

Google, data voids, and the dynamics of the politics of exclusion

Big Data & Society
January–June:1–14
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517221149099
journals.sagepub.com/home/bds



Ov Cristian Norocel¹ and Dirk Lewandowski^{2,3}

Abstract

This study deploys a critical approach to big data analytics to gauge the tentative contours of data voids in Google searches that reflect extreme-right dynamics of exclusion in the aftermath of the 2015 humanitarian crisis in Europe. The study adds complexity to the analysis of data voids, expanding the framework of investigation outside the USA context by concentrating on Germany and Sweden. Building on previous big data analytics addressing the politics of exclusion, the study proposes a catalogue of queries concerning the issue of migration in both Germany and Sweden on a continuum from mainstream to extreme-right vocabularies. This catalogue of queries enables specific and localized queries to identify data voids. The results show that a search engine's reliance on source popularity may lead to extreme-right sources appearing in top positions. Furthermore, using platforms for user-generated content provides a way for localized queries to gain top positions.

Keywords

Critical data studies, data void, Germany, politics of exclusion, search engines, Sweden

This article is a part of special theme on The State of Google Critique and Intervention. To see a full list of all articles in this special theme, please click here: <https://journals.sagepub.com/page/bds/collections/stateofgooglecritiqueandintervention>

Introduction

Among general-purpose web search engines, Google is the most popular in the Western hemisphere, with a total of approximately 90.8 billion visits in early 2022 (SimilarWeb, 2022). This ubiquity of Google search and its seeming ease of use in people's daily lives (Haider and Sundin, 2019; Sundin et al., 2017) obfuscates the low level of search engine knowledge among most ordinary users, which opens the door for the manipulation of propaganda and disinformation. One important vulnerability is described in research as a 'data void' (Golebiewski and Boyd, 2019). A data void becomes possible when, in response to a specific query, search engines return skewed and manipulatory content, which, from the point of view of societal relevance (Haider and Sundin, 2019), is of low quality because there is hardly any high-quality content corresponding to that search. Of interest here are two types of data voids. First are problematic queries, which concern

search results for highly contested or fraught terms that yield hits only from extreme-right outlets. Second are strategic new terms, which denote new terminologies specially created at the centre of extreme-right information ecosystems before being amplified to reach wider audiences with the aim of introducing (unaware) ordinary users to problematic content and polarizing frames. With this in mind, the article addresses the following research question:

¹Department of Gender Studies, Lund University, Lund, Sweden

²Department of Information, Hamburg University of Applied Sciences, Hamburg, Germany

³Department of Computer Science and Applied Cognitive Science, University Duisburg-Essen, Duisburg, Germany

Corresponding author:

Ov Cristian Norocel, Department of Gender Studies, Lund University, Box 117, SE-221 00 Lund, Sweden.

Email: ov_cristian.norocel@genus.lu.se



What are the tentative contours of data voids reflecting the extreme-right dynamics of exclusion?

To assess this issue, we locate our study outside the USA context and concentrate on the topic of migration in Germany and Sweden. In the European context, the two countries were the most generous in the 2015 humanitarian crisis in which hundreds of thousands of people sought refuge from war-torn countries. In addition, extreme-right environments experienced explosive growth in both Germany¹ and Sweden² in the context of this crisis. Given that data voids are difficult to identify and that, once detected, they are filled rapidly with mainstream sources, we quickly noticed that seeking data voids pertaining to the topic at the national level in the two countries would not yield significant results.

Consequently, we concentrated on examining queries at the subnational level, namely, on municipalities in both Germany and Sweden, in the following manner. The article first discusses the theoretical scaffolding, which is structured in three steps, allowing us to articulate the theoretical concepts that aid our analysis. Thereafter, we present the methodological implications of our approach, discussing the importance of specific and localized queries for the purpose of our study and connecting this to the catalogue of queries that we designed to enable us to collect our data. This is followed by a detailed presentation of our results. Finally, we focus on discussing the implications of our research findings and reflect critically upon the study's inherent limitations, which we integrate into a concluding reflection concerning the wider intellectual conversation in the field.

Theoretical scaffolding

The theoretical infrastructure of this study is underpinned by a critical approach to big data analytics. Notwithstanding this, it is solidly anchored in the research field of critical examinations of big data, also known as Critical Data Studies (henceforth, CDS) (Boyd and Crawford, 2012; Hargittai, 2015; Iliadis and Russo, 2016; Kitchin, 2022; Metcalf and Crawford, 2016). First, we explain the theoretical scaffolding of the present study in three steps. In the first step, we succinctly explain some of the key conceptual underpinnings of CDS, including the critical and reflexive approach to big data, the emergence of data politics, and the issue of digital inequality. We then connect these concepts to the opportunities provided by big data analytics to identify the dynamics of the extreme-right framing of such polarizing issues as ethnicity, nationalism, and religion in increasingly diverse European societies, which are part of a wider push to normalize the politics of exclusion. Second, we discuss the manner in which search algorithms, among which Google is of particular interest here, open up for the existence of data voids. In the third and final step in our theoretical scaffolding, we address

the issue of CDS research ethics, which we embed in feminist data ethics.

Critical data studies, digital inequality, and the dynamics of the politics of exclusion

The key endeavour of CDS is to move beyond the initial allure of big data studies that have promised methodological objectivity, whereby scientifically derived hypotheses are tested dispassionately, the results push knowledge forwards, and they offer opportunities to assess social spaces in a quantitative manner by means of searching, aggregating, and cross-referencing large datasets (Boyd and Crawford, 2012: 663, 667). As such, CDS scholarship aims to 'interrogate all forms of potentially depoliticized data science and to track the ways in which data are generated, curated, and how they permeate and exert power on all manner of forms of life' (Iliadis and Russo, 2016: 2). At the same time, researchers in this field acknowledge that 'claims to objectivity are necessarily made by subjects and are based on subjective observations and choices' (Boyd and Crawford, 2012: 667) and consequently argue that 'scholarship needs to be reflexive, situated and nuanced, countering transcendent narratives about data by making clear data's contingent, contextual nature' (Kitchin, 2022: 40).

Critiquing the initial reluctance in the field to acknowledge the political aspect of big data studies, some researchers have warned that big data generate 'data politics' (Bigo et al., 2019; Ruppert et al., 2017), which pertains to both the politics behind data collection, the way in which data are interpreted, and the way the results are deployed for political ends as well as the crystallization of new relations of power and political articulations that are novel in both manner and scale. Illustrating the negative consequences of data politics, the opaque process of algorithm writing that underpins data paves the way to insidious 'weapons of math destruction' (O'Neil, 2016), also labelled 'technological redlining' (Noble, 2018), whereby bad software design and poorly conceived models impact discriminatory effects on already vulnerable and disadvantaged users. In other words, big data are 'not only shaping our social relations, preferences, and life chances but our very democracies' (Ruppert et al., 2017: 2).

In this context, there is a pressing need for users to be able to discern the risks and benefits of the growing complexity of contemporary datafied societies and to recognize the systemic and structural changes brought about by big data (D'Ignazio, 2017; Haider and Sundin, 2019). Indeed, big data generates new power asymmetries and inequalities that exacerbate existing 'digital inequalities' (Dimaggio et al., 2004; Hargittai, 2021), as the 'actors who collect, store and process the data are very different from those whose data are collected, stored and processed' (D'Ignazio, 2017: 15). These digital inequalities become

entrenched, as users whose data are harvested are generally neither encouraged to improve their technical knowledge nor made aware of the datafied intricacies of contemporary societies. Rather, simple answers are provided to them, in part crafted within the extreme-right environment that attempts to weaponize these answers for its political ends (Norocel, 2022: 7). These answers are packaged into emotional appeals and calls to ‘do your own research’, which are part of a wider changing public discourse marked by ‘uncivility, dogmatic politicization and ideologization as well as fact denial’ (Krzyżanowski, 2020: 432). The usual techniques involve ‘practices of recontextualization and reframing’ that are performed by means of ‘a few textual amendments: using certain naming strategies, selective extraction, reformulation of paragraphs or omission of explanatory factors in the story’ (Ekman, 2019: 608).

