramas long after the story breaks.

Michael Jackson's death produced much saturation in the news media and in query volume, yet unlike Tiger Woods, this story was fairly front-loaded and was slightly quicker to dissipate. Query volume was only significantly different from broadcast networks ( $D=0.4559$ ,  $p\text{-value}<.001$ ): the gamma shape parameter for broadcast coverage was 0.99 ( $se=0.149$ ), which essentially indicates that attention decreased at a fairly constant rate within the time period. Search volume, newspaper, and Web coverage all ceased sooner than did broadcast, indicating the networks were covering this issue proportionately more than any other media.

Finally, aggregate parameters for all hard news and soft news events were separately computed (Figure 3). These graphs depict the consistent patterns of attention given to hard and soft news by the public, the printed press, broadcast networks, and Web bloggers. Newspapers always exhibit the slowest rate of decline for hard news issues, and broadcast media always has

the quickest. Search volume most closely aligns with broadcast coverage for hard news, while Web coverage most closely aligns with newspapers. For soft news events, broadcast networks and search volume still decline fairly quickly, yet the rate of decline of search traffic is now slower than all other forms of media. While newspapers and Web content also exhibit a much shorter attention span to these items, interest, as measured by search volume, lasts longer.



**Figure 2: News coverage and search interest for soft news events**

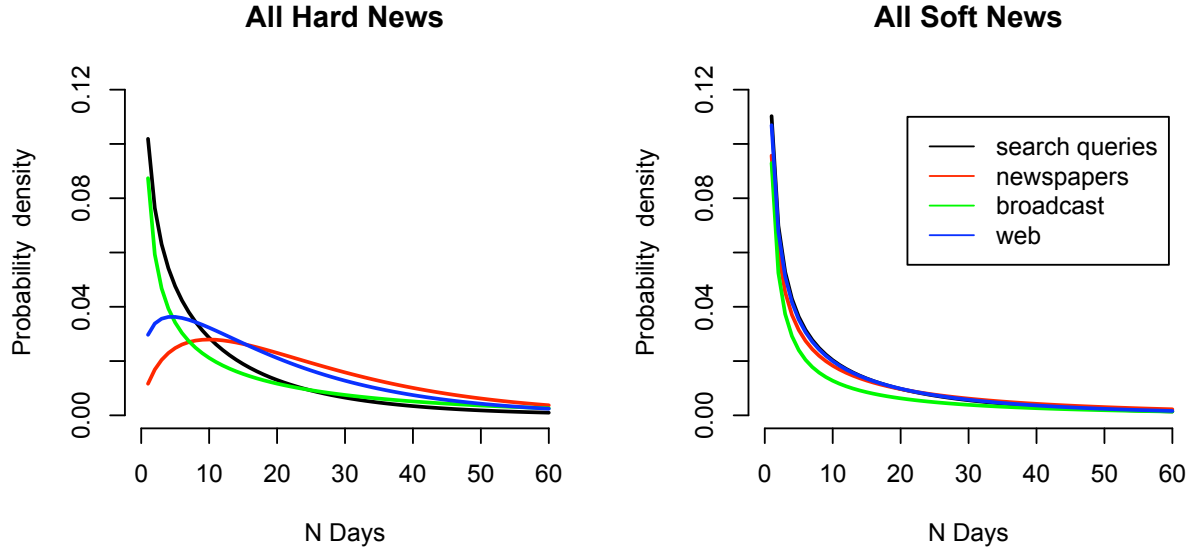
Query volume is routinely sustained longer than traditional news coverage with two exceptions: Joe Wilson (where broadcast and newspaper covered the issue for a longer time), and Tiger Woods (newspapers had a pattern of decay similar to online search traffic). Search volume and Web coverage were notably higher in the instances of Climategate and Wikileaks, indicating that searchers are following up to find some of the classified documents released on the Web.

