Events and Controversies: Influences of a Shocking News Event on Information Seeking

Danai Koutra*
Carnegie Mellon University
Pittsburgh, PA
danai@cs.cmu.com

Paul N. Bennett
Microsoft Research
Redmond, WA
pauben@microsoft.com

Eric Horvitz
Microsoft Research
Redmond, WA
horvitz@microsoft.com

ABSTRACT

It has been suggested that online search and retrieval contributes to the intellectual isolation of users within their preexisting ideologies, where people's prior views are strengthened and alternative viewpoints are infrequently encountered. This so-called "filter bubble" phenomenon has been called out as especially detrimental when it comes to dialog among people on controversial, emotionally charged topics, such as the labeling of genetically modified food, the right to bear arms, the death penalty, and online privacy. We seek to identify and study information-seeking behavior and access to alternative versus reinforcing viewpoints following shocking, emotional, and large-scale news events. We choose for a case study to analyze search and browsing on gun control/rights, a strongly polarizing topic for both citizens and leaders of the United States. We study the period of time preceding and following a mass shooting to understand how its occurrence, follow-on discussions, and debate may have been linked to changes in the patterns of searching and browsing. We employ information-theoretic measures to quantify the diversity of Web domains of interest to users and understand the browsing patterns of users. We use these measures to characterize the influence of news events on these web search and browsing patterns.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications—*Data mining.*

Keywords: Controversies, Filter bubble, Log / behavioral analysis.

1. INTRODUCTION

How do people navigate webpages on polarizing topics? Are they isolated in their echo chambers? Do shocking news events burst their ideological bubbles and make them more likely to seek information on opposing viewpoints? These are the key questions we investigate.

With advances in personalization methods, search engines and recommendation systems increasingly adjust results to users' preferences, as inferred from their past searches and choices. In addition, users often input biased queries [38], which reflect their own positions, while personalized results have the potential to reinforce these opinions, acting as echo chambers. As a result, according to several recent studies [38, 15, 29], users remain within informational bubbles. The phenomenon is sometimes referred to as the "filter bubble" effect, where people get exposed only to opinions that align with their current views. This effect, where the world of viewpoints that people are exposed to on the web does not reflect the richness of views in the real world, may be especially strong for polarizing topics. We take as polarizing or controversial topics those linked to opposing perspectives, such as abortion, gun control vs. rights, labeling of genetically modified food, and death penalty.

To understand users' information seeking behaviors on polarizing issues, we focus on a highly controversial topic in the US: gun control and rights. At one end of the spectrum, extreme gun rights supporters argue an interpretation of the 2nd Amendment to the US Constitution that would prohibit any regulation of firearms. On the other side of the spectrum, extreme gun control supporters advocate the total ban of any private citizen ownership of firearms. Beyond these two extreme opinions are a spectrum of variations that lay between them (*e.g.*, more background checks, ban of fully automatic firearms). For our study we use web browser toolbar logs from November and December 2012, and primarily consider two time periods: before and after the Sandy Hook Elementary School Shooting (S.H.) in Newtown, Connecticut (December 14th), an event with broad news coverage and nationwide impact.

For a historical perspective, we summarize the event facts in [1] to aid readers in understanding why the event might be expected to broadly influence information consumption. The Sandy Hook shooting is the most deadly shooting in US history at a high, middle, or elementary school and the second deadliest in US History by a single perpetrator. The casualties included 20 children ages 6-7, six staff members, the perpetrator's mother (offsite), and the perpetrator – a 20-year-old male with no motive ever determined and a history of several psychological conditions. In a span of five minutes, the shooter entered the building and fired 156 rounds (one bullet every other second) causing all but one fatality during that span. The perpetrator committed suicide as the police arrived on the scene (five minutes after the shooter entered the building). Given the complexity and nature of the event, there was considerable political debate and media discussion following the event. Our focus here is on how this event may have influenced the general US population's information search and retrieval.

The event clearly had considerable influence on information seeking about gun control related topics as signified by the increased user activity in the days following the event (see Fig. 1). The first big spike in the figure, which corresponds to visits to on-topic websites on the day of the shootings, and other important spikes have been annotated. The effect on the quantity of information seeking is indisputable; so our focus is *not* on the increase in user activity, *but* on whether (and how) the event changed the *type* of activity.

*Work done during an internship at Microsoft Research.

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741099 .

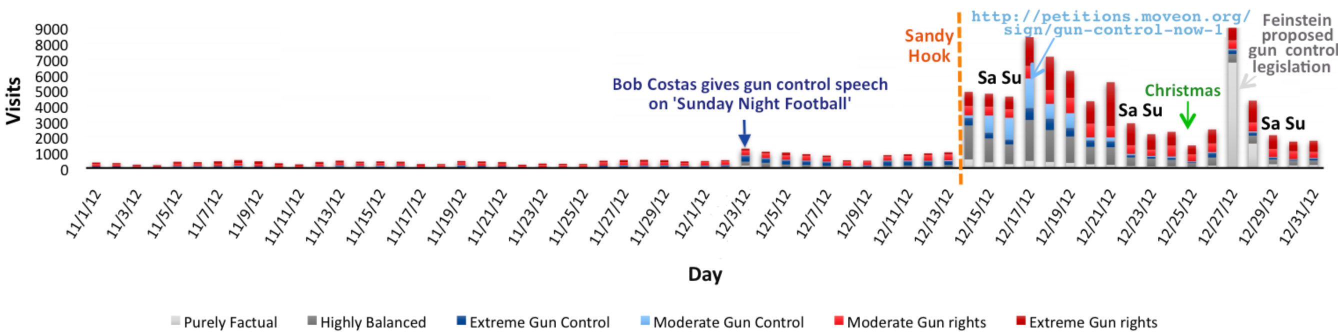

Figure 1: Number of visits to gun control/rights related webpages over time (November-December 2012). The colors correspond to webpage categories: gray for factual and balanced pages; blue for pages supporting gun control; and red for pages supporting gun rights. The categories and the labeling process are described in the Appendix and Sec. 3.2 respectively.

For the following analysis, we use raw web browser visitation logs from Internet Explorer, where users have given consent to logging all non-https URLs from URLs visited from search and those reached by direct entry or browsing. By employing techniques such as a two-step random walk on the query-click graph [11] and white-list and keyword-based classifiers, we extract –from this broad set of visitations– and label a large-scale dataset of user interaction data that is relevant to the gun debate, constituting about 61K users visiting 378K on-topic websites (Sec. 3). We first present evidence that websites are polarized with respect to individual topics in terms of their webpage content (Sec. 4). Then, we evaluate the diversity of the users and investigate to what extent ideological bubbles exist before and after the shootings (Sec. 5). Moreover, we explore the click trails of the users to understand how people transition among webpages of opposing views, and how news about the shootings influences such transitions (Sec. 6). Finally, we categorize users based on their browsing behavior (Sec. 7), and discuss the dynamics of the communities over the course of the time.

Contributions. This paper presents a case-study which is both interesting in its own right, but also highlights the computational tools and analysis methodology to answer questions such as : What type of websites offer the most diverse opinions? (Sec. 4); Do users desire diversity in opinions? (Sec. 5.1); Does a shocking event impact the user's desire for diversity? (Sec. 5.2); Is the polarity of a web page predictive of the polarity of the next domain on topic that a user will read a page from? (Sec. 6.1); Does a shocking event change the predictability of the polarity of the next domain on topic conditioned on the current one? (Sec. 6.2); Does a shocking news event permanently shift the user's topical view? (Sec. 7)

As a case-study the answers to these questions for this topic have implications for ranking scores for websites on polarizing topics based on predicted diversity, when diversity should be incorporated into search and recommendation results, how that diversity should change in the face of events, and for what duration of time a user's view should be persisted for personalization.

2. RELATED WORK

We first place our work in the context of related research, which includes studies on political controversies, conjectures about the so-called filter bubble, and the temporal evolution of knowledge.

Political Controversies. Munson et al. [25] focus on blog posts to study if people seek diverse information, while Balasubramanyan et al. [8] use an LDA-based methodology to predict how different communities respond to political discourse. Aktogla and Allan [6] show that diversification of search results in terms of sentiments to an explicit bias improves user satisfaction. The authors in [13] propose a model to mine contrastive opinions for political issues, and many research groups devise methods for polarity detection and political leaning classification [10, 36, 40, 30] or for understanding event dynamics and their relation to sentiment shifts [35]. In [23] and [39], the authors present work on extending sentiment analysis to match political text to parties. Awadallah et al. [7] mine the web to automatically map well-known people to their opinions on political controversies.

