The Web Is Flat: The Inflation of Uncommon Experiences Online

Danaë Metaxa-Kakavouli
Stanford University
metaxa@stanford.edu

Gili Rusak
Stanford University
gili@stanford.edu

Jaime Teevan
Microsoft Research
teevan@microsoft.com

Michael S. Bernstein
Stanford University
msb@cs.stanford.edu

CHI'16 Extended Abstracts, May 07-12, 2016, San Jose, CA, USA.
ACM 978-1-4503-4082-3/16/05.
http://dx.doi.org/10.1145/2851581.2892424

Abstract
People populate the web with content relevant to their lives, content that millions of others rely on for information and guidance. However, the web is not a perfect representation of lived experience: some topics appear in greater proportion online than their true incidence in our population, while others are deflated. This paper presents a large scale data collection study of this phenomenon. We collect webpages about 21 topics of interest capturing roughly 200,000 webpages, and then compare each topic's popularity to representative national surveys. We find that rare experiences are inflated on the web (by a median of 7x), while common experiences are deflated (by a median of 0.7x). We call this phenomenon *novelty bias*.

Author Keywords
Computational social science; world wide web; novelty bias.

ACM Classification Keywords
H.5.3 [Group and Organization Interfaces]: Web-based interaction

Introduction
The web may be the most expansive record of human experience to date, yet there are striking examples of the

ways in which we have authored a distorted reflection of our lives [2, 5]. Consider chest pain, for example: if someone browses the web and looks at pages about chest pain, they might conclude that chest pain will signal an imminent heart attack rather than a temporary annoyance [20]. These deviations exist because people do not author content equally across all experiences [11]. Instead, we choose what to create [12, 13, 14], and what to hold back [6]. Yet, prior work has analyzed information on the web as faithful representations of naturalistic social interactions [4], information behaviors [16], and evidence of language differentiation [7]. If the web differs significantly from lived experience, we must modify our tactics for using the web to inform our research as well as personal decisions.

In this paper, we seek to measure the difference between what people actually experience and what experiences are given visibility on the web. We propose a novel method (online crawl vs. offline survey) for quantifying the difference between human experiences and what people share on the web. First, we perform a geographically-restricted web crawl from a neutral seed query. Then, we annotate a sample of the resulting pages to quantify how the internet represents each component, and compare the proportion to representative offline statistics from the same region (e.g., Pew surveys in the United States).

Using this method, we collected data on 21 topics (e.g., religion, smartphone ownership, opinion on same-sex marriage) with 74 total components (traits, e.g., being Christian, Jewish, or Muslim; owning an iPhone or Android; supporting same-sex marriage or not), using a breadth of topics for which nationally representative survey data was available. We then observed that uncommon experiences ($< 10\%$ of offline) are inflated by a factor of 7 relative to their offline value, dominant experiences ($> 60\%$) are deflated to about 0.7 times their offline value, and components in the middle region appear roughly in proportion to their offline counterparts, at a factor of 0.9. We also show that these differences remain robust regardless of potential confounds including page visitation rates and temporal recency. We call this observed phenomenon *novely bias*.

Related Work

In many domains, information on the web performs well as a reflection of lived experience [3]. In some domains, however, information collected by the web compares less favorably with offline reality. Cultural and social biases affecting content creation are well-documented on Wikipedia [2, 5], as well as on social media sites [10]. Search engines are a similarly imperfect representation of the world, as are the queries users issue to them, which can carry subtle biases [19, 9].

The implications of the web's reflection of human experience are varied and meaningful, particularly because web users do not understand these discrepancies [8], and because individuals can mistake the frequency of encountering an opinion for a proxy of its offline frequency [18]. For individual web users, this can have serious consequences: the overrepresentation of a risky behavior may increase the perception of that behavior as normal [17]. For researchers, web data are increasingly being used to make predictions about the real world, conjectures that may not hold if the web misrepresents reality [10]. Understanding the web's ability to reflect or distort reality is critical both for consumers and producers of information online.

