

# Local news in Google News: a response to Fischer et al. (2020)

Magnusson, Evan

Google LLC

25 February 2021

Online at https://mpra.ub.uni-muenchen.de/106431/ MPRA Paper No. 106431, posted 05 Mar 2021 03:52 UTC

## Local News in Google News: A Response to Fischer et al. (2020)

Evan Magnusson\*

February 2021

#### Abstract

Prior research has shown news aggregators, including Google News, to be beneficial to news outlets. In this article, I analyze why Fischer et al. (2020) produce contradictory results – essentially claiming that local news outlets are disadvantaged in Google News relative to national ones. I find that the authors' Gini coefficient measuring inequality in articles returned for local and national outlets to be highly biased by an error in its computation, and correcting this error reduces the value by up to 80% for some topics. I also show that the authors fail to highlight their own results showing the degree to which Google News returns local outlets in local searches, and that they further contradict their own results with numerous statements in the text.

## 1 Introduction

In Fischer et al. (2020), the authors seek to uncover whether the struggles of local news outlets are due to how, and with what frequency relative to national outlets, digital platforms choose to display their articles. Specifically, the authors take the search results of the aggregator Google News as a case study, querying the site with search terms and analyzing the resulting news articles and their corresponding outlets. Their first finding is a high level of inequality in the distribution of news outlets returned in the search results, implying that Google News is favoring national outlets over local ones. To deepen the analysis, query terms are divided into two groups: local topics (police, weather, emergency services, traffic, etc) and general-interest topics (Syria, election, FBI, abortion, etc). The authors claim that unless local search terms are used, the resulting news articles overwhelmingly come from national news outlets.

<sup>\*</sup>Evan Magnusson is an Economist at Google. The views represented here, however, are personal and do not bind Google. Any errors are my own.

Google News, according to the authors, is thus rerouting ad money from local news to national outlets. They even go as far as to state that "Google News may be directing individuals away from crucial local reporting and, as such, reducing the viability of these local news operations."

The impact of news aggregators on publishers has received great attention from both policymakers and researchers over the course of the past decade. Athey et al. (2017) are able to show that the 2014 shutdown of Google News in Spain reduced news consumption by 20% for the treated group, with a 10% reduction in publishers' page views. These reductions in readership are particularly felt by breaking news, hard news, and news not covered by readers' favorite publishers. Similarly, the 2015 and 2017 NERA reports studying the shutdown of Google News in Spain found that following the event, online publishers lost significant revenue from advertising (a finding corroborated by Calzada and Gil (2020) and others). This loss was particularly heavy for small publishers, with additional negative effects on competition due to the increased barriers to entry as a result of the event. Turning to the USA, Chiou and Tucker (2017) found a decrease in traffic due to a brief suspension of Google News. Thus, it appears that there is substantial empirical evidence that aggregators (in particular Google News) support publishers – and small, local news outlets in particular.

In this context, the conclusions of Fischer et al. (2020) are surprising. If aggregators, namely Google News, are beneficial to local news, how does one make sense of the authors' statement that "platforms may be diverting web traffic and desperately needed advertising dollars away from local news?" In the rest of this paper, I examine the authors' main results and highlight issues in their analysis.

First, I show numerous problems in both the authors' interpretation and computation of outlet inequality. These include: not accounting for differences in article volume due to publisher resources and focus; failing to account for a data corruption issue that undercounts results for over half of their local topic results; and computing a single Gini coefficient over the entire US, rather than averaging over counties. Using the authors' data, after correcting for the last issue, the average Gini coefficient for results from local topics and the average from general interest topics fall by 58% and 44%, respectively (with even more pronounced decreases for particular terms, up to 80%), to levels far below what one would consider harmful inequality. Second, I show that the authors fail to highlight many results from their own analyses of whether Google News is responsive to a user's search intent. Not only are the vast majority of results for local topics arising from local outlets, this is even true (for most local topics) when only looking at the first five results. The authors relegate some of these findings to their Supplementary Information (despite it answering one of their main hypotheses), misdirect readers on the relationship between search depth and local outlet share, and state that "[their] findings suggest that Google News users are exposed to far more national news outlets than local outlets" – even though they show not only that local searches result in chiefly local results, and that local searches may be more frequent than general topic searches. Sadly, many of the authors' results are also irreproducible using their publicly provided code and data.