Filter Bubble. Pariser [26] points out the existence of the filter bubble, which he defines as "this unique, personal universe of information created just for you by an array of personalizing filters", and many works propose ways to mitigate its effects [15, 29]. For example, Munson et al. [24] build a browser widget that encourages the users to read diverse articles on political issues in order to avoid the selective exposure of users to political information. Yom-Tov et al. [38] focus on news outlet sites that people visit, quantify the filter bubble and study whether users browse webpages supporting disagreeable information when opposing views are introduced in their search results.

Temporal Evolution of Knowledge. White et al. [37] focus on the temporal search behavior of users to quantify the differences between experts and non-experts in terms of vocabulary, sites visited and search strategies. Kotov et al. [20] model and analyze user search behavior that spans multiple sessions in order to improve search for complex needs and support tasks which require cross-section searches. In a similar context, Liu et al. [22] study how the acquired user knowledge changes over time through performing multi-session information tasks.

This Work. In contrast to most previous work, which considers primarily news outlets and blogs, and studies whether people access sources of different political categories to get informed [24, 38], we put a particular topic under the microscope and study how *that* affects the browsing behavior of the users. Another major difference from prior efforts is that we separate the political orientation of the users from their orientation to the gun debate. For example, a Gallup poll in 2005 [2] indicated that 23%/27%/41% of, respectively, Democrats/Independents/Republicans own a gun for an overall average of 30% of US adults. Thus, while gun ownership correlates with political leanings, there is significant ownership in each population. Given that, it is quite likely that views toward the gun debate may differ from party affiliation as well. Thus, we do not engage in the common practice of characterizing websites as liberal and non-liberal. Rather, we define our own *content*-oriented labels (Appendix A). Finally, although our work is motivated by the findings of prior studies on the existence of the filter bubble, our focus is not limited to corroborating or opposing this view. Our goal is to understand the types of webpages people visit, as well as how they transition among content expressing different viewpoints.

We contribute an analysis of the temporal evolution of the users' *browsing* behaviors, and especially the *influence of specific external events with nationwide impact* on the shaping of the users' stances and their overall polarity. We also analyze the transitions of the users among webpages of different viewpoints.

3. THE GC-DEBATE DATA

We now present the dataset that we used for our study, focusing mainly on the data extraction and annotation.

3.1 Data Extraction

The data comes from users' anonymized search and browse behavior logged through Internet Explorer's instrumentation during November and December 2012. The data covers queries issued to a variety of search engines, as well as non-encrypted URLs that were visited, for more than 29 million users in the US-English market. While the sample of users in the log may not perfectly represent the distribution of the US population, independent studies [3] demonstrate that the user population of Internet Explorer contains significant representation from both genders and nearly all age and income levels of the US population. Thus, the changes we discuss at least indicate broad patterns of change across demographics and with respect to our user base.

Beyond the analysis of interaction on this particular topic, we seek to identify computational approaches to analyzing changes in patterns of information browsing given typical constraints on observation. To that end, we do not assume that our logs capture all of the user's on-topic activity across all devices but rather a random sample of the user's activity with respect to the topical content orientation. By random, we mean, that the user's selection of browser or device is independent of the topical polarity; for example, the user does not perform all of their browsing of gun rights on an alternative browser or device for which we would not have log information while all of their gun control activity on an instrumented browser on the desktop.

We consider primarily two time periods: before and after the Sandy Hook Elementary School shooting on December 14th. We note that we consider logs from a longer period of time before the event to develop a more robust estimate of users' habitual activity—a similar quantity of activity is observed in the period after the shooting because information seeking is more frequent after the event (Fig. 1). For the purposes of our study, we consider the URLs that are *on-topic*, i.e., websites that discuss gun control/rights issues. Hence, our first goal is to extract the relevant data with techniques that can be re-used in a programmatic manner for the analysis of other topics.

A naïve approach to obtaining a corpus of on-topic data is to consider all webpages containing the word "gun". Such an approach leads to numerous false positives, including websites about toys, video games, glue guns *etc*. We took an alternate approach that yielded a corpus with many fewer false positives. The extraction process focused on identifying on-topic seed queries with high precision and then expanding these to related URLs and queries to obtain high coverage of all of the on-topic activity. Specifically, the multi-step procedure, as illustrated in Fig. 2, does the following:

STEP 1. Identification of Relevant Queries: We start with easy to identify relevant queries through keyword matching, and automatically expand them to as many relevant queries as possible by exploiting usage data.

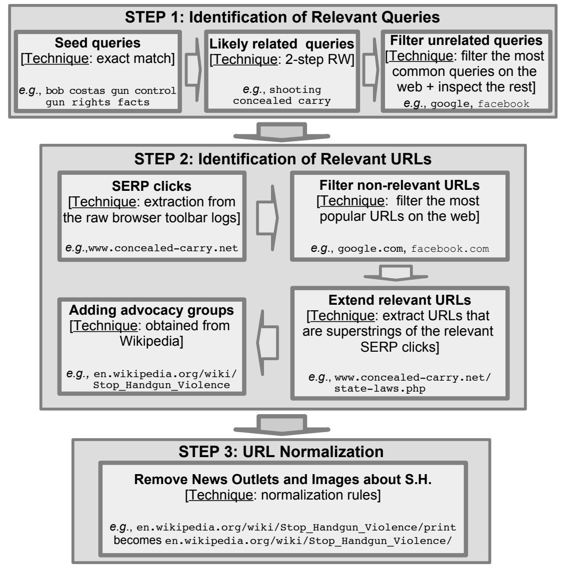

Figure 2: Illustration of the data extraction process.

Table 1: The most popular seed queries (col. 1), and relevant queries before and after the Sandy Hook shootings (col. 2 and 3).

Top 15 seed queries	Top 15 relevant queries before Sandy Hook	Top 15 relevant queries after Sandy Hook
Bob Costas gun control	Bob Costas gun control	Connecticut shooting
gun control petition	shooting	school shooting in Connecticut
Rupert Murdoch gun control	2nd amendment	school shooting
Piers Morgan gun control	gun control	Connecticut school shooting
gun control	nutnfancy	shooting in Connecticut
Feinstein gun control	Oregon shooting	elementary school shooting
gun control debate	second amendment	gun control petition
Rahm Emanuel gun control	concealed carry	Rupert Murdoch gun control
Murdoch gun control	National Rifle Association	Sandy Hook shooting
gun control laws	Obama gun control	shooting
obama gun control	home invasion	piers Morgan gun control
boehner on gun control	jason whitlock gun control	nra statement
ted nugent gun control	illinois gun laws	gun control
white house gun control petition	gun news	obama gun ban
piers morgan gun control debate	the second amendment	conneticut shooting

1A. Seed Queries. First, we identify seed queries by extracting those queries that contain the phrases "gun control" or "gun rights", but that are not related to electronic games. By doing this, we automatically filter out the queries that have an exact match with "xbox", "wii", "gun controller", "game", or "playstation". The resulting set consists of 6,878 queries, the 15 most popular of which are given in Table 1 (col. 1).

1B. Identifying Likely Related Queries. The second step consists of expanding the set of seed queries to relevant queries (misspelled, different expressions of the same intent, *etc.*). For this purpose, we create the query-click graph, a bipartite graph, where each query in the web logs is connected to the impression URLs that some user clicked; queries linked via a clicked URL are referred to as co-clicked. Starting from the seed set, we perform a two-step random walk [11], and expand the seed set to all the similar co-clicked queries, as evaluated by their character-trigram cosine similarity with the seed queries. The threshold for similarity is set to 0.5 to require relatively high similarity. Intuitively, the new queries are connected to the same URLs as the seed queries and have textual overlap. Thus, they are likely on-topic, and probably represent alternative ways of querying for highly related results.