Research studies like those discussed above demonstrate that there is evidence of both the web's impressive record

Component	Offline	Online
Religion		
Christian	78.4	62.0
Atheist	4.0	14.0
Jewish	1.7	5.0
Buddhist	0.7	7.0
Muslim	0.6	9.0
Hindu	0.4	3.0
Politics		
Republican	24	29.5
Democrat	28.0	21.9
Independent	46.0	48.6
Airlines		
Delta	23.5	13.2
JetBlue	5.9	30.4
U.S. Airways	11.8	1.9
United	17.6	14.1
Southwest	23.5	30.1
American	17.6	10.3
Smartphones		
Android	48.0	30.0
iPhone	43.0	34.0
Blackberry	7.0	29.0
Windows	2.0	7.0
Abortion		
Pro-	57.4	44.7
Anti-	42.6	55.3
Marijuana		
Pro-legal.	56.2	62
Anti-legal.	43.8	38

Table 1: Offline and online percentages for a selected subset from the total 74 components across 21 that were collected.

of human experience, and also of several point examples of the ways in which that record is inaccurate compared to the offline in the form of targeted studies. However, we know little about the web as a whole — when it is accurate, when it is not, and why these patterns arise.

Online Crawl vs. Offline Survey

We begin by looking at how closely the volume of experiences reported on the web match offline data. To investigate this relationship we compared content from web crawls to population-representative surveys such as Pew.

To cover a broad sample of topics, we began with three major categories of information:

- *Identity topics*, reflecting affiliations: e.g., religion, political party
- *Experience topics*, describing people's actions: e.g., smartphone ownership, sport viewership
- *Opinion topics*, reflecting personal views: e.g., same-sex marriage, marijuana legalization

We chose 7 topics in each of the above types for investingation, yielding a total of 21 topics (Table **??**). Some, such as abortion, are hotly debated; others, such as airline popularity, are less active. Each category had between two and six components ($\mu = 3.5, \sigma = 1.5$), for a total of 74 components.

Method

After identifying the offline level for each topic, we (i) collect a set of on-topic webpages through a crawl that simulates the averaged behavior of random web users. We then (ii) label a random sub-sample of on-topic webpages according to each topic's components. Finally, we (iii) compare the statistical estimates produced via classification to nationally-representative survey statistics.

Offline Surveys

For each topic, we first identified an offline reflection of human experience. We focused on the United States, which has many publicly-available, reputable, nationally-representative survey sources such as Pew and Gallup. These offline metrics established our components for each topic. (We are limited to topics where recent offline data is available; further work in this area is needed to generalize our results further, but we attempted to cover a broad and varied set of topics in our sample.) For example, a 2013 Pew survey covering the smartphone experience topic produced the components Android (48%), iPhone (43%), Blackberry (7%) and Windows Phone (2%). Likewise, a 2014 Gallup poll on the same-sex marriage opinion topic produced 57% in favor and 43% against. We treat these numbers as a representation of the world offline, and aim to measure how closely the web mirrors them.

Collecting Relevant Web Content

To measure the web's coverage of each component within a topic, we built a web crawler using the Scrapy library for Python. Our goal was for the crawling algorithm to roughly simulate the aggregate search-based browsing behavior of many non-expert users, expanded in scale for a wide coverage of on-topic pages. Random walk metaphors are common as such research strategies, for example forming the basic conceptual model behind PageRank [15].

For each topic, we used the name of the topic (e.g. "marijuana legalization") as a general seed for each topic. We generated a whitelist of terms that were divided into subgroups of synonyms. For example, "religion," "religious," and "faith" were one subgroup of keyword terms for the

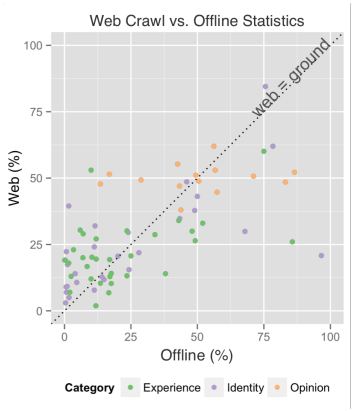

Figure 1: Online vs. offline percentages for each of 74 components across 21 topics, colored by topic type.

religion topic, and "Jewish," and "Judaism," were in another.

