Topical biases in local news curation: an audit of Google News

Sean Fischer[*]
University of Pennsylvania
sean.fischer@asc.upenn.edu

Kokil Jaidka
National University of Singapore

Yphtach Lelkes
University of Pennsylvania

ABSTRACT

Local news outlets have struggled to stay open in the more competitive market of digital media. The factors contributing to this struggle are not entirely clear. Demand-side preferences certainly play a role, but supply-side decisions may also divert readership in ways that are harmful to local news outlets. To gain a better understanding of how one major kind of gatekeeper – an online news aggregator – may be affecting the ability of media consumers to access local news outlets, we conduct an audit of Google News. We find evidence that the amount of local news returned by Google News depends heavily on the actual query used and not geographic or market-specific features.

CCS CONCEPTS

• **Information systems → Web and social media search**; *Page and site ranking*;

KEYWORDS

local news, search platforms, search bias

ACM Reference format:
Sean Fischer, Kokil Jaidka, and Yphtach Lelkes. 2019. Topical biases in local news curation: an audit of Google News. In *Proceedings of Computation + Journalism 2020, Boston, MA, 2019,* 5 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The widespread decline of local news has been associated with changes in the advertising business brought about by online news consumption [4]. Almost 90% of Americans get some of their local news digitally, with almost half of local news consumption occurring on mobile devices. Digital consumption has also nearly surpassed television as the preferred medium for accessing local news (37% vs. 41%) [8]. Because the Internet offers a broad set of media options, individuals rely on cues from platforms like search engines, which are now central conduits through which individuals access news on the Internet. Almost a quarter of traffic to online-news sites comes from search engines [2], and it falls on them to rank and prioritize news outlets on the World Wide Web to help their users determine what content to consume [7]. By highlighting specific stories and outlets over others, these platforms can operate as gatekeepers to advertising dollars.

In this paper, we take a first step in assessing whether an online news aggregator platform – specifically, Google News – might contribute to the observed declines in local news. It might do so by favoring large national outlets at the expense of smaller local outlets, or by varying the local-national composition of results varies according to local socioeconomic factors. To systematically audit whether (the decline in) local news' online traffic can be attributed to Google News, we must show that Google News is more likely to feature national news outlets than local news outlets (Hypothesis 1). If Google's selection of news outlets overwhelmingly favors national newspapers, then we have evidence that Google could be reducing the traffic to local news webpages, and potentially reducing the likelihood that readers would be exposed to local outlets.

However, individuals' browsing behavior can be confounded by other factors, such as their search interests, as we have seen in other audits of Google and its search platform. We may find that depending on the nature of the search query, Google News may appear to prioritize national news outlets for topics of national importance, and local news outlets for hyperlocal issues (Hypothesis 2). If that were to be the case, this would lead to concerns about essential local information gaps in regards to political issues and policy, as well as the appropriateness of national newspapers for reporting on and following stories of local interest.

Recent speculation has also suggested that Google News results are sorted in response to local demands (Hypothesis 3). Evidence of this would be a geographical or a demographic variation in the proportion of local or national outlets returned in search results. If our analysis finds this

[*]All authors contributed equally. Full paper available upon request.

Computation + Journalism 2020, 2019, Boston, MA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

to be true, then it would be credible to suspect that Google News' curation algorithms would promote regional or local filter bubbles of news content.

METHODS AND MATERIALS

Selection of Geographic Units and Search Terms

For this project, we set our geographic area of interest to that of the entire United States. To assess geographic variation, we followed the lead of [6] and considered counties as our unit of interest. For each county in the U.S., we identified the county seat and created a list of latitude-longitude coordinate pairs to be fed to our data collection algorithm.