**Table 3: Sensational News Fitted Scale and Shape Parameters**

<b>Event</b>	<b>Queries</b>	<b>Broadcast</b>	<b>Newspapers</b>	<b>Web</b>
<b>Climategate</b>				
	Shape	0.4014 (0.0683)	0.1378 (0.0218)	0.2890 (0.0478)
	Rate	0.0112 (0.0032)	0.0093 (0.0040)	0.0088 (0.0028)
	KS D-stat		0.3778**	0.1778
<b>Wikileaks</b>				
	Shape	0.4457 (0.0735)	0.0931 (0.0139)	0.1074 (0.0161)
	Rate	0.0432 (0.0117)	0.0187 (0.0092)	0.0146 (0.0067)
	KS D-stat		0.7755**	0.6327**
<b>Joe the Plumber</b>				
	Shape	0.5821 (0.0995)	0.2899 (0.0465)	0.5443 (0.0924)
	Rate	0.0528 (0.0135)	0.0136 (0.0042)	0.0323 (0.0083)
	KS D-stat		0.3333*	0.2708
<b>Joe Wilson</b>				
	Shape	0.1131 (0.0162)	0.2248 (0.0334)	0.4348 (0.0682)
	Rate	0.0268 (0.0115)	0.0179 (0.0058)	0.0251 (0.0065)
	KS D-stat		0.6667**	0.7037**
<b>Sniper Fire</b>				
	Shape	0.3943 (0.0664)	0.2431 (0.0394)	0.4025 (0.0678)
	Rate	0.0213 (0.006)	0.0181 (0.004)	0.0144 (0.004)
	KS D-stat		0.413*	0.3043
<b>Michael Jackson</b>				
	Shape	0.9190 (0.1377)	0.9903 (0.1493)	0.8262 (0.1226)
	Rate	0.0823 (0.0161)	0.0452 (0.0088)	0.0560 (0.0112)
	KS D-stat		0.4559	0.1765
<b>Tiger Woods</b>				
	Shape	1.2154 (0.2461)	0.6418 (0.1226)	1.2797 (0.2604)
	Rate	0.0330 (0.0082)	0.0186 (0.0051)	0.0498 (0.0123)
	KS D-stat		0.1538	0.3077
<b>Sandra Bullock</b>				
	Shape	1.2476 (0.1650)	0.2993 (0.0348)	0.7646 (0.0968)
	Rate	0.0898 (0.0145)	0.0357 (0.0080)	0.0755 (0.0131)
	KS D-stat		0.3913**	0.1739
<b>Roman Polanski</b>				
				0.3261**

Shape	1.2154 (0.2461)	0.6418 (0.1226)	1.2797 (0.2604)	0.7947 (0.1551)
Rate	0.0330 (0.0082)	0.0186 (0.0051)	0.0498 (0.0123)	0.0363 (0.0096)
KS D-stat	0.4394**		0.50**	0.303*

**Table 3:** Standard errors are reported in parentheses below each parameter estimate. K-S test D values follow. P-values less than 0.001 are indicated with \*\*; p-values less than 0.05, with \*.



**Figure 3.** Fitted gamma distributions, aggregating all hard and soft news events

**Table 4: Aggregate Hard and Soft News Parameters**

Hard News	Queries	Broadcast	Newspapers	Web
	0.6445 (0.0918)	0.4736 (0.1252)	1.6257 (0.2515)	1.2781 (0.1935)
	0.0554 (0.0117)	0.0229 (0.0052)	0.0628 (0.0113)	0.0617 (0.0114)
Soft News	Queries	Broadcast	Newspapers	Web
	0.6237 (0.1012)	0.3167 (0.0496)	0.5213 (0.0854)	0.4795 (0.0762)
	0.0455 (0.0109)	0.0212 (0.0065)	0.0302 (0.0073)	0.0409 (0.0100)

**Table 4:** Standard errors are reported in parentheses below each parameter estimate.

## NATURAL DISASTERS

Finally, the news coverage and query volumes following the four natural disasters were also fitted to gamma distributions; these four events show the most marked within-event-group difference (Table 4). Specifically, the December 26, 2004 tsunami off of Indonesia, the April 6, 2009 earthquake in L'Aquila Italy, the 2008 Hurricane Katrina hitting Louisiana, and the January 2010 earthquake striking Port au Prince, Haiti. These distributions are all fairly unique, likely dependent on the amount of damage and devastation caused. It is quite possible that there will be no one single trend to predict how news media will cover a natural disaster depending on its scope and devastation, and when analyzing coverage and attention to natural disasters, it may not be best to lump them in one given category.

