Measuring the Filter Bubble: How Google is influencing what you click

GOOGLE HAS OVER 2000 TECHNOLOGY TRICKS TO CHANGE YOUR MOOD, YOUR MIND, YOUR POLITICS AND YOUR SEXUAL PERCEPTIONS

Filed under Privacy Research

Over the years, there has been considerable discussion of Google's "filter bubble" problem. Put simply, it's the manipulation of your search results based on your personal data. In practice this means links are moved up or down or added to your Google search results, necessitating the *filtering* of other search results altogether. These editorialized results are informed by the personal information Google has on you (like your search, browsing, and purchase history), and puts you in a *bubble* based on what Google's algorithms think you're most likely to click on.

The filter bubble is particularly pernicious when searching for political topics. That's because undecided and inquisitive voters turn to search engines to conduct basic research on candidates and issues in the critical time when they are forming their opinions on them. If they're getting information that is swayed to one side because of their personal filter bubbles, then this can have a significant effect on political outcomes in aggregate.

Back in 2012 we ran a study showing Google's filter bubble may have significantly influenced the 2012 U.S. Presidential election by inserting tens of millions of more links for Obama than for Romney in the run-up to that election. Our research inspired an independent study by the Wall Street Journal (paywall):

> A Wall Street Journal examination found that the search engine often customizes the results of people who have recently searched for "Obama"—but not those who have recently searched for "Romney."

Now, after the 2016 U.S. Presidential election and other recent elections, there is justified new interest in examining the ways people can be influenced politically online. In that context, we conducted another study to examine the state of Google's filter bubble problem in 2018.

Summary of Findings

Google has claimed to have taken steps to reduce its filter bubble problem, but our latest research reveals a very different story. Based on a study of individuals entering identical search terms at the same time, we found that:

1. Most participants saw results unique to them. These discrepancies could not be explained by changes in location, time, by being logged in to Google, or by Google testing algorithm changes to a small subset of users.
2. On the first page of search results, Google included links for some participants that it did not include for others, even when logged out and in private browsing mode.
3. Results within the news and videos infoboxes also varied significantly. Even though people searched at the same time,

people were shown different sources, even after accounting for location.
4. Private browsing mode and being logged out of Google offered very little filter bubble protection. These tactics simply do not provide the anonymity most people expect. In fact, it's simply not possible to use Google search and avoid its filter bubble.

 Visualization of three differing Google search results for the query 'gun control'.

For those interested in more details, we've written out everything below, as well as provided the underlying data and code. We hope this work encourages further study of this important issue.

Methodology

We asked volunteers in the U.S. to search for "gun control", "immigration", and "vaccinations" (in that order) at 9pm ET on Sunday, June 24, 2018. Volunteers performed searches first in private browsing mode and logged out of Google, and then again not in private mode (i.e., in "normal" mode). We compiled 87 complete result sets — 76 on desktop and 11 on mobile. Note that we restricted the study to the U.S. because different countries have different search indexes.

During analysis of the search results, we only looked at websites' top-level domains, for example *www.cdc.gov/features/vaccines-travel* and *www.cdc.gov/vaccines/adults* would both be treated as just *cdc.gov*.

Finding #1: Most people saw results unique to them, even when logged out and in private browsing mode.

To count variants of results, we noted the order of the major elements: the organic (regular) links, the news (Top Stories) infobox, and the videos infobox. We ignored ads, sections containing related searches, and other infoboxes. There were variations in these too, but we didn't consider them.

A quick note on ordering of links: You might think that as long as the same links are shown to users, the ordering of them is relatively unimportant, but that's not the case. A given link gets only about *half as many clicks* as the link before it and *twice as many clicks* as the link after it. In other words, link ordering matters a lot because people click on the first link much more than the second, and so on.

The amount of variations we saw for each search term is listed below. For this part of the study, we excluded mobile results because the number of infoboxes displayed can vary significantly between mobile and desktop. That's why it says 76 participants instead of the overall total of 87. We also controlled for location (more on that below).

Private browsing mode (and logged out):

- "gun control": 62 variations with 52/76 participants (68%) seeing unique results.
- "immigration": 57 variations with 43/76 participants (57%) seeing unique results.
- "vaccinations": 73 variations with 70/76 participants (92%) seeing unique results.

Normal mode:

- "gun control": 58 variations with 45/76 participants (59%) seeing unique results.
- "immigration": 59 variations with 48/76 participants (63%) seeing unique results.
- "vaccinations": 73 variations with 70/76 participants (92%) seeing unique results.

 Visualization of unique search results shown for the search query 'gun control'.

With no filter bubble, one would expect to see very little variation of search result pages — nearly everyone would see the same single set of results. That's not what we found.

Instead, most people saw results unique to them. We also found about the same variation in private browsing mode and logged out of Google vs. in normal mode.

Now, some search result variation is expected due to two factors that we controlled for. First, search results can change over time, such as the inclusion of time-sensitive links. We controlled for this factor by having everyone search at the same time.

Second, search results can change by location, such as the inclusion of local news articles. We controlled for this factor by checking all links by hand for this possibility, comparing them to the city and state of the volunteer. We saw very few local links for *gun control* (1 organic link, 1 news infobox link) and *immigration* (0), though more for *vaccinations* (15 organic links, 4 news infobox links).

To control for these local links, we replaced all of them with the same placeholder — localdomain.com for organic links and "Local Source" for infoboxes — in all of our analysis. This adjustment means two users whose results only differed by a different local domain in the same slot would not count as different. Interestingly, this adjustment didn't affect overall variation significantly.

