



HAL
open science

An inquiry into Search Engine Neutrality: the case of movements against police violence in France and the U.S.

Romain Badouard, Inna Lyubareva, Patrick Maillé, Bruno Tuffin

► **To cite this version:**

Romain Badouard, Inna Lyubareva, Patrick Maillé, Bruno Tuffin. An inquiry into Search Engine Neutrality: the case of movements against police violence in France and the U.S.. 2024. hal-04690241

HAL Id: hal-04690241

<https://inria.hal.science/hal-04690241v1>

Preprint submitted on 6 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

An inquiry into Search Engine Neutrality: the case of movements against police violence in France and the U.S.

Romain Badouard Inna Lyubareva
Université Paris-Panthéon-Assas IMT Atlantique

Patrick Maillé Bruno Tuffin
IMT Atlantique Centre Inria de l'Université de Rennes

Abstract

The digital transformation has revolutionized information consumption, with search engines playing a pivotal role in shaping user access to diverse media. Employing algorithms, these engines influence content visibility and aggregate news sources, significantly molding public opinion. As gatekeepers of information, search engines impact media outlet visibility, affecting online traffic, revenue, and journalistic diversity. In breaking news and societal issues, search engines expedite information dissemination, influencing initial narratives. Understanding their role is crucial for transparency and user access to diverse information. Focusing on movements against police violence, our paper conducts a comparative analysis across 12 search engines for terms “Black Lives Matter” and “Justice pour Adama”. Our innovative methodology identifies biases in information diversity, providing insights into the dynamics shaping visibility of societal issues.

1 Introduction

The ongoing digital transformation has wrought a profound impact on the information consumption patterns of individuals. Within this digital milieu, search engines assume a pivotal role in facilitating access to the extensive array of available media. Operationally, these search engines employ algorithms to discern and rank search results, thereby influencing their visibility. Moreover, they aggregate content from diverse news sources, exerting a notable influence on public opinion by amplifying certain voices while marginalizing others. Functioning as gatekeepers of information, search engines wield significant authority in determining content visibility. The user tendency to predominantly engage with the top search results perpetuates a hierarchical information access model, potentially sidelining content beyond the top results and consequently diminishing its exposure.

The visibility of media outlets in search results directly impacts their online traffic and revenue, thereby influencing the economic viability of news organizations. This, in turn, may have cascading effects on the diversity and quality of journalism accessible to the public. Consequently, media outlets commonly employ search engine optimization (SEO) strategies to enhance their content visibility in search results. As a corollary, the discernibility of information is shaped not solely by its intrinsic value or relevance but also by the efficacy of employed SEO tactics.

In the context of breaking news events and societal issues, search engines assume a critical role in expeditiously disseminating information and presenting news stories that contribute to shaping the initial narrative of an event.

In summary, comprehending the role and function of search engines as information gatekeepers is imperative for fostering transparency and ensuring user access to a diverse range of information. A focused examination of search engine-driven media coverage pertaining to social movements, such as those against police violence in various countries, holds particular significance. Media coverage is instrumental in shaping public perceptions, and analysis thereof aids in ensuring the accuracy, objectivity, and impartiality of the information disseminated. Additionally, different media outlets may offer distinct perspectives on events; analyzing coverage facilitates a more comprehensive understanding of the varied viewpoints and narratives surrounding sensitive issues like police violence. This nuanced comprehension contributes to informed public discourse.

Furthermore, media coverage often acts as a catalyst for public awareness and activism. By scrutinizing how events are portrayed, individuals and communities can gain insight into the social and political context, fostering informed discussion and mobilization for change. Lastly, media analysis serves as a platform for dialogue among communities, law enforcement, and policymakers. Understanding how events are reported facilitates constructive conversations about the multifaceted aspects of complex issues, such as the role of law enforcement and the systemic implications of racism in the context of police violence, thereby paving the way for envisioning potential solutions. In the realm of global issues like police violence, transcending national or community boundaries in media coverage enables the identification of shared challenges and facilitates the exchange of ideas and strategies for addressing common problems associated with law enforcement practices.

The aim of this paper is two-fold. Initially, it conducts a comparative analysis of information sources across 12 search engines for two sensitive search terms related to movements addressing the complex issues of police violence and racism. The terms under examination are "Black Lives Matter" (BLM), predominantly relevant to the United States, and "Justice pour Adama" (JPA), originating in France in 2016 following the death of Adama Traoré. We simulate the information search as if conducted by a user located in France, making BLM the "foreign" case and JPA the French-specific information search.

Subsequently, the paper introduces an innovative methodology designed for broader applicability across diverse case studies. The objective is to identify potential biases and deficiencies in the diversity of information presented to

users by search engines. This approach allows for a nuanced understanding of the dynamics shaping the visibility of information related to sensitive societal issues in different contexts.