Indeed, one of the main takeaways from the Fischer et al. (2020) study – for the reader who is attentive to the figures and Supplementary Information (and not only what the authors choose to write about the figures) – is in fact how supportive Google News is of local outlets. From their results, the platform also appears to be highly responsive to individual users' search interest, aiming to return results that match what they are looking for; puzzlingly, the article does not seem to treat this as a good thing. Indeed, the authors seem to advocate for Google News almost always returning local outlets on all topics, regardless of user interest. This fails to account for user reactions: if Google News does not provide the content users ask for, they will likely decrease their use, which will hurt all outlets – local ones inclusive.

In the rest of this article, I go into greater detail on the aforementioned issues with Fischer et al. (2020). I first focus on their Gini coefficient computations in Section 2, then in Section 3 turn to their measures of the responsiveness of Google News to user intent, and then in Section 4 point out a prominent feature on Google News that is composed mainly of local outlets. Section 5 concludes.

## 2 Inequality in Search Results

The authors begin by investigating whether the selection of outlets in Google News' search results is composed mainly of national, rather than local, outlets. To this end, the authors compute the Gini coefficient for the distribution of articles from local and national outlets. The higher the Gini coefficient, the more inequality in the outlets; namely, a few news outlets are dominating Google News results. The authors find an overall Gini coefficient of .825, which they claim supports their hypothesis.

As a brief note, the authors mistakenly refer to their computation of the Gini coefficient as the Gini Index throughout their article. As the Gini Index is 100 times the Gini coefficient (and expressed as a percent), such a mistake is important to correct, as a Gini Index of .825 is minuscule. For accuracy, I will refer to what they compute as the Gini coefficient, as that is what their (as well as my) code is actually computing.

#### 2.1 Gini Limitations

The use of the Gini coefficient here has a few limitations generally. First, this measure will not pick up on absolute levels of search results. Google News could be providing immense additional viewership to local news outlets, and this impact could even be increasing over time, but the value of the Gini coefficient will not be correlated with this. As such, it is unclear how this line of inquiry will result in an answer to how Google News harms local news – rather than merely not being as beneficial as it could be if it steered all users to only local news.

Second, what the authors are assuming is inequality in Google News' algorithms could instead be inequality in the volume and breadth of articles across news outlets. For example, CNN likely has many more journalists (and greater funding) than most local outlets. As such, it can cover a wider number of topics per week, and even produce multiple articles per topic, where a local news outlet may be more constrained and forced to focus on fewer topics. CNN will therefore have more articles that can appear in search results simply by producing more of them, and thus the Gini coefficient will be high due not to unfair algorithms, but to the diverse nature of what different outlets can publish.

Third, due to data issues, the results for ten out of the sixteen local search terms (compared to only one out of the sixteen general-interest search terms) have only been scraped one or two times, compared to the three data scrapes for the other terms. The authors make no attempt to account for this in computing the aggregate Gini coefficient, which biases the result toward what is seen on average for generalinterest topics. In some cases, local outlets will only be seen as occurring one third as often as national outlets, simply because local searches are scraped only one third as often. Multiplying search query terms by actual captures, then, there are 47 captures of general-interest topics, almost twenty five percent more than the 38 captures of local topics. To get a sense of this error, it would be like trying to get a sense of income inequality in the United States, but instead of randomly sampling all counties, oversampling wealthy or poor zip codes.

#### 2.2 Gini Miscalculations

The greatest limitation of the authors' Gini coefficient, however, comes from how they are computing it. To obtain their data, the authors perform Google News searches for thirty two search terms (such as "police" or "weather"). In each search query, they also add the text "near" plus the name of a county and state. In this way, they perform a search for each term in every county in the United States. They then record which news outlets (which are then classified as either local or national outlets) appear in the corresponding results. Finally, they compute the Gini coefficient over the full set of observed news outlets in these county-by-county searches, based on how often each outlet is returned in these searches. However, by doing so, they ignore a key part of how they obtained their data: including location specific context in each query. Their Gini coefficient, then, ignores what location the search query was looking for content from, and what outlets are relevant to such a search.

