Capturing the Aftermath of the *Dobbs v. Jackson Women's Health Organization* Decision in Google Search Results across the U.S.

Brooke Perreault, Lan Dau, Anya Wintner, Eni Mustafaraj

Department of Computer Science
Wellesley College, MA, USA
bp101, ldau, aw105, emustafaraj@wellesley.edu

Abstract

How do Google Search results change following an impactful real-world event, such as the U.S. Supreme Court decision on June 24, 2022 to overturn Roe v. Wade? And what do they tell us about the nature of event-driven content, generated by various participants in the online information environment? In this paper, we present a dataset of more than 1.74 million Google Search results pages collected between June 24 and July 17, 2022, intended to capture what Google Search surfaced in response to queries about this event of national importance. These search pages were collected for 65 locations in 13 U.S. states, a mix of red, blue, and purple states, with respect to their voting patterns. We describe the process of building a set of circa 1,700 phrases used for searching Google, how we gathered the search results for each location, and how these results were parsed to extract information about the most frequently encountered web domains. We believe that this dataset, which comprises raw data (search results as HTML files) and processed data (extracted links organized as CSV files) can be used to answer research questions that are of interest to computational social scientists as well as communication and media studies scholars.

Introduction

Few political events were more talked about in 2022 than the overturn of Roe v. Wade by the U.S. Supreme Court. Roe v. Wade was a 1973 court decision that conferred to pregnant people a constitutional right to choose abortion, making abortions legal in all 50 U.S. states. However, the conservative legislatures of several U.S. states, where prior to Roe v. Wade abortion was illegal, repeatedly drafted laws that restricted or banned access to abortion, with the explicit strategy to erode the power of Roe v. Wade. Once the U.S Supreme Court reached a strong conservative majority, due to three court vacancies filled during Trump's presidency, it overturned Roe v. Wade, in a decision known as *Dobbs v. Jackson Women's Health Organization*, which upheld the constitutionality of an abortion ban after 15 weeks by the Mississippi state legislature, and explicitly overruled two prior cases, Roe v. Wade and Planned Parenthood v. Casey. A draft of this decision was first leaked by Politico,[1] on May 3, 2022, with the formal court decision announced to the public on June 24, 2022.[2]

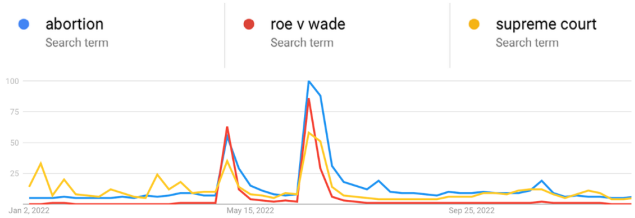

Figure 1: In 2022, the U.S. Supreme Court had a memorable year, but its most consequential decision was the overturning of Roe v. Wade, which treated abortion as a constitutional right. The screenshot shows the Google Trends for the three terms: abortion, roe v wade, and supreme court. They reached their peak when the *Dobbs v. Jackson Women's Health Organization* decision was made public on June 24, 2022, capturing the high public interest on the topic.

In Figure 1, we show a screenshot from the 2022 Google Trends for three query phrases: *abortion*, *roe v wade*, and *supreme court*. The graph shows two pronounced peaks on May 3rd (when the leak came out) and June 24th (the day of the formal decision), with the latter event being more popular. As it was reported in the media, Google searches started to skyrocket 45 minutes after the decision was announced,[3] with people in various U.S. states asking questions relevant to their particular situations. Concretely, people in the so-called red states (that had trigger abortion ban laws in place) asked for "abortion clinics near me", while people in the blue states asked for "planned parenthood clinics near me".

In light of the immense interest in this topic, what did Google Search results look like? In what ways did they change to reflect the amount of content that was being written about this landmark decision? How well were they serving the needs of the public for localized information, given the overturn of Roe v. Wade affected differently Americans living in different parts of the country, based on who had

[1]https://www.politico.com/news/2022/05/02/supreme-court-abortion-draft-opinion-00029473

[2]https://www.npr.org/2022/06/24/1102305878/supreme-court-abortion-roe-v-wade-decision-overturn

[3]https://www.axios.com/2022/06/26/abortion-supreme-court-google-searches

the political power in their state? To answer these questions, we set out to collect a dataset that captures this historical moment as reflected in the Google Search results pages for 65 locations across the U.S. We are making this dataset of more than 1.7 million search pages available to researchers, with the hope that it will be used to answer a multitude of questions in various disciplines. In addition to providing the "raw data" (HTML pages of search results), we also provide data of interest that we have extracted in the meantime: all organic search results (URLs, domain names, titles); all top stories (URLs, domain names, titles); and a list of tweets (only tweet IDs) embedded in the search pages. Our data are currently accessible through two public repositories: Harvard Dataverse,[4] and our own website.[5]

This dataset makes multiple important contributions. We provide a comprehensive dataset of entire Search Engine Result Pages and parsed information that capture Google's coverage of a significant national event, collected in many diverse locations and for a substantial amount of search queries. By auditing Google Search results for the most popular queries related to abortion and the overturning of Roe v. Wade, our dataset can help researchers better understand and characterize the algorithmically-mediated information environment surrounding abortion. Given that access to truthful and accurate information about reproductive health is an important component of the path towards reproductive freedom, our dataset holds the promise to providee significant insights into the state of this critical information environment.

Summary of Dataset Characteristics

- Searches were conducted for 65 locations in 13 U.S. States.
- There are 26 snapshots collected in 21 days from June 24, 2022 (day of the Supreme Court decision) to July 17, 2022. (On five days we collected data in the morning and the evening.)
- 1,744,299 HTML pages of Google Search results pages.
- 1,698 search phrases collected with different methods and labeled on their relevance to the topics of the search.
- 19,342 unique URLs from 5,216 websites that occurred 13,307,766 times in the organic search results.
- 17,503 unique URLs from 2,198 websites that occurred 2,803,419 times in the Top stories results.
- 15,183 unique tweet IDs that occurred 232,769 times in the Twitter panel embedded in Google's search results.