Some contextualization is necessary here. The dynamics of racism and xenophobia have gained visibility in the public discourse around Europe in the context of the 2015 refugee (reception) crisis and in the aftermath of terrorist attacks in several major European cities between 2015 and 2017 sponsored by various Islamist and Jihadist terror organizations. These events have evidenced the complicated intersections between the politics of preserving national culture, the politics of safeguarding the European welfare model, and the politics of regulating migration flows and promoting integration (Hellström et al., 2020; Norocel, 2017; Wodak, 2015). Researchers have nonetheless noted that in addition to the ‘genuine societal and political debate Islamophobic, xenophobic, and nationalist views have also gained ground, claiming that especially the Muslim minorities constitute a danger to European societies and the Western lifestyle’ (Laaksonen et al., 2020: 3). Deep lines of demarcation have been drawn between ‘the native us’ sharing such ‘civic virtues’ as gender equality and tolerance and those ‘migrant them’, which in its most extreme manifestation coalesces into ‘an identitarian logic of “ethnopluralist” difference which recasts the racist distinctions of racial biology into insurmountable cultural distinctiveness, which privileges “our culture” as both separate and, importantly, discretely better than “their culture”’ (Hellström et al., 2020: 5). The tragic events of early 2022, when people who sought refuge from war-torn Ukraine were welcomed across Europe on grounds of being preponderantly defenseless women with children from a culturally close (and white) European country, seemed only to confirm these assertions. In this context, studies have evidenced the process of ‘normalization of the politics of exclusion [which] enables uncivil, racist, and populist discourse to be recontextualized into – as well as anchored within – the mediated and political spaces of the mainstream that at least nominally were once viewed as largely civil in nature’ (Krzyżanowski et al., 2021: 4).

A growing body of scholarship has made use of big data analytics to provide critical insights into the dynamics of racism and xenophobia emerging across Europe in the past decade (Åkerlund, 2020; Ekman, 2019; Farkas et al., 2018; Laaksonen et al., 2020; Mahoney et al., 2022; Monnier et al., 2021; Nikunen, 2021; Pöyhtäri et al., 2021; Siaperä et al., 2018). For example, research has helped unveil how commercial social media platforms ‘provide spaces for xenophobic, racist and nationalistic discourse online, and they shape antagonistic [...] attitudes towards immigrants’ (Ekman, 2019: 607). More widely, it has enabled the identification of ‘the dynamics and networks of racism’ across both legacy media and social media platforms (Nikunen, 2021: 9). In addition, it has helped researchers map out the manner in which ‘hate speech targeting various ethnic, migrant, and religious minorities flourishes especially in general social media discussions [...], but it has also long taken the forms of organized propaganda, hate groups, and hate sites’ (Laaksonen et al., 2020: 3). Generally, these studies have pointed out that social media platforms contain hostile, negative, and stereotypical framings of migrants (particularly those of Islamic faith), with strong racist undertones, whereby accusations of crimes (especially sexual assault) allegedly committed by both asylum seekers and other migrants are a recurring frame (Åkerlund, 2020; Ekman, 2019; Monnier et al., 2021; Nikunen, 2021; Norocel, 2022; Pöyhtäri et al., 2021). We build and expand on the findings of these studies, which have chosen various social media platforms as their preponderant loci of analysis. In turn, we examine the opportunities for the politics of exclusion that are provided by search engines.

Google search algorithms and data voids

Search engines, particularly Google, have generally not experienced the same level of scholarly scrutiny as social media platforms for actively spreading or at least tacitly enabling misinformation (such as fake news and political and religious manipulations) aimed at normalizing the politics of exclusion (Noble, 2018; Torres and Rogers, 2020). Notwithstanding this, there is a significant body of scholarship that examines commercial search engines such as Google. Critical research has scrutinized the ideological underpinnings of search algorithms (Mager, 2014) and how people attempt to make themselves ‘algorithmically recognizable’ (Gillespie, 2017); unveiled the ‘bent testimony’ of search algorithms (Narayanan and De Cremer, 2022), which is commonly disguised by the seamless ubiquity of search engines in users’ daily lives (Haider and Sundin, 2019; Sundin et al., 2017); or provided a cultural retrospective of the process of Googlization and its information politics (Rogers, 2018; Vaidhyanathan, 2011).

Whenever we search for information, it seems that ‘Google dominates all aspects of the search engine

market, largest share of searches on all devices, most visitors to its site, most users of its browser, Chrome, which in effect is a search engine, and so on' (Haider and Sundin, 2019: 11). Researchers have shown that whether at leisure or 'doing our own research', whenever searching on Google, we find ourselves in the situation of 'asking someone about whom to ask for an answer. And any act of googling [...] by its very nature is dependent upon the beliefs and actions of other people. In that way, googling is more like testimony' (Gunn and Lynch, 2018: 43). Those testifying to our queries are those producing online content, be it news websites, personal blogs and videos, or discussion forums. Consequently, they have an impact on the beliefs we hold about the matter at hand by means of the content they produce and make algorithmically recognizable (Gillespie, 2017: 65), in a sense generating a collaborative framework for 'the validation of truth by public opinion' (Oleinik, 2022: 186). 'As such, googling resembles testimony because of our epistemic reliance on those who produce the content that gets presented to us when we perform a Google search' (Narayanan and De Cremer, 2022: 3).

Notwithstanding this, the impact of the search results is mediated by and through Google's own mix of content-based and collaborative algorithms (Oleinik, 2022: 186; Rogers, 2018: 7), which play a directive role in ranking, structuring, and presenting the information to which we are exposed (Pan et al., 2007; Schultheiß et al., 2018). This is important for several reasons. First, search algorithms compile and rank all results pertaining to a query. This ranking has a great influence on what sort of information users access and read, with most users focusing on links higher up in the search results (Pan et al., 2007; Schultheiß et al., 2018). Second, users are generally unaware of the commodification of their searches, and a large portion are unable to distinguish between organic results and paid advertisements on Google's search engine results page (SERP) (Lewandowski and Schultheiß, 2022). Some laboratory studies have even shown that users choose the top results even when they are less relevant to the query (Pan et al., 2007; Schultheiß et al., 2018) or less credible than results that are displayed lower in a ranked list (Haas and Unkel, 2017; Unkel and Haas, 2017).

This brings us to the issue of how the information provided by Google at the end of the complex process described above is received by users. It seems that users trust the information they find in their searches (European Commission, 2016; Purcell et al., 2012), although many are not familiar with the practices of influencing search engine results by means of search engine optimization (SEO) (Lewandowski and Schultheiß, 2022). Furthermore, users with lower levels of search engine knowledge are more likely to trust and use Google than those with more knowledge (Schultheiß and Lewandowski, 2021). Notably, in this context, Google has recently become the centre of an intense debate about

misinformation and the search engine's role in 'the appearance of misogynistic and extremist content that the company previously defended as "reflective" of societal concern rather than the product of algorithmic error or "culture hacking"' (Torres and Rogers, 2020: 100).