The distribution of queries was significantly different than all news types for the both the Italy earthquake and the Haiti earthquake. This difference is distinctly more obvious in the case of the Haiti earthquake, as while search volume had a shape parameter of 1.029 (se=0.185), all media had distinctly larger shape parameters (newspapers: 2.073 (se=.393); broadcast: 1.393 (se=0.357); Web: 1.699 (se=0.318)). Similarly, the other international disaster, the December 26, 2004 tsunami off of Indonesia, showed significant differences between search queries and traditional media (newspapers and broadcast), yet searches were consistent with what might be predicted from Web coverage (difference between queries and transcripts/ newspapers – Tsunami:  $D = 0.4359$ ,  $p\text{-value} = 0.003$ ; Italy earthquake:  $D = 0.50$ ,  $p\text{-value} = 0.003$ ). Searches related to Hurricane Katrina were not distributed significantly differently than transcripts or web content, but did show a distinct difference from newspapers; newspapers slowly tapered off coverage in comparison to the other media types. Again, we would expect this of newspaper and Web content, as there is simply more space for coverage and follow-up.

**Table 5. Natural Disasters, Fitted Shape and Scale Gamma Parameters**

Event	Queries	Broadcast	Newspapers	Web
Haiti				
Shape	1.0295 (0.1854)	1.3936 (0.2574)	2.0731 (0.3936)	1.6990 (0.3185)
Rate	0.0560 (0.0129)	0.0567 (0.0125)	0.0645 (0.0138)	0.0741 (0.0161)
KS D-stat		0.3958*	0.5625**	0.4792**
Tsunami				
Shape	1.5071 (0.3105)	0.3889 (0.0710)	0.6075 (0.1153)	0.6430 (0.1229)
Rate	0.0446 (0.0109)	0.0171 (0.0054)	0.0130 (0.0036)	0.0204 (0.0056)
KS D-stat		0.4359*	0.4359*	0.1282
Italy				
Shape	0.1298 (0.0182)	0.1069 (0.0148)	0.1921 (0.0275)	0.1366 (0.0192)
Rate	0.0293 (0.0115)	0.0156 (0.0066)	0.0150 (0.0050)	0.0198 (0.0076)
KS D-stat		0.5439**	0.5439**	0.3333*
Katrina				
Shape	0.6752 (0.1297)	0.9409 (0.1865)	0.5483 (0.1031)	0.7515 (0.1458)
Rate	0.0220 (0.0060)	0.0455 (0.0117)	0.0136 (0.0039)	0.0272 (0.0073)
KS D-stat		0.2821	0.3846*	0.2051

**Table 5:** Standard errors are reported in parentheses below each parameter estimate. K-S test D values follow. P-values less than 0.001 are indicated with \*\*; p-values less than 0.05, with \*.

## DISCUSSION

This research seeks to answer two questions: first, can online search queries be used as a measure of media agenda setting; and second, how well does public interest and attention in a news event match what the media coverage might predict? Specifically, does the news coverage and public attention vary based on whether the news event is substantive or sensational?

From the results of the paper, the answer to all questions is affirmative. Clearly, there are unique correlations between search queries and news coverage, as news coverage must be one of the key triggers that prompt an individual to search about the given event (in order to perform a



search about a given topic, an individual first needs to hear *something* about it). From comparing the distributions of the media types across specific events, it is clear that search queries do not always follow the path of one media type, though in the case of hard news (both international and domestic), the distribution of search queries most closely resembles that of broadcast network coverage. In the case of sensational stories, search query volume is most closely aligned with the news mediums that can afford covering the issue longer after the story breaks – Web content, and in some instances, newspaper content.

One possible explanation for the relationship between search volume and news coverage is that broadcast news may be the most effective media in setting the agenda for the public when it comes to substantive or international news, and in a way that may diminish public perceptions of importance of these events. Broadcast coverage of hard news events drops off more quickly than any other news media, in some cases before the public attention span has declined. It is possible that this lack of coverage discourages individuals from searching about these topics, as agenda setting suggests that the public takes a cue from the volume (or lack thereof) of news coverage. It is conceivably a bit harder for the average American to fully digest or form an opinion on nuanced international issues, and they may rely most on broadcast news to inform their perceptions of event importance. It should also be noted that broadcast network coverage of international and substantive news is at a low, whereas news that is thought to “sell” (e.g., “if it bleeds it leads”) has been given more airtime. These findings seem in accordance with this trend, as there is only brief broadcast coverage of hard news events, yet more sustained coverage of celebrity events.

