Market or Markets? Investigating Google Search's Market Shares Under Vertical Segmentation

Desheng Hu,[*1] **Jeffrey Gleason,**[*2] **Muhammad Abu Bakar Aziz,**[*2] **Alice Koeninger,**[2]
Nikolas Guggenberger,[3] **Ronald E. Robertson,**[2, 4] **Christo Wilson**[2]

[1]University of Zurich, Zurich, Switzerland
[2]Northeastern University, Boston MA, USA
[3]University of Houston, Houston TX, USA
[4]Stanford University, Stanford CA, USA
desheng@ifi.uzh.ch, gleason.je@northeastern.edu, aziz.muh@northeastern.edu,
nguggenb@central.uh,edu, rer@acm.org, cbw@ccs.neu.edu

Abstract

Is Google Search a monopoly with gatekeeping power? Regulators from the US, UK, and Europe have argued that it is based on the assumption that Google Search dominates the market for horizontal (a.k.a. "general") web search. Google disputes this, claiming that competition extends to all vertical (a.k.a. "specialized") search engines, and that under this market definition it does not have monopoly power.

In this study we present the first analysis of Google Search's market share under vertical segmentation of online search. We leverage observational trace data collected from a panel of US residents that includes their web browsing history and copies of the Google Search Engine Result Pages they were shown. We observe that participants' search sessions begin at Google greater than 50% of the time in 24 out of 30 vertical market segments (which comprise almost all of our participants' searches). Our results inform the consequential and ongoing debates about the market power of Google Search and the conceptualization of online markets in general.

1 Introduction

Google is one of largest corporations in the world. In 2022, it reported $282.8B in revenue and a 26% profit margin (Alphabet Inc. 2022). Its products are ubiquitous—for example, it owns the world's most popular video streaming service (S. Dixon), web browser (StatCounter a), online display advertising platform (Haggin and Dapena 2019; U.S. District Court Southern District of New York 2020; Srinivasan 2019), navigation and mapping application (L. Ceci), and smartphone operating system (StatCounter c). Google operates in numerous additional markets, including email and cloud computing (Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary 2020).

Among Google's many products, its original product—Google Search—continues to be its most lucrative. In 2022, 57.4% of Google's revenue ($162 billion) came from ad sales in Google Search (Alphabet Inc. 2022). Google Search's market share is estimated to be 61–80% of desktop web searches in the US, and has been stable for over 14 years (Joseph Johnson; StatCounter b). The durability of Google Search's position in the market is due, at least in part, to its position as the default web search engine on the vast majority of smartphones, tablets (e.g., it pays Apple $12B per year to be the default on iOS devices (Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary 2020)), and web browsers (e.g., it pays Mozilla an estimated $450M per year to be the default in Firefox (Lyons 2020)). Indeed, the word "google" is synonymous with the act of searching the web (Merriam-Webster).

Because of its conduct towards competitors, Google Search has been the focus of numerous antitrust inquiries and litigation over the last decade. Currently, the US Department of Justice (DoJ) and 46 US states are suing Google for unlawfully maintaining a monopoly over the *horizontal search* market (U.S. District Court for the District of Columbia 2020c,a)—defined as the market for "general search services" that index the public web and return results for any query. The DoJ and the attorneys general allege that Google uses "exclusionary default agreements" with third parties like browser vendors, mobile device manufacturers, and cellular service providers to help maintain its monopoly. Other regulators and legislators have made the same allegations in the past (Federal Trade Commission 2012; Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary 2020; Competition & Markets Authority 2020b), and collectively they argue that Google's conduct harms consumers—especially when Google preferences its own products in search results. Indeed, in 2017, the European Commission (EC) found that Google abused its dominance in horizontal search by favoring its own comparison shopping service over those of competitors (European Commission a).

A critical facet of antitrust jurisprudence is the definition of the market for a good or service (Federal Trade Commission). Regulators and legislators have argued that Google Search dominates the market for horizontal search, within which, they claim, it competes with products like Microsoft

[*]These authors contributed equally.

Bing and DuckDuckGo (U.S. District Court for the District of Columbia 2020c,a; Competition & Markets Authority 2020b; Federal Trade Commission 2012; European Commission a; Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary 2020). Google, however, disputes this market definition, claiming that competition extends to all horizontal and *vertical search engines*—defined as search engines that specialize in one particular category of information or data from one particular service (Federal Trade Commission 2012; Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary 2020). Google has argued that people search for "news on Twitter, flights on Kayak and Expedia, restaurants on OpenTable, recommendations on Instagram and Pinterest", and products on Amazon (Walker 2020a). Similarly, in documents produced for the US House Subcommittee investigation, Google argued that estimates of its share of online search "do not capture the full extent of Google's competition in search" (Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary 2020). To date, however, Google has provided no convincing evidence to back up its claim that it faces significant competition from vertical search engines.