One of the explanations for the presence of such content in search results is the existence of data voids, which 'occur when obscure search queries have few results associated with them, making them ripe for exploitation by media manipulators with ideological, economic, or political agendas' (Golebiewski and Boyd, 2019: 2). Indeed, when the pool of results that are potentially relevant to a query is limited, the search algorithm simply ranks what is available, even in circumstances in which the links in the pool lead to information that, from the point of view of societal relevance, is of low quality (Haider and Sundin, 2019: 9–10; Sundin et al., 2022: 640). Here, we distinguish information that may have societal relevance, such as, on our topic of concern, the number of migrant families with small children seeking refuge in a certain location, which may be relevant from the point of view of organizing local reception efforts. There is also skewed and manipulative information, which is thus of low quality from the point of view of its societal relevance writ large but may have relevance for the extreme-right environment specifically. For example, the arrival of migrant men in the aforementioned location may be presented as an impending danger of sexual assault for local women, which may serve as an impetus for local extreme-right mobilization. This means that the ranking of results is not an absolute measure of quality; rather, the ranking reflects the relevance of query results *in relation to each other*. Therefore, a major obstacle in researching data voids is detecting them. More often than not, data voids are filled quickly when they are detected. For instance, in a search for the conspiracy-related term 'chemtrails', the query initially yielded results directing the user to conspiracy theory sources (Ballatore, 2015). As the mainstream news media started to report on the matter, the initial results were demoted in rankings and news sources, and links to Wikipedia and mainstream news outlets eventually appeared at the top of the results. This had a major impact on the design of our study, which is detailed below in the 'Methods and data' section.

Research ethics and situated knowledge

Given that this study is anchored in the field of CDS, we are acutely aware that 'how data are ontologically defined and delimited is cast not as a value-free, technical process, but a normative, ideological and ethical one that has concrete consequences for subsequent analysis, interpretation, and action' (Kitchin, 2022: 16). Having clarified the normative and ideological underpinnings of this study, we now approach the interrelated issues of research ethics and

situated knowledge (Boyd and Crawford, 2012; Daniels, 2015; D'Ignazio and Klein, 2020; Nikunen, 2021).

Concerning research ethics, we are aware that big data analytics are reliant on datasets that are historically and socially contingent, generated and imbued with political and ethical values (Daniels, 2015; Metcalf and Crawford, 2016). There are two issues to be discussed here. First, as computational skills are highly prized in big data analytics, this generates new digital hierarchies centred around researchers (most often men) who possess these skills. Nonetheless, we pursue this research project as a multidisciplinary research team, 'recognizing that computer scientists and social scientists both have valuable perspectives to offer' (Boyd and Crawford, 2012: 674). Second, we consider issues of accountability, carefully weighing the possible ramifications of big data analytics in this project and the differentiated effects that our study may have in terms of reproducing discrimination and enforcing inequalities along such intersectional axes of social structuring as gender and sexuality, race and/or ethnicity, and social class (Hill Collins, 2019). Regarding this matter, we carefully assess how, in the process of generating the dataset for our study, as described in the 'Methods and data' section below, our repeated queries may impact future search results (Lewandowski and Sünkler, 2019) by potentially allowing extreme-right sources to ascend in the overall ranking. As we sent each query only once, we are confident that the overall effect on future results and query recommendations is low, if it exists at all.

With this in mind, we subscribe to the calls for thorough ethical reflection, especially on how big data analytics may cause both individual and networked harm (Metcalf and Crawford, 2016: 3), and embrace the seven imperatives of data feminism: 'examine power, challenge power, elevate emotion and e[...].mbodiment, rethink binaries and hierarchies, embrace pluralism, consider context, and make labor visible' (D'Ignazio and Klein, 2020: 213). Armed with this ethical awareness, we have striven, programmatically and consistently, to avoid recirculating discriminatory framing and stereotyping of vulnerable and disadvantaged people (Nikunen, 2021: 3).

Methods and data

In this section, we detail the methods that contribute to our approach to big data analytics. First, we explain the principles of search engine queries and the role of specific and localized queries in identifying data voids. Second, we design a catalogue of queries pertaining to the issue of migration on a continuum from mainstream to extreme-right vocabularies. Third, we describe how we construed our specific and localized queries. Fourth, we explain the data collection and detail the characteristics of the dataset.

First, given that the more general a search query, the greater the pool of sources a search engine such as

Google may use for its ranking, we decided to generate a dataset for our study in which the queries become both more specific (by adding a longer string of keywords) and localized (by adding the names of all German and Swedish municipalities). When more keywords are added, the query yields fewer results. This follows from Boolean logic: The number of results to $A \wedge B$ must be smaller or equivalent to A . Generally, this means that the more keywords are added to a search string, the fewer results it produces (assuming that a search engine sets the Boolean AND as a standard). An exception may occur when a search engine rewrites the query by adding synonyms or related terms. This can be especially rewarding when the query yields a low number of results, and the results set can be enriched through this strategy. On a more general level, one may even view this as a strategy for filling data voids. When the name of a municipality is added, the query yields results specific to the location in question. As such, we assume not only that these queries yield fewer results but also that there is greater potential for identifying data voids with these queries, especially for smaller municipalities.

Second, based on previous big data analytics addressing the politics of exclusion (Mahoney et al., 2022; Monnier et al., 2021; Nikunen, 2021; Pöyhtäri et al., 2021; Siapera et al., 2018), we designed a catalogue of queries concerning the issue of migration in both Germany and Sweden on a continuum from mainstream to extreme-right vocabularies. We ranked them from innocuous queries containing words/concepts that would commonly occur in a casual conversation on the topic, informed by ongoing debates in mainstream media, to queries containing extreme-right 'red pills' – words or combinations of words that have been developed as part of a specific extreme-right vocabulary used to describe this topic. To develop this catalogue, we turned to previous analyses of the media debates in the two countries (Ekman, 2019; Klawier et al., 2022) as well as expert reports on the politics of exclusion by Non-Governmental Organisations (NGOs) working to strengthen civil society and fight against extremism, racism, and other forms of bigotry: Amadeu Antonio Foundation in Germany (2015; 2016) and Expo in Sweden (2014; 2016a; 2016b). In total, we chose nine queries per country (three from each category) for the empirical part of our research. Our catalogue sorts queries into three discretely distinct categories:

- *Level A queries.* This category contains keywords used in the mainstream media to frame the events. An ordinary (moderate) user in either Germany or Sweden with an interest in the issue of migration may search for these terms. In German, we chose 'Flüchtlingskrise' (refugee crisis), 'Masseneinwanderung' (mass migration), and 'Flüchtlingswell' (refugee wave). In Swedish, we chose

- ‘flyktingskris’ (refugee crisis), ‘flyktingsvåg’ (refugee wave), and ‘asylvåg’ (asylum wave).
- *Level B queries.* This category contains keywords that move beyond the mainstream political debate and that an ordinary user may identify as specific to a radical-right populist manner of describing the issue of migration and therefore use in their searches as a means to become acquainted with the radical-right populist political agenda. In German, we chose ‘Asylschmarotzer’ (asylum parasite), ‘Asylmissbrauch’ (asylum abuse), and ‘Asylflut’ (asylum wave). In Swedish, we chose ‘asylparasit’ (asylum parasite), ‘massinvandring’ (mass migration), and ‘rasförrädare’ (race traitor).
 - *Level C queries.* This category contains keywords specifically developed within the extreme-right environment that aim to further polarize and radicalize attitudes towards migration and that are rarely known to an ordinary user. In German, we chose ‘Rapefugees’ (a combination of rape and refugees), ‘Krimigranten’ (a combination of crime and migrants), and ‘Moslemterror’ (Muslim terror). In Swedish, we chose ‘invandrarvåldtäkt’ (a combination of migrant and rape), ‘kulturberikare’ (a juxtaposition of culture and enricher, with derogatory meaning), and ‘skäggbarn’ (bearded child, an allegation of false documentation of adults as children).

We are aware that the distinctions between these categories are rather fluid. This fluidity is due to the continuous process of mainstreaming radical-right populist points of view on these matters (Hellström et al., 2020; Krzyżanowski, 2020; Krzyżanowski et al., 2021; Norocel, 2017; 2022) and to the continuous efforts of extreme-right entities to manipulate the mainstream debate towards polarized and extremist views in general (Åkerlund 2020; Ekman, 2019).