2 Background

If search engines, social networks, and application portals are recognized as the new gatekeepers in the online information market (Bozdag, 2013; Gillespie, 2018), there is a growing scrutiny into how their design introduces bias (Maillé et al., 2022; Urman et al., 2022a). The criteria employed by these platforms to sort, curate, and prioritize content directly influence how users access information, thereby attributing ethical and political implications to these ranking algorithms. The role of “infomediaries” (Rebillard and Smyrnaio, 2010) in shaping exposure to news sources and their impact on information pluralism has been a longstanding debate, prompting calls for new forms of regulation in the information market (Van Hoboken, 2015; Lyubareva and Rochelandet, 2017). Among these “infomediaries,” search engines have been a focal point of academic research (Ulloa and Kacperski, 2023), particularly regarding their editorial responsibility (Grimmelmann, 2014). At the heart of this debate is the tension between the perceived neutrality of the criteria used to rank websites and the effects of this ranking on the exposure to news sources in the context of controversial issues. More precisely, if search engines could potentially facilitate the discovery of unknown news outlets and increase the variety of information available for readers, the question arises on the capacity of niche and alternative information sources to appear in top positions in search engines results. This depends on how different search engines configure their algorithms and distribute visibility between small or less popular sources and established outlets.

As early as the late 1990s, A.L. Barabasi highlighted the hierarchical nature of the web as an information network. Hypertext links, which connect pages in one direction (referred to as “directed links”), convey a form of recognition and authority between pages, resulting in an uneven distribution of links across the web, commonly following a power law (Barabasi, 2002): a small minority of sites benefit from a huge amount of incoming links, while the vast majority of sites receive very few links. This power law has been depicted through web mapping tools (Ghitalla, 2021), revealing the partition of the web based on ideological factors (Fouetillou, 2008; Cardon et al., 2011) and the phenomena of authority between sites within thematic clusters (Benkler, 2006).

The uneven distribution of links has a direct impact on the online information practices of Internet users. Search engines like Google employ link distribution as a criterion for ranking websites; the more links a site receives, the more relevant and popular it is considered, influencing its ranking in search engine results. While this ranking formula raises significant economic questions concerning abuse of dominant position, it also poses important political and ethical issues. (Hindman et al., 2003) demonstrated that, in the context of controversial issues, sites with numerous incoming links were more likely to be visible and

consulted by Internet users, irrespective of their relevance and political content. Among the subjects under scrutiny in this study were gun control, abortion, and the death penalty. Given the sensitive nature of these issues in the US, strategies are being implemented to disseminate inbound links between supporters of opposing positions. The objective was to optimize search engine rankings and potentially influence public opinion.

Another issue related to the political and social implication of website ranking by search engines is the personalization of results. This personalization is intended to guide users toward information that confirms their opinions, producing a form of ideological confinement and polarization, popularized by Eli Pariser as “filter bubbles” (Pariser, 2012). If the formation of such bubbles, on search engines or social networks, poses a threat to public debate by limiting users’ exposure to opposing arguments (Sunstein, 2002), some authors point out that personalized search technology could moot the problematic “winner-take-all” effect of search engine bias (Goldman, 2005).

Empirical studies on “filter bubbles” tend to contextualize these cognitive bubbles, highlighting the limits of the “bubble effect” (Haim et al., 2018; Fletcher and Nielsen, 2018; Puschmann, 2019) and describing the conditions under which they form (Robertson et al., 2018). Contradictory results across studies may be attributed to methodological differences: while some studies profile users based on socio-demographic criteria, observing a low level of results personalization (Haim et al., 2018), others rely on purely political profiling, noting a strong personalization of results (Robertson et al., 2018). (Puschmann, 2019) adds that the degree of personalization varies according to the issue about which the user is seeking information, and the controversial nature of this issue, but also according to the general political context. Not all countries are as polarized as the U.S. and their results, which relate to politics in Germany, show that search engines personalize responses less because public debate there is less polarized. (Robertson et al., 2023) have also shown that the effects of filter bubbles were more closely tied to the beliefs of internet users than to the rankings of search engines. In other words, even if a search engine ranks websites in a more or less neutral manner, the user’s beliefs will lead them to click on the information most likely to reinforce their opinions. In this context, algorithmic curation appears secondary with regard to user beliefs.

Going deeper, certain empirical studies, such as (Puschmann, 2019), question the validity of the filter bubble concept and its effectiveness in mitigating search monopolies. (Nechushtai and Lewis, 2019) also argue that the primary concern with ‘algorithmic gatekeeping’ is not the polarization of news sources but rather the concentration of the media market. Through an empirical study of information queries on Google News in the U.S., they showed that the five most recommended media outlets comprised between 49% and 69% of all results. A similar study in France demonstrated that the specific functioning of Google News led to a form of “editorial hype,” where a small number of subjects were highlighted, reducing users’ exposure to news diversity (Marty et al., 2012). According to (Diaz, 2008), the main factor contributing to this concentration phenomenon lies in the design of Google’s algorithm, which prioritizes

popularity over other journalistic criteria, such as deontological considerations prevalent in news production by journalists.