In the following, we describe in detail how the dataset was constructed; we provide some preliminary analysis of the nature of the websites that occur more frequently in the organic results and the Top stories; and compile a list of potential questions for future analysis. We invite other researchers to extend and expand the research on datasets like this one, which provide a window into the functionality of a complex sociotechnical system such as the Google Search engine.

[4]https://doi.org/10.7910/DVN/YFAH9X

[5]https://cs.wellesley.edu/~credlab/icwsm2023/

Localized Search on Google

When the term "filter bubble" was coined by (Pariser 2011), it helped popularize the concern that Google Search was personalizing search results in ways that limited access to important information. However, carefully constructed studies such as (Hannak et al. 2013) or (Kliman-Silver et al. 2015) found that personalization depends on the query and it affects a small number of searches. One of the most consistently used criteria for "personalization" was found to be the geolocation, which can be typically inferred from the IP address of one's browsing device.

A 2022 longitudinal study by (Mejova, Gracyk, and Robertson 2022) used a Google Search's feature to set the location of the search to a desired one and utilized it to collect data from 467 locations in the United States about abortion clinics. The study relies on the open-source library, WebSearcher,[6] used in other search engine audits (Robertson, Lazer, and Wilson 2018; Robertson et al. 2018) as well, which automatically parses the components of a search results page. Given Google's tendency to change the structure of a page and introduce new panels, our approach is to store the HTML pages as they appear and do the parsing later, to ensure we can find things that we had not anticipated. In the past, this has been a strategy for capturing Google's short-lived features, such as "Reviewed Claims" (Lurie and Mustafaraj 2020), aimed to support news literacy efforts.

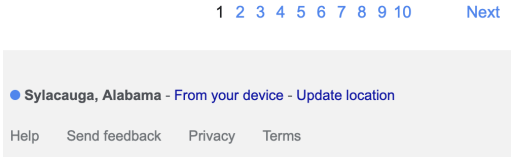

Figure 2: A screenshot from a Google Search result page showing the changed location to one of the 65 locations used for data collection, Sylacauga, Alabama.

Our method for localizing search results makes use of Google Search's feature for updating the location of the browser by providing latitude and longitude values for a different location from the one of the device. We use the automated browser control through Selenium[7] to perform this process. We can test that the location was changed by inspecting that the web page contains the name of the desired location. For an example, refer to Figure 2, showing a location of results, different from where the authors are located. All the HTML pages in our dataset contain the names of one of the 65 locations used for the search.

Figure 3 is a good example of what localization does for the search page. Given the query "abortion clinic near me medicaid" for the location Temple, TX, Google tailored the ads, the Places section, and the organic search results to the location. Furthermore, the ads are labeled based on whether they provide abortion or not, as well as the entities listed in Places. In the past, the lack of labeling was a concern, and

[6]https://github.com/gitronald/WebSearcher

[7]https://www.selenium.dev

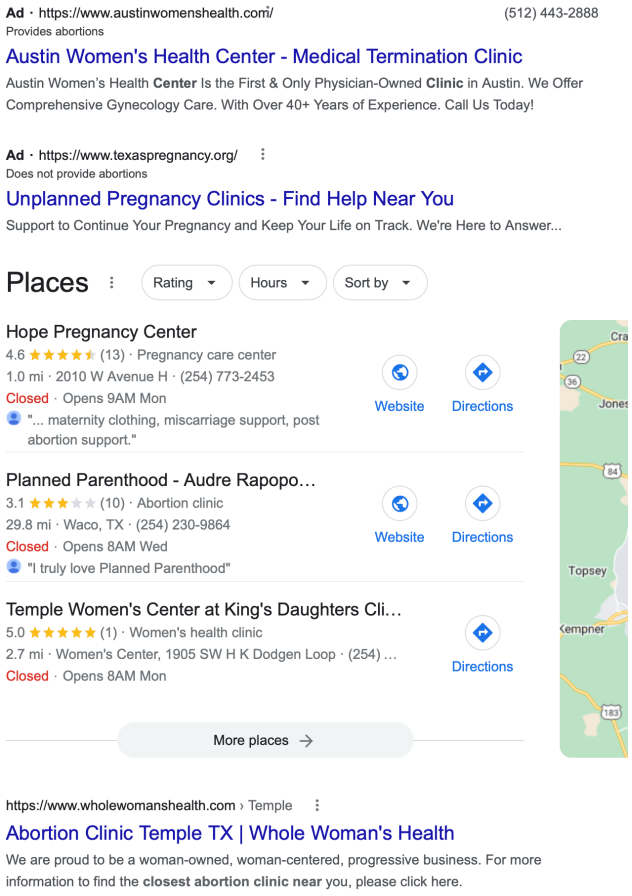

Figure 3: Localization of search results is evident when queries contain the phrase "near me". This screenshot belongs to one such search page for Temple, TX. All visible elements: the ads, the Places panel, and the organic results are specific to this location.

prior research (Mejova, Gracyk, and Robertson 2022) has focused on auditing this aspect of Google Search.

Related Work

Communication and media scholars, applied philosophers, and social scientists were among the first to take a critical stance toward the rise of search engines. Their (mostly qualitative) research was then followed by quantitative research from computer scientists, especially in the field of algorithm audits. In the following we summarize published studies from these two perspectives.

The Politics of Search

When Google Search started to gain popularity in the late '90s, its emergence as a new powerful gateway to online information immediately drew the interest of applied social scientists (law, technology ethics, etc.), for example, (Introna and Nissenbaum 2000; Elkin-Koren 2000), who outlined the risks associated with so much power. The critical examination of the potential search engine bias was constant through the 2000's, as evidenced by the work of many researchers (Pasquale 2006; Hargittai 2007; Diaz 2008). Calls for regulations were also quick to arise, though the focus was on establishing norms of engagements, for example, the values proposed in (Gasser 2005): informational autonomy, diversity, and, information quality.