1C. Filtering Non-Relevant Queries. Finally, from the likely relevant set of queries, after ordering them in decreasing order of popularity, we inspect and filter out the most common overall and seasonal queries, such as the navigational queries, *google* and *facebook*. Moreover, by manually inspecting the queries without the word "gun", we remove those queries that are not directly related to gun control, and lead to retrieval of numerous URLs unrelated to gun control (high recall/low precision) – e.g., *what do democrats and republicans stand for*, *conservative viewpoint*. The final, extended set, to which we will refer as set of *relevant* queries, consists of 7,778 queries. The most popular queries before and after Sandy Hook are given in Table 1 (col. 2 and 3).

STEP 2. Identification of Relevant URLs: Users reach URLs through many ways (*e*.g., browsing, search, bookmarks). Our objective is to use the resulting on-topic *queries* to identify sessions of information-seeking behavior, which according to IR research tend to be topically coherent. Again, a naïve approach would be to extract any clicked URL from the search engine result page (SERP) of a topical query, as well as the pages browsed subsequently by consecutive clicks (click trail). However, users may click on ads and other contextual links (some of which may be topically relevant, but often not), and browse from a topical article to a non-topical one as they drift to a different topic. Therefore, similar to identifying relevant queries, we developed a semi-automated way of expanding to a broad topically relevant set without incorporating significant amounts of off-topic search and browsing:

2A. SERP Clicks. Starting from the relevant queries of the previous step, we obtain only the URLs users clicked directly from a topically relevant query's SERP.

2B. Filtering Non-Relevant URLs. Then, we filter URLs that are among the most popular URLs worldwide (e.g., `google.com`, `yahoo.com`, `youtube.com`, `facebook.com`), which reflect the way the users reached the on-topic URLs, but are not on-topic themselves. Although media analysis is interesting, we focus primarily on non-video web pages (*i.e.*, mainly text). We refer to this set of filtered URLs on gun control and gun rights as *seed* URLs.

2C. Extend Relevant URLs. We continue by extending the set of seed URLs to include more webpages that might not belong to the SERP clicks of relevant queries. To this end, we consider relevant all URLs that are superstrings of the seed URLs. The intuition is that those were either reached from or led to a seed URL, and have high overlap in the site organization – implying a topical relationship. Moreover, this process leads to higher recall, as it also extracts URLs entered in the toolbar, or saved as bookmarks.

2D. Adding Advocacy Groups. The method described above is *not* guaranteed to extract *all* the URLs that are relevant to gun control and rights. However, the procedure attempts to extract as many, highly related websites as possible, while maintaining neutral criteria with respect to the topic of study. Extracting all the webpages that are on-topic is challenging and is a distinct research problem. We seek to make sure that we capture visits to webpages for the most prevalent gun control and rights advocacy groups. Thus, we take compiled lists from Wikipedia[1], and explicitly extract user visits to both the advocacy group websites and their Wikipedia pages.

STEP 3. URL Normalization: Finally, we normalize the URLs so that different webpages with the same content, mobile versions of the websites, print requests of a page, user id encoding pages, *etc.* are considered the same.

The resulting dataset, GC-DEBATE, consists of records <user-id, session-id, URL, timestamp> (Table 2). In the following sections, we refer to the intersection between the sets of users before and after the shootings as **common users**. Studying them enables us to directly compare changes in user behavior by controlling for the set of users.

[1]Gun control / rights advocacy groups in the United States:
http://en.wikipedia.org/wiki/Category:Gun_control_advocacy_groups_in_the_United_States
http://en.wikipedia.org/wiki/Category:Gun_rights_advocacy_groups_in_the_United_States

Table 2: GC-DEBATE dataset. The last column holds the number of common users, URLs and domains between the two time periods.

	Before S.H.	After S.H.	Total	Overlap
Users	12,919	56,293	61,276	7,936
Unique URLs	6,081	20,788	25,201	1,668
Unique Domains	340	682	803	219
Total Visits	123,596	253,994	377,590	N/A

Table 3: Inter-rater agreement for the high-level labels (col. 1), and the expanded set of labels (col. 2). Overall agreement is simply the percent of labels on which the raters agree.

	Labels	
	High-level	**Expanded**
Overall agreement	86.10%	73.61%
free-marginal κ	82.64%	66.21%
fixed-marginal κ	77.53%	69.84%
chance-expected agreement	19.69 %	10.30%

3.2 Data Annotation

Answering the questions we have posed is not possible unless the webpages are labeled based on their stance on gun control/rights. Rather than focusing on alignment with a political party, we focus on the disposition of the content itself. Visits to a site that is predominantly affiliated with one party (*e.g.*, Democratic/Republican) or a particular pundit, does not by itself imply a lack of diversity in content; sites may contain content discussing a broad range of material. Considering the content also enables us to measure that extent to which sites provide information representing diverse views.

Manually labeling *all* the webpages is difficult. Our attempts to automate the labeling process by building *content*-centric classifiers failed to achieve high accuracy, revealing the challenges of classifying controversial pages by their stance. We could not apply the extensive work on detecting and labeling controversial topics [10, 40, 30], as our setting is different: we seek to characterize the presented *viewpoints* in documents on a *given controversial topic*. To overcome these challenges, we judged all webpages that had more than two unique visitors and sampled from the remaining webpages, obtaining this way 99.5% coverage of total visitations. The on-topic and accessible pages were initially judged by their content and classified into three high-level categories: balanced, gun control, gun rights. Then, they were further classified into expanded categories that reflect the stance of the webpages at a finer granularity: purely factual and highly balanced, extreme and moderate gun control, and extreme and moderate gun rights. Details about the labels are provided in the Appendix.

Three expert assessors were provided with a subset of over 2,100 popular webpages, and were asked to classify them. One assessor self-identified as "moderate gun rights", while the other two self-identified as "moderate gun control". The inter-rater agreement [28], which already accounts for the chance-expected proportion of agreement between the assessors, is 82.64% for the high-level classification, and 66.21% for the expanded labels that reflect further key category distinctions. We note that these inter-rater agreements are high, since the chance-expected agreement [17] using the marginal distribution is 19.69% for the high-level labels, and 10.30% for the expanded labels.

By using a white-list and keyword-based classifier, we obtain all the URLs that correspond to news outlets. Among these, the news articles that are labeled as "Purely Factual" are not taken into account in the following analyses because they merely report news about the incident without discussing gun-related issues and policies, and do not serve the purpose of our study on exploring how users access information in reaction to a news event (versus how they are informed about the event). Although one can argue that some news sites are representative of specific ideological views, we do not rely on the latter, because often the political orientation differs from the orientation to the gun control issue [2].

4. DIVERSITY OF DOMAINS

Our first study seeks to characterize websites with respect to the diversity of opinions they present. Our findings help us label the large corpus we extracted, but also have broader implications on search: for example, they may be useful when considering how to rank search results to ensure that diversity is present. To evaluate the diversity of web domains, we use an information-theoretic measure, Shannon's entropy.

We identify domains with at least eight labeled webpages and give their label distribution in Fig. 3. It is evident that most of the web domains are one-sided, with almost all their webpages expressing similar opinions (e.g., supporting only gun rights). An exception to this finding is that user-generated content, such as that found on `wikipedia.org` and `answers.yahoo.com`, tends to be either balanced or diverse respectively.

To quantify the heterogeneity of the available information per domain in a principled way, we use Shannon's entropy [31], an information-theoretic measure of the uncertainty for a random variable. The higher the entropy associated with a random variable, the higher the uncertainty about its value, or, equivalently, the more diverse it is. Formally, for each domain d with entropy

$$H(X_d) = E[-\log P(X_d)] = -\sum_i P(X_d = x_i) * \log P(X_d = x_i),$$

we compute the *normalized entropy* for its webpage labels:

$$H^{norm}(X_d) = H_t(X_d)/H_t(X'_d|X'_d \sim \mathcal{U}),$$

where X_d, X'_d are the labels of the webpages with domain d, X'_d is uniformly distributed, and log is the base-2 logarithm. We note that $H_t(X'_d|X'_d \sim \mathcal{U})$ corresponds to the maximum entropy where the labels occur with equal probability.

We compute the *normalized* entropy for the labels of the URLs instead of the entropy for two reasons: (1) the normalized entropy handles comparisons across different event space sizes, which is needed when comparing high-level and expanded labels and (2) the normalized entropy ensures that comparisons between domains with different number of observations are at the same basis. Normalizing the measure helps to handle estimation error, as the entropy can have high variance when there are only a few observations.