Research presents divergent views on the concentration of online news readership and its association with search engines. Some studies suggest online concentration linked to the market model adopted by search engines, favoring access to mainstream sites and national news providers (Hong and Kim, 2018; Muddiman, 2013), rather than diverse, less familiar sources. In the same vein, (Jiang, 2014) observes a high degree of information concentration on search engines in China, compared to western countries.

Contrary insights are also presented. For instance, (Fletcher and Nielsen, 2018) have shown, through an international study in four different countries, that search engines users are more likely to be exposed to a wide variety of news sources, and to more diverse political news, than those who use other means of information. They even talk about “automated serendipity” to describe the fact that through search engines, users access sources they would never consult otherwise.

In light of these academic debates, a scientific challenge arises in developing tools to audit search engine algorithms and measure their biases. (Mowshowitz and Kawaguchi, 2005) suggest that such measurement can be achieved by comparing search engines, observing how the distribution of URLs by a specific search engine deviates from the average distribution of these URLs by various search engines based on the same query. (Maillé et al., 2022) propose a “meta search engine,” the SNIDE tool, as a means to operationalize this comparison. The analysis of the current paper relies on this tool.

3 Research Methods

This paper combines qualitative and quantitative methods to provide a comprehensive and nuanced comparative analysis of media coverage by different search engines. Each approach brings unique strengths to the study, addressing different aspects of the research question. To compare the results of several search engines, we use the publicly available platform SNIDE (Maillé et al., 2022), standing for uSiNg all search engINes results to reduce and DETect bias¹.

The motivation for the SNIDE tool is the so-called *search neutrality debate* (Odlyzko, 2009; Maillé and Tuffin, 2022) highlighted in 2009 by the co-founder of Foundem complaining that his company, a Vertical Search & Price Comparison website, was intentionally ranked low by Google, and therefore commercially penalized, with respect to Google’s own service, Google Shopping.

The SNIDE platform allows a user to perform a search with a term of their choice, over several search engines. The results are then collected and compared through a series of statistical tests aimed at detecting outliers. For example, the tool can detect whether the page that tends to be the most visible in average is well ranked by each search engine. Recall that a statistical test is a rejection test, designed to reject a hypothesis: failing a test means that scrutiny may

¹<https://snide.irisa.fr/>

Table 1: List of search engines considered in this paper

• AllTheInter-	• Duckduckgo	• Lilo	• Startpage
net	• Ecosia	• Lycos	• Yahoo
• AOL	• Google	• Qwant	• Yandex.
• Bing			

be needed and a deeper investigation carried out, but not necessarily that a malicious intention was there. SNIDE in its search results disable cookies and browsing history. However it keeps geolocation. For the purposes of this paper, all searches are located in France, which certainly has an impact on the results we obtain (several search engine take into account the user location to build their ranking). This permits us to clearly distinguish between the national (JPA) and the international (BLM) cases, as they can be approached by a Lambda user based in France.

The outputs from the following 12 search engines given in Table 1 are collected using a web crawler written in Python. Note that the tool may sometimes imperfectly collect results, in part due to the presentation of the result pages by search engines being frequently modified. To the best of our knowledge the results presented here accurately correspond to those provided by the search engines.

Since our goal here is *not* to point to search engines displaying unusual pages but rather focus on the type of displayed media to investigate the diversity of sources, we adapt the work of (Maillé et al., 2022) to this specific context, by manually associating a category of media with each displayed web pages for a given query: e.g. the web page of Le Monde corresponds to the “Mainstream news” category. This permits focusing on the diversity of sources for specific news queries, i.e., comparing the *types* of sources in the results rather than the specific web pages.

The categories of media contributing to the information coverage of both national and international movements that we considered are listed below (from (Lyubareva et al., 2021)):

1. Mainstream News outlets (national and international)
2. Advocacy Media, such as the official web sites of the Black Lives Matter (BLM) and Justice Pour Adama (JPA) movements and other organizations supporting social action against police violence.
3. Dictionaries, such as Wikipedia pages in both languages
4. Online Social Networks, such as Twitter (X) and Facebook (Meta)
5. Left-wing and right-wing “partisan media” marked by extreme political positioning
6. Cultural content, such as books, films, concerts and video documentaries

7. Reports, research papers and other independent analyses
8. Subject-specific press covering only certain subjects or geographical areas
9. Institutional Sources and Archives, such as web sites of administrative court or the Library of Congress.

Formally, let us define \mathcal{M} as the set of those $m = 10$ media types, and \mathcal{I} as the set of search engines considered in the study. For each research term and each search engine $i \in \mathcal{I}$, we compute for each category $\ell \in \mathcal{M}$ a score corresponding to that category’s visibility on the search engine. The visibility of a result is assumed to vary with the position in the ranking, and we estimate it through the click-through-rate (CTR) or probability to be clicked: the higher in the ranking, the more likely to be seen and clicked. We extrapolate it as a good proxy to the visibility of a position. Let θ_k be the visibility/CTR for position k . The values considered in this paper are given in Table 2, taken from (Dejarnette, 2012) and obtained from extensive analysis.