Each crawl began at webpages returned from a Google search of the seed. A crawler then collected on-topic webpages based on the prevalence of whitelist terms appearing on each page. On-topic pages were those containing at least a threshold number of keywords from a threshold of at least k subgroups, where k was determined per crawl. After crawling each page, the crawler followed a link on the page with a probability of 0.7, and otherwise chose a related search (suggested by Google). In each crawl, roughly one tenth of crawled pages qualified as on topic, and for each topic our crawler ran until 10,000 on-topic page URLs were collected.

Labeling Web Content

For each topic, after collecting 10,000 relevant pages, we sent a randomly chosen subset of 500-600 on-topic pages to crowdworkers from Amazon Mechanical Turk, where microtask workers labeled the relevance of each page to the given topic, and to all components within that topic. For example, for the same-sex marriage topic, we asked crowdworkers whether the webpage contained content related to same-sex marriage, and whether the page reflected a pro-same-sex marriage viewpoint, an anti-same-sex marriage viewpoint, or both. We chose to rate only 600 pages per topic since we noticed that components' proportions were fairly stable by 600 pages and did not change with more annotated pages.

To verify the quality of our annotations, we hand-annotated 20 pages for a subset of 6 topics (2 identity, 2 experience, 2 opinion) and calculated Cohen's kappa as a metric of inter-rater reliability. The average unweighted Cohen's kappa across this subset of 6 topics was 0.784, indicating good agreement between our ratings and the Turkers'.

Analysis

We compared the percentage of our collected pages discussing each component for a topic to the offline percentages from the topic's nationally-representative survey. For both the offline statistics and the crowdlabeled data, we develop relative percentages for each of the 2-6 components within each topic. For example, 78% of the United States population identifies as Christian according to Pew, and 62% of web pages in our religion sample significantly discussed Christianity.

Results

In total, our crawler collected approximately 200,000 web pages across the 21 topics, and crowdworkers manually annotated over 12,000 of them. Webpages collected by our crawler ranged broadly in type and content, and included articles from individuals' personal webpages and blogs, news sites, organizations' webpages, and social media content.

Figure 1 presents each topic component's offline prominence against its representation online. In the case of a null result in which every topic component is proportionally represented, we expect to see all data points fall on the line $y = x$; deviations from this line in our data reflect discrepancies between the web and reality. Opinion topics displayed the strongest bias (Figure 2).

Uncommon Inflated, Dominant Deflated

In our data, *uncommon components* with offline percentages less than 10% were over-represented at a median of 6.7 times the offline rate. For instance, while only 4% of the American population self-identifies as Atheist or Agnostic, 14% of religion pages mentioned atheism or agnositicism. Similarly, while heart attacks account for only 10% of chest pain incidents, over half of the pages

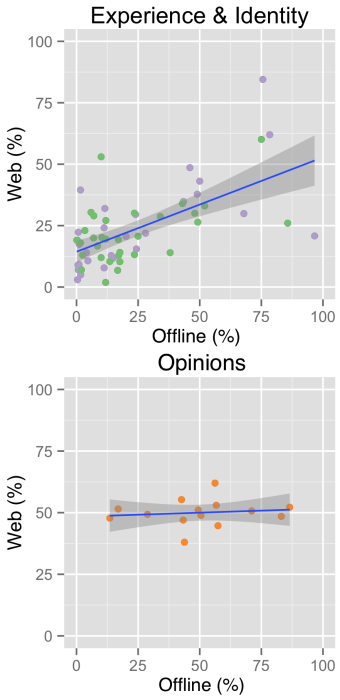

Figure 2: Experience and Identity topics (top) inflate uncommon components and deflate common ones, but components of opinion topics (bottom) are equally represented.

we crawled (53%) mentioned heart attack as a cause for chest pain.

Dominant components are those whose offline representation is greater than 60%. The median dominant component appeared online at 0.66 times its offline rate. For instance, while 85% of the American population watches football, only 26% of the rated sports pages discussed it.