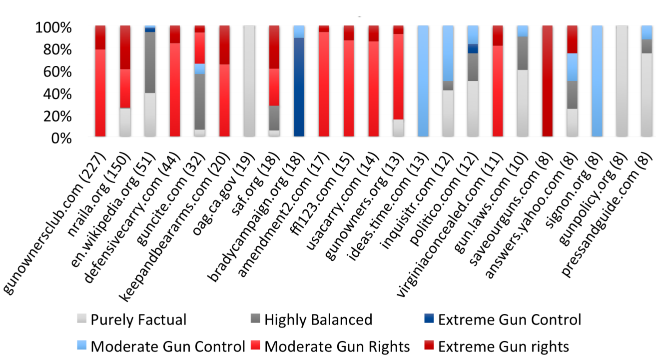

Figure 3: Label distribution per domain. The domains are in decreasing order of manually characterized URLs (in parentheses).

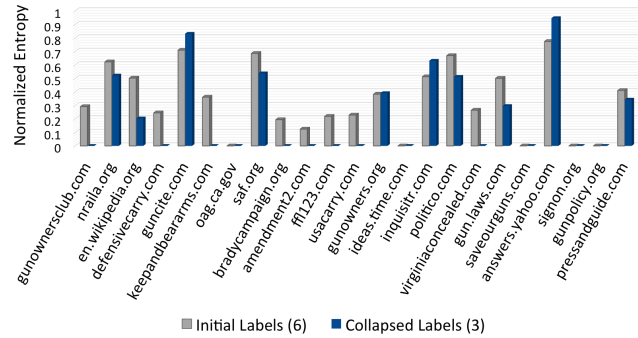

Figure 4: GC-DEBATE: Diversity of domains in terms of label entropy (for the manually labeled URLs).

Figure 4 depicts the normalized entropy in the labels of the webpages per domain, where we consider the 2,100 webpages that were manually labeled by expert assessors (Sec. 3.2). For each domain, the left and right bars correspond to the normalized label entropy for the expanded, and the high-level labels respectively. Overall, for the high-level labels, the normalized entropy is 0 (no diversity) for 54% of the domains, and smaller than or equal to 0.5 for 63% of the domains. The median normalized entropy is 0, and the mean 0.27. Similarly, for the expanded set of labels, 34% of the domains have entropy 0, and 73% have normalized entropy smaller than or equal to 0.5. The median and mean normalized entropy are 0.36 and 0.30 respectively.

The main finding is that the domains offer to the users mostly a single myopic view on gun control issues. Based on this observation, we were able to automatically label the remaining ~$23K$ webpages that were not labeled manually by the assessors, and obtain a rich, annotated dataset that can serve the purposes of our next analyses. We note that among those 23K webpages that we label automatically, 10,525 AND 4,398 URLs belong to the gun rights forums `gunownersclub.com` and `defensivecarry.com` respectively, while 4,221 URLs belong to the gun control petition page `signon.org`. That is, 82% of the webpages that we label automatically correspond to three domains with very clear stances. For the automatic labeling, we apply a label propagation approach from the webpages to their domains:

- **Forums.** We replace URLs that belong to a forum with its main page, and classify the latter based on the overall stance of its labeled webpages, (*i.e.*, the dominant category of the manual labeling).
- **Advocacy groups.** We label each domain based on the identified stance using Wikipedia's characterization.
- **Domains.** For the domains with normalized entropy smaller than 0.5, we first assign the dominant high-level category, and then the stance (moderate, extreme) of the majority of the labels. If we have a tie among the possible categories, we do not classify the domain, and keep the initial URLs and their labels for our analysis.

By following these rules, we obtain the final labeling of the *domains*, as well as the remaining URLs whose domain's stance could not be summarized succinctly by a single label. The distribution of the final labels is: 4% purely factual, 2% highly balanced, 58% and 16% moderate and extreme gun rights respectively, and 18% and 2% of moderate and extreme gun control.

Overall, this study indicates sites with user-generated content, such as Wikipedia and Q&A websites, are more diverse. In contrast, forums about controversial topics tend to be very polarized.

5. WITHIN-USER DIVERSITY

Our second study focuses on the diversity of information consumed by each user browsing controversial topics, and how the diversity in the information sought is influenced by a shocking news event. The within-user diversity can be expressed in terms of the number of different domains that a user browses, as well as the number of different *categories* (e.g., gun control, balanced webpages) of pages that she visits.

5.1 Examining the Existing Theories

We start by evaluating whether users desire diversity in opinions. As in Sec. 4, we use Shannon's entropy to quantify the diversity in the categories of webpages that each user visits. We note that this study may indicate whether recommendation systems and search results should be composed of diverse opinions in order to satisfy the user.

In the prior literature we find two contradictory theories, which we consider regarding the implications of using entropy to capture variance:

THEORY 1. "People use the web to access a variety of sources, and become more aware of political news and events." [18, 34].
IMPLICATION 1. If this assertion is true, we would expect users to visit domains with different labels regarding perspective, and that the associated normalized label entropy of the domains that the users browse would be high. In the case where users visit only a few domains, we would expect the label entropy of the domains to be high.

THEORY 2. "People use the web to access mostly agreeable political information" [5, 16, 14].
IMPLICATION 2. If this assessment is true, we would expect most users to access mostly domains supporting one side of the topic. Thus, the label entropy of the domains that the users access should be low.

We now analyze the diversity of information that users consume to explore the two assertions. We first focus on all users who visited at least three relevant domains during November and December.

From all the users, only 5% visited at least 4 relevant domains and news articles; the vast majority of these users, 50%, accessed exactly 4 domains, 24% browsed 5 domains, and 11% visited 6 domains. This observation, in combination with the low diversity of domains (Sec. 4), provide evidence that Theory 1 is unlikely. Most users appear too "narrow-minded" as far as the number of web domains is concerned, and the domains themselves are mostly one-sided.

To evaluate Theory 2, we need to examine the diversity of each user's consumed information by computing the entropy in the labels of the domains accessed. The intuition behind this need is that the number of domains does not provide enough information about the diversity of a user's exposure, as it does not fully characterize the type of information consumed. Two extreme cases would be a user who visited three websites supporting gun control, and another user who visited a website of each category: gun control, gun rights and balanced. Clearly, the second user's information stream is more diverse. Thus, for each user u who visited at least three different relevant domains during November and December, we compute a normalized label entropy

$$H^{norm}(X_u) = H(X_u)/H(X'_u|X'_u \sim \mathcal{U}) \quad (1)$$

where X_u, X'_u are the labels of the domains visited by user u (thus, X_u, X'_u take values in $\{$gun control, gun rights, balanced/factual$\}$), and X'_u is uniformly distributed rendering $H(X'_u|X'_u \sim \mathcal{U})$ the maximum entropy.

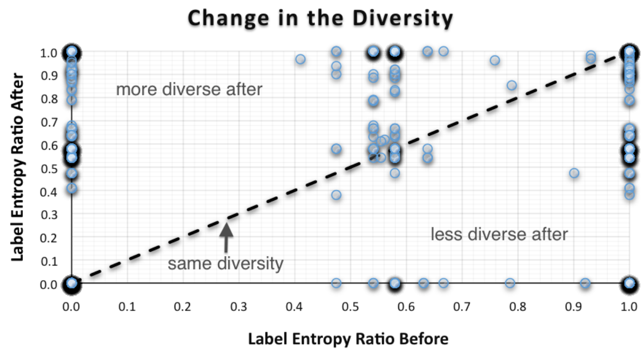

Figure 5: Change in the user diversity after S.H. for users who visited at least two different domains both before and after the event. Every point in the plot corresponds to a user.

Among the 5% of the users who accessed at least 4 domains, the average normalized entropy is 0.70, and the median is 0.66. Hence, from the users who access a considerable number of domains, the majority neither access webpages of a single stance (H^{norm} should be 0), nor websites of all possible labels (H^{norm} should be 1).

All in all, 5% of the users appears to seek diversity (Theory 1), while the vast majority seem to primarily get their information on a topic from one source and therefore are only exposed to that view.

5.2 Event: Change in Within-User Diversity

Although the diversity of information consumed by each user is low overall, we seek to understand if a *shocking event* with nationwide impact influences *a change* in the diversity of the information that users access. For this study, we use Shannon's entropy and assess the *change* in the label entropy of the information sought before and after the event. As in Sec. 5.1, this study may imply whether recommendation systems should adjust the diversity of their results after a shocking event to keep the user satisfied.