Third, to assess the existence of data voids, we designed a series of Google searches that combine the selected keywords from this catalogue of queries with the names of municipalities in the two countries, such as ‘Asylschmarotzer Schwetzingen’ in German or ‘asylparasit Västerås’ in Swedish. To do so, we used a list of all 2055 German municipalities that are independent towns/cities (comprising 61 million inhabitants) from the German Federal Statistical Office (Destatis³). We cleaned the data by removing identifiers from the official municipality names to obtain a form that a ‘common person’ would use (such as ‘Frankfurt’ instead of ‘Frankfurt an der Oder’ or ‘Bernau’ instead of ‘Bernau bei Berlin’). In Sweden, we collected data on all 290 municipalities from Statistics Sweden (SCB⁴). This process resulted in a total of 18,495 queries for Germany and 2610 queries for Sweden.

Our choice was motivated by the fact that search engine algorithms, particularly Google’s, seem to be trained to prioritize mainstream news content, authoritative sources, and Wikipedia entries (Lewandowski and Spree, 2011; Sundin

et al., 2022). At the same time, search engines continuously fine-tune their algorithms to demote extreme-right content; consequently, ordinary users may not be exposed to extreme-right attempts to manipulate them (Torres and Rogers, 2020). However, searches with a local focus, such as those in which the name of a specific municipality is added, are seemingly open to data voids, as these searches typically result in fewer results – or, in the case of smaller municipalities, in only a handful of results. When there is not much competition on a query, it is much easier for external actors to make their content visible in the search results. Particularly, when there is little or no mainstream content in response to a query, extreme-right radical content may fill these data voids.

Fourth, to collect the data, we used Google. It is by far the most popular search engine in the two countries; it has a market share of 90% in Germany and 93% in Sweden (StatCounter, 2022). To query Google using the selected keyword and municipality combinations, we employed software that automatically sends queries to Google and scrapes the results pages (Lewandowski et al., 2012). Scraping is a method whereby data from web pages are collected by a machine using a list of URLs as a starting point. This allows for easy and large-scale collection of data from the web. When screen scraping is applied to search engines, the process is more complicated, as the input data are a list of queries; thus, the scraper needs to query the search engine, grab the SERP for each query, and then extract the resulting URLs and their positions from the SERP. As search engine providers change the structure of these pages quite frequently and often take measures to prevent automatic querying, sophisticated software is needed that simulates real users and schedules queries so as not to query a particular search engine too frequently.

For scraping, we used the query of search engines based on GET parameters to process our search queries and the desired result positions. The typical search query URL from Google with the search query and position appears as follows: <https://www.google.de/search?q=QUERY&start=POS>. In this URL, both the location (here Google Germany, google.de) and the search query with the parameter q and the start position are specified. If we now query Google, we replace q with a search query from our database and set a start value that begins at 0. We generate a total of three scraper jobs per query, each collecting ten results (1–10; 11–20; and 21–30), leading to the top 30 results per query. Based on findings regarding user focus on top results and the fact that especially for smaller municipalities, there are often not many results, we deem 30 results to be a sufficiently large number for our analysis. Furthermore, it is more efficient to scrape a smaller number of results per query for the reasons given above.

The technical implementation of scraping and saving the source code of the search results is done using the Selenium

web driver and the HTML library Beautiful Soup⁵. This combination allows us to save web pages as they are displayed in a web browser, i.e., by capturing dynamically generated content and executing JavaScript. Selenium is a cross-browser test suite for websites and therefore provides an interface with various web browsers. It practically simulates users of web browsers and their interactions. Additionally, it is able to call up web pages as they are displayed in a browser. The scraping process continues until all results for the queries stored in the list are scraped. The identified URLs and their positions are stored in a database. Data were collected between 8 and 24 June 2021, which yielded a dataset with the characteristics described in Table 1. Notably, the German queries produced a larger average number of results per query, pointing to more content being available in response to those queries. The German queries clearly showed that the average number of results decreased by query level A > level B > level C, whereas the average number of results was relatively stable in the case of the Swedish queries.

To classify the top sources, we use ad hoc classification, namely, categories developed from the data. This hands-on approach allows us to gain initial insights into the types of sources shown in the search results. We choose this approach because we are confident that it is sufficient for the rather low number of different sources that we consider in the classification.

Results

Top sources

The top sources (domains) found in our dataset are shown in Table 2 for Germany and Table 3 for Sweden. The tables show the top 10 sources per query level, the number of occurrences, the percentage of results accumulated by the respective sources, the cumulative percentage of occurrences of the top N sources, and the domain source types. In the analysis of the top sources, we notice that the source concentration differs between query levels. The source concentration is higher for level C > level B > level A. This holds true for both Germany and Sweden.

For level A queries, the top 10 sources account for 11.31% of the German results and 22.61% of the Swedish results. For level B queries, the top 10 sources account for 20.23% of the German results and 33.12% of the Swedish results. For level C queries, the top 10 sources account for 31.51% of the German results and 40.28% of the Swedish results. We find a variety of source types in the top results, ranging from mainstream (news and governmental) websites to dedicated extreme-right websites. An interesting case is social media platforms, document-sharing platforms, and blog hosts, as these may provide content from a whole range of political standpoints, both mainstream and more extreme. For instance, within our dataset, we find that WordPress hosts, on the one hand, blogs with an extreme-right agenda (such as ‘Widerworte: Spotlights aus Absurdistan’ [Answering back: Spotlights from Absurdistan]⁶ in Germany or ‘Petterssons gör Sverige lagom!’ [Pettersson’s (blog) makes Sweden fair]⁷ in Sweden) and, on the other hand, blogs that oppose the extreme right (such as ‘Jugend gegen Rassismus Hannover’ [Youth against Racism Hanover]⁸).

Top sources in Germany

The top sources for level A queries show a mix of open platforms (such as Facebook, DocPlayer, and Wikipedia), news sites, and governmental websites. While open platforms may host content of any kind, the rest of the top sources are traditional mainstream sites from either the government or publishing houses. The top sources for level B queries are a mixture of open platforms, governmental websites, blog hosts, and an extreme-right news platform. We can see that as the query level moves beyond the mainstream political debate, the distribution of sources changes to a mix of mainstream and extreme-right fringe sources. The top source for level C queries is pi-news.net, an extreme-right news website, which accounts for 10.03% of all results. Another extreme-right website, rapefugees.net, adds another 2.77% to the resulting total. The other top sources in this category are mainly websites that host content, such as social media platforms, blog hosts, and

Table 1. Characteristics of the dataset.

	Germany	Sweden
Number of municipalities	2055	290
Number of queries	18,495	2610
Number of results	341,315	37,266
Average number of results per query, level A	Mean: 28.16 (STD 4.46) Median: 29	Mean: 14.87 (STD 12.80) Median: 8
Average number of results per query, level B	Mean: 20.33 (STD 9.23) Median: 23	Mean: 17.72 (STD 11.88) Median: 21
Average number of results per query, level C	Mean: 9.93 (STD 9.28) Median: 6	Mean: 17.97 (STD 10.70) Median: 10

Table 2. Top domains per query level (Germany).