One feature that is found in some of these critical studies is the focus on a very small set of queries: e.g., (Grimmelmann 2008) explores only five search queries: "mongolian gerbils", "talentless hack", "jew", "search king", and "tiananmen", or (Gillespie 2017) selects a single search phrase "Rick Santorum", to discuss issues of what it means to be a platform with big power and responsibility, a concept that Gillespie explores in details in his "politics of platforms" (Gillespie 2010). Such cases provide examples of "algorithmic breakdowns" (Mulligan and Griffin 2018) that expose an opening to understand their priorities and values.

However, given the scale of searches on Google, there is also value in testing its algorithmic behavior for larger sets of queries, and in experimental setups to measure variability across time and space. That is what the field of algorithm audits does.

Algorithm Audits

Algorithm audits (Sandvig et al. 2014) have emerged as an important direction in the interdisciplinary domain of fairness, accountability, and transparency of algorithmic systems, with the goal of exposing bias in such systems. Search engines, as one of the most used algorithmic systems in the world, are often the focus of audits by the research community, especially in the context of political elections or other high profile political events (Robertson, Lazer, and Wilson 2018; Robertson et al. 2018; Metaxas and Pruksachatkun 2017; Diakopoulos et al. 2018; Metaxa et al. 2019). The audits are not limited to search engine platforms, and we refer to a recent systematic literature review (Bandy 2021) and synthesis article (Metaxa et al. 2021) for a broader and deeper look at this field.

In prior work, (Lurie and Mustafaraj 2019; Kawakami, Umarova, and Mustafaraj 2020), we have discussed the kind of audit methodology we use to collect data: blank instances of the Chrome browser that are automated by Selenium and do not retain the prior search activity. Collecting Search Engine Result Pages (SERPs) in such a way removes the influence of most variables that might influence the search results, leaving as relevant to the search: the query phrase, the time of the collection, and the geolocation. By varying the geolocation programatically, and keeping the query phrases and collection time the same, we can study the localization of results for a desired topic, as we do for this project.

Data Collection Process

We collected SERPs for abortion related queries between June 24, 2022, the day of the *Dobbs v. Jackson Women's Health Organization* decision, and July 17, 2022. In this section, we describe the details of this process, including the development of search queries, selection of locations, and automated Google searches for these queries and locations.

Source	# of Queries in Final List	Example Queries
Our seed list	6	abortion, abortion law, roe v. wade, Dobbs vs Jackson Women's Health Organization, pro life*, pro choice*
Google Trends - United States	102	abortion clinic, abortion ban, abortion near me, supreme court leak, is abortion illegal now, supreme court abortion ruling
Google search autocomplete	898	abortion clinic near me cost, leaked supreme court opinion on abortion, medical abortion vs surgical, overturn roe v wade ivf
Google Trends - 50 States*	522	what states will ban abortion, abortion pill, pro choice protest near me, 6 week abortion ban, heartbeat bill
Expert Panel*	49	reproductive health restrictions, reproductive justice, anti-choice legislation, abortion funds, maternal health crisis
Pro-choice websites*	29	reproductive freedom, birth control, pregnancy discrimination, anti-choice agenda, legal abortion
Pro-life websites*	87	abortion industry, preborn baby, pro-life movement, post-roe america, abolish abortion
Related Search*	5	pro life groups near me, is abortion murder, when does life begin, anti abortion organization, fetal rights

Table 1: The sources for the queries, counts, and example queries that compose our query dataset. An * indicates that the query or query group was part of the updated query dataset, beginning collection on July 8, 2022.

State	Localities	% of population voting D and R in 2020	Abortion legality before Dobbs Ruling	Abortion legality after Dobbs ruling
AL	Andalusia, Bessemer, Gulf Shores, Opp, Sylacauga	D: 36.6; R: 62.0	Very restricted	Total abortion ban
AZ	Apache Junction, Ash Fork, Casa Grande, Mesa, Topock	D: 49.4; R: 49.1	Somewhat restricted	Fifteen-week abortion ban
CA	Arcata, Bakersfield, King City, San Bernardino, San Gabriel	D: 63.5; R: 34.3	Legal	Legal: protected in the state constitution
GA	Camilla, LaGrange, Lovejoy, Marietta, Waycross	D: 49.5, R: 49.3	Somewhat restricted	Six-week abortion ban
MA	Agawam, Boston, Needham, North Adams, Salem	D: 65.6; R: 32.1	Legal	Legal; protected by state legislation
NY	Glenmont, Jamestown, Lackawanna, Putnam Valley, Rochester	D: 60.9; R: 37.7	Legal	Legal; expanded access
NC	Eastover, Lumberton, Mount Pleasant, Shallotte, Southern Pines	D: 48.6; R: 49.9	Somewhat restricted	Tweenty-week ban; other restrictions
OH	Belpre, Coshocton, Mason, Monroe, Youngstown	D: 45.2; R: 53.3	Very restricted	Six-week ban
OK	Ada, Bethany, Marlow, Seminole, Watonga	D: 32.3; R: 65.4	Very restricted	Total abortion ban
TX	Celina, McAllen, McKinney, Temple, Winters	D: 46.5; R: 52.1	Very restriced	Total abortion ban
VA	Afton, Aylett, Berrys, Bristol, Lorton	D: 54.1; R: 44.0	Legal	Legal; not protected
WA	Big Lake, Mount Vista, Port Orchard, Walla Walla, Waller	D: 58.0; R: 38.8	Legal	Legal; protected by state legislation
WV	Bethany, Charleston, Elkins, Petersburg, Westo	D: 29.7; R: 68.6	Somewhat restricted	Total abortion ban

Table 2: The list of 13 states and 65 localities for which SERPs were collected. The states were chosen to be a mix of states with different positions towards abortion, dependent on their voting behavior. States with high support for the Democratic Party (D) kept abortion legal, while states with strong support for the Republican Party (R) moved to total abortion bans. The voting data are available from: https://www.cookpolitical.com/2020-national-popular-vote-tracker. The legality of abortion information is as of January 15, 2023 and available at: https://reproductiverights.org/maps/abortion-laws-by-state/.