For a fair comparison of accesses before and after the shocking news event, we need to control for the users. Thus, we focus on the common users, *i*.e., those who were active *both before and after* Sandy Hook (Table 2), as these are the only users whom we can characterize before and after the event. For each user u who visited at least two domains both before and after Sandy Hook, we compute per time period the normalized label entropy, $H^{norm}(X_u)$, as defined in Eq. (1). Then, to quantify the change in the within-user diversity, we compute the difference between the normalized entropies

$$H^{norm}_{AFTER}(X_u) - H^{norm}_{BEFORE}(X_u).$$

Next, we consider two cases for the data: (1) *all* the websites, and (2) all the webpages *except for* Feinstein's page about the new gun control legislation (gray spike on Dec. 27 in Fig. 1), which was heavily visited the day of its creation and, thus, corresponds to an "outlier" webpage. The reason for this two-fold analysis is to understand how much of the user diversity is attributed to a single page, which attracted the interest of many users the day of its creation.

With Feinstein's website. Figure 5 shows the change in the user diversity after the Sandy Hook shootings. On average, the normalized entropy increased by 12.54%. However, the results vary across users: for 30.20% of the users, the diversity remained the same; for 41.40% of the users, the entropy increased by 65.02%, and for 28.76% it decreased by 49.18%.

To understand what drove so many users to this site, we investigated more deeply. On Dec. 27, 2012, the Drudge Report, a primarily conservative link aggregation site, featured a gun rights re-

Table 4: Markov chain states for browsing history: Abbreviations for the high-level and expanded labels.

High-level States (3)		Expanded States (6)	
GC	Gun Control	**EC**	Extreme Gun Control
		MC	Moderate Gun Control
BF	Balanced/Factual	**HB**	Highly Balanced
		PF	Purely Factual
GR	Gun Rights	**MR**	Moderate Gun Rights
		ER	Extreme Gun Rights

action grabbing headline "Senate to Go For Handguns"[2]. Additionally, a summary underneath mentioned key sensitive points for gun rights supporters (gun registries and owner fingerprinting). 84% of the users who reached the Feinstein list came via the Drudge Report front page – nearly all of them had primarily consumed gun rights information prior to this. The implication is that when users consume information outside of their typical viewpoint the link comes from *within* their own sphere. Moreover, it suggests future research to redesign contextual link recommendation – which displays related-content to a page – may have a greater potential for changing the diversity of information a user browses than the composition of search results.

Without Feinstein's website. By repeating the analysis described above after removing the website which was heavily visited when the gun ban list was announced, we find that, on average, the normalized entropy of the users increased by 11.92%. That is, the increase is slightly smaller than in the previous analysis where we consider the outlier webpage. Specifically, for 32.87% of the users, the diversity remained the same; for 39.38% of the users, the entropy increased by 49.07%, and for 27.75% it decreased by 52.13%.

Based on our study, on average users do slightly increase the diversity of opinions they consume and peek outside of their bubble after a *shocking* event. One large motivator seems to be the the uncertainty and potential for change, which prompts speculatory reading on that change (*e.g.*, Feinstein's page).

6. WEB TRANSITIONS

Our third study addresses the way users navigate polarizing topics. Specifically, we seek insights about how the stance of content on the current website influences the type of webpage that users browse to next, when exploring polarizing political topics. What webpages are users more likely to visit after browsing a site supporting extreme gun control or rights? Although most of the users are not diverse in terms of the label entropy of the domains they visit (Sec. 5), many of them *do* transition between pages supporting opposing views. We seek to understand the most common transitions, as well as possible changes in the transitions due to the news on the Newtown shootings. By focusing on the influences of the stance of the current page on transitions to next pages, we obtain a micro-level view of information consumption patterns in distinction to the user-level view presented in Section 5.

For each user we represent her browsing history as a Markov chain with either the high-level or expanded labels as states X_i (Table 4). Then, we describe the distribution of the transition probabilities by an n-state transition matrix $\mathbf{P}^n$, with elements $p_{ij} = Prob(X_{t+1} = j | X_t = i)$. We note that the row-wise sums are equal to 1, $\sum_j p_{ij} = 1$. To make sense of the underlying trends of this matrix, we employ mobility indices that have been widely used in economics and sociology (e.g., credit mobility [19], social status mobility [9]).

[2]See `http://www.drudgereportarchives.com/data/2012/12/27/20121227_160126.htm` for an archive.

Table 5: *All* users: 6-state transition matrix $\mathbf{P}^6$ for November-December.

	EC	MC	HB	PF	MR	ER
EC	13.58%	18.57%	31.60%	12.45%	8.27%	15.53%
MC	10.67%	18.54%	33.06%	9.41%	7.70%	20.62%
HB	9.33%	15.10%	28.41%	16.94%	10.15%	20.08%
PF	6.06%	10.26%	34.46%	9.70%	12.55%	26.97%
MR	5.57%	9.39%	24.59%	13.00%	19.27%	28.17%
ER	5.18%	8.81%	19.40%	8.95%	12.08%	45.58%

Table 6: *All* users: 3-state transition matrix $\mathbf{P}^3$ for November-December.

	GC	BF	GR
GC	30.48%	43.15%	26.37%
BF	22.88%	45.12%	31.99%
GR	14.31%	31.34%	54.36%

6.1 Overall Transition Patterns

We start by studying whether the polarity of a web page can be predictive of the polarity of the next on-topic site that a user will browse, which might have implications for contextual search and recommendation. To that end, we use a specific type of mobility indices, called *Summary Mobility Indices*.

We consider the ~35,000 transitions of *all* users during November and December. We note that these are the users who visited at least two different domains, and, hence, we record for them at least one transition. The transitions are given in the form of a transition matrix in Tables 5 and 6.

We employ the Summary Mobility Indices, which describe the direction of the mobility:

- Immobility Ratio: $IR = \sum_{i=1}^{n} p_{ii}/n$
- Moving Up: $MU = \sum_{i<j} p_{ij}/n$
- Moving Down: $MD = \sum_{i>j} p_{ij}/n$,

where n is the number of states. The indices take values in $[0, 1]$. The immobility ratio represents the percent of same-state transitions (higher for more *im*mobility), while the other two indices give the percent of transitions from one extreme to the other, *i.e.*, the MU index captures the transitions from extreme gun control towards extreme gun rights, and the MD index describes the opposite directionality. The higher the MU and MD indices are, the more mobility is observed in the system.

The Summary Mobility Indices for all users during November and December are: (a) for the high-level states $IR = 0.4332$, $MU = 0.3384$, and $MD = 0.2284$, and (b) for the extended states $IR = 0.2251$, $MU = 0.4535$, and $MD = 0.3214$. Firstly, we observe that the overall system is characterized by mobility ($IR \ll 1$). Specifically, for the extended states, about 23% of the transitions are same state (and, thus, predictive), and 45% of the transitions occur in the direction from extreme gun control towards extreme gun rights. From the transitions in the opposite direction, the most dominant transitions are towards the "middle" states: from factual to balanced webpages (34.46%), from extreme gun rights to balanced pages (19.40%), and from moderate gun rights to balanced pages (24.59%).

6.2 Event: Change in Transition Patterns

We continue by studying whether a shocking event *changes* the predictivity of the polarity of the next on-topic domain conditioned on the current one – and, thus, implies changes in contextual search and recommendation. To evaluate the change, we restrict our analysis to the common users, and create two transition matrices, $\mathbf{P}_{before}$ and $\mathbf{P}_{after}$ (Table 7). For our analysis we employ both mobility indices and matrix distances.

Throughout our studies, we considered two different versions of our dataset: (i) GC-DEBATE: which contains all of the on-topic URLs that users visited during the two-month period that we are studying, and (ii) NONNEWS: contains the on-topic URLs excluding the news articles, which we separated from the rest URLs using a white list and keyword-based classifier (Sec. 3.2). The reason for this separation is to understand whether there are differences in the behaviors of the users when we consider different types of engagement; short-term interests in the topic (GC-DEBATE) vs. only long-term interests beyond news articles (NONNEWS). All the results we have presented up to this point refer to the dataset GC-DEBATE, because the results were very similar in NONNEWS and the conclusions were consistent. However, in this study, we observed that there are differences in the transitions of the users depending on the level of their engagement. Therefore, we present the results for both cases, and discuss the differences. Unless we explicitly state that the results consider the NONNEWS URLs, we refer to all the URLs (GC-DEBATE).