Before moving ahead with a naive computation of the Gini coefficient, one must first consider whether it will inform what we're trying to understand. The Gini coefficient computation performed by the authors will not give a good indication of inequality between local and national outlets. First, this is due to the fact that by looking at results from thirty two different search terms, they will naturally find a perceived bias against outlets which are narrowly specialized in terms of topics. A local newspaper focusing on politics should not have its total number of articles compared to local newspapers that focus on politics and the weather – there will appear to be high inequality in aggregate when looking at results for searches on politics and weather, even if Google News was returning them equally as often on a per-topic basis. The authors do remedy this by performing the Gini calculation separately for each topic, which allows them to account for some of the inequality arising due to natural differences in what topics different newspapers choose to write about – but this factor is obviously a problem for the headline Gini coefficient results.

A second and more problematic issue is that by performing the Gini calculation over all counties at once, the authors conflate differences in relevance with inequality. It stands to reason that local news outlets are likely only or mostly relevant in their local geographic area. For example, when searching for "weather near Boston, MA", results from the Boston Herald would likely be very relevant, but articles from the Sacramento Bee would not. It is a very strong claim that the Sacramento Bee's weather-related news should be appearing at the same frequency as the Boston Herald's, CNN's, and every other news network's weather articles when someone searches for their own local weather. However by computing the Gini coefficient over all counties at once, this is exactly what the authors are imposing in their definition of inequality. This greatly inflates the Gini coefficient, since Google News not showing articles about Sacramento's weather to those searching for "weather near Boston, MA" is construed by the authors as amounting to unequal treatment of the Sacramento Bee. This is clearly nonsensical.

To account for this second issue, I implement the same fix that the authors applied to the first issue: computing the Gini coefficient over the results for each county. I can then analyze the distribution of Gini coefficients, or even just take its average. Thus, instead of comparing how often a newspaper from Boise appears in searches across the country to how often CNN appears, which does little to inform how "equally" an algorithm is treating local and national news outlets due to the relevance issues discussed above, I compare how often the Boise newspaper appears compared to CNN in Boise county searches. This more granular index gives exactly what is desired: in a search within a specific county (which is how the authors obtained the data), how equitable were the results between national outlets and local ones? And from there one can easily move to how the inequality measure looks like for the average county-level search.

This is precisely what I have performed using the authors' publicly available data: I compute the Gini coefficient for each county, and then average the results across counties. I do this using all topics and separately for each individual topic (with averages for the two groups of topics). The more topics are pooled together, the greater the Gini coefficient will be inflated due to differences in what news outlets may specialize in, but I still include them for comparison.<sup>1</sup>

#### 2.3 Gini Corrections

When performing the above calculation on all search topics, I obtain a Gini coefficient (averaged over counties) of 0.69, a 16% drop from the authors' measure of 0.825. This value is much more in line with the Gini coefficient of 0.65 for overall newspaper circulation computed by the University of North Carolina Center for Innovation and Sustainability in Local Media's Database of Newspapers (Abernathy (2018)). This result is very informative for whether the distribution of content on Google News is greatly different, or specifically more unfair, than what is seen outside of Google News for local newspapers.

Next, I compute the Gini coefficient using only search terms from each local topic. In doing so, I get an average Gini coefficient value of 0.289, a 58% drop from the authors' average value of 0.668 for the same topics. In Table 1, I show the values for each individual search term for local topics. I include the value for the corresponding term computed by Fischer et al. (2020), the updated Gini coefficient which corrects for (i.e., accounts for) differences in relevance across counties, as well as the percentage difference. As one can see, the values are substantially smaller, with some as low as 0.11 and with absolute percent differences as great as 80%.

Lastly, I do the same analysis for each general-interest search term. I obtain an average value of 0.456, a 44% drop from the authors average Gini coefficient of 0.807. Table 2 shows the values for each specific general-interest term, with a minimum value of 0.22 and a maximum absolute percent difference of nearly 70%.

As one can see, correcting for how the Gini coefficient is computed yields stunningly different results. Once one removes the inflation of the coefficient caused by local outlets

<sup>&</sup>lt;sup>1</sup>Also, if a news outlet never appears in any of the results for any queries in a county, that outlet will not be used in computing the county's Gini coefficient. While this is beneficial for the purpose of isolating each analysis to the relevant set of outlets for each county-specific query, one could imagine that a relevant outlet may be unfairly excluded from all search results – and thus should be included in the computation. However, in such a scenario, unless one believes that this outlet would be included when they don't have relevant content, they will also be excluded from the authors' Gini computation.