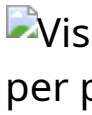
Another reason you might expect some variation is testing of the search algorithm, where you show slightly different results to different people. In that case, you'd expect to see most people seeing the same results, with a few people seeing slight differences. What we saw, by contrast, was most people seeing different results.

Finding #2: Google included links for some participants that it did not include for others.

Google search results typically have ten organic links. While the ordering of those links really matters (i.e. link #1 gets ~40% of clicks, link #2 ~20%, link #3 ~10% and so on), we also wanted to know how many different domains were being displayed.

With no filter bubble, one would expect to see this total to be around ten. We saw significantly more. In private browsing mode, logged out of Google, and with local domains replaced with *localdomain.com*, here are the totals:

- "gun control": 19 different domains
- "immigration": 15 different domains
- "vaccinations": 22 different domains

 Visualization of domains appearing in organic search results per person.

As you can see this clearly in the visualization above, some people were shown a very unusual set of results relative to the other participants, offered some domains seen by no-one else. If you were one of these people, you would have no way of knowing what you're missing.

Finding #3: We saw significant variation within the News and Videos infoboxes.

We also wanted to look at variation within the news (Top Stories) and videos infoboxes. We also saw significant variation within those, even though there are only three slots available. Again, these are for private browsing mode, logged out of Google, and with local domains replaced with "Local Source".

News infobox:

- "gun control": 3 variations from 5 sources, appearing for 75/76 people. The most common variation was seen by 69 people (90%).
- "immigration": 6 variations from 7 sources, appearing for 76/76 people. The most common variation was seen by 35 people (46%).
- "vaccinations": 2 variations from 3 sources, appearing for 2/76 people. Each variation was seen by one person (1%).

Videos infobox:

- "gun control": 12 variations from 7 sources, appearing for 75/76 people. The most common variation was seen by 24

people (32%).
- "immigration": 6 variations from 6 sources, appearing for 75/76 people. The most common variation was seen by 42 people (55%).
- "vaccinations": Not shown in the search results.

As an example, the Videos infobox for the "immigration" query showed the following six variations. As with organic search results, the ordering matters here because the second and third slots get far fewer clicks.

- Today, MSNBC, NBC News (shown to 42 participants)
- MSNBC, Today, NBC News (shown to 26 participants)
- Today, MSNBC, MSNBC (shown to 4 participants)
- MSNBC, Today, Today (shown to 1 participant)
- New York Times, CNN, MSNBC (shown to 1 participant)
- Today, MSNBC, RealClearPolitics (shown to 1 participant)

Remember, we had people search at the same time, and we changed all local-links to the be same, so this variation is not explained by time or location. And again, some people were real outliers; in fact, some didn't see the infoboxes at all.

Finding #4: Private browsing mode and being logged out of Google offered almost zero filter bubble protection.

Finally, we saw the variation in private browsing mode (also known as incognito mode) and logged out of Google as about the same as in normal mode. Most people expect both being logged out and going "incognito" to provide some anonymity. Unfortunately, this is a common misconception as websites use

IP addresses and browser fingerprinting to identify people that are logged out or in private browsing mode.

If search results were more anonymous in these states, then we would expect everyone's private browsing mode results to be similar. That's not what we saw.

To test this more rigorously, we took the organic results, excluding ads and infoboxes, and:

1. Assigned each domain a letter (e.g. A for nytimes.com, B for wsj.com, etc.).
2. Made a string of letters for each person's results, e.g. ABDFJKMSL.
3. Compared these strings to see how similar they were to each other.

To do this comparison we counted domain changes between different sets of search results, reducing the differences to a number. For example, ABC -> ACB is one change. (Technically, we used a letter to represent each domain within each search result and calculated the Damerau-Levenshtein edit distance between them.)

 Visualization showing how edit distances are calculated to measure the difference between strings.

We saw that when randomly comparing people's private modes to each other, there was more than double the variation than when comparing someone's private mode to their normal mode:

gun control:

- Average of normal and private browsing mode (same user): 1.03
- Average of private browsing mode (random user): 2.89
- Average of private browsing mode (five closest users): 2.65

immigration:

- Average of normal and private browsing mode (same user): 1.38
- Average of private browsing mode (random user): 3.28
- Average of private browsing mode (five closest users): 2.80

vaccinations:

- Average of normal and private browsing mode (same user): 2.23
- Average of private browsing mode (random user): 4.97
- Average of private browsing mode (five closest users): 4.25

Visualization showing that there's little difference in results between searching in normal mode and private browsing mode.

We often hear of confusion that private browsing mode enables anonymity on the web, but this finding demonstrates that Google tailors search results regardless of browsing mode. People should not be lulled into a false sense of security that so-called "incognito" mode makes them anonymous.

Study Data and Code

The data is available for download in two parts: Basic non-identifiable participant data, and raw data from the search results.

- *duckduckgo-filter-bubble-study-2018_participants.xls* contains the instructions we sent to each participant, as well as basic anonymized data for each participant.
- *duckduckgo-filter-bubble-study-2018_raw-search-results.xls* contains a separate sheet for search results per query and per mode (private and non-private). The results are listed as they appeared on the screen for each participant, showing both organic domains and infoboxes such as Top Stories (news), Videos, etc.

The code that we wrote to analyze the data is open source and available on our GitHub repository.

For more privacy advice, follow us on Twitter & get our privacy crash course.