We wanted our assessment of the variation of Google News search results to cover as many news and policy topics as possible, and be the least biased by the specificity of search terms. Accordingly, we developed two lists of search terms. The first list referred to policy issues, or were associated with current events in the U.S. during the period from January, 2019 to February, 2019. This list comprised *abortion*, *caravan*, *climate*, *conservative*, *corruption*, *election*, *fbi*, *gun*, *immigration*, *liberal*, *politics*, *president*, *scandal*, *shutdown*, *syria*, and *taxes*. The second list was aimed at capturing local heterogenerity and was based on the set of local terms used in [6]. This list comprised *accident*, *college*, *crime*, *death*, *emergency services*, *governor*, *high school*, *hospital*, *mayor*, *obituary*, *police*, *school board*, *traffic*, *transit*, *university*, and *weather*. These data were collected over the course of June, 2019.

Data Collection

We developed a script to collect data using the Python package Selenium. We leveraged Selenium's ability to set the location of the browser in order to collect search results from each county in the United States. Every query was instituted on a new browser window opened in incognito mode. Next, query order was randomized for each search term, so that time-related biases did not confound the local results for each county. Three full sets of results were collected for each search term over the data collection period, which should adjust for any local maxima in news attention over six months, from January, 2019 to June, 2019.

Finally, after instituting all these checks, we repeatedly ran a Google News search for each term in our national and local sets, setting the location to each of the 3,143 counties. The headline, outlet, URL, and timestamp for the first 100 results of each county-specific search were scraped. The county order was randomized.

Coding Outlets

We filtered the collected data to retain the outlets and their position in the search results, which we refer to as the *rank* of the result. We then classified the outlets as being local, regional, national, or international media outlets by delineating websites by the scope of their reporting. Outlets covering stories in small finite areas would be classified as local outlets, while those covering stories in larger, but still finite areas would be classified as regional outlets. National outlets would be those located in the United States covering news across multiple areas of the United States.

Many of our outlet classifications came from data provided by the SMaPP Lab at NYU [10]. The outlets included in this dataset were scraped from databases maintained by the United States Newspaper Listing and Station Index. This dataset identified local newspapers, magazines, and broadcast television stations across the United States. However, given the variety of outlets observed during our searches, the dataset provided minimal coverage of our outlets. To supplement this existing resource, we performed a classification procedure on Amazon Mechanical Turk (AMT) for 1,078 additional outlets. As part of our procedure, we showed AMT workers an unclassified outlet, linked them to a general Google search for the outlet, and asked them to decide whether the outlet was local, regional, national, or international media. This classification question was split into logical branches, so that workers first identified whether the given outlet was located in the United States, then, if so, classified the outlet as either local, regional, or national. If the outlet was located outside the United States, workers identified the country within which the outlet was located. We collected classifications for each outlet from two workers. The inter-coder agreement was about 73%. If the workers agreed, the outlet was added to the database with that classification. In cases where the workers disagreed, instead of forcing hard boundaries, we logged both classifications as a boundary category, like local-regional or regional-national.

There are many potential ways to define what constitutes a local news outlet, and this decision framework impacts all of our results. Our definition focuses on the scope of coverage, but in many cases, this framework produces disagreement, as when dealing with major metro papers, like *The Boston Globe*, *Houston Chronicle*, and *LA Times*. These papers were classified as being local or regional, but under definitions based on circulation or readership, could be seen as national outlets. Our approach has the advantage that it is based on the actual publishing decisions and intentions of the outlets. This allows us to separate the metro papers listed above from *The New York Times* and *Washington Post*, both of which ostensibly serve as metro papers, but devote substantial attention and resources to covering national political news.

RESULTS

Distribution of Returned Outlets

An important characteristic of the results of our Google News searches is the distribution of outlet occurrences, which is heavy-tailed. The three most frequently featured outlets are the *Washington Post*, *New York Times*, and *CNN*, each of which had over 750,000 appearances and which together account for 17.22% of stories across all searches (N = 2,435,956, $N_{WaPo} = 851,145$, $N_{NYT} = 809,599$, $N_{CNN} = 775,212$). The next closest outlet, *The Guardian*, was returned only just over 500,000 times. Of the rest of the top-10 most frequently returned outlets, most are nationally-oriented American outlets and are generally clustered around 250,000 occurrences. The majority of outlets, though, were returned very few times, with this distribution having a median of 1,648 occurrences across all searches, compared to mean of 9,915 occurrences. The most heavily populated portion of the distribution is in the range from 1 occurrence to 5 occurrences, with 45 outlets falling in this range; the next closest (2-6) has only 32 outlets.