In this study we present the first analysis of Google Search's market share under vertical segmentation of online search. We leverage ecologically-valid, observational trace data collected from a panel of US residents over a five month period in 2020 that includes their web browsing history and complete copies of the Google Search Engine Result Pages (SERPs) they were shown. To quantify vertical market share, we identify searches carried out by participants on *all* websites within our corpus, group all searches (on Google, on Bing, and on all other vertical search engines) into 90 vertical segments (e.g., Shopping, Health and Wellness, News and Media), and compare participants' search behavior on and off Google products[1] within each vertical segment. We also examine participants' propensity to switch between competing Google products, Microsoft products, and independent vertical competitors.

We find that Google's products receive over 50% of participants' searches across 21 of the top 30 market segments (which account for 94.1% of all searches performed by our participants). We also find that Google holds significant power as a *gatekeeper* to independent vertical search engines (Competition & Markets Authority 2020b). In 24 of the top 30 market segments, participants began their search activity on a Google product more than 50% of the time, sometimes followed by additional searches on an independent vertical search engine. Contrary to Google's assertions, our data suggests that participants do not treat Google Search and independent vertical search engines as substitutable.[2] Further, our results highlight Google's power to steer users towards their own vertical search engines (Jeffries and Yin 2020; Gleason et al. 2023).

[1] "Google products" includes participants' searches on Google's vertical search engines, e.g., GMail and YouTube.

[2] "Substitutable" products are equivalent to consumers. A Toyota sedan is substitutable for a Honda sedan, but a truck is not.

In summary, our work presents novel methods and analyses that inform consequential, ongoing debates about the market power of Google Search in particular, and the conceptualization of online markets in general. Our results speak to the prospects of ongoing antitrust litigation and the need for regulators to consider structural remedies—e.g., separating Google Search from Google's vertical search engines, Android, and Chrome—and behavioral remedies—e.g., prohibiting Google from signing exclusionary contracts with third-parties—to curtail the power of online intermediaries (Khan 2019; Heidhues et al. 2021).

2 Definitions

In its lawsuit against Google, the DoJ et al. define horizontal search engines as "'one-stop shops' consumers can use to search the internet for answers to a wide range of queries" (U.S. District Court for the District of Columbia 2020c). Others have used similar language to define horizontal search engines (Federal Trade Commission 2012; European Commission a; Competition & Markets Authority 2020b; Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary 2020).

In contrast, the DoJ et al. explain in their lawsuit that vertical search engines "are not 'one-stop shops' and cannot respond to all types of consumer queries, particularly navigational queries" (U.S. District Court for the District of Columbia 2020c). The staff at the US Federal Trade Commission (FTC) defined vertical search engines similarly, in 2012, as "search engines focus[ed] on more narrowly-defined categories of content, such as product words" (Federal Trade Commission 2012). The US House Subcommittee concurred with these definition (Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary 2020).

We adopt this broad definition of vertical search and consider any website that supports search functionality, but is not a horizontal search engine, to be a vertical search engine. We chose this inclusive conceptualization of vertical search because it comports with Google's own assertions that it competes with all manners of websites (e.g., social media, retailers, travel agencies, etc.) that support search functionality (Walker 2020a; Competition & Markets Authority 2020a; U.S. District Court for the District of Columbia 2020b). We describe how we identified vertical search engines and quantified their usage in § 4.2.

In this study we examine Google and Bing's market share under vertical segmentation of the market for search, as well as their ability to function as gatekeepers to independent vertical search engines. The UK's Competition & Markets Authority (CMA) defined "gatekeepers" as online platforms that "mediat[e] relationships between consumers and businesses in a wide variety of markets" (Competition & Markets Authority 2020b). The European Union's Digital Markets Act designates an online platform as a gatekeeper if "(a) it has a significant impact on the internal market; (b) it provides a core platform service which is an important gateway for business users to reach end users; and (c) it enjoys an entrenched and durable position" (Digital Markets Act 2022). Bipartisan antitrust legislation that includes a similar

definition of "covered platforms" has been proposed in the US (American Innovation and Choice Online Act).

In their lawsuit, the DoJ argues that Google Search is a gatekeeper to third-party websites in general, and vertical search engines in particular, due to its market share and its ability to answer navigational queries (U.S. District Court for the District of Columbia 2020c). Navigational queries describe searches for a specific website, often by name, followed by a click on the result link that points to this website (Broder 2002; Jansen, Booth, and Spink 2008). Vertical search engines cannot answer navigational queries because they do not index the entirety of the web. We describe our approach for measuring gatekeeping power in § 4.3.

3 Background

We now introduce related work that has studied online search engines and how our study draws from this literature.