Queries for the Search

As it is typical in many audits, we started with an initial set of seed queries relevant to the topic. Our list contained: "abortion", "abortion law", "Dobbs vs Jackson Women's Health Organization", and "roe v. wade". To capture how average users were searching for this topic at this time, we downloaded the "Rising" and "Top" related queries in the United States from Google Trends for each seed query as of June 15, 2022. We then collected the first five Google Autocomplete suggestions for each unique query using the open-source tool `suggests`.[8] This resulted in a list of 1,004 unique queries.

Part way through data collection, on July 8, 2022, we augmented our query list in the following ways.

1. Google Trends from All 50 States. Since we collected Google Trends data for the entire United States on June 15, we augmented the list with the "Rising" and "Top" related queries from Google Trends in all 50 states on July 1, 2022 using the initial seed queries and adding two more seeds, "pro choice" and "pro life". We removed duplicate queries between states, and we also excluded queries that contained the names of specific locations or states, which we did not do in the first phase. At this point, we did not filter for relevancy of queries to our topic. This process resulted in 522 additional unique queries.

2. Biased Queries. To capture how biased language in queries may impact search results (something that the literature supports, e.g., (Hu et al. 2019)), we wanted to ensure that the queries we used reflected pro-life and pro-choice viewpoints. Review of our initial 1,004 queries revealed a lack of queries reflecting a pro-life viewpoint. To resolve this, in addition to using "pro life" as a seed query for Google Trends as described above, we automatically extracted (and then manually curated) noun phrases from the websites of the pro-life organizations Americans United for Life,[9] Prolife Across America,[10] Students for Life of America,[11] and Focus on the Family.[12] We also extracted noun phrases from the United States Conference of Catholic Bishops' webpage[13] describing their pro-life stance, as well as from the Related Search suggestions for the queries "pro life" and "pro life organizations" until reaching a saturation point. This process resulted in 92 additional pro-life biased queries. We similarly extracted noun phrases from the websites of Planned Parenthood[14] and NARAL Pro-Choice America Foundation[15] to develop 29 additional pro-choice biased queries. As a result of these processes, the final query list included a balanced number of pro-life and pro-choice queries (98 and 86, respectively).

3. Expert Panel. We considered how the language that experts used to discuss abortion following the *Dobbs v. Jackson Women's Health Organization* ruling would differ from the language of non-experts, such as average Search users, and how this may impact Google Search results for different queries. On June 30, 2022, one of the authors attended a panel discussion of the *Dobbs v. Jackson Women's Health Organization* Supreme Court ruling hosted by Wellesley College, featuring attorneys, medical professionals, and reproductive health advocates as panelists and wrote down relevant phrases used frequently during the discussion to create 49 additional queries. The final list included 1,698 search queries, summarized in Table 1.

Locations for the Search

We collected data in 5 locations in 13 states, for a total of 65 locations. These locations are detailed in Table 2. Initially, we non-randomly selected states while ensuring that we included a mix of political leanings (as documented by the popular vote in the 2020 U.S. presidential elections) as well as stance on abortion. The locations within each state were selected randomly from among a list of locations with latitude and longitude coordinates available online.[16]

Time Period of Collection

Data collection started on June 24, 2022, at 6:00 PM ET. On June 25, 2022, data was collected once, starting at 6:00 AM. Between June 26 and June 30, 2022, data was collected twice a day, starting at 2:00 AM and 6:00 PM. Between July 1 and July 17, 2022 (excluding July 6, 7, and 9) data was collected once a day, starting at 6:00 PM. We used 7 computers, with each computer collecting SERPs for 2 states. It took approximately 1.5 to 3 hours to collect SERPs for all queries in one location, since we used pauses between searches to respect Google's limits.We used the original set of 1,004 queries between June 24 and July 5, and we used the updated list of 1,698 queries between July 8 and July 17, 2022. In summary, we collected 26 snapshots over 21 days, for an anticipated 2,102,750 total SERPs. Given that software or hardware failure sometimes interrupted or prevented data collection, we successfully collected 1,744,299 SERPs.

Automated Searches

We use a custom Python script and Selenium to automate the process of querying Google Search in a desired location: for each query from our list, the script opens a new instance of a Chrome browser in a blank-slate (with no user history or cookies), enters the query in the search box, waits for the page to load, scrolls down to the bottom of the page, updates the location using latitude and longitude coordinates and waits for Google to refresh the content of the page for the new location. The resulting page is then stored as an HTML file. We consider such files as "raw data", as they store information in the format, position, and order decided by Google's algorithms.