In order to compare to prior results of outlet concentration, we calculated the Gini Index for this distribution. The Gini Index is a measure of inequality that ranges between 0 and 1, with low levels indicating equal distribution across outlets and high levels indicating more inequality between outlets. In our context, high Gini Indices indicate that a few outlets are returned much more frequently than all others in a set of results. For the whole set of aggregated results, we find that the Gini Index signals a very high level of inequality ($G = 0.83$). This value is in line with the highest levels of concentration observed in other search audits [9]. These results, then, provide evidence in support of Hypothesis 1, that Google News is favoring national outlets at the expense of local outlets, as we see a high degree of concentration in which outlets are commonly returned.

As visualized in Figure 1, the variation across terms is minimal. We can see between the two panels that the set of local terms appears to produce a lower level of inequality, on average, than the set of national terms. This intuition is confirmed via a one-tailed Wilcoxon test ($W = 24, p = 1.81\times10^{-5}$). However, it is not obvious from these results whether the levels of concentration are substantively different between local and national terms, leaving us with inconclusive initial evidence for Hypothesis 2.

Composition of Search Results by Query

While the level of inequality in the number of times an outlet is returned may not be substantively different between sets of queries, the same might not be valid for the composition of the returned results. That is, queries may produce sets of results that vary quite a bit in the volume of local news returned, even if the same few national outlets appear across all sets of results and produce the observed levels of inequality. Based on previous audits of Google's search platform, we expect Google News to return more local media outlets when users search for news about topics of local interest and national media outlets when users search for news about national issues [3, 9].

This pattern of results is precisely what we find. As we can see in Panels A and B of Figure 2, there is a substantial difference in the prevalence of classes between the sets of local and national search terms. The cases at the extremes provide additional evidence for this mechanism: *obituary* produces sets of results drawn almost entirely from local media outlets, while *shutdown* produces almost entirely national results.

What is most interesting, though, are the cases in the middle. Searches for *police*, *college*, and *gun* produced results notably different from the average for their set of terms. *Police*, *crime*, and *gun* all produced results that were close to achieving equal shares of local and national outlets. *College*, a term we anticipated as being a locally oriented term, produced results composed of over 50% national outlets and only about 25% local outlets.

Even with these in-between cases, the two sets of terms are statistically different in their average composition. The set of local terms return local news outlets at a rate that is 36 percentage points higher than the national terms ($p = 2.63 \times 10^{-7}$). On the other hand, local terms return results that feature national outlets at a rate of 37 percentage points lower than the national-term results ($p = 1.82 \times 10^{-7}$).

These results, though, provide credible support for our second prediction: that the choice of the query will affect the composition of results. We can see that depending on a user's interests, the composition of results, in regards to the volume of local and national news outlets, will change in dramatic and substantive ways. In turn, we can expect users to be nudged towards differing classes of outlets based on the topic of their search.

Spatial variation and determinants of local news

Given that we have already seen that the variation in the types of outlets returned in Google News search results appears to be related to the term being queried, we should also consider whether there is evidence for similar variation across geography. Prior audits have found evidence of geographic effects [3, 6]. To test this possibility, we first calculated the percentage of results from local or regional outlets for each county in the United States.

However, there is minimal variation between counties in the amount of local or regional outlets returned. The unweighted percentage of stories coming from local or regional outlets ranges between 41% and 46%. If we weight results by their rank on the results page, then the percentage range