Between uncommon and dominant components lie *moderate components*, whose offline representation is between 10% and 60%. Overall, these components appear proportionally, comparing online to offline representation. The median moderate component appeared at 0.9 times its offline rate. For example, support for gun regulation is presented in the American population exactly as it is our webcrawled data, at 51%.

We fit a line to the identity and experience components ($online = 0.38 * offline + 14.3$, $R^2 = 0.37$), and tested whether this pattern was significantly different from total agreement with the offline, $y = x$. In order to do so, we tested against a null hypothesis of $\beta = 1$. The result was significant ($p < 0.001$), confirming that the representation of these topics online was not proportional to offline representation.

Opinion topics deviated from $y = x$ even more strongly, with points clustering near the line $y = 50\%$ (Figure 2). Regardless of offline representation, each side of these debates was represented equally relative to the other—for example, topics as different in public opinion as same-sex marriage (56% in favor to 43% against offline) and human cloning (15% in favor to 85% against) both displayed this 50%-50% balance.

As before, we fit a line to opinion datapoints ($online = 0.033 * offline + 48.3$, $R^2 = 0.02$) and tested whether this pattern was significantly different from $y = x$ ($\beta = 1$). We found that opinion topics, like experience and identity, deviated significantly online ($p < 0.001$).

Exploring Possible Confounds

To test the robustness of our data, we explore possible confounds and find our results to be robust.

Production vs. Visitation

Despite our collection of webpages stemming from search engine results, which implicitly adapt based on web page visitation, another criticism might note the difference between production of content and visitation to content online: while a variety of content exists online, web users only visit a small subset of that information. This criticism would allege that the patterns in Figure 1 do not reflect any meaningful information about an individual user's experience online, and that the patterns in our data will differ dramatically when weighting pages by viewership.

To address this concern, we gathered visitation data for each URL (5,000-10,000 per topic) using web browser logs gathered from opt-in users of a popular internet browser. This data represents the frequency of viewership of each webpage relative to all others in its topic for one week in August 2014. For confidentiality reasons, all visitation numbers were given as an order of magnitude (e.g., 1, 10, 100, 1000). As expected, most pages were not visited over the course of a week [1]. Pages with no pageviews during that week were removed from the dataset, and pages in the topic that received pageviews but had not been previously annotated by crowdworkers were newly annotated. We removed 1-2 outliers from

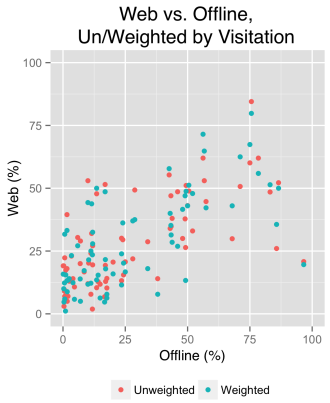

Figure 3: Topic components weighted and unweighted by visitation.

each topic that represented navigation pages that change often, such as http://games.espn.go.com/frontpage because a large number of page viewers would not have seen topic-relevant content there.

We tested whether the coefficient differs significantly across the weighted and unweighted groups by building a linear model to predict web percentage, adding a binary indicator variable to this predictor representing whether the datapoint was visitation-weighted, and an interaction term between offline and the indicator. The resulting term represents the difference in the weighted and unweighted datasets' coefficients, and is not significant: $\beta = -0.04, t(144) = -0.48, p = 0.63$.

Temporality

Since the Web acts as an archive, retaining all information even after that information is no longer up-to-date, another critique might note that results of crawling the web might reflect old information patterns that are not represented in recent national surveys. For example, pages about Blackberry phones (7% offline, 29% online) might be plentiful but outdated, artifically inflating its presence. If this is the case, we should expect to see different patterns in our data when considering only recently viewed pages online. As mentioned above, our visitation data weighted pages by viewership in a recent, one-week period; Figure 3 reflects, again, that this did not result in significantly different patterns in our data, though it did impact some specific points, like Blackberry (5% when weighted by recent visitation).

Discussion

The results from this paper suggest that there are systematic biases in the extent to which information is represented online. Both lay-persons who rely on the web for general information, and computational social science researchers leveraging observational data from the Web in their work are affected by this phenomenon. Many users that rely on the web are unaware of its shortcomings [8]. As the web is increasingly integrated into everyday lives, individuals' perceptions of the world and their place in it may be skewed by the Web's biases. These skewed contexts can have dramatic results on their behaviors [17].