[8] https://github.com/gitronald/suggests

[9] https://aul.org/advocacy/

[10] https://prolifeacrossamerica.org/learn/blog/

[11] https://studentsforlife.org/blog/

[12] https://www.focusonthefamily.com/pro-life/abortion/pro-life-pro-choice/

[13] https://www.usccb.org/prolife

[14] https://www.plannedparenthood.org/

[15] https://www.prochoiceamerica.org/

[16] https://www.latlong.net/country/united-states-236.html

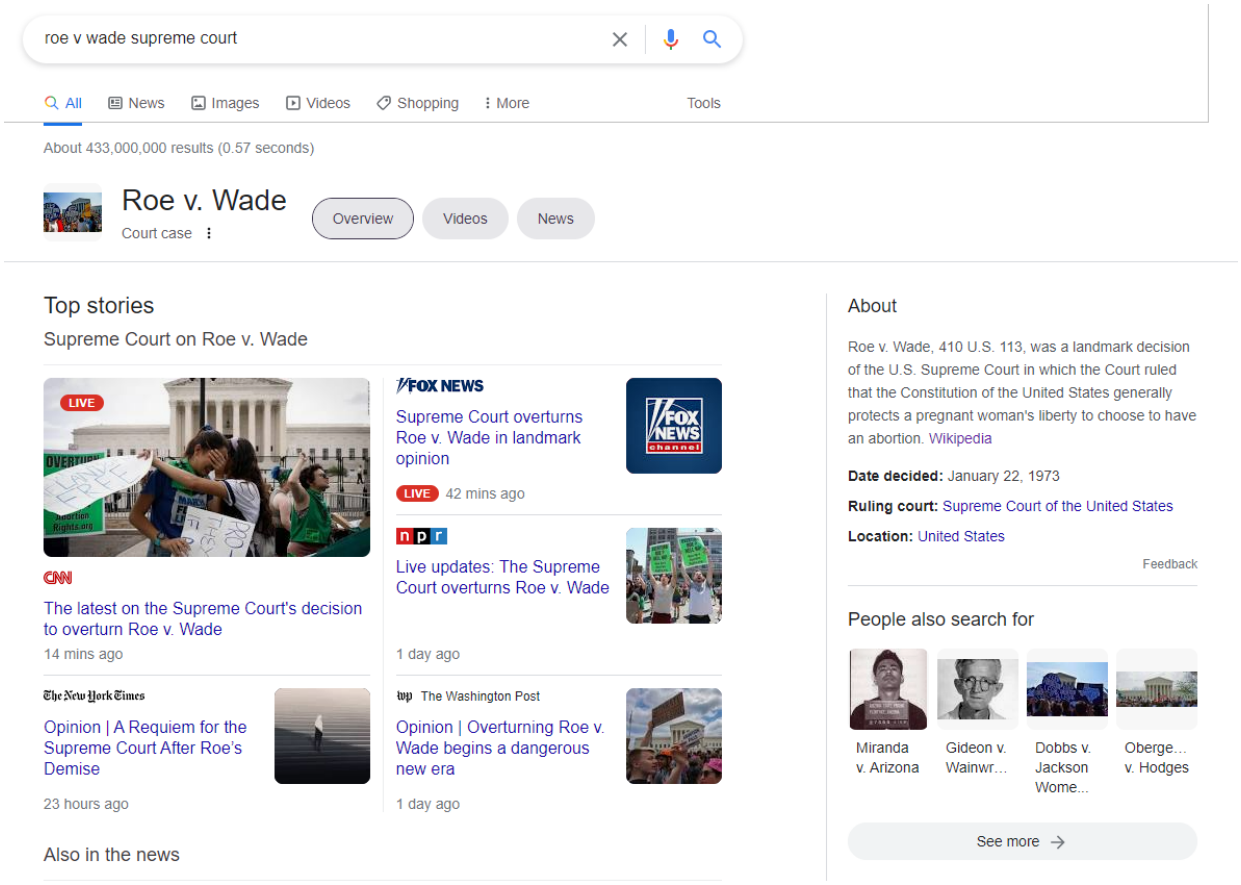

Figure 4: A screenshot from a Google Search result page for the query "roe v wade supreme court". Notice the prominence of the Top stories section. The screenshot also depicts a knowledge panel specific to supreme court cases.

Parsing of HTML Files

To extract information from the SERPs, we rely on BeautifulSoup[17] to parse the HTML files for various components, including organic results, Top stories, top and bottom ads, featured snippets, people also ask, maps, knowledge panels, related searches, Google Scholar results, and dictionary, image, video, and Twitter blocks. Per SERP, a corresponding JSON file is saved with the parsed information, which can then be used to compile data about specific components, as we have done for organic results, Top stories, and Twitter panels. While we capture many components with our parser, it may not capture all possible information contained in a SERP. Therefore, we include the raw HTML files in our dataset to allow for future parsing or other analysis.

Dataset Characteristics

Raw Data: HTML Search Engine Result Pages

In total, we collected 1,744,299 SERPs. Table 3 summarizes the number of SERPs collected per state and the average number of SERPs collected per location within each state.

Aggregated Data: Organic Results, Top Stories, Tweets

As described, we parse the HTML files and extract information about SERP components. Using parsed information from the JSON files, we can aggregate data about specific SERP components. In this dataset, we aggregate data on the Organic Results, Top stories, and Twitter panels. In particular, there were 13,307,766 total instances of organic results across all SERPs, corresponding to 19,342 unique links, and 5,216 domains. There were 2,803,419 Top stories, corresponding to 17,503 unique links and 2,198 domains. Lastly, there were 15,183 unique tweet IDs that occurred 232,769 times in the Twitter panel embedded in Google's search results. These data are included as CSV files as part of our dataset.[18]

Preliminary Analysis

There are a multitude of questions to pursue, but for starters, we are interested in exploring the nature of content generated following an impactful event, such as the *Dobbs v. Jackson Women's Health Organization* decision, as reflected by Google's algorithms; as well as the extent to which various

[17] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[18] https://doi.org/10.7910/DVN/YFAH9X

State	# SERPs	Avg. # SERPs per Location
Alabama	146,658	29,331
Arizona	131,237	26,247
California	136,988	27,397
Georgia	149,577	29,915
Massachusetts	156,515	31,303
New York	119,949	23,989
North Carolina	122,175	24,435
Ohio	119,284	23,856
Oklahoma	114,277	22,855
Texas	133,878	26,775
Virginia	122,313	24,462
Washington	144,760	28,952
West Virginia	146,688	29,337

Table 3: The number of SERPs collected per state and the average number of SERPs collected per location within each state. Software or hardware failure that interrupted data collection caused some states to collect data fewer than others.

components of a Google's SERP are localized. In the following, we offer some statistics from our dataset with these interests in mind.

Nature of Frequently Occurring Websites

Top Stories: As evident in Figure 4, Top stories is the most prominent element in a SERP, and therefore, important for this type of analysis. Table 4 summarizes the top ten domains for Top stories. It is notable that Top stories from these 10 domains constitute almost half (49.7%) of all Top stories instances in our dataset. The top 8 domains are those of national news organizations, including The Washington Post, The New York Times, and Politico. The ninth top domain is that of the international news organization, Reuters, and the tenth top domain is that of news organization The Texas Tribune, a state-level news organization that also partners with The Washington Post to share their content nationally.[19] These statistics suggest that national news organizations dominate Google's Top stories panel.