Level		Domain	Count	%	% Cumulative	Source type
A	1	facebook.com	3777	2.29	2.29	Social media platform
	2	docplayer.org	3491	2.11	4.40	Document sharing platform
	3	wikipedia.org	2248	1.36	5.76	Online encyclopedia
	4	bayern.de	1635	0.99	6.75	Government website
	5	sachsen.de	1594	0.97	7.72	Government website
	6	faz.net	1483	0.90	8.62	National newspaper
	7	genios.de	1412	0.86	9.47	News archive
	8	bundestag.de	1028	0.62	10.09	Government website
	9	springer.com	1010	0.61	10.71	Academic publisher
	10	nrw.de	1003	0.61	11.31	Government website
B	1	facebook.com	5301	4.36	4.36	Social media platform
	2	docplayer.org	3970	3.27	7.62	Document sharing platform
	3	blogspot.com	3134	2,58	10.20	Blog host
	4	bundestag.de	2517	2.07	12.27	Government website
	5	pi-news.net	2247	1.85	14.12	Extreme-right news platform
	6	wordpress.com	1777	1.46	15.58	Blog host
	7	issuu.com	1573	1.29	16.88	Document sharing platform
	8	unionpedia.org	1420	1.17	18.04	Concept map generated from Wikipedia
	9	sachsen.de	1371	1.13	19.17	Government website
	10	bayern.de	1284	1.06	20.23	Government website
C	1	pi-news.net	5477	10.03	10.03	Extreme-right website
	2	wordpress.com	3442	6.30	16.34	Blog host
	3	rapefugees.net	1511	2.77	19.10	Extreme-right website
	4	docplayer.org	1419	2.60	21.70	Document sharing platform
	5	twitter.com	1392	2.55	24.25	Social media platform
	6	idz-jena.de	1071	1.96	26.22	Research centre
	7	blogspot.com	780	1.43	27.64	Blog host
	8	facebook.com	734	1.34	28.99	Social media platform
	9	rssing.com	719	1.32	30.31	RSS feed aggregator
	10	nrw.de	659	1.21	31.51	Government website

an RSS feed aggregator. What these websites have in common is that they do not have an editorial focus. Only two websites shown for level C queries are governmental (nrw.de) and research sources (idz-jena.de), accounting for 1.21% and 1.96% of occurrences, respectively. IDZ Jena is a research centre for democracy and civil society that provides information on extreme-right activity. In our view, this shows that if mainstream sources produce content relevant to extreme-right queries, they have a good chance of appearing in search engine results.

Top sources in Sweden

The top sources for level A queries show a mix of open platforms (such as DocPlayer, Facebook, and an RSS feed aggregator), governmental websites (ranging from local government organizations to local and regional administrative entities), and state-owned news sites (radio and TV). As in the German case, while open platforms may host any kind of content, the remainder of the sources are traditional mainstream sites. The top sources for level B queries, as in Germany, show a mixture of blog hosts, open platforms (such as Facebook, DocPlayer, Twitter), governmental websites (ranging from regional

libraries to the state agency for crime prevention), and a self-styled anti-establishment internet forum (flashback.org) that hosts discussion threads about controversial issues, such as extreme-right mobilization in Sweden (Blomberg and Stier, 2019). Given that this query level moves beyond the mainstream, we notice a change in the distribution of sources, which are now a mix of mainstream and extreme-right fringe sources, as in Germany. The level C queries show a significant change in the top sources compared to the previous level. Blog hosts, open platforms, and the aforementioned internet forum (flashback.org) appear higher up among the top sources. Notably, a self-titled ‘immigrant critical’ personal blog (petterssonsblogg.se) accounts for 2.89% of the total results. In turn, specialist governmental and inter-governmental websites (namely, the state agency for crime prevention and the antidiscrimination information system) and the partial archive of a university research project in linguistics (svn.spraakdata.gu.se) appear lower in the ranking. The remainder of the top sources in this category are an open-source platform provider and an RSS feed aggregator. As in the German case, we interpret the presence of governmental and intergovernmental websites among the top sources at this level as an indication that they produce content countering

Table 3. Top domains per query level (Sweden).

Level		Domain	Count	%	% Cumulative	Source type
A	1	docplayer.se	448	3.58	3.58	Document sharing platform
	2	facebook.com	381	3.05	6.63	Social media platform
	3	pressen.se	377	3.01	9.64	RSS feed aggregator (Swedish news)
	4	skr.se	325	2.60	12.24	Local government organization
	5	mala.se	274	2.19	14.43	Governmental website (municipality)
	6	sverigesradio.se	241	1.93	16.36	Governmental website/radio (state-owned)
	7	gotland.se	225	1.80	18.15	Governmental website (region)
	8	diva-portal.org	198	1.58	19.74	Institutional repository (university)
	9	riksdagen.se	193	1.54	21.28	Governmental website
	10	svt.se	166	1.33	22.61	Governmental website/TV (state-owned)
B	1	wordpress.com	1021	8.55	8.55	Blog host
	2	facebook.com	553	4.63	13.18	Social media platform
	3	bra.se	510	4.27	17.45	Governmental website
	4	docplayer.se	444	3.72	21.17	Document sharing platform
	5	barnensbibliotek.se	292	2.44	23.61	Governmental website (library)
	6	twitter.com	280	2.34	25.95	Social media platform
	7	lagen.nu	249	2.08	28.04	Private website (legal lexicon)
	8	flashback.org	211	1.77	29.81	Internet forum (in Swedish)
	9	blogspot.com	202	1.69	31.50	Blog host
	10	diva-portal.org	194	1.62	33.12	Institutional repository (university)
C	1	wordpress.com	1358	10.60	10.60	Blog host
	2	flashback.org	704	5.49	16.09	Internet forum (in Swedish)
	3	blogspot.com	516	4.03	20.12	Blog host
	4	facebook.com	457	3.57	23.69	Social media platform
	5	bra.se	416	3.25	26.93	Governmental website
	6	tandis.odihl.pl	379	2.96	29.89	Intergovernmental information platform
	7	petterssonsblogg.se	370	2.89	32.78	Self-titled 'immigrant critical' personal blog
	8	huggingface.co	359	2.80	35.58	Open-source platform provider
	9	rssing.com	323	2.52	38.10	RSS feed aggregator
	10	svn.spraakdata.gu.se	279	2.18	40.28	Research project (university)

that is produced by extreme-right sources and therefore are relevant to these queries.

Source concentration

To measure source concentration, we have adapted the Gini coefficient, which is a measure of statistical dispersion used to measure income or wealth inequality (Gini, 1936). The Gini coefficient is a single number ranging from 0 to 1, where 0 represents perfect equality and 1 represents maximum inequality. Adapting this measure to the distribution of sources in search result sets allows us to easily compare the distributions between query levels as well as between individual queries. In the case of source distribution in search results, the lower the Gini coefficient, the more equally the results are distributed over all sources that contribute results to a certain query. It should be noted that the Gini coefficient considers only the distribution of sources, not the total number of sources that contribute to the results. Therefore, we add the total number of sources per query and per level to our analysis.

Figures 1 and 2 show the source distributions for Germany and Sweden, respectively. In the first column (a, e, i), the

aggregated data for the query levels are shown, while the rest of the plots show results for individual queries. Each diagram shows the Gini coefficient, the distribution of sources, and the number of sources that contribute to the respective result set. Within each diagram, the Lorenz curve shows the proportion of sources (y-axis) and the proportion they cumulatively contribute to the total results (x-axis). The line at 45 degrees represents perfect equality of sources.

The Gini coefficient on level A is 0.78 for Germany and 0.76 for Sweden. For the German results, it increases to 0.85 at both level B and level C. In contrast, for the Swedish results, it increases to 0.78 at level B and 0.83 at level C. This shows that at least level A < level C for both countries. The number of sources decreases from level A to C in Germany, with the total number of sources decreasing from more than 150,000 to less than 6000. For Sweden, the total number of sources is lower, with approximately 12,000 sources on all query levels.