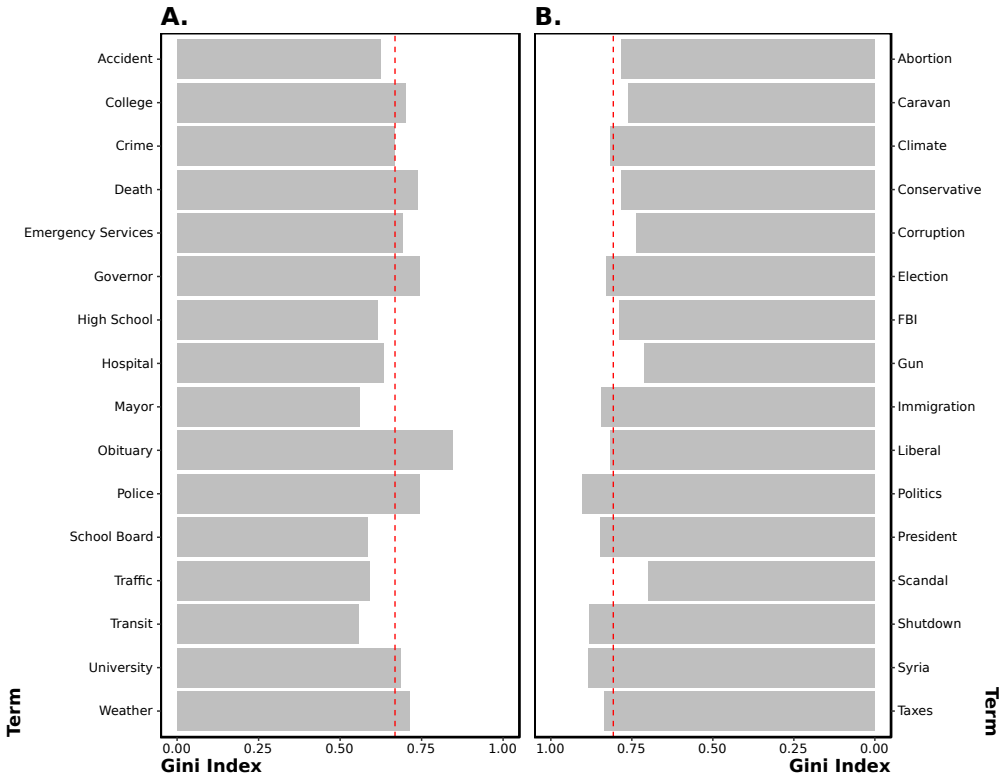

Figure 1: A. The Gini Index for each local search term. The dashed red line marks the mean value. B. The Gini Index for each national search term. The dashed red line marks the mean value.

shifts from 31% to 41%. Further evidence for a lack of geographic effects comes from the fact that the intercounty correlation is effectively zero (Moran's I = 0.0082, $p = 0.22$), indicating that the minimal variation observed does not follow a spatial pattern across the country.

Even with no evidence of spatial patterns, geography may play a role in shaping results. Studies of news production have found that media located in state capitals have a higher likelihood of focusing on state politics in their reporting [5]. Such a difference may extend to the results that Google News returns and lead to more non-national outlets appearing in results. However, we find weak evidence to support this specific conjecture. Results for searches located in counties with a state capital produced sets of results with local or regional outlets making up 0.21 percentage points less of the returned results than those from counties without state capitals ($p = 0.047$). In comparison, the effect size produced by a change in search term ranges from only .10 percentage points to 86.12 percentage points.

We expected that for those counties which are already classified as news deserts, or have a small, non-zero number of local newspapers [1], Google News would have the least likelihood of returning local news results. Nevertheless, our result suggests that Google News did not take such factors into account when serving results. Instead, local news stories are shared between counties.

These results undermine our Hypothesis 3 that geography and local community features would meaningfully affect the composition of results returned by Google News. Counties that encompassed a state capital did not produce a meaningfully different composition of search results, even though we had reason to expect that local news would benefit from an increased focus on state politics and related issues. Furthermore, there were no consistent patterns of high or low levels of local news in results across the different counties.

Models regressing the occurrence of local or regional outlets on on its rank in the search results, and statistics both about the number of local newspapers being published, as well as the county demographics under consideration confirmed the results presented here that the type of query is the only substantial influence on the composition of search results.