A concern that is often discussed is whether Google's algorithm manifest a political bias with their choice of domains. This is why it is important, to the extent this is possible, to analyze the political leaning of the selected domains, especially the ones that occur most frequently. To this effect, we utilize a dataset of Partisan Audience Bias (PAB) scores, compiled by (Robertson et al. 2018). Each of the 19,022 websites in their dataset received a score between -1 (far left) and +1 (far right) to indicate their perceived political leaning as reflected by the audience that shares their content on Twitter. The PAB datset contained scores for 1,427 out of 2,198 domains (65%) that appeared as Top stories in our dataset. Figure 5 visualizes the distribution of PAB scores by domain, number of top stories links, and number of top stories appearances. Center-left domains, with a PAB score between 0 and -0.5, produced 53% of Top stories unique links and constituted 55% of all Top stories appearances overall, while center-right domains, with a PAB score between 0 and +0.5, produced 33% of Top stories unique links and constituted 25% of Top stories overall.

An important aspect that Figure 5 depicts is that the subset of 1,427 out of 2,198 news domains itself is pretty balanced from a political bias perspective (top chart), with slightly more center-right than center-left news outlets. However, when looking at the production of news from these outlets (middle chart) center-left outlets produce more news than those from center-right. As discussed elsewhere (Kawakami et al. 2020), what is perceived as center-left are the main national news outlets in the country that still follow the "fairness doctrine."

[19]https://www.texastribune.org/about/

Domain	# Queries	# Links	Total Occ.
washingtonpost.com	800	475	196,071
nytimes.com	747	364	218,586
npr.org	744	246	202,180
cnn.com	698	423	224,286
politico.com	601	217	97,371
cnbc.com	573	99	109,325
nbcnews.com	536	245	76,957
cbsnews.com	486	365	60,632
reuters.com	470	169	62,230
texastribune.org	469	60	145,470

Table 4: The top 10 domains appearing as Top stories. They occurred as Top stories in all 65 locations, produced 15% of all unique Top stories links, and constitute 49.7% of all Top stories. All these domains are news organizations.

Domain	# Queries	# Links	Total Occ.
npr.org	897	270	658,016
nytimes.com	787	253	424,420
washingtonpost.com	729	274	346,189
cnn.com	709	235	418,613
en.wikipedia.org	676	257	326,943
politico.com	657	211	562,565
supremecourt.gov	563	117	547,204
cnbc.com	538	82	275,547
guttmacher.org	500	97	288,705
texastribune.org	492	108	250,020

Table 5: The top 10 domains appearing as organic results. These 10 domains occurred as organic results in all 65 locations, produced 9.8% of all unique organic result links, and constitute 30.7% of all organic results. Seven out of these ten domains belong to news organizations.

Organic Results: Table 5 summarizes the top ten domains among organic results. Notably, seven of the top ten Top stories domains also appear among the top ten organic results domains; the top organic results domains also see the addition of the Supreme Court's government website, Wikipedia, and the Guttmacher Institute, a leading sexual and reproduc-

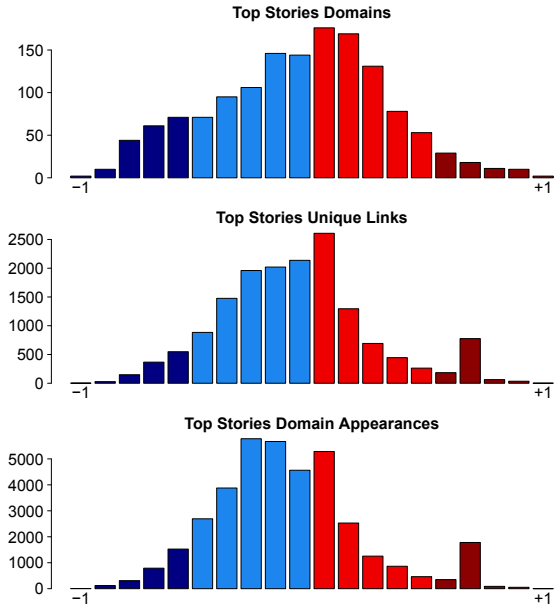

Figure 5: The distribution of Top stories domains, Top stories unique links, and Top stories appearances by Partisan Audience Bias score. Although Google samples from an almost normal-distribution-like set of news sources (top graph), they do not produce news at the same pace, leading to more news stories from the center-left news outlets.

tive health policy and research organization. While these domains produced just under 10% of all unique organic links, they constitute over 30% of all organic results.

While we are similarly interested in the political leaning of websites that constitute organic results, the PAB dataset only includes a score for 1,529 out of 5,180 domains (29.5%) that appeared as organic results. Future work with our dataset should further investigate characteristics of organic results, including political leaning.

In addition to political leaning, localization of organic results is also of interest: to what extent are Google's organic results localized to a user's location? To begin to explore this question, we calculated the number of locations in which each unique organic link appeared. Figure 6 visualizes the results. It is clear that local and national sources compose the majority of organic results, as 40% of organic results appear in five or fewer locations, and 33% of organic results appear in all 65 locations. Some methods for identifying and measuring the locality of search results have been developed, such as (Hagar et al. 2020). In future work, we will measure the locality of search results and quantify the extent to which localized search results match the locale of the search.

Other Uses for the Dataset

In the following, we describe potential research directions and questions that can be explored using our dataset.