On the level of the individual queries, we find that for the German results, there is a clear distinction between query level A vs. levels B and C. All queries on level A have a Gini coefficient < 0.8, whereas all queries on levels B and C have a Gini coefficient ≥ 0.80 . However, there is no difference

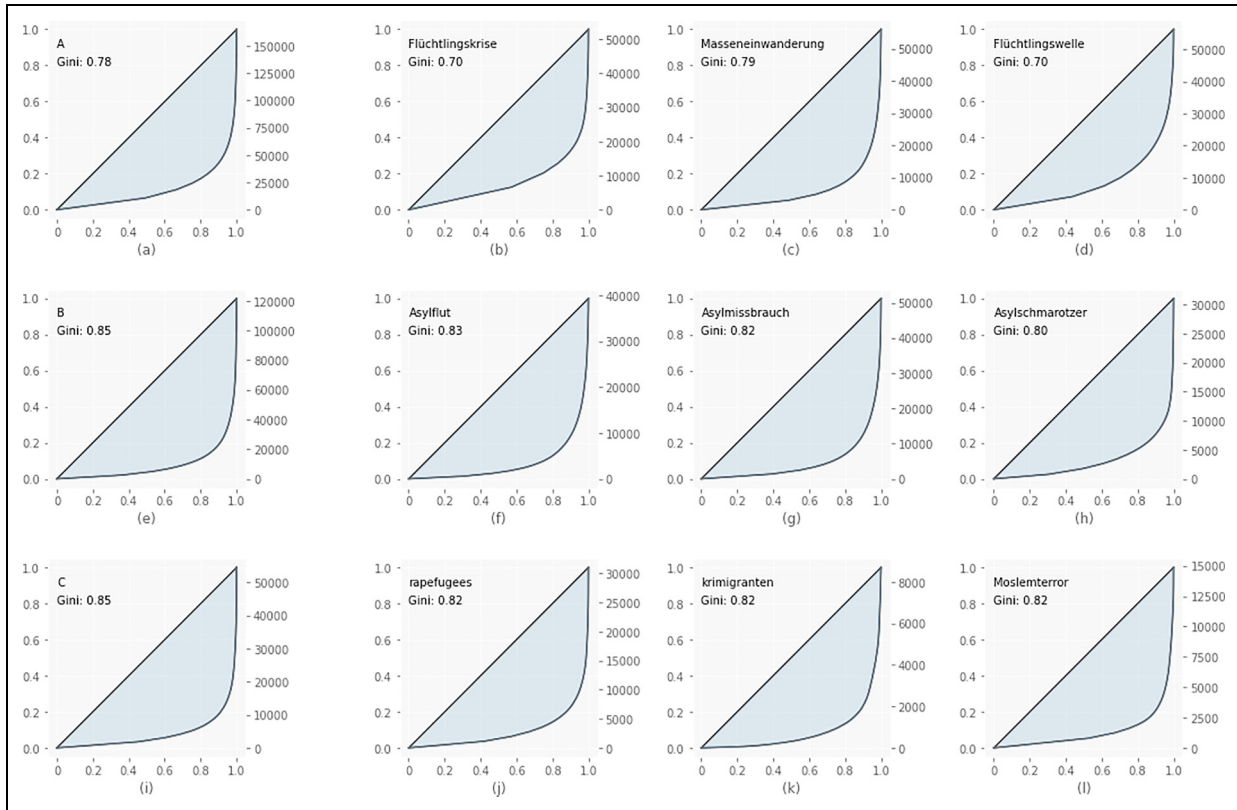


Figure 1. Source distribution for Germany – aggregated data for level A queries in (a), presented per keyword in (b), (c), (d); aggregated data for level B queries in (e), presented per keyword in (f), (g), (h); aggregated data for level C queries in (i), presented per keyword in (j), (k), (l).

between queries on level B and level C. The Swedish results are more dispersed and do not allow for a clear distinction between queries on different levels. For instance, the Gini coefficient for the query ‘invandrarvåldtäkt’ (level C) is 0.54, which is much lower than for query level C overall (0.83). It should be noted, however, that the two queries with the lowest Gini coefficient (‘asylparasit’ and ‘invandrarvåldtäkt’) also have the lowest numbers of sources (157 and 243, respectively).

The analysis shows that the distribution of sources contributing results differs between queries – and, at least in part, between query levels. This notwithstanding, we cannot make a clear distinction showing that the Gini coefficient rises from query level to query level. The assumption that when more mainstream sources are available, these (few) sources contribute more results ($A > B > C$) does not hold true. However, neither does the opposite ($C > B > A$) hold true, meaning that the search engine algorithms aggregate results to the few sources available or to the few mainstream sources available for problematic queries.

Discussion and conclusion

In this study, we deploy a critical approach to big data analytics to assess the tentative contours of data voids reflecting

extreme-right dynamics of exclusion by concentrating on the topic of migration in Germany and Sweden. We argue that by concentrating on the role of search engines in normalizing the politics of exclusion (Noble, 2018; Torres and Rogers, 2020), our study adds further complexity to big data analyses that have examined the dynamics of racism and xenophobia in the European context (Åkerlund, 2020; Ekman, 2019; Farkas et al., 2018; Laaksonen et al., 2020; Mahoney et al., 2022; Monnier et al., 2021; Nikunen, 2021; Pöyhtäri et al., 2021; Siapera et al., 2018). Furthermore, our study expands the methodological framework to investigate data voids outside the USA context (Golebiewski and Boyd, 2019) by proposing a catalogue of queries that enables specific and localized queries to identify data voids in the German and Swedish contexts.

Our analysis shows the dominance of mainstream sources for level A queries. This can be interpreted as Google prioritizing such sources whenever available, mainly by measuring source popularity in its ranking algorithms (Lewandowski, 2012; Sundin et al., 2022). However, it should be kept in mind that such correlations do not mean causality: It may well be other reasons that lead to such a distribution; for instance, mainly mainstream

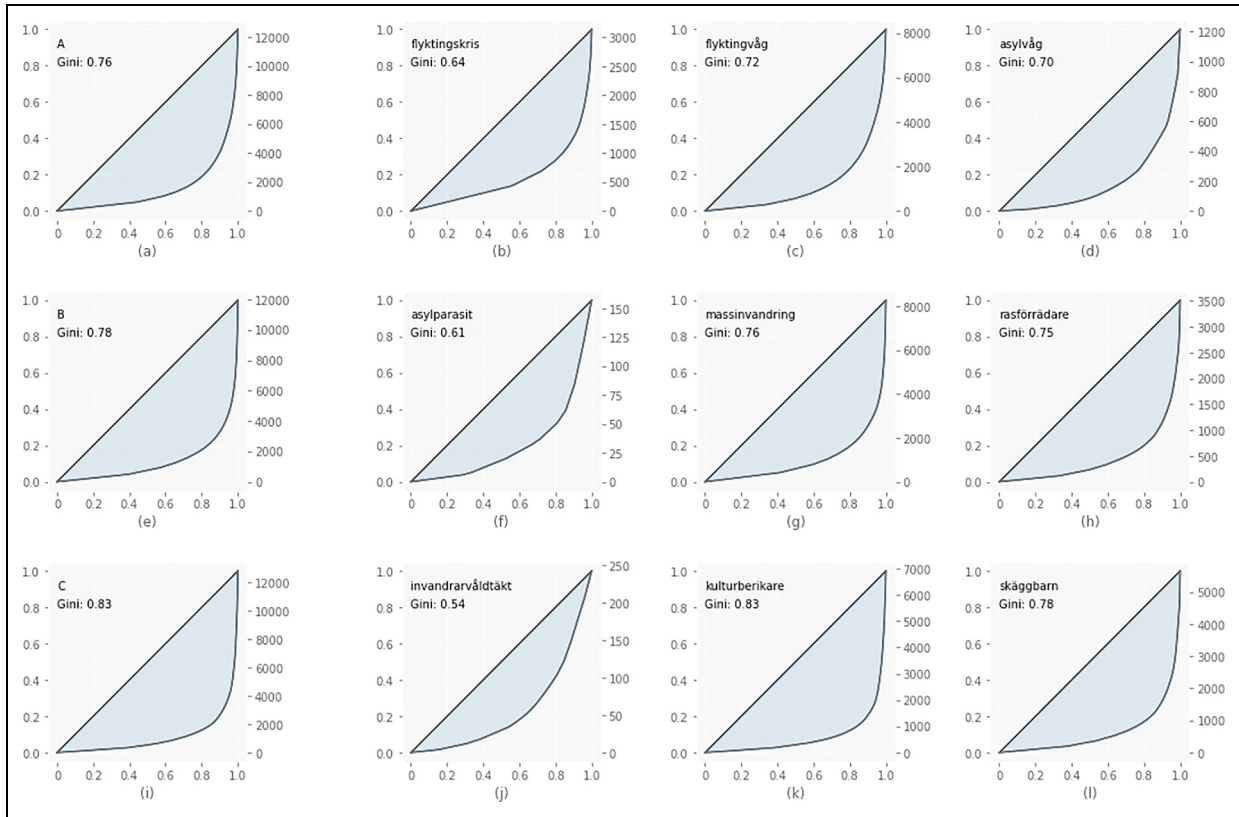


Figure 2. Source distribution for Sweden – aggregated data for level A queries in (a), presented per keyword in (b), (c), (d); aggregated data for level B queries in (e), presented per keyword in (f), (g), (h); aggregated data for level C queries in (i), presented per keyword in (j), (k), (l).