Limitations & Future Work

The novelty bias battern we pbserved was robust to several possible confounds, but in collecting pages we examined only a limited subset of the web reachable through a Google search-based crawler. We aim to develop an automated pipeline to of crawl and classify all webpages related to a particular topic, using the record of the web recorded by Common Crawl. Toward that goal, we are currently using crowdworker-generated data to train machine learning classifiers to annotate webpages automatically. We will then perform the same analysis on this exhaustive dataset to confirm the aforementioned results.

Conclusion

The web is an emergent product of millions of authors. So it is striking that we have, collectively, transformed the relative volumes of our lived experiences so consistently online. Through a large-scale web crawl across 21 topics and 74 components, we see that unpopular experiences are overrepresented and popular experiences underrepresented. As a community, we are learning more about *why* people share content on the web (e.g., [12, 13, 14]), but our results make clear that more attention can be paid to naturalistic investigations of *when* and *under what conditions* this occurs.

References

[1] Eytan Adar, Jaime Teevan, and Susan T. Dumais. 2008. Large Scale Analysis of Web Revisitation Patterns. In *Proc. CHI '08*.

[2] Aaron Shaw Benjamin Hill. 2013. Mapping The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation. *PLoS ONE* (2013). http://hdl.handle.net/1721.1/80697

[3] danah boyd. Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life. In *MacArthur Foundation Series on Digital Learning*, David Buckingham (Ed.).

[4] danah boyd. 2010. Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications. In *A Networked Self: Identity, Community, and Culture on Social Network Sites*, Zizi Papacharissi (Ed.). New York, NY, USA.

[5] E. S. Callahan and S. C. Herring. 2011. Cultural bias in Wikipedia content on famous persons. *JASIS* (2011).

[6] Sauvik Das and Adam Kramer. 2013. Self-Censorship on Facebook. In *ICWSM*.

[7] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proc. EMNLP '10*.

[8] Leah Graham and Panagiotis Takis Metaxas. 2003. "Of Course It's True; I Saw It on the Internet!": Critical Thinking in the Internet Era. *Commun. ACM* 46, 5 (May 2003), 70–75.

[9] Lucas D. Introna and Helen Nissenbaum. 2000. Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society* 16, 3 (2000), 169–185. DOI:http://dx.doi.org/10.1080/01972240050133634

[10] Andreas Jungherr, Pascal Jürgens, and Harald Schoen. 2012. Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions. *Soc. Sci. Comput. Rev.* 30, 2 (May 2012), 229–234.

[11] Emre Kıcıman. 2012. OMG, i have to tweet that! a study of factors that influence tweet rates. In *Proc. ICWSM '12*.

[12] Alice Marwick and danah boyd. 2010. I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. *New Media and Society* (September 2010).

[13] Bonnie A. Nardi, Diane J. Schiano, Michelle Gumbrecht, and Luke Swartz. 2004. Why We Blog. *Commun. ACM* 47, 12 (Dec. 2004), 41–46.

[14] Oded Nov. 2007. What Motivates Wikipedians? *Commun. ACM* 50, 11 (Nov. 2007), 60–64.

[15] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report.

[16] Peter L. T. Pirolli. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press.

[17] DA Prentice and DP Miller. 1993. Pluralistic ignorance and alcohol use on campus. *J Pers Soc Psychol* 64, 2 (2 1993), 243–56.

[18] Kimberlee Weaver, Stephen M Garcia, Norbert Schwarz, and Dale T Miller. 2007. Inferring the popularity of an opinion from its familiarity: A repetitive voice can sound like a chorus. *J Pers Soc Psychol* 92, 5 (2007), 821–33.

[19] Ryen White. 2013. Beliefs and Biases in Web Search. In *Proc. SIGIR '13*.

[20] Ryen W. White and Eric Horvitz. 2009. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. *ACM TOIS* 27, 4, Article 23 (Nov. 2009), 37 pages.