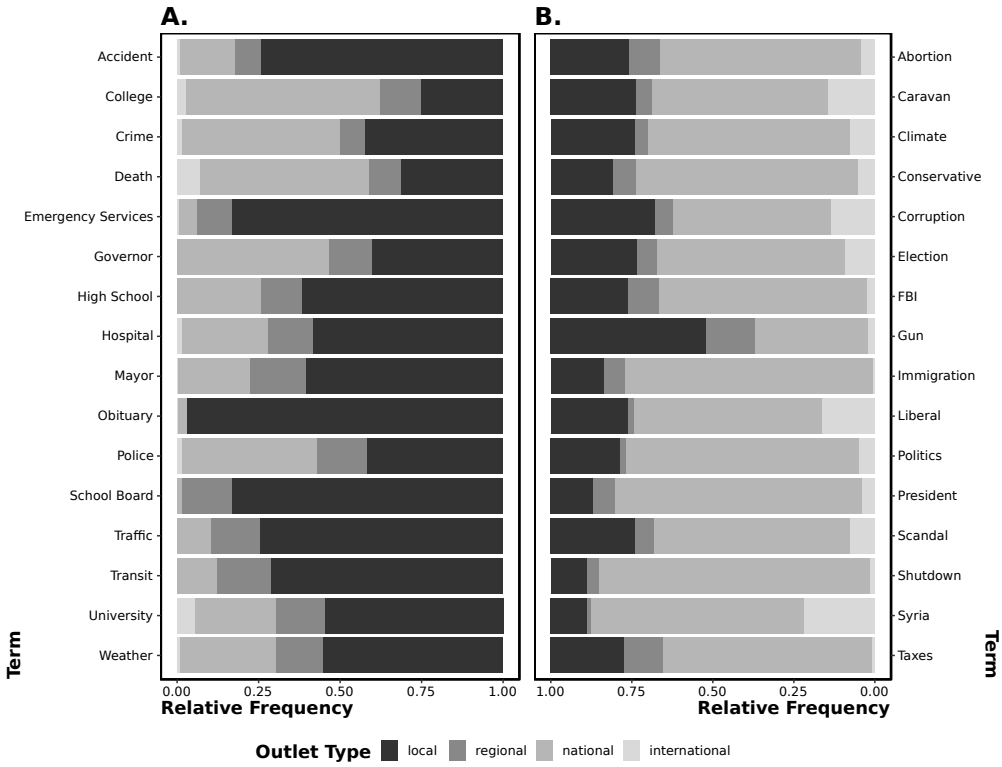

Figure 2: A. Relative frequencies of local, regional, national, and international outlets for local search terms. B. Relative frequencies of local, regional, national, and international outlets for national search terms.

REFERENCES

[1] Penelope Muse Abernathy. 2018. *The Expanding News Desert.* resreport. UNC: The Center for Innovation and Sustainability in Local Media.

[2] Nicholas Diakopoulos. 2019. Audit suggests Google favors a small number of major outlets. *Columbia Journalism Review* (May 2019). https://www.cjr.org/tow_center/google-news-algorithm.php

[3] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring Personalization of Web Search. In *Proceedings of the 22nd International World Wide Web Conference (WWW 2013).* Rio de Janeiro, Brazil.

[4] Matthew Hindman. 2018. *The Internet trap: How the digital economy builds monopolies and undermines democracy.* Princeton University Press.

[5] Daniel J Hopkins. 2018. *The increasingly United States: How and why American political behavior nationalized.* University of Chicago Press.

[6] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. 2015. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of the Internet Measurement Conference (IMC 2015).* Tokyo, Japan.

[7] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In Google we trust: UsersâĂŹ decisions on rank, position, and relevance. *Journal of computer-mediated communication* 12, 3 (2007), 801–823.

[8] Pew Research Center. 2019. *For Local News, Americans Embrace Digital but Still Want Strong Community Connection.* Technical Report. Pew Research Center. 123 pages.

[9] Daniel Trielli and Nicholas Diakopoulos. 2019. Search as News Curator: The Role of Google in Shaping Attention to News Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, 453.

[10] Leon Yin. 2018. Local News Dataset. https://doi.org/10.5281/zenodo.1345145