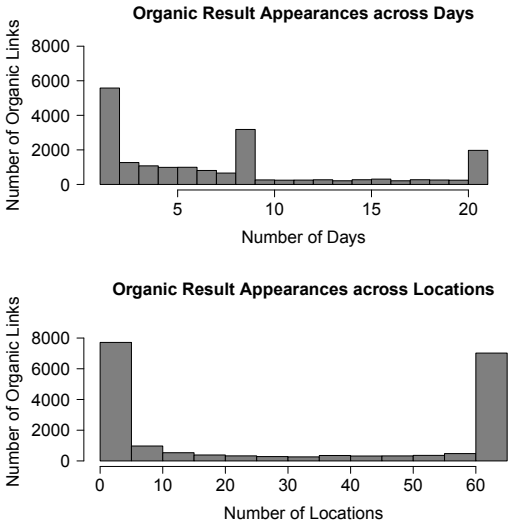

Figure 6: Distributions show how organic results appeared across dates and locations. Notably, the majority of organic results appeared in less than 5 locations or all 65 locations, indicating that most organic results are either local or national in nature.

Qualitative Analysis of News Coverage

Our dataset captures a snapshot in time that is otherwise unavailable and can be used to analyze how national and local news sources discussed the *Dobbs v. Jackson Women's Health Organization decision*. Given that search interest for abortion-related queries spiked drastically following the *Dobbs v. Jackson Women's Health Organization* decision (Figure 1), our dataset can help researchers analyze the content that average Google Search users encountered when searching for abortion during this time. Moreover, given that our dataset includes 65 locations across the United State, diverse in their political leaning and legal or social stance on abortion, it allows for comparison of results across locations and has potential to give an insight into how Google Search facilitates access to abortion-related information in different locations. More generally, our dataset can be used to discover the nature of event-driven content generated by local and national news outlets and other participants in the online information ecosystem.

Impact of Query Formulation on Google Search Results

How does query formulation impact Google Search results? While the top domains among Top stories and organic results appeared for a large portion of our query list, 91.6% of organic results appeared for five or fewer queries, with 67% of organic results appearing for only one query. What features of a query phrase impact the composition of Search results? How do the search results for query "abortion" differ from the search results for the query "abortion near me"? With 1,698 queries from a variety of sources, our dataset provides ample opportunity to explore such questions.

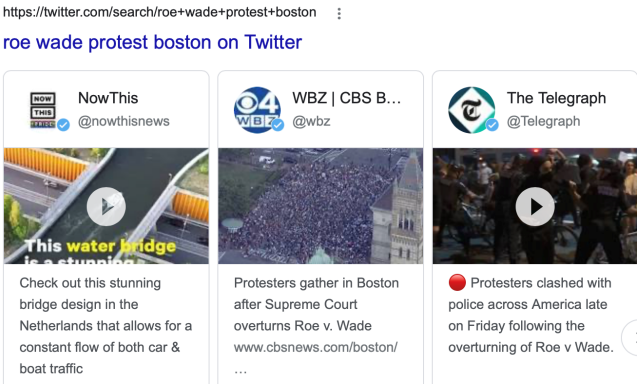

Figure 7: A small proportion of the collected SERPs contain embedded tweets. The only information we have scraped are tweet IDs, so that adopters of this dataset can check with the Twitter API if these tweets are still publicly available.

Rate of Change of Search Results

Figure 6 shows that some organic links only appeared on a few days, while others appeared for all 21 days of data collection. At what rate does Google update its organic results following an impactful real-world event that is covered widely? Which links and domains are "sticky," or, in other words, which links and domains consistently appear in Google Search results, and which are replaced rapidly? We hypothesize that increased coverage of a topic following an impactful event will cause Google to more rapidly change the content it shows in Search results for relevant queries. Given that our dataset contains 26 snapshots of information across 21 days, statistics such as day-to-day changes in organic results can be calculated to explore the rate at which Google Search results change. Here also, the impact of query formulation can be considered: for what queries do organic results change more frequently than others?

Ethical Considerations and FAIRness

Our dataset contains HTML pages that are automatically generated by Google Search algorithms in response to supplied search phrases. Google handles billions of such searches daily and the corresponding web pages with the results are not considered content that belongs to Google. Furthermore, courts have ruled that the use of text snippets or images from the indexed websites (or other copyrighted work) within Google's results is "fair use," in accordance with U.S. copyright laws.[20] In rare instances, when Google receives legitimate requests for copyright infringement, it removes a link to the infringing website from the results and replaces its snippet with a copyright infringement notice. If this happens in the future for the pages we have collected, we will not be able to know or take actions. However, given that our dataset will not be used to visit any websites, we hope that this risk is small.

[20]https://www.flaglerlawgroup.com/a-new-era-for-fair-use-court-changes-fair-use-law-in-google-decision/

Some search pages (less than 2%) contain a panel composed of fresh tweets from Twitter, an example of which is shown in Figure 7. As it can be noticed, these tweets contain the names of the accounts from which they were sent. While often tweets are from news organizations, they also include tweets from regular users. If some of these tweets are deleted in the future by the sender, a copy will remain in these pages. Among our parsed results, we only provide the list of the tweet IDs contained in the "raw data', but no user-specific information. We request that whoever wants to use the tweets from the SERPs should use the Twitter API to find out if these tweets are still publicly available, before scraping them from the SERPs.

Our dataset is hosted on the Harvard Dataverse, making it findable and accessible. We have provided CSV files capturing the content of the "raw data," making the dataset interoperable. Our paper clearly describes the processes followed to collect and extract the data, making it reusable. As such, we believe that we follow the FAIR principles.[21]

Limitations

There are a few limitations to our dataset. The query list was automatically extracted from Google Trends and Google Autosuggest, including queries that are not always relevant to the topic we are interested in. Additionally, because we augmented our query list in the middle of data collection, we miss a key period of time, the immediate aftermath of the Dobbs decision, for 694 queries that were added on July 8. We collected data for 65 locations, a dataset which is sufficient for statistical analyses, but is still small with respect to the size of the United States. We did not test that changing the location using coordinates exactly simulates physically being in a location; however, we did ensure that the location for each SERP was changed correctly, and we tested and manually validated that local results were returned for localized queries (such as test queries for "pizza" and "taxes near me"), consistent with prior studies (e.g., (Mejova, Gracyk, and Robertson 2022)). Lastly, due to the large amount of space needed to store the results as HTML pages, the files were periodically moved from the seven computers used for the collection to a central server with more available space. This caused the timesteamps of the file creation to be lost, so we do not have the exact timestampe (in terms of hours and minutes) that each SERP was collected.