We start with the Summary Mobility Indices, as well as the eigenvalue-based indices [27, 32, 33, 19, 21] that quantify the amount of mobility in the system. This category includes the eigenvalue M_E, second eigenvalue M_2, determinant M_D indices. A value of 0 means to total immobility, and a value of 1 to perfect mobility.

The first observation on Table 9 is that, in all cases, the immobility ratio (IR) decreases after Sandy Hook signifying that users transition between different states more often after than before the event. The same can also be drawn from the eigenvalue-based indices, all of which increase. The direction of the transitions depends on the type of interests that we focus on. When we consider the transient interests (GC-DEBATE), the transitions towards extreme gun rights (MU) decrease, while the transitions towards extreme gun control (MD) increase. On the other hand, when we consider the long-term interests (NONNEWS), the system moves mainly towards extreme gun rights.

The indices described above are used to assess the underlying mobility behaviors in an *individual* transition matrix $\mathbf{P}^n$, but *not* the similarities between different transition matrices. To compute the latter, we need to introduce the notion of comparison between matrices. The first step towards this goal is to have both matrices at the same base, which is achieved by computing their deviation from a perfectly immobile system described by the identity matrix $\mathbf{I}$. Among the matrix distances in the literature, we use the L1-norm and L2-norm:

$$||\mathbf{P}-\mathbf{Q}||_1 = \sum_i \sum_j (p_{ij} - q_{ij})$$

$$||\mathbf{P}-\mathbf{Q}||_2 = \sqrt{\sum_i \sum_j (p_{ij} - q_{ij})^2},$$

where $\mathbf{P} = \mathbf{I}$ and $\mathbf{Q} = \mathbf{P}_{before}$ or $\mathbf{P}_{after}$. In addition to these standard matrix distances, we also use two "risk"-adjusted difference indices, D_1 and D_3, which have the advantage of comparing the two transition matrices directly and also detecting the direction of the transition –while weighing proportionally "close" and "far" transitions by the factor $(i-j)$:

$$D_1(\mathbf{P},\mathbf{Q}) = \sum_i \sum_j (i-j)(p_{ij} - q_{ij})$$

$$D_3(\mathbf{P},\mathbf{Q}) = \sum_i \sum_j (i-j) sign(p_{ij} - q_{ij})(p_{ij} - q_{ij})^2,$$

where $\mathbf{P} = \mathbf{P}_{after}$ and $\mathbf{Q} = \mathbf{P}_{before}$. The distances are given in Table 10. The L_1 and L_2 distances increase after Sandy Hook, showing that the users transition between different states more often after than before the event. As far as the D_1 and D_3 distances are concerned, we observe an interesting pattern: the two versions of data yield distances of different signs. The negative values in the NONNEWS data suggest that the long-term interests move from extreme gun control towards extreme gun rights after the event. On the other hand, the positive values in the GC-DEBATE data show that the transient interests due to the news about the event tend toward balanced and gun control stances. These conclusions are also corroborated by the MU and MD indices.

All in all, the event leads to more mobility in the system, and the way it affects the latter depends on the type of user interest (long-term vs. short term).

Table 7: *Common* users: 6-state transition matrices $\mathbf{P}^6_{before}$ (top) and $\mathbf{P}^6_{after}$ (bottom).

	EC	MC	HB	PF	MR	ER
EC	28.35%	0.00%	21.13%	3.09%	30.93%	16.49%
MC	8.33%	0.00%	16.67%	4.17%	29.17%	41.67%
HB	12.53%	1.36%	28.07%	7.63%	25.61%	24.80%
PF	12.75%	1.96%	30.39%	3.92%	29.41%	21.57%
MR	8.33%	1.00%	12.83%	6.83%	40.33%	30.67%
ER	7.61%	0.60%	10.91%	2.70%	22.42%	55.76%

	EC	MC	HB	PF	MR	ER
EC	11.71%	20.42%	32.28%	9.76%	8.86%	16.97%
MC	9.06%	17.21%	28.91%	9.21%	11.17%	24.45%
HB	7.58%	16.00%	23.84%	17.56%	11.38%	23.64%
PF	4.61%	10.22%	26.45%	8.02%	13.83%	36.87%
MR	5.03%	11.95%	23.40%	15.72%	12.58%	31.32%
ER	4.04%	9.12%	17.41%	10.43%	12.06%	46.94%

Table 8: *Common* users: 3-state transition matrices $\mathbf{P}^3_{before}$ and $\mathbf{P}^3_{after}$.

$\mathbf{P}^3_{before}$				$\mathbf{P}^3_{after}$			
	GC	**BF**	**GR**		**GC**	**BF**	**GR**
GC	26.15%	23.85%	50.00%	**GC**	28.23%	39.43%	32.35%
BF	14.07%	35.39%	50.53%	**BF**	21.78%	39.98%	38.24%
GR	8.63%	15.88%	75.48%	**GR**	14.17%	30.83%	55.00%

Table 9: *Common* users: Summary Mobility and Eigenvalue-based Indices for $\mathbf{P}^3_{before}$ and $\mathbf{P}^3_{after}$. The top (bottom) rows correspond to transitions between the 3 high-level (6 expanded) states. The rows annotated by NONNEWS refer to the version of the data that excludes news articles, while the rest refer to the whole dataset, GC-DEBATE.

	IR	MU	MD	$\mathbf{M}_E$	$\mathbf{M}_2$	$\mathbf{M}_D$
Before-3	0.4567	0.4146	0.1286	0.8149	0.7483	0.9702
After-3	0.4107	0.3667	0.2226	0.8840	0.7994	0.9937
Before-3 NONNEWS	0.5342	0.3772	0.0886	0.6987	0.5780	0.9238
After-3 NONNEWS	0.4405	0.4362	0.1233	0.8393	0.7670	0.9794
Before-6	0.2607	0.5050	0.2342	0.8871	0.6868	<1
After-6	0.2005	0.4944	0.3051	0.9594	0.7642	1
Before-6 NONNEWS	0.3118	0.4988	0.1894	0.8259	0.5051	<1
After-6 NONNEWS	0.2197	0.5713	0.2091	0.9364	0.7042	1

Table 10: *Common* users: Distances of transition matrices from the immobility matrix $\mathbf{I}$. The top (bottom) rows correspond to transitions between the 3 high-level (6 expanded) states. The rows annotated by NONNEWS refer to the version of the data that excludes news articles, while the rest refer to the whole dataset, GC-DEBATE, introduced in Sec. 3.

	$\mathbf{L}_1$	$\mathbf{L}_2$	$\mathbf{D}_1$	$\mathbf{D}_3$
Before-3	3.2594	1.2797	0.6575	0.0876
After-3	3.5359	1.2833		
Before-3 NONNEWS	2.7947	1.1386	-0.1546	-0.0202
After-3 NONNEWS	3.3571	1.3138		
Before-6	8.8713	2.1333	2.1419	0.4262
After-6	9.5941	2.2204		
Before-6 NONNEWS	8.2586	2.0464	-0.8122	-0.1632
After-6 NONNEWS	9.3639	2.2484		

7. TIME-EVOLVING COMMUNITIES

In our last study we seek to understand whether a shocking news event permanently shifts the user's topical view, as the answer may be useful in conditioning personalization. To do that, we employ community analysis: we characterize the users based on their search and browsing patterns, and track the temporal evolution of their communities.

We introduce an approach that assigns to each user a score reflecting her stance for each time period, depending on the stance of the sites she visited. Let $X_{u,t}$ be the set of labels x of the webpages that user u visited during time period t, and $\mathcal{L}$ be the set of unique labels. Algorithm 1 finds the communities to which each user belongs at time t. For example, according to the algorithm, a user who visited 2 balanced, and 5 gun rights webpages before S.H., belongs to the gun rights community.

ALGORITHM 1: User Characterization and Assignment to Communities.