Table 2: Position-related considered CTR (visibilities): θ_k represents the visibility of position k in a ranking

θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
0.364	0.125	0.095	0.079	0.061	0.041	0.038	0.035	0.03	0.022

Considering a given query y on a search engine $i \in \mathcal{I}$, let $r_{i,k}(y)$ be the media type (in \mathcal{M}) of the page presented at place k . The score (or again, visibility) of a media type $\ell \in \mathcal{M}$ on engine i and for query y is computed as the aggregated visibility of this type of media over all 10 first displayed pages (i.e., the top ten positions in search engine i ’s ranking):

$$v_{\ell,i}(y) = \sum_{k=1}^{10} \theta_k \mathbb{1}_{r_{i,k}(y)=\ell}, \quad (1)$$

where $\mathbb{1}$ is the indicator function.

The average visibility of media type ℓ over all 12 search engines is then

$$v_{\ell}(y) = \frac{1}{12} \sum_{i \in \mathcal{I}} v_{\ell,i}(y).$$

Once the media type visibilities are computed, we can compute a score for each search engine $i \in \mathcal{I}$, describing how media types that are visible (over all engines) are displayed by this specific engine. That search engine i score, which we denote by $S_i(y)$ for query y is defined as the weighted sum of the visibilities of all displayed media types at each position, with position visibilities as weights:

$$S_i(y) = \sum_{k=1}^{10} \theta_k v_{r_{i,k}(y)}(y).$$

Since our goal in this paper is to measure the level of diversity of media types, for each search engine $i \in \mathcal{I}$ we normalize the vector of visibilities $(v_{\ell,i})_{\ell \in \mathcal{M}}$ to interpret a ranking as the engine allocating its total visibility (of 1) among the different types. From that normalized vector, akin to a probability distribution, we compute the entropy of search engine $i \in \mathcal{I}$, to quantify the diversity of displayed sources:

$$e_i(y) = \sum_{\ell \in \mathcal{M}} \left(\frac{v_{\ell,i}(y)}{\sum_{j \in \mathcal{M}} v_{j,i}(y)} \right) \log_2 \left(\frac{v_{\ell,i}(y)}{\sum_{j \in \mathcal{M}} v_{j,i}(y)} \right) \quad (2)$$

with the convention $x \log_2(x) = 0$.

Displaying visibilities, scores and entropies will allow to point to differences. It may be hard though to formally point to an outlier due to the limited number of existing search engines. However, in statistics there exists a test devoted to outlier detection in small sample sizes, called Dixon’s Q test (Dean and Dixon, 1951; Dixon, 1953). The test assumes Gaussian-distributed data, which we do not have here, but can be used as a quite good approximation. It is a rejection test of a null hypothesis of the form H_0 : “the minimum value (or the maximum value depending on the context) is not an outlier” against H_1 : “it is an outlier”. We can specify the level of confidence for the rejection test (again, valid when data are independent and identically distributed, with a Gaussian distribution). We will use Dixon’s Q test to address the four following questions, testing as H_0 for a given query y :

- no search engine has an abnormally low entropy (showing an abnormal lack of diversity), by looking at values $(e_i(y))_{i \in \mathcal{I}}$;
- no search engine score has an abnormally low score (showing a very different choice of media types), by looking at values $(S_i(y))_{i \in \mathcal{I}}$;
- for the most-shown type of media ℓ^* , no search engine has an abnormally low score, by looking at values $(v_{\ell^*,i}(y))_{i \in \mathcal{I}}$;
- no search engine has a most-shown source far less visible than the others, by looking at values $(\max_{\ell \in \mathcal{M}} v_{\ell,i}(y))_{i \in \mathcal{I}}$.

4 Findings

The qualitative analysis of the results of all search engines shaping the information for the search terms “Black Lives Matter” (BLM) and “Justice pour Adama” (JPA), puts in light the categories of media covering these movements. In total, across the twelve search engines in our sample, we found 106 media sources for BLM and 108 sources for JPA, each belonging to one of the media categories (mainstream news outlets, advocacy media, dictionaries, etc.). In both cases, mainstream news outlets represent the main source of information proposed by the search engines. They correspond to 41,5% of information sources for BLM and 58,3% of sources for JPA.

The second and third positions of the most frequent sources of information in the search engines of our sample are shared by the dictionaries and advocacy media. For “Black Lives Matter,” dictionaries, mainly Wikipedia pages, represent 26.4% of all sources and advocacy media 11.3% (mainly represented by the official website of the BLM movement and its related organisations). For the “Justice pour Adama” term, the ranking is inverted and the difference between advocacy sources and dictionaries is smaller: 10,2% of advocacy media and 9,25% of dictionaries.