Search Term	Fischer et al. (2020) Gini	Average County Gini	Percent Difference
Accident	0.625	0.227	-63.68%
College	0.700	0.264	-62.34%
Crime	0.668	0.277	-58.59%
Death	0.738	0.370	-49.86%
<b>Emergency Services</b>	0.691	0.386	-44.21%
Governor	0.745	0.336	-54.97%
High School	0.615	0.172	-72.12%
Hospital	0.635	0.246	-61.28%
Mayor	0.560	0.109	-80.50%
Obituary	0.845	0.599	-29.16%
Police	0.744	0.386	-48.17%
School Board	0.584	0.148	-74.56%
Traffic	0.592	0.195	-67.08%
Transit	0.556	0.236	-57.49%
University	0.684	0.304	-55.65%
Weather	0.714	0.366	-48.72%

Table 1: Gini Coefficients for Local Topic Search Terms

Table 2: Gini Coefficients for General-Interest Topic Search Terms

Search Term	Fischer et al. (2020) Gini	Average County Gini	Percent Difference
Abortion	0.782	0.393	-49.72%
Caravan	0.761	0.413	-45.77%
Climate	0.817	0.470	-42.47%
Conservative	0.783	0.456	-41.69%
Corruption	0.735	0.345	-53.06%
Election	0.827	0.475	-42.56%
FBI	0.790	0.449	-43.20%
Gun	0.711	0.283	-60.16%
Immigration	0.845	0.508	-39.89%
Liberal	0.816	0.476	-41.46%
Politics	0.902	0.629	-30.25%
President	0.847	0.524	-38.05%
Scandal	0.699	0.222	-68.27%
Shutdown	0.881	0.593	-32.69%
Syria	0.885	0.587	-33.63%
Taxes	0.835	0.465	-44.28%

not being returned in search results for counties on the other side of the country on topics such as local weather, one sees a sharp reduction in inequality. For local interest topics, the inequality is incredibly small. The overall inequality is roughly on par with what is found in the aforementioned study Abernathy (2018), and for individual topics (and even the groups of local and general interest topics) Google News exhibits a fraction of the inequality alleged by the authors.

## **3** Responsiveness to User Intent

Fischer et al. (2020) next move to evaluating whether Google News is responsive to the queries users enter. Does Google News provide local outlets when the searches are on inherently local topics? To this end, they compute the fraction of outlets returned for each of the thirty two search terms that are local, regional, national, and international. Despite this being the main test for the second hypothesis in their paper, the results are relegated to their online Supplementary Information (more specifically, Supplementary Figure 2), with the authors only commenting in the main text that "[it] is the case" that Google News provides local content when users search for it. But the results in Supplementary Figure 2 are actually quite stunning<sup>2</sup>: the minimum share of results from local outlets for local terms appears to be roughly two-thirds. When including regional outlets, this share looks to be essentially 80-90%. Even for general-interest topics, many of the terms have a local share greater than 25%.

Instead of focusing on these results, which highlight the attention that Google News does give to local outlets, the authors instead choose to focus on a different metric: namely, what fraction of the first N results come from local outlets. In a way this is fair – if users pay more attention to the first few articles in results, then understanding what fraction of the first N results are local is an interesting question. In Figure 2 of Fischer et al. (2020), the authors plot the fraction of the first N results from local and national outlets, for N ranging from one to one hundred, for each of their search terms. However, the authors do not comment (or report directly) on the level of the share of local news in the first few results – instead commenting that "the relative

<sup>&</sup>lt;sup>2</sup>While the code and data used to create this plot are publicly available, running the authors' code puzzlingly yields materially different values to what is presented in Fischer et al. (2020). As such, I only comment on what the values appear to be in their published article.

frequency of local outlets declines as we isolate only the top results for many terms in the set" and that "local news is available on Google News, but users need to scroll through the results to find it." As a solitary example, they state that "even for topics as contextually dependent as weather, the share of local news outlets in the first ten hits is lower than that for national outlets, and the trend changes only after the tenth result."