INPUT: set of labels $\mathcal{L}$
time period t
set of labels, X_t, of visited webpages
label function L

// 1: `define the weight` w `for each label` $l \in \mathcal{L}$

$$w(l) = \begin{cases} -1 & \text{if } l = \text{gun rights} \\ 0 & \text{if } l = \text{balanced} \\ +1 & \text{if } l = \text{gun control} \end{cases}$$

for each user u **do**
for each label $l \in \mathcal{L}$ **do**
// 2: `compute the number of webpages of label` l `browsed by user` u
$n_{u,t}(l) = \sum_{x \in X_{u,t}} \mathbb{1}_{x=l}$
end for
// 3: compute user's u score for time period t
$s_t(u) = \sum_{l \in \mathcal{L}} w(l) \cdot n_{u,t}(l)$
// 4: find user u's community for time period t

$$c_t(u) = \begin{cases} \text{gun rights} & \text{if } sign(s_t(u)) < 0 \\ \text{balanced} & \text{if } sign(s_t(u)) = 0 \\ \text{gun control} & \text{if } sign(s_t(u)) > 0 \end{cases}$$

end for
return the vector of user communities $\mathbf{c}_t$

To obtain the communities of the users who visited at least two on-topic webpages both before and after the event, we applied Algorithm 1 for $\mathcal{L}$ ={gun control, gun rights, balanced/factual} and t equal to the time period before and after Sandy Hook. As in Section 6.2, we analyze both GC-DEBATE and the NONNEWS data, because they exhibit differences. Our baseline is GC-DEBATE—we explicitly mention when we refer to NONNEWS.

First, we analyze the dynamics of the communities that we found. Table 11 shows that 76% and 75% of the users supporting gun control and gun rights respectively, stand firm after the event. The users who were part of the balanced community before the event split almost evenly among the three communities after Sandy Hook. This observation about the movement of users across communities naturally leads us to the question: Do the communities change size after the event? As shown in Fig. 6, the gun control community has almost 10% more users after S.H., while the gun rights community lost the same percent of users. The balanced community increased in size slightly, by about 2%.

If we focus on the users with long-term interest in the topic (NONNEWS), we observe that more of them remain loyal to their communities, consisting 79% and 91% of the gun control and gun rights community respectively. Moreover, the gun control and rights communities shrink by about 1%, while the balanced community

Table 11: Transition matrix capturing the mobility between communities before (rows) and after (columns) Sandy Hook.

	GC	BF	GR
GC	76.32%	14.21%	9.47%
BF	29.94%	32.20%	37.85%
GR	12.25%	13.19%	74.56%

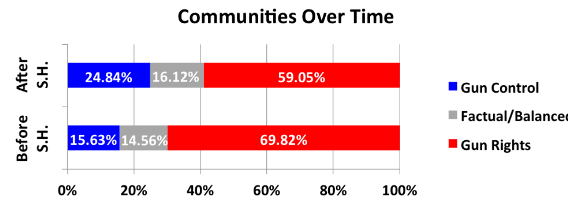

Figure 6: The gun control community increases in size, while the gun rights community shrinks after S.H.. For each community we give its size before and after S.H. in terms of the percentage of users belonging to them.

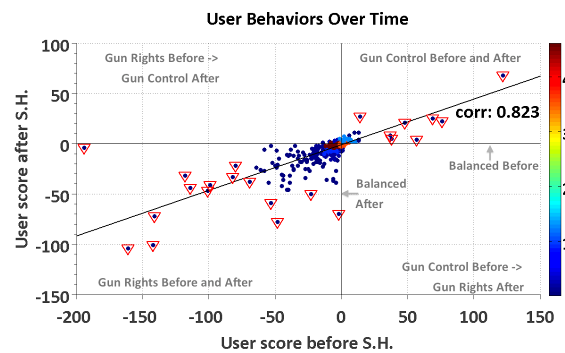

Figure 7: Heatmap of user scores before vs. user scores after Sandy Hook. Each point corresponds to one user. The points withing red triangles correspond to anomalous users (spotted by DBSCAN, a density-based spatial clustering method) with significantly different behavior from the rest.

expands by 2%. Therefore, the communities consisting of users with long-term interests are less prone to changes than users with transient interest in the topic.

We go beyond the aggregate user transitions and the community dynamics, and study the users' individual transitions to understand if and how they changed after the event. Figure 7 shows the score of each user, $s_t(u)$ (in Algorithm 1), before (x-axis) and after (y-axis) S.H. The bigger the absolute value of a score is, the more active and devoted is the user to the corresponding community. While most of the users have small or moderate activity, there are some very heavy users who are faithful supporters of their communities (either gun rights or gun control). Among them, we found some outlier users (noted in red triangles), using a density-based spatial clustering method, DBSCAN [12]. The outlier users that remain in the gun rights community after S.H. (third quadrant) are mostly members of some gun rights forums, including `gunownersclub.com` and `defensivecarry.com`. On the other hand, most of the outlier users who remained in the gun control community after the shootings, visited several times the webpage `childrensdefense.org`. About one third of the total users crossed communities after Sandy Hook (points in the second and fourth quadrant in Fig. 7). From the users with long-term interest in the topic, only 15% crossed communities, supporting our argument that they are less prone to changes than users with transient interests. In the data, there is no evidence for significant difference between the cross-over and loyal users with respect to their news consumption.

To sum up, it appears that the event affected people's browsing behavior: about 88% of the total users (and 61% of the NONNEWS users) browsed more balanced and factual webpages discussing gun control issues after the event, which indicates that the Sandy Hook shootings sparked their interest in the topic. As for the size of the communities, the change depends on the type of interest (long term vs. short term), as mentioned earlier.

8. LIMITATIONS

To understand the information-seeking behaviors of people about a controversial topic such as gun control, as well as the effect of an external, shocking event, we focused on the search and browsing patterns of 29 million users of Internet Explorer (IE) in the US-English market. The first limitation of our study is that we consider users of only one browser. However, during the time period that we studied (November-December 2012), the majority of users worldwide (54%) were using IE to browse the web [4]. Moreover, independent studies [3] show that the user population of IE is a broad representation of the US population covering both genders, and nearly all age and income levels. The second limitation of our study arises from the inherent differences in the browsing behavior of gun control and rights supporters. Before the Sandy Hook shootings, the observed activity comes mainly from users who support gun rights and are members of related forums. To compensate for the difference in browsing activity, we draw conclusions both for: a) the system as a whole, which includes users who have transient interests in the topic, and become active after the event, and b) the users who are active during the whole time period (mainly gun rights supporters), and have long-term interest in the topic.

9. CONCLUSIONS

We have examined the browsing behavior of searchers for the controversial and polarizing topic of gun control. We focused on the influence of a disruptive and shocking event involving the tragic mass shooting at the Sandy Hook Elementary School in December 2012. By starting from a large corpus of web logs from November and December 2012, we extracted a footprint of user information-seeking behavior on the URLs that are germane to the topic, and followed a multi-step labeling procedure. Our key findings are as follows:

- We find that people use the web to largely access information they agree with, as signified by the low diversity of labels capturing viewpoints expressed in visited domains.
- Domains provide a myopic view on the polarizing topic, showing low diversity in stances that are presented.
- When the external event threatens to influence users directly, they explore content outside their filter bubble.
- The overall system, including transient interests, largely moves mainly towards extreme gun control. However, the long-term interests in the topic, which are captured by visits to non-news webpages, tend to move towards the opposite direction and support extreme gun rights.
- The gun control and balanced communities grow after S.H., while the gun rights community shrinks.

We believe that the methods and results presented are a step toward leveraging log data to better understand how people navigate webpages on controversial topics. Future directions include devising studies exploring whether and how ranking and presentation procedures that expose users to a greater diversity of viewpoints can lead to increased user satisfaction. Other directions include predicting the changes in polarity in the information accessed via search and retrieval systems.

Acknowledgements. Thanks to Rebecca Hanson for her extensive support in labeling web pages.

APPENDIX

A. CATEGORIES OF WEBPAGES

We seek to label every page that is not “Off-Topic” or “Not Accessible”. Thus, we define symmetric and objective categories:

A. Purely Factual: The page is on-topic, but only presents facts with no obvious interpretation or commentary on politics. This may include pages that give statistics regarding guns, laws in different locales about guns, or reporting on news events involving guns without additional commentary.