The further less frequent media categories differ depending on whether a user is looking for information about the national movement (i.e., JPA) or the international one (i.e., BLM). Regarding the BLM movement, search engines put forward such sources as “Reports and independent analyses” (8%), “Subject-specific press” (6%), “Online Social Networks” (3%), “Partisan media” and “Archives” (less than 1%).

Regarding the French search term (JPA), online social networks occupy a more important place in the information landscape (more than 8%). They are followed by partisan media and, contrary to the BLM movement, various cultural contents (5%). Finally, the lowest visibility belongs to the reports and independent analyses (3%), subject-specific media and institutional sources with less than 1%.

These results can be interpreted in light of the fact that the BLM movement is older, more globally structured, and has generated a larger body of publications in dictionaries and related official sources and reports. In contrast, JPA is predominantly covered by news media, with fewer mentions in other sources such as dictionaries, advocacy media, or online social networks. Intriguingly, in the case of JPA, which is specific to France, the pathway to alternative and civil society discourse appears more pronounced compared to BLM. This is evident through the higher presence of online social networks, followed by partisan media and cultural content. This nuanced distinction highlights the varying sources shaping discourse around these socio-political movements. Notably, it is important to consider that these findings are based on searches conducted from France.

Tables 3 and 5 show the relative visibility of information sources on different search engines, for both requests. More specifically, the values in the tables are $\frac{v_{\ell,i}(y)}{\sum_{s \in \mathcal{M}} v_{s,i}(y)}$ for each search engine $i \in \mathcal{I}$ and media type $\ell \in \mathcal{M}$, where $v_{\ell,i}(y)$ for each request y is computed as in (1). That normalization allows to have a total of 100 percent, even if the search engine displayed less than 10 results, in order to reflect how each engine allocated its “visibility budget” among sources. We also show in Tables 4 and 6 the results to the four outlier detection tests (based on Dixon’s Q test) evoked earlier.

For the BLM search term, Table 3 shows that:

- Dictionaries represent a dominant source of information on Yahoo and AOL, then on Ecosia, Bing, Duckduckgo, and then on others (Lilo, AllTheInternet, Google, etc).

Table 3: Repartition of visibilities among source types for each search engine (percentages), for the query “Black Lives Matter” and associated entropy. Values in bold identify the most-shown source type for each search engine.

	AllTheInternet	AOL	Bing	Duckduckgo	Ecosia	Google	Lilo	Lycos	Qwant	Startpage	Yahoo	Yandex	Average
Dictionaries	48	79	54	52	60	41	49	15	52	34	84	25	49
Mainstream News	37	0	39	13	20	35	39	77	37	14	0	9	27
Subject-specific Press	4	0	7	0	4	10	9	9	11	0	0	0	5
Partisan Media - Right Wing	0	0	0	0	0	0	0	0	0	0	0	4	0
Partisan Media - Left Wing	0	0	0	0	0	0	0	0	0	0	0	0	0
Advocacy Media	11	21	0	18	11	9	2	0	0	41	16	46	15
Report and independent analyses	0	0	0	17	4	5	0	0	0	7	0	7	3
Social Networks	0	0	0	0	0	0	0	0	0	0	0	10	1
Cultural content	0	0	0	0	0	0	0	0	0	0	0	0	0
Archives	0	0	0	0	0	0	0	0	0	4	0	0	0
Institutional Source	0	0	0	0	0	0	0	0	0	0	0	0	0
Entropy	1.6	0.7	1.3	1.8	1.6	1.9	1.5	1.0	1.4	1.9	0.6	2.1	1.5

Table 4: Results of outlier detection tests (Dixon test) for the query “Black Lives Matter”: a “1” indicates a positive test, i.e., an outlier is detected.

	Entropy	SE score	SE visibility for most-shown source	Visibility of SE’s most-shown source
Dixon test with $\alpha = 0.1$	0	1 (Yandex)	1 (Lycos)	0
Dixon test with $\alpha = 0.05$	0	0	0	0

Table 5: Repartition of visibilities among source types for each search engine, for the query “Justice pour Adama” and associated entropy. Values in bold identify the most-shown source type for each search engine.

	AllTheInternet	AOL	Bing	Duckduckgo	Ecosia	Google	Lilo	Lycos	Qwant	Startpage	Yahoo	Yandex	Average
Dictionaries	41	0	14	4	14	43	11	14	11	41	0	11	17
Mainstream News	27	84	45	78	44	45	74	77	2	51	87	0	51
Subject-specific Press	0	0	0	0	0	0	0	0	14	0	0	0	1
Partisan Media - Right Wing	0	0	0	0	0	0	0	0	0	0	0	0	0
Partisan Media - Left Wing	7	0	0	0	0	0	0	0	12	0	0	3	2
Advocacy Media	0	11	41	11	42	0	15	9	8	0	13	48	16
Report and independent analyses	0	5	0	7	0	0	0	0	5	0	0	0	1
Social Networks	18	0	0	0	0	11	0	0	0	8	0	7	8
Cultural content	4	0	0	0	0	0	0	0	48	0	0	0	8
Archives	0	0	0	0	0	0	0	0	0	0	0	0	8
Institutional Source	3	0	0	0	0	0	0	0	0	0	0	0	0
Entropy	2.1	0.8	1.4	1.1	1.5	1.4	1.1	1.0	2.3	1.3	0.6	1.8	1.4

Table 6: Results of outlier detection tests (Dixon test) for the query “Justice pour Adama”: a “1” indicates a positive test, i.e., an outlier is detected.