sources may contain relevant content. We also found that popular sites for user-generated content, such as Facebook, WordPress, and DocPlayer, are highly ranked on all query levels, thereby allowing such user-generated content to rank highly, presumably not because of its merit but because of the popularity of the source where it is published. This allows extreme-right fringe entities to easily make their content visible in search engines. Our analysis of source concentration shows an increase in the Gini coefficient from level A to level C, indicating that sources are distributed more unevenly when there is less diversity in them. An example is the extreme-right news website *pi-news.net* in Germany. However, one should remember, first, that a search engine can rank only the documents it finds on the web and second, that the resources for crawling and storing content may limit a search engine’s coverage. When there is not much content for a given query, it simply ranks these documents in relation to each other, regardless of the actual ranking score they receive. One way to address this problem can already be seen in Google, which in some cases shows a notice warning of low-quality results if a query yields only a few results that have low relevance according to Google’s ranking function.

It is worth noting that in pursuing this study, we have paid attention to ethical issues specific to big data analytics (Daniels, 2015; D’Ignazio and Klein, 2020; Metcalf and Crawford, 2016), especially striving to avoid discriminatory framing of vulnerable and disadvantaged people (Nikunen, 2021). We are nonetheless aware that we are not able to assess how often users would actually search for these queries, especially in combination with the name of any given municipality. This is a general problem with researching search results since there is no standard for modeling query sets on actual user querying behaviour, although some researchers have tried to develop more systematic query sets (see Lewandowski and Sünkler, 2019). Concomitantly, we acknowledge the ethical ramifications of deploying this catalogue of queries to generate the dataset for our study. We were able to navigate these ethical complexities because of our heterodox approach and the multidisciplinary character of our research team – combining extensive expertise in data science approaches to information research and information retrieval (e.g., Lewandowski, 2012; Lewandowski et al., 2012; Lewandowski and Sünkler, 2019; Schultheiß et al., 2018) with knowledge in the political and media mechanisms of exclusionary politics from a feminist perspective (e.g.,

Hellström et al., 2020; Norocel, 2017; 2022) – which were guided by the shared ethical imperative to avoid individual and networked harm (Daniels, 2015; D’Ignazio and Klein, 2020; Metcalf and Crawford, 2016).

From a big data analytics perspective, then, our study analyzes a somewhat small dataset containing only combinations of nine preselected keywords and names of municipalities from Germany and Sweden, resulting in a grand total of 21,105 queries. Considering the recent developments across Europe, with a growing number of people seeking refuge from Ukraine, future research could endeavour to expand the number of countries and enlarge the set of keywords. Another avenue for further research could be to expand the scope of the analysis, which considered only the sources in the results of our queries, and delve more deeply into the content of the results that are not immediately obvious, particularly of the social media platforms, document sharing platforms, and blog hosts that we have identified as being among the top sources.

Acknowledgements

We thank the Pufendorf Institute for Advanced Studies at Lund University, the Theme ‘In Search of Search and Its Engines’ (PIs Olof Sundin and Alison Gerber), for introducing the coauthors to one another. We also thank Sebastian Sünkler and Nurce Yagci for their help in collecting and analyzing the data.



Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The work for this study was supported by the Swedish Research Council (*Vetenskapsrådet*, VR) (grant number 2019-03363) (Norocel) and the German Research Foundation (*Deutsche Forschungsgemeinschaft*, DFG) (grant number 417552432) (Lewandowski).

ORCID iDs

Ov Cristian Norocel  <https://orcid.org/0000-0002-7349-4000>
Dirk Lewandowski  <https://orcid.org/0000-0002-2674-9509>

Notes

1. <https://www.dw.com/en/germany-right-wing-extremists/a-54105110> (Accessed 21 November 2022).
2. <https://www.buzzfeednews.com/article/lesterfeder/how-sweden-became-the-most-alt-right-country-in-europe> (Accessed 21 November 2022).
3. https://www.destatis.de/DE/Home/_inhalt.html
4. <https://scb.se/>
5. For full details on the implementation, see <https://osf.io/vzehr/> (Accessed 21 November 2022).
6. <https://widerworte.wordpress.com/> (Accessed 21 November 2022).
7. <https://petterssonsblogg.se/> (Accessed 21 November 2022).

8. <https://jgrhannover.wordpress.com/> (Accessed 21 November 2022).

References

- Åkerlund M (2020) The importance of influential users in (re)producing Swedish far-right discourse on Twitter. *European Journal of Communication* 35(6): 613–628.
- Amadeu Antonio Foundation (2015) Geh Sterben! Umgang mit Hate Speech und Kommentaren im Internet [Go Die! Dealing with Hate Speech and Comments on the Internet]. Berlin. <https://www.amadeu-antonio-stiftung.de/w/files/pdfs/hatespeech.pdf>.
- Amadeu Antonio Foundation (2016) Monitoringbericht 2015/2016: Rechtsextreme und menschenverachtende Phänomene im Social Web [Monitoring Report 2015/2016: Right-wing extremist and dehumanising phenomena on the social web]. Berlin. <https://www.amadeu-antonio-stiftung.de/w/files/pdfs/monitoringbericht-2015.pdf>.
- Ballatore A (2015) Google chemtrails: A methodology to analyze topic representation in search engine results. *First Monday* 20(7). <https://doi.org/10.5210/fm.v20i7.5597>.
- Bigo D, Isin E and Ruppert E (2019) Data politics. In: Bigo D, Isin E and Ruppert E (eds) *Data Politics: Worlds, Subjects, Rights*. London: Routledge, pp. 1–17.
- Blomberg H and Stier J (2019) Flashback as a rhetorical online battleground: Debating the (Dis)guise of the nordic resistance movement. *Social Media + Society* 5(1). <https://doi.org/10.1177/2056305118823336>.
- Boyd D and Crawford K (2012) Critical questions for big data: Provocations for a cultural, technological and scholarly phenomenon. *Information, Communication & Society* 15(5): 662–679.
- Daniels J (2015) “My brain database doesn’t see skin color”. Colorblind racism in the technology industry and in theorizing the web. *American Behavioural Scientist* 59(11): 1377–1393.
- D’Ignazio C (2017) Creative data literacy: Bridging the gap between the data-haves and data-have nots. *Information Design Journal* 23(1): 6–18.
- D’Ignazio C and Klein LF (2020) *Data Feminism*. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/11805.001.0001>
- Dimaggio P, Hargittai E, Celeste C, et al. (2004) Digital inequality: From unequal access to differentiated use. In: Neckerman KM (ed) *Social Inequality*. New York: Russell Sage Foundation, pp. 355–400.
- Ekman M (2019) Anti-immigration and racist discourse in social media. *European Journal of Communication* 34(6): 606–618.
- European Commission (2016) *Special Eurobarometer 447 – Online Platforms*. Brussels: European Commission. <https://doi.org/10.2759/937517>
- Expo (2014) Folk, familj och fosterland – Nationalismens konsekvenser för jämställdhet [People, Family, and Fatherland – The consequences of nationalism for gender equality]. <https://expo.se/fakta/resurser/test-folk-familj-och-fosterland>.
- Expo (2016a) Moralisk kollaps [Moral collapse]. *Expo: Demokratisk tidskrift* 1.
- Expo (2016b) 0,01 procent [0,01 percent]. *Expo: Demokratisk tidskrift* 3.
- Farkas J, Schou J and Neumayer C (2018) Cloaked Facebook pages: Exploring fake Islamist propaganda in social media. *New Media & Society* 20(5): 1850–1867.