B. Discusses Policies and Issues: The page is on-topic and discusses gun policies and issues regarding legislation on gun ownership and usage, or ethical and historical justifications for gun control/rights. This includes pages that discuss how laws have been interpreted for application in court cases, as well as the personal/official pages of politicians, other persons, organizations, and entities whose stance on gun-related policy is well known even if the page does not feature content currently discussing the policy. The pages in this category are further classified into:

Extreme Gun Control: These pages present a view which favors extreme changes to the current gun laws in an area. This includes viewpoints that support laws banning any private citizen ownership of guns, as well as what would be viewed as major legislation changes relative to a locale that are not as sweeping. Pages that use derogatory and insulting language toward those supporting gun rights belong to this category. Discussion forums and blogs where most comments support this view, and webpages giving contact information about only anti-gun organizations belong here.

Moderate Gun Control: These pages present views that favor some to moderate changes to the current gun laws in an area. This includes views that may view private citizen ownership of guns as acceptable with appropriate conditions and limitations, but argue that the current laws are not sufficient in defining these conditions and limitations. Discussion forums and blogs where the preponderance of comments support this view fall in this category.

Highly Balanced: These pages either discuss both sides with no obvious bias, or present a straightforward discussion of how laws and policy have been interpreted in the past. For example, pages that discuss of court case reasoning involving guns would fall into this category. Likewise, educational texts that appear to fairly present both sides would also belong here.

Moderate Gun Rights: These pages present a view which favors little to no changes to the current gun laws in an area. This includes viewpoints that generally support private citizen ownership of guns with appropriate conditions and limitations, and argue that the current laws are generally sufficient. This includes pages selling guns that likely would be viewed acceptable for private ownership under appropriate limitations by a moderate gun control viewpoint. Discussion forums and blogs where the preponderance of comments support this view belong here.

Extreme Gun Rights: These pages present a view which favors no changes to the current gun laws in an area and argue that current laws may be overly restrictive and intrusive. This includes viewpoints that claim current laws are an intrusion on individual rights and argue for lessening of any current gun control policies. Pages that use derogatory and insulting language toward those supporting gun control are included in this category. This includes pages selling guns or providing information on guns limited not only to those guns viewed acceptable for private ownership under appropriate limitations by a moderate gun control viewpoint but also those falling under currently debated or proposed legislative control. Discussion forums and blogs where the preponderance of comments support this view, and webpages giving contact information about only pro-gun organizations belong to this category.

B. REFERENCES

[1] Sandy Hook elementary school shooting. `http://en.wikipedia.org/wiki/Sandy_Hook_Elementary_School_shooting`. Retrieved 6/10/14.

[2] Gun Ownership and Use in America. `http://www.gallup.com/poll/20098/gun-ownership-use-america.aspx`, 2005. Retrieved 7/10/14.

[3] Firefox vs. Internet Explorer. `http://www.comscore.com/Insights/Blog/Firefox-vs.-Internet-Explorer,Firefox-vs-Internet-Explorer-Part-2`, 2007. Retrieved 11/6/14.

[4] Usage share of web browsers. `http://thenextweb.com/apps/2012/10/01/internet-explorer-8-falls-25-market-share-firefox-15-passes-10-mark-chrome-loses-users`, 2012. Retrieved 3/7/15.

[5] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 U.S. election: divided they blog. In *LinkKDD*, pages 36–43. ACM, 2005.

[6] Elif Aktolga and James Allan. Sentiment diversification with different biases. In *SIGIR*, pages 593–602. ACM, 2013.

[7] Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. Harmony and dissonance: organizing the people's voices on political controversies. In *WSDM*, pages 523–532. ACM, 2012.

[8] Ramnath Balasubramanyan, William W. Cohen, Doug Pierce, and David P. Redlawsk. Modeling polarizing topics: When do different political communities respond differently to the same news? In *AAAI ICWSM*, 2012.

[9] P.M. Blau. *The American occupational structure*. Wiley, 1967.

[10] Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise van Raaij. Network analysis of collaboration structure in wikipedia. In *WWW*, pages 731–740, 2009.

[11] Nick Craswell and Martin Szummer. Random walks on the click graph. In *ACM SIGIR*, pages 239–246, 2007.

[12] Martin Ester, Hans peter Kriegel, Jörg S, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[13] Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. Mining contrastive opinions on political texts using cross-perspective topic model. In *WSDM*, 2012.

[14] R Kelly Garrett. Politically motivated reinforcement seeking: Reframing the selective exposure debate. *Int. J. Comm.*, 59(4):676–699, 2009.

[15] R Kelly Garrett and Paul Resnick. Resisting political fragmentation on the internet. *Daedalus*, 140(4):108–120, 2011.

[16] Eric Gilbert, Tony Bergstrom, and Karrie Karahalios. Blogs are echo chambers: Blogs are echo chambers. In *HICSS*, pages 1–10, 2009.

[17] Kilem Li Gwet. *Handbook of Inter-Rater Reliability (3rd Edition): The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters.* Advanced Analytics Press, 3 edition.

[18] J.B. Horrigan, K. Garrett, and P. Resnick. *The Internet and Democratic Debate*. PIALP, 2004.

[19] Yusuf Jafry and Til Schuermann. Measurement, estimation and comparison of credit migration matrices. *J. Bank. Financ.*, 28(11):2603–2639, November 2004.

[20] Alexander Kotov, Paul N. Bennett, Ryen W. White, Susan T. Dumais, and Jaime Teevan. Modeling and analysis of cross-session search tasks. In *ACM SIGIR*, pages 5–14, 2011.

[21] Vasileios Koutras and Konstantinos Drakos. A migration approach for USA banks' capitalization: Are the 00s the same with the 90s? *IRFA*, 30(0):131 – 140, 2013.

[22] Jingjing Liu, Nicholas J. Belkin, Xiangmin Zhang, and Xiao-Jun Yuan. Examining users' knowledge change in the task completion process. *Inf. Process. Manage.*, 49(5):1058–1074, 2013.

[23] R. Malouf and T. Mullen. Taking Sides: User Classification for Informal Online Political Discourse. *Internet Res.*, 18(2):177–190, 2008.

[24] Sean A. Munson, Stephanie Y. Lee, and Paul Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *AAAI ICWSM*, 2013.

[25] Sean A. Munson and Paul Resnick. Presenting diverse political opinions: how and how much. In *CHI*, pages 1457–1466, 2010.

[26] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. The Penguin Group, 2011.

[27] S. J. Prais. Measuring Social Mobility. *J. Roy. Statist. Soc. Ser. A*, 118(1):56–66, 1955.

[28] Justus J. Randolph. Free-Marginal Multirater Kappa (multirater K[free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. In *Joensuu Univ. Learn. Instruct. Symp.*, 2005.

[29] Paul Resnick, R. Kelly Garrett, Travis Kriplean, Sean A. Munson, and Natalie Jomini Stroud. Bursting your (filter) bubble: strategies for promoting diverse exposure. In *CSCW*, pages 95–100. ACM, 2013.

[30] Hoda Sepehri-Rad and Denilson Barbosa. Identifying controversial articles in wikipedia: A comparative study. In *8th Int. Symp. on Wikis & Open Collab.*, 2012.

[31] Claude E. Shannon. A mathematical theory of communication. *ATT Thec. J.*, 27:379–423, 623–656, July, October 1948.

[32] A. F. Shorrocks. The Measurement of Mobility. *Econometrica*, 46(5):1013–1024, 1978.

[33] PM Sommers and J Conlsk. Eigenvalue immobility measures for markov-chains. *J. Math. Sociol.*, 6(2):253–276, 1979.

[34] Jennifer Stromer-Galley. Diversity of political conversation on the internet: Users' perspectives. *J. Computer-Mediated Comm.*, 8(3), 2003.

[35] Mikalai Tsytsarau, Themis Palpanas, and Malu Castellanos. Dynamics of news events and social media reaction. In *KDD*, pages 901–910. ACM, 2014.

[36] Mikalai Tsytsarau, Themis Palpanas, and Kerstin Denecke. Scalable discovery of contradictions on the web. In *WWW*, pages 1195–1196, 2010.

[37] Ryen W. White, Susan T. Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In *WSDM*, pages 132–141, 2009.

[38] E. Yom-Tov, S.T. Dumais, and Q. Gao. Promoting civil discourse through search engine diversity. In *QPOL*, 2013.

[39] B. Yu, S. Kaufmann, and D. Diermeier. Classifying party affiliation from political speech. *JITP*, 5(1):33–48, 2008.

[40] Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. Classifying the political leaning of news articles and users from user votes. In *AAAI ICWSM*, 2011.