	Entropy	SE score	SE visibility for most-shown source	Visibility of SE’s most-shown source
Dixon test with $\alpha = 0.1$	0	0	0	0
Dixon test with $\alpha = 0.05$	0	0	0	0

- Mainstream media are particularly visible and put forward on Lycos, which is followed by Bing, Lilo, AllTheInternet, Google and others.
- Qwant gives priority to the subject-specific media.
- Yandex, in comparison to other search engines, makes more visible partisan media and online social networks.
- Yandex and Startpage are also leaders in terms of visibility of advocacy media websites; they are followed by AOL and Duckduckgo.
- Finally, Duckduckgo is the search engine most showing reports and research work as information sources.

As for the JPA request (Table 5), the ranking appears quite different:

- Dictionaries represent a dominant source of information on Google and then, to a lesser extent, on Startpage and AllTheInternet.
- Mainstream news media are mostly visible on Yahoo and AOL and, then, on Duckduckgo, Lycos, Lilo and others.
- Qwant as in the BLM case gives particular priority to the subject-specific media; but this time this search engine is also a leader in partisan media and in cultural content.
- Exactly as for BLM, Yandex is the engine giving the largest visibility to online social networks.
- As well, Yandex and, then, Bing and Ecosia are also leaders in terms of visibility of advocacy media websites.
- Finally, as with the BLM request Duckduckgo shows reports and research work more than the other engines.

Among the statistical tests whose results are shown in Tables 4 and 6, only two are positive, with a confidence level of 90%:

- In terms of diversity of sources (measured by the entropy computed as per (2)), the values range from 0.6 to 2.3, with no outlier detected.
- The “SE Score” test aims at identifying a search engine showing very different sources of information from the other engines, here only for the “Black Lives Matter” search do we have a positive test, identifying Yandex as an outlier. Note however that if the sensitivity parameter α (that would correspond to the probability a false positive under Gaussian assumptions) is lowered to 5%, the test is made less sensitive and here does not identify any outlier.

- Still for the “Black Lives Matter” search, Lycos is identified by the test as not showing enough the most-shown type of information (here, dictionaries). Given the outlined type, as well as the type shown instead by Lycos (Mainstream news), Lycos could hardly be suspected of bias. But the test could help detect such biases. Again, making the test more conservating by setting α to 5% turns the result into a negative: more data would be needed to have more confidence into whether there is a significant discrepancy with the other engines.
- Finally, the type of information that is shown the most by each engine (visible in bold in Tables 3 and 5) is always sufficiently shown by at least some other search engines, so that no outlier is found.

Therefore, some search engines in our sample give priority to particular sources regardless of the information sought, while others change their information coverage between the BLM and JPA movements. In this respect, Yandex, for example, clearly prioritizes online social networks and partisan media marked by extreme political positioning for any subject. Similarly, Qwant puts forward subject-specific media specializing on particular topics, whereas Duckduckgo frequently proposes information from the reports and research papers on JPA and especially on the BLM movement. In contrast, for instance, Google and Startpage structure their information offer differently depending on the request: among other sources, they put forward the information coming from online social networks for JPA, but not for the BLM request.

Finally, some differences among the search engines were found also regarding the main media categories that concentrate information on the BLM and JPA subjects. For example, mainstream media – the most visible source for many search engines – is slightly represented by Yandex in the case of BLM, but is completely absent for the JPA movement. On the contrary, AOL ranks it well for JPA, but not for BLM. Concerning the advocacy media – another very popular source – Bing, Lycos and Qwant do not put these sources for BLM (contrary to JPA); whereas AllTheInternet, Google, and Startpage do the same for JPA, but prioritise these information sources for BLM.

5 Conclusions

Search engines consistently rank among the most widely utilized contemporary digital services for seeking products, obtaining answers, and accessing information across diverse topics. Existing research (Puschmann, 2019; Kleis Nielsen and Ganter, 2018) has established their significant impact on information production. This influence is manifested through their role as digital intermediaries between news organizations and citizens, and as tools employed by journalists for research (Puschmann, 2019). While the acknowledgment of search engines wielding considerable political and social influence is not novel, the empirical findings from research exhibit variations contingent upon the applied methodology, studied subject matter, or the specific country under scrutiny. This en-

during influence has led to the establishment of search engines as a one of the central focus of research over the past two decades.