An alternate explanation presumes more weight on the type of event rather than the influence of news media. Specifically, the international substantive events studied in this

research are inherently complex and multi-faceted, and are in fact best suited for much post-hoc analysis and even editorializing in newspapers and blog/Web content. Thus, there is much more topical news to be discussed and elaborated upon in newspapers and Web, and the decay curves for these mediums may be mere artifacts of each news mediums' inherent ability to report nuance surrounding the news event. Despite this, the fact that individuals are no longer using the Internet to search for these further details of substantive events indicate that the public attention span is not as great as the newspaper coverage allows. Further research needs to be conducted with a larger sample of events to determine if these effects are in fact consistent, and also more closely analyze the between-media differences in news coverage. In addition to assessing the relative attention spans of different news media and the public, a time-series analysis would more closely indicate from which media the public takes their cues.

#### ACKNOWLEDGMENTS

Thank you to Shanto Iyengar, Julie Granka, and Kieron Wesson for valuable ideas, discussions, and contributions regarding this analysis.

## REFERENCES

- Adamic, L., & Glance, N. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery* (p. 36–43). ACM New York, NY, USA.
- Althaus, S. L. 1998. Information effects in collective preferences. *American Political Science Review*, 92, 545–558.
- Baum, M. A. 2002. Sex, lies, and war: How soft news brings foreign policy to the inattentive public. *American Political Science Review*, 96, 91–110.
- Choi, H. & Varian, H. 2009. Predicting the Present through Google search queries. April 2. Retrieved: [http://google.com/googleblogs/pdfs/google\\_predicting\\_the\\_present.pdf](http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf)
- Delli Carpini, M. X. 2000. In search of the informed citizen: What Americans know about politics and why it matters. *Communication Review*, 4, 129–164.
- Delli Carpini, M. X., & Keeter, S. 1996. *What Americans know about politics and why it matters*. New Haven, CT: Yale University Press.
- Delli Carpini, M. X., & Williams, B. A. 2000. Let us infotain you: Politics in the new media environment. In W. L. Bennett & R. M. Entman (Eds.), *Mediated politics: Communication in the future of democracy* (pp. 160–181). Cambridge, England: Cambridge University Press
- Eveland, William P., Jr., and Dietram A. Scheufele. 2000. “Connecting News Media Use with Gaps in Knowledge and Participation.” *Political Communication* 17(3):215–37.
- Fallows, Deborah. Search Engine Use. Pew. Aug 2008.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L., et al. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012–4. doi: 10.1038/nature07634.
- Goel, S., Hofman, J., Lahaie, S., Pennock, D., & Watts, D. 2010. What can search predict. In *WWW 2010, April 26–30, 2010, Raleigh, North Carolina*. Retrieved from <http://cam.cornell.edu/~sharad/papers/searchpreds.pdf>.
- Google Insights for Search. <http://www.google.com/insights/search/>
- Hindman, M., Johnson, J. A., & Tsioutsouliklis, K. 2003. Googlearchy: How a few heavily-linked sites dominate politics on the web. In *Annual Meeting of the Midwest Political Science Association* (Vol. 4, pp. 1–33). American Political Science Association. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.7.7520&rep=rep1&type=pdf>.
- Iyengar, S. & Kinder. 1984. *News that Matters: Television and American Opinion*. Chicago: U. of Chicago Press.
- Iyengar, S., Hahn, K. S., Krosnick, J. A., & Walker, J. 2000. Selective Exposure to Campaign Communication: The Role of Anticipated Agreement and Issue Public Membership. *Journal of Politics*, 70(1), 186–200. doi: 10.1017/S0022381607080139.

- Iyengar, S., & Hahn, K. S. 2009. Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication*, 59(1), 19-39. Blackwell Publishing Inc. doi: 10.1111/j.1460-2466.2008.01402.x.
- Iyengar, S., & Bennett, W. L. 2008. A new era of minimal effects? The changing foundations of political communication. *Journal of Communication*, 58(4), 707–731. Blackwell Publishing. Retrieved from <http://pcl.stanford.edu/research/2008/bennett-minimaleffects.pdf>.
- Krosnick, J., 1999. Survey Research. *Annual Review of Psychology*, 50: 537-67.
- Leskovec, J., Backstrom, L., & Kleinberg, J. 2009. Meme-tracking and the Dynamics of the News Cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (p. 497–506). Paris: ACM New York, NY, USA.
- Mccombs, M., Shaw, D. L., & Research, O. 1972. The Agenda-Setting Function of Mass Media. *Public Opinion Quarterly*, 36(2), 176-187.
- Price, V., & Zaller, J. 1993. Who gets the news? Alternative measures of news reception and their implications for research. *Public Opinion Quarterly*, 57, 133–164.
- Prior, M. 2005. News vs. Entertainment: How Increasing Media Choice Widens Gaps in Political Knowledge and Turnout. *American Journal of Political Science*, 49(3), 577-592.
- Prior, M. 2003. Any good news in soft news? The impact of soft news preference on political knowledge. *Political Communication*, 20, 149-72.
- Prior, M. 2009. Improving Media Effects Research through Better Measurement of News Exposure. *The Journal of Politics*, 71(3), 893. doi: 10.1017/S0022381609090781.
- Shoemaker, P. 1991. *Communication Concepts 3: gatekeeping*. Newbury Park, CA: Sage.
- Sunstein, C. R. 2007. *Republic 2.0*. New York: Princeton University Press.
- Wu, F., & Huberman, B. a. 2007. Novelty and collective attention. *Proceedings of the National Academy of Sciences of the United States of America*, 104(45), 17599-601. doi: 10.1073/pnas.0704916104.