Conclusion

We present and share publicly a dataset of more than 1.74 million Google SERPs collected in the aftermath of a historic event for the United States, the U.S. Supreme Court's overturning of a 50-year constitutional right to abortion. The dataset comprises results for 1,698 search phrases that were searched daily on 65 U.S. locations during a 21-day period. A preliminary analysis of the dataset reveals that the results were dominated by links from news organizations, which appeared both in Top stories and organic search results. We believe that the dataset can be useful to researchers interested in how Google's algorithms shape the online news

[21]https://www.go-fair.org/fair-principles/

ecosystem by choosing some domains over others. Additionally, the dataset contains a set of 17,503 unique news article links that appeared in Top stories, which will be of interest to communication and media scholars studying the coverage of the *Dobbs v. Jackson Women's Health Organization* decision.

Acknowledgments

We are grateful to the members of the Wellesley Cred Lab for their support, and acknowledge funding from the National Science Foundation, under grant IIS 1751087.

References

Bandy, J. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5(CSCW1): 1–34.

Diakopoulos, N.; Trielli, D.; Stark, J.; and Mussenden, S. 2018. I Vote For—How Search Informs Our Choice of a Candidate. In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, 121–133. Springer.

Diaz, A. M. 2008. *Through the Google Goggles: Sociopolitical Bias in Search Engine Design*, 11–34. Berlin, Heidelberg: Springer Berlin Heidelberg.

Elkin-Koren, N. 2000. Let the Crawlers Crawl: On Virtual Gatekeepers and the Right to Exclude indexing. *U. Dayton L. Rev.*, 26: 179.

Gasser, U. 2005. Regulating search engines: Taking stock and looking ahead. *Yale JL & Tech.*, 8: 201.

Gillespie, T. 2010. The politics of 'platforms'. *New Media & Society*, 12(3): 347–364.

Gillespie, T. 2017. Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information, Communication & Society*, 20(1): 63–80.

Grimmelmann, J. 2008. The Google Dilemma. *NYL Sch. L. Rev.*, 53: 939.

Hagar, N.; Bandy, J.; Trielli, D.; and Wang, Y. 2020. Defining Local News: A Computational Approach. In *Proceedings of the Computation + Journalism Symposium*.

Hannak, A.; Sapiezynski, P.; Molavi Kakhki, A.; Krishnamurthy, B.; Lazer, D.; Mislove, A.; and Wilson, C. 2013. Measuring personalization of web search. In *Proceedings of the 22nd WWW International Conference*, 527–538.

Hargittai, E. 2007. The Social, Political, Economic, and Cultural Dimensions of Search Engines: An Introduction. *Journal of Computer-Mediated Communication*, 12(3): 769–777.

Hu, D.; Jiang, S.; E. Robertson, R.; and Wilson, C. 2019. Auditing the partisanship of Google search snippets. In *The World Wide Web Conference*, 693–704.

Introna, L. D.; and Nissenbaum, H. 2000. Shaping the Web: Why the Politics of Search Engines Matters. *The information society*, 16(3): 169–185.

Kawakami, A.; Umarova, K.; Huang, D.; and Mustafaraj, E. 2020. The'Fairness Doctrine'Lives on? Theorizing about the Algorithmic News Curation of Google's Top Stories. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, 59–68.

Kawakami, A.; Umarova, K.; and Mustafaraj, E. 2020. The Media Coverage of the 2020 US Presidential Election Candidates through the Lens of Google's Top Stories. In *Proceedings of the 14th AAAI ICWSM*, volume 14, 868–877.

Kliman-Silver, C.; Hannak, A.; Lazer, D.; Wilson, C.; and Mislove, A. 2015. Location, location, location: The impact of geolocation on web search personalization. In *Proceedings of the 2015 internet measurement conference*, 121–127.

Lurie, E.; and Mustafaraj, E. 2019. Opening Up the Black Box: Auditing Google's Top Stories Algorithm. In *32th AAAI FLAIRS*.

Lurie, E.; and Mustafaraj, E. 2020. Highly Partisan and Blatantly Wrong: Analyzing News Publishers' Critiques of Google's Reviewed Claims. In *Proceedings of the 2020 Truth and Trust Online Conference*, TTO '20, 64–72.

Mejova, Y.; Gracyk, T.; and Robertson, R. E. 2022. Googling for Abortion: Search Engine Mediation of Abortion Accessibility in the United States. *Journal of Quantitative Description: Digital Media*, 2.

Metaxa, D.; Park, J. S.; Landay, J. A.; and Hancock, J. 2019. Search Media and Elections: A Longitudinal Investigation of Political Search Results. 3(CSCW).

Metaxa, D.; Park, J. S.; Robertson, R. E.; Karahalios, K.; Wilson, C.; Hancock, J.; Sandvig, C.; et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, 14(4): 272–344.

Metaxas, P. T.; and Pruksachatkun, Y. 2017. Manipulation of search engine results during the 2016 US congressional elections. In *Proceedings of the 12th ICIW*.

Mulligan, D. K.; and Griffin, D. S. 2018. Rescripting Search to Respect the Right to Truth. *Georgetown Law Technology Review*, 2(2): 557–584.

Pariser, E. 2011. *The Filter Bubble: How the new personalized web is changing what we read*. Penguin.

Pasquale, F. 2006. Rankings, reductionism, and responsibility. *Clev. St. L. Rev.*, 54: 115.

Robertson, R. E.; Jiang, S.; Joseph, K.; Friedland, L.; Lazer, D.; and Wilson, C. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 148.

Robertson, R. E.; Lazer, D.; and Wilson, C. 2018. Auditing the personalization and composition of politically-related search engine results pages. In *WWW'18*, 955–965.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22.