In this paper, we introduce a methodology to analyze that diversity, and in particular to pinpoint discrepancies among twelve search engines, through statistical tests that can be performed for each individual search or for a set of searches. We illustrate the use of the method, as well as what conclusions can be driven, on the example of two highly controversial search terms: “Black Lives Matter” and “Justice pour Adama”.

Aligned with prior research, our analysis yields crucial insights indicating a predominant influence of mainstream news sources and dictionaries across all search engines, commanding nearly the entirety of visibility space. This scenario restricts the exposure available to “challengers,” encompassing subject-specific press, partisan, and advocacy media. These entities strive to secure a portion of media visibility through distinctive approaches, either grounded in expertise within a specific domain or rooted in a particular ideological or political standpoint. Notably, alternative information sources, including online social networks, exhibit minimal visibility in the majority of cases. The paper also shows a difference in information coverage between the french-specific social movement (JPA) and the international BLM movement: if online social networks and partisan media seem to be still part of the media coverage, as it constructed by the search engines, for JPA, they are almost nonexistent for the BLM movement.

Essentially, search engines perpetuate offline media hierarchies by favoring traditional sources with visibility bonuses. While this approach undeniably possesses merits, it runs counter to the concept of civil society actively participating in public discourse, especially concerning contentious topics such as police violence. At the same time, some search engines, as Russia’s leading state-controlled search engine Yandex, distinguish themselves by prominently featuring the sources as online social networks, advocacy or partisan media, occasionally at the expense of mainstream news outlets. This makes Yandex potentially risky in terms of information quality and reliability. Interestingly, this result obtained in our paper for the social movements against police violence in France and U.S., corroborates the insights from prior work showing that Yandex is the search engine with the highest proportion of the ‘conspiracy-dedicated’ websites and links to social media (Urman et al., 2022b).

Future work encompasses two primary extensions. First, emphasis should be placed on conducting more extensive campaigns, incorporating a broader array of controversial subjects and an increased number of search terms. This approach will facilitate the identification of subtle differences with heightened statistical significance, allowing for the formulation of comprehensive conclusions regarding the functioning and impact of search engines. The exploration of a wider scope will enhance the robustness of findings.

The second avenue of research pertains to the systematic and thorough comparison between information coverage provided by different search engines and that generated by advanced AI tools. These generative AI tools have demonstrated the potential to disrupt the longstanding search engine market. Inves-

tigating and contrasting the outputs of traditional search engines with those of emerging generative AI technologies will contribute valuable insights into the evolving landscape of information retrieval and dissemination.

References

- Albert Laslo Barabasi. *Linked: The New Science of Networks*. Basic Book, 2002.
- Yochai Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, 2006.
- Engin Bozdag. Bias in algorithmic filtering and personalization. *Ethics Inf Technol*, 15:209–227, 2013. doi: 10.1007/s10676-013-9321-6.
- Dominique Cardon, Guilhem Fouetillou, Clémence Lerondeau, and Christophe Prieur. Esquisse de géographie de la blogosphère politique (2007-2009). *Continuer la lutte. com*, 2011.
- R. B. Dean and W. J. Dixon. Simplified statistics for small numbers of observations. *Analytical Chemistry*, 23(4):636–638, 1951. doi: 10.1021/ac60052a025.
- R. Dejarnette. Click-through rate of top 10 search results in Google, 2012. <http://www.internetmarketingninjas.com/blog/search-engine-optimization/click-through-rate>, last accessed June 28, 2017.
- A. Diaz. Through the Google goggles: Sociopolitical bias in search engine design. In Amanda Spink and Michael Zimmer, editors, *Web Search: Multidisciplinary Perspectives*, pages 11–34. Springer, 2008. doi: 10.1007/978-3-540-75829-7_2.
- W. J. Dixon. Processing data for outliers. *Biometrics*, 9(1):74–89, 1953. ISSN 0006341X, 15410420. doi: 10.2307/3001634.
- R. Fletcher and R. K. Nielsen. Automated serendipity: The effect of using search engines on news repertoire balance and diversity. *Digital Journalism*, 6(8):976–989, 2018. doi: 10.1080/21670811.2018.1502045.
- Guilhem Fouetillou. Le web et le traité constitutionnel européen. Écologie d’une localité thématique compétitive. *Réseaux*, 147(1):229–257, 2008.
- Franck Ghitalla. *Qu’est-ce que la cartographie du web ? : Expéditions scientifiques dans l’univers des données numériques et des réseaux*. OpenEdition Press, 2021.
- Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