- Gillespie T (2017) Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information, Communication & Society* 20(1): 63–80.
- Gini C (1936) On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series* 208(1): 73–79.
- Golebiewski M and Boyd D (2019) Data voids. *Data & Society*. <https://datasociety.net/library/data-voids/>
- Gunn H K and Lynch M P (2018) Googling. In: Coady D and Chase J (eds) *The Routledge Handbook of Applied Epistemology*. London: Routledge, pp. 41–53.
- Haas A and Unkel J (2017) Ranking versus reputation: Perception and effects of search result credibility. *Behaviour & Information Technology* 36(12): 1285–1298.
- Haider J and Sundin O (2019) *Invisible search and online search engines: the ubiquity of search in everyday life*. Abingdon: Routledge.
- Hargittai E (2015) Is bigger always better? Potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science* 659: 63–76.
- Hargittai E (2021) Introduction. In: Hargittai E (ed) *Handbook of Digital Inequality*. Cheltenham: Elgar, pp. 1–7.
- Hellström A, Norocel O C and Jørgensen M B (2020) Nostalgia and hope: Narrative master frames across contemporary Europe. In: Norocel OC, Hellström A and Jørgensen MB (eds) *Nostalgia and Hope: Intersections Between Politics of Culture, Welfare, and Migration in Europe*. Cham: Springer, pp. 1–16.
- Hill Collins P (2019) *Intersectionality as Critical Social Theory*. Durham: Duke University Press.
- Iliadis A and Russo F (2016) Critical data studies: An introduction. *Big Data & Society* 3(2): –7.
- Kitchin R (2022) *The data revolution: A critical analysis of big data, open data & data infrastructures*. London: Sage.
- Klawier T, Prochazka F and Schweiger W (2022) Comparing frame repertoires of mainstream and right-wing alternative media. *Digital Journalism* 10(8): 1387–1408.
- Krzyżanowski M (2020) Normalization and the discursive construction of “new” norms and “new” normality: Discourse in the paradoxes of populism and neoliberalism. *Social Semiotics* 30(4): 431–448.
- Krzyżanowski M, Ekman M, Nilsson P, et al. (2021) Uncivility, racism, and populism: Discourses and interactive practices in anti- & post-democratic communication. *Nordicom Review* 42(S1): 3–15.
- Laaksonen SM, Haapoja J, Kinnunen T, et al. (2020) The datafication of hate: Expectations and challenges in automated hate speech monitoring. *Frontiers in Big Data* 3: 3.
- Lewandowski D (2012) Credibility in web search engines. In: Folk M and Apostel S (eds) *Online Credibility and Digital Ethos*. Hampshire: IGI Global, pp. 131–146. <https://doi.org/10.4018/978-1-4666-2663-8.ch008>
- Lewandowski D, Drechsler J and Mach S (2012) Deriving query intents from web search engine queries. *Journal of the American Society for Information Science and Technology* 63(9): 1773–1788.
- Lewandowski D and Schultheiß S (2022) Public awareness and attitudes towards search engine optimization. *Behaviour & Information Technology*. <https://doi.org/10.1080/0144929X.2022.2056507>.
- Lewandowski D and Spree U (2011) Ranking of Wikipedia articles in search engines revisited: Fair ranking for reasonable quality? *Journal of the American Society for Information Science and Technology* 62(1): 117–132.
- Lewandowski D and Sünkler S (2019) What does google recommend when you want to compare insurance offerings? *Aslib Journal of Information Management* 71(3): 310–324.
- Mager A (2014) Defining algorithmic ideology: Using ideology critique to scrutinize corporate search engines. *Triple C: Communication, Capitalism & Critique* 12(1): 28–39.
- Mahoney J, Le Louvier K, Lawson S, et al. (2022) Ethical considerations in social media analytics in the context of migration: Lessons learned from a horizon 2020 project. *Research Ethics* 18(3): 226–240.
- Metcalf J and Crawford K (2016) Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society* 3(1): 1–14.
- Monnier A, Seoane A, Hubé N, et al. (2021) Discours de haine dans les réseaux sociaux numériques [hate speech on social media]. *Mots Les Langages du Politique* 125: 9–14.
- Narayanan D and De Cremer D (2022) “Google told me so!” on the bent testimony of search engine algorithms. *Philosophy & Technology* 35(2): 22.
- Nikunen K (2021) Ghosts of white methods? The challenges of big data research in exploring racism in digital context. *Big Data & Society* 8(2): 2053951721110489.
- Noble SU (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.
- Norocel OC (2017) Åkesson at Almedalen: Intersectional tensions and normalization of populist radical right discourse in Sweden. *NORA: Nordic Journal of Feminist and Gender Research* 25(2): 91–106.
- Norocel OC (2022) Gendering Web2.0 sociotechnical affordances of far right metapolitics. *Social Media + Society* 8(3): 205630512211080.
- Oleinik A (2022) Relevance in web search: Between content, authority and popularity. *Quality & Quantity* 56: 173–194.
- O’Neil C (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishers.
- Pan B, Hembrooke H, Joachims T, et al. (2007) In Google we trust: Users’ decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication* 12(3): 801–823.
- Pöyhkäri R, Nikunen K, Nelimarkka M, et al. (2021) Refugee debate and networked framing in the hybrid media environment. *International Communication Gazette* 83(1): 81–102.
- Purcell K, Brenner J and Raine L (2012) Search Engine Use 2012. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/media/Files/Reports/2012/PIP_Search_Engine_Use_2012.pdf.
- Rogers R (2018) Aestheticizing Google critique: A 20-year retrospective. *Big Data & Society* 5(1): 205395171876862.
- Ruppert E, Isin E and Bigo D (2017) Data politics. *Big Data & Society* 4(2): 205395171771774.
- Schultheiß S and Lewandowski D (2021) Misplaced trust? The relationship between trust, ability to identify commercially influenced results and search engine preference. *Journal of Information Science*. <https://doi.org/10.1177/01655515211014157>.
- Schultheiß S, Sünkler S and Lewandowski D (2018) We still trust in Google, but less than 10 years ago: An eye-tracking study.

- Information Research* 23(3). <http://www.informationr.net/ir/23-3/paper799.html>
- Siapera E, Boudourides M, Lenis S, et al. (2018) Refugees and network publics on Twitter: Networked framing, affect, and capture. *Social Media + Society* 4(1): 1–21.
- SimilarWeb (2022) Google.com Traffic Statistics. Retrieved 7 April 2022 <https://www.similarweb.com/website/google.com/#ranking>.
- StatCounter (2022) Search Engine Market Share. <https://gs.statcounter.com/search-engine-market-share/>.
- Sundin O, Haider J, Andersson C, et al. (2017) The searchification of everyday life and the mundane-ification of search. *Journal of Documentation* 73(2): 224–243.
- Sundin O, Lewandowski D and Haider J (2022) Whose relevance? Web search engines as multisided relevance machines. *Journal of the Association for Information Science and Technology* 73(5): 637–642.
- Torres G and Rogers R (2020) Political news in search engines: Exploring google’s susceptibility to hyperpartisan sources during the Dutch elections. In: Rogers R and Niederer S (eds) *The Politics of Social Media Manipulation*. Amsterdam: Amsterdam University Press, pp. 97–122.
- Unkel J and Haas A (2017) The effects of credibility cues on the selection of search engine results. *Journal of the Association for Information Science and Technology* 68(8): 1850–1862.
- Vaidhyathan S (2011) *The Googlization of Everything (And Why We Should Worry)*. Oakland: University of California Press.
- Wodak R (2015) *The politics of fear: What right-wing populist discourses mean*. London: Sage.