From these comments, an inattentive reader would get the idea that there are very few local outlets in the first few results, even if there are more later on. However, this could not be further from the truth. For local topics, the share of local outlets for the first five results appears<sup>3</sup> to be at least 75% for half of the topics. Almost all topics appear to have a share of at least 50% for the first five results; far from being a representative example, weather is actually the only topic to have a share less than 50% for the first five results, and then just barely so. One also wonders if some of the reason the shares are not even higher is due to the choice of queries themselves: the terms with the most obvious local intent (like emergency services) see the highest shares. The other local queries might be suffering from (at least partial) misclassification errors in the paper as to whether they really are local or not. For national interest topics, most topics have a local outlet share of at least 25% for the first five results, with the exception of some obviously national topics such as president, shutdown, and Syria. So while it is the case that the share of local outlets is increasing with search depth, the majority of the initial results for local topics are from local outlets.

#### 3.1 Additional Evidence

In addition to these results, in the aforementioned Gini coefficient analysis it was shown that when using the corrected computation, the Gini coefficient for local topics is 0.289, compared to 0.456 for national topics and 0.69 for all topics. This is another indication that when the user is searching for a local topic specifically, there is much less inequality in results.

In their paper, Fischer et al. (2020) also perform a regression of whether a result is

 $<sup>^{3}</sup>$ Again, likely due to the issues mentioned in the prior footnote, I am unable to recreate this plot to obtain exact numbers. This is compounded as the authors do not provide code for this plot. For this reason, I again only comment on what the values appear to be in the published article.

from a local or regional outlet on a host of variables. The only feature which is both statistically and meaningfully significant in magnitude is whether the search term is a general topic, an indication that Google News is strongly (and chiefly) responsive to what is searched. Using a general-topic term makes it roughly 63 percentage points less likely of seeing a local outlet; or in other words, using a local-topic term makes it roughly 63 percentage points more likely of getting a local outlet. Given that the intercept is 0.94, this means that there is a non-trivial probability of local outlets being returned for a non-local query.

#### **3.2** Contradictory or Inaccurate Statements

The authors make a number of statements which either contradict or otherwise do not follow the findings from their own tests of whether Google News responds to a user's query. They state "our findings suggest that Google News users are exposed to far more national news outlets than local outlets." Such a claim cannot be supported by any of the results in their article. While they show the prevalence of outlets varies by topic, this statement depends on how often users search for one topic over another, which they lack data on. However, the authors do have search data from Google Trends, where they compare the popularity of local vs general-interest searches. They find that local searches are more popular: if this holds within Google News as well (which is an extrapolation the authors are making to other ends), then this casts doubt on their claim that national news outlets are being shown more than local ones. They have already shown that local outlets are shown more than national outlets for local searches: if local searches are occurring more frequently than general-interest ones, then it is likely that local news are being shown more than national outlets.

They further state "This relationship is driven by the fact that readers who search for general policy issues are almost exclusively directed to reports by national newspapers. Only the readers who explicitly search for keywords related to local governance and public services are directed to local and regional newspapers, and only if they look past the first few results—an unlikely phenomenon." The first sentence here is immediately contradicted by the second: if users are shown what they search for, then what they are shown in aggregate across their searches will depend on what they search most (which, from their Google Trends results, is actually local searches). The second sentence is also incorrect: the majority of the first five search results, according to the authors' own findings, is from local outlets.

#### 3.3 What We Learn

So what have we learned from the authors' analyses of the responsiveness of Google News to user intent? We learn that, as they hypothesized, when users search for local topics, they chiefly get local outlets returned to them. This is true even for the first few articles in the list of results (and the share increases as the user gets deeper in the results). When a user searches for a general interest term, they see lower (but still substantial) shares of local outlets. This could partly be due to misclassifications in whether a query is really local or general, but even so, is such a result undesirable? Should a user be directed to a local paper, that is potentially more limited in resources, when a user is searching for an event going on in Syria? The authors show that Google News responds well to the interests of the user, and when local outlets have less to say on a topic than national news outlets, Google News returns results from outlets which can comment on the issue of interest.

We also learn that Google News may be more supportive of local outlets than other media. In their conclusion, the authors cite the following: "researchers at Duke University reported that while local newspapers accounted for only one-quarter of the total media outlets in a random sample, they produced 50% of all original news stories reporting local news" (Mahone et al. (2019)). Given that we see most local searches yielding over half of their results from local outlets, even in just the first few articles, Google News may be offering a more favorable setting for local news, which corresponds with past researchers' findings that local news is uniquely harmed when Google News (or the set of news aggregators in general) is shut down.