- Eric Goldman. Search engine bias and the demise of search engine utopianism. *Yale JL & Tech.*, 8:188, 2005.
- James Grimmelmann. Speech engines, 98 minn. *L. Rev.*, 868(874):893–95, 2014.
- Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. Burst of the filter bubble?: Effects of personalization on the diversity of Google News. *Digital Journalism*, 6(3):330–343, 2018. doi: 10.1080/21670811.2017.1338145.
- Matthew Hindman, Kostas Tsioutsoulouklis, and Judy A Johnson. Googlearchy: How a few heavily-linked sites dominate politics on the web. In *annual meeting of the Midwest Political Science Association*, volume 4, pages 1–33. Chicago, IL, 2003.
- Sounman Hong and Nayeong Kim. Will the internet promote democracy? search engines, concentration of online news readership, and e-democracy. *Journal of Information Technology & Politics*, 15(4):388–399, 2018.
- M. Jiang. Search concentration, bias, and parochialism: A comparative study of Google, Baidu, and Jike’s search results from china. *Journal of Communication*, 64(6):1088–1110, 2014. doi: 10.1111/jcom.12126.
- Rasmus Kleis Nielsen and Sarah Anne Ganter. Dealing with digital intermediaries: A case study of the relations between publishers and platforms. *New media & society*, 20(4):1600–1617, 2018.
- Inna Lyubareva and Fabrice Rochelandet. Economic models, uses and the pluralism of news online. *Rezeaux*, 205(5):9–19, 2017.
- Inna Lyubareva, Julien Mesangeau, Nadira Boudjani, Imad El Badisy, and Laurent Brisson. La plateforme des médias français et le ton du débat public. Exemple de YouTube. *Communication. Information médias théories pratiques*, 38(2), 2021.
- Patrick Maillé, Gwen Maudet, Mathieu Simon, and Bruno Tuffin. Are search engines biased? detecting and reducing bias using meta search engines. *Electronic Commerce Research and Applications*, pages 101–132, 2022.
- Patrick Maillé and Bruno Tuffin. *From Net Neutrality to ICT Neutrality*. Springer International Publishing, 2022. doi: 10.1007/978-3-031-06271-1.
- Emmanuel Marty, Franck Rebillard, Stéphanie Pouchot, and Thierry Lafouge. Diversité et concentration de l’information sur le web. Une analyse à grande échelle des sites d’actualité français. *Réseaux*, 176(6):27–72, 2012. doi: 10.3917/res.176.0027.
- Abbe Mowshowitz and Akira Kawaguchi. Measuring search engine bias. *Information Processing & Management*, 41(5):1193–1205, 2005. doi: 10.1016/j.ipm.2004.05.005.

- Ashley Muddiman. Searching for the next us president: differences in search engine results for the 2008 us presidential candidates. *Journal of Information Technology & Politics*, 10(2):138–157, 2013.
- E. Nechushtai and S. C. Lewis. What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior*, 90:298–307, 2019.
- A. Odlyzko. Network neutrality, search neutrality, and the never-ending conflict between efficiency and fairness in markets. *Review of Network Economics*, 8(1):40–60, 2009.
- Eli Pariser. *The Filter Bubble: What the Internet Is Hiding From You*. Penguin, 2012.
- Cornelius Puschmann. Beyond the bubble: Assessing the diversity of political search results. *Digital Journalism*, 7(6):824–843, 2019.
- Franck Rebillard and Nikos Smyrniotis. Les Infomédiaires, au Cœur de la Filière de l’Information en Ligne. Les Cas de Google, Wikio et Paperblog. *Réseaux*, 160-161(2-3):163–194, 2010. doi: 10.3917/res.160.0163.
- Ronald E. Robertson, David Lazer, and Christo Wilson. Auditing the personalization and composition of politically-related search engine results pages. In *Proceedings of the 2018 World Wide Web Conference*, pages 955–965. International World Wide Web Conferences Steering Committee, 2018. doi: 10.1145/3178876.3186143.
- Ronald E Robertson, Jon Green, Damian J Ruck, Katherine Ognyanova, Christo Wilson, and David Lazer. Users choose to engage with more partisan news than they are exposed to on google search. *Nature*, pages 1–7, 2023.
- Cass Sunstein. *Designing Democracy: What Constitutions Do*. Oxford University Press, 2002.
- Roberto Ulloa and Celina Kacperski. Search engine effects on news consumption: Ranking and representativeness outweigh familiarity in news selection. *New Media & Society*, 2023. doi: 10.1177/14614448231154926. online first.
- Aleksandra Urman, Mykola Makhortykh, and Roberto Ulloa. The matter of chance: Auditing web search results related to the 2020 us presidential primary elections across six search engines. *Social science computer review*, 40(5):1323–1339, 2022a.
- Aleksandra Urman, Mykola Makhortykh, Roberto Ulloa, and Juhi Kulshrestha. Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results. *Telematics and informatics*, 72:101860, 2022b.

Joris Van Hoboken. Search engines, pluralism and diversity: What is at stake and how to move policy forward? In Peggy Valcke et al., editors, *Media Pluralism and Diversity: Concepts, Risks and Global Trends*, pages 341–357. Palgrave Macmillan UK, 2015. doi: 10.1057/9781137304308_21.