## APPENDIX A: EVENTS USED FOR ANALYSIS

Following is a list of events and in brackets, the exact specification used to query news volume from the Lexis Nexis database and in Google Insights for Search. Two bracketed terms are sometimes specified for Lexis Nexis data to indicate that the two words should be found within 5 words of each other, or within the same sentence. (An additional note: If Lexis Nexis finds more than 3,000 matching news documents, it truncates them to the 1,000 most relevant documents, so multiple batches of news data had to be collected to avoid this problem.)

### INTERNATIONAL HARD NEWS

1. Iran election: 6.22.2010  
[iran election]  
The additional specification of [election] “within the same sentence” as [Haiti] was used to query the Lexis Nexis database.
2. David Cameron (UK election 5.6.2010)  
[David Cameron]
3. Gaza [Israel-Gaza war]:12.27.2008 - 1.18.2009  
[Gaza]  
The additional specification of [Israel] within the same sentence as [Gaza] was used to query the Lexis Nexis database.
4. Iraq Election: 3.7.2010  
[Iraq election]  
The additional specification of [election] “within the same sentence” as [Iraq] was used to query the Lexis Nexis database.

### DOMESTIC HARD NEWS

5. Arizona immigration: 4.23.2010 signed into law  
[Arizona immigration]
6. Sonia Sotomayor nomination: May 26, 2009  
[sotomayor] in queries; [sonia sotomayor] for Lexis Nexis
7. Health care bill passage: 3.23.2010  
[health care bill]
8. Cash for clunkers program: started 7.1.2009  
[cash for clunkers]

### CELEBRITY EVENTS

9. Michael Jackson’s death: 6.25.2009  
[Michael Jackson]
10. Tiger Woods: 11.25.2009 (National Enquirer story); 11.27.2009 (car accident)  
[Tiger Woods]
11. Sandra Bullock divorce: 3.23.2010  
[Sandra bullock]
12. Roman Polanski arrest 9.26.2009

### SENSATIONAL EVENTS

13. Joe the Plumber (mention in 3rd presidential debate): 10.15.2008  
[Joe the plumber]
14. Sniper fire: 3.17.2008 (speech); 3.24.2008 (retracted)

- [sniper fire]  
The additional specification of [Hilary Clinton] “within the same paragraph” as [sniper fire] was used to query the Lexis Nexis database.
15. Joe Wilson; 9.9.2009  
[Joe Wilson]
16. Wikileaks; 4.5.2008 (Iraq documents released)  
[Wikileaks]
17. Climategate; 11.19.2009 (documents uploaded)  
[climategate]
18. Hudson plane crash; 1.15.2010  
[Hudson plane crash]  
The additional specification of [Hudson] “within the same sentence” as [plane crash] was used to query the Lexis Nexis database.

#### DISASTERS

19. Hurricane Katrina; 8.23.2005 (formed); 8.29.2005 (struck)  
[Hurricane Katrina]
20. Italy earthquake; 4.06.2009  
[italy earthquake]  
The additional specification of [earthquake] “within the same sentence” as [Italy] was used to query Lexis Nexis.
21. Tsunami; 12.26.2004  
[tsunami]
22. Haiti earthquake; 1.12.2010  
[Haiti earthquake]  
The additional specification of [earthquake] “within the same sentence” as [Haiti] was used to query the Lexis Nexis database.

## APPENDIX B

### COMPARISON OF RAW DATA AGAINST THE FITTED WEIBULL AND GAMMA DISTRIBUTIONS

The two functions exhibit similar results, so only the results of the gamma model were reported in the paper.