Lastly, we have evidence against the hypothesis that Google News is prioritizing national over local outlets. In their conclusion, the authors ask: "is there a general platform preference for large national outlets, with local outlets being included when there are no more national outlet stories to include?" If this were the case, then we should see such a preference on all topics, not just non-local ones. And for local topics, national outlets should always dominate the first few results. The fact that this "preference" only occurs for general interest topics implies that there is no such overt prioritization, as the simpler and more consistent explanation is that the algorithms are responding to users with the articles that they are searching for. Perhaps the reason the authors obfuscate their results from this section, and move on from it so quickly, is because they realize it directly contradicts their main premise. Indeed: local outlets are shown, in large numbers, when they are relevant and desired.

### 4 Missing the Forest for the Trees

In their quest to understand whether Google News is failing to give ample space to local news, the authors have focused solely on what Google News provides in search results. However, it is an unstated and unproven assumption that this is the main way users find articles in Google News, particularly local ones. When one navigates to news.google.com, aside from a number of suggested articles, one sees a series of tabs with various topics, such as "Business", "Sports", and "World." The third such tab, existing since at least 2018, is called "Your local news." Clicking on this tab yields a page full of articles entirely from local news outlets, as well as the ability to choose which geographical area one wants news from. Accordingly, it does not seem to be particularly challenging for users to be connected with local news.

It is odd that the authors – in an article whose purpose is to understand whether Google News is giving preference to national outlets – fail to mention a feature which gives specific preference to local news.

## 5 Conclusion

In this article, I have shown that the claims made in Fischer et al. (2020) are greatly weakened if not flatly contradicted by two errors in how Gini coefficients are calculated, as well as the authors' own results. There appears to be remarkably little inequality across outlets in search results, especially when looking at individual topics or groups of topics. Contrary to what the authors imply, the majority of the first five results in local topic searches are from local outlets, and there is strong additional evidence that Google News provides chiefly local outlets when that is what is being searched for. Indeed, it appears that Google News is quite supportive of local news, including its "Your local news" tab on the main page, which the authors fail to mention. These fundamental corrections put Fischer et al. (2020) in greater harmony with past research showing aggregators to be beneficial to local publishers.

### **Competing Interests**

The author is currently – as well as at the time of writing – employed by Google, owner of Google News.

#### Author Contributions

E.M. drafted and revised the manuscript, as well as performed the data analyses.

#### Data Availability

Data for the original analysis in Fischer et al. (2020), used for replication in this study, are available via the associated OSF repository for the original authors' project (https://osf.io/hwuxf/?view\_only=3fa7499661df487689031e11b8ea20b4).

#### Code Availability

For the code used to obtain the values in Tables 1 and 2, please email the author (evan.magnusson@alumni.stanford.edu).

## References

Abernathy, P. M. (2018). The Expanding News Desert. UNC Press.

- Athey, S., M. M. Mobius, and J. Pal (2017). The Impact of Aggregators on Internet News Consumption. Stanford University Graduate School of Business Research Paper.
- Calzada, J. and R. Gil (2020). "what do news aggregators do? evidence from google news in spain and germany". Marketing Science 39(1), 134–167.

- Chiou, L. and C. Tucker (2017). Content Aggregation by Platforms: The Case of the News Media. Journal of Economics & Management Strategy 26(4), 782–805.
- Fischer, S., K. Jaidka, and Y. Lelkes (2020). Auditing Local News Presence on Google News. Nature Human Behaviour 4(12), 1236–1244.
- Mahone, J., Q. Wang, P. Napoli, M. Weber, and K. McCollough (2019). Who's Producing Local Journalism? Technical report, Duke University.
- NERA (2015). Impacto del Nuevo Artículo 32.2 de la Ley de Propiedad Intelectual Informe para la Asociación Española de Editoriales de Publicaciones Periódicas (AEEPP).
- NERA (2017). Impact On Competition and On Free Market of the Google Tax of AEDE Fee: Report for the Spanish Association of Publishers of Periodical Publications (AEEPP).