3.1 Quantifying Search Behavior

There is a long history of scholars studying peoples' behavior on search engines. Early studies made foundational contributions to our understanding of how people (re)formulate queries and interact with Search Engine Result Pages (SERPs) using query logs from engines like AltaVista, Yahoo, and AOL (Silverstein et al. 1999; Huang and Efthimiadis 2009; Teevan et al. 2007). Unfortunately, these studies were limited to studying behavior on single search engines in isolation. Further, studies of query logs are rare today because search engines stopped sharing them in the wake of the AOL query log deanonymization debacle (Barbaro and Jr. 2006). That said, eye-tracking approaches continue to refine our understanding of how people interact with search engines (Papoutsaki, Laskey, and Huang 2017).

Navigational Search One influential study that emerged from the early search engine literature was a taxonomy of web search that included navigational, informational, and transactional queries (Broder 2002). Multiple approaches have been proposed to identify navigational queries. Jansen et al. identify navigational searches using a rules-based approach that checks whether a query contains (1) a domain suffix or (2) a company/organization name (Jansen, Booth, and Spink 2008). In contrast, Teevan et al. identify navigational searches using a query's click entropy (Teevan, Liebling, and Ravichandran Geetha 2011). Previous studies have labeled 10–21% of queries as navigational (Jansen, Booth, and Spink 2008; Teevan, Liebling, and Ravichandran Geetha 2011).

Navigational queries are a key facet of regulators' concerns about horizontal search engines. We adopt the Jansen et al. approach to identify navigational queries in this study (Jansen, Booth, and Spink 2008) (see § 4.2).

Search Sessions Jansen et al. define a search session as a "series of interactions by the user toward addressing a single information need" (Jansen et al. 2007). The authors propose two approaches for identifying sessions: (1) 30 minutes without a search, and (2) query reformulation patterns. The temporal approach produces a smaller number of sessions with a longer average length than the query reformulation approach. Many subsequent studies that model and analyze search behavior have used a 30 minute temporal cutoff to define a session boundary (Downey, Dumais, and Horvitz 2007; Downey et al. 2008; White and Dumais 2009; Hassan et al. 2014). One relevant finding from these studies is that only 4% of sessions involve switching between horizontal search engines (White and Dumais 2009). We adopt these methods to analyze our participants' search sessions.

3.2 Competition in Search

Several studies have focused on competition issues in the design of Google SERPs. Edelman and Lai exploited a natural experiment in which idiosyncratic differences in user queries determined whether Google displayed its Flight service on the SERP (Edelman and Lai 2016). They found that Google Flights increased paid click volume to travel agencies (e.g., Expedia) by 65% and decreased organic click volume by 55%. Kim and Luca designed a controlled experiment to evaluate Google's decision to only include reviews from its own platform in the Local "Onebox" on the SERP (Kim and Luca 2019). They found that users preferred a Onebox that included reviews from competitors (e.g., Yelp). Gleason et al. leveraged an observational dataset to find similar pairs of queries that triggered different SERP components (Gleason et al. 2023). They found that Google's local, shopping, and image components decreased organic click-through rate (CTR) to third-party websites and that local and image components increased organic CTR to Google's own services.

Our study relies on web browser extension-based data collection techniques that have been successfully used in many prior studies of horizontal search engines (Robertson, Lazer, and Wilson 2018; Robertson et al. 2018, 2023).

3.3 Domain Classification

Automatically classifying websites and domains into topics or categories is a long-standing challenge. Numerous studies have proposed algorithms (Zhang and Lee 2004; Kwon and Lee 2003; Sun et al. 2014; Buber and Diri 2019; López-Sánchez, Corchado, and Arrieta 2017) and features (Mladenic 1998; Qi and Davison 2009; Golub and Ardö 2005; Shih and Karger 2004; Utard and Fürnkranz 2006; Camastra et al. 2015; López-Sánchez, Arrieta, and Corchado 2019) for this task. Given the large number of approaches and datasets that are available for this task, recent studies have focused on comparing the relative accuracy of different classification approaches (Bruni and Bianchi 2020; Hodžić, Kevrić, and Karadag 2016; Do et al. 2021). In this work, we adopt Fortiguard's domain to category mapping, based on the comprehensive evaluation in Vallina et al. (2020) (see § 4.2).

4 Data and Methods

In this section we present the datasets and methods that we used in our study. First, in § 4.1, we introduce the participant data we use throughout this study. Next, in § 4.2, we present

our methodologies for identifying search queries on independent vertical search engines and grouping search queries into vertical market segments. Finally, in § 4.3, we discuss our approach for clustering individual search queries into search sessions.

4.1 Participant Data

Beginning in August 2020, we engaged the survey company YouGov to recruit a panel of US residents to take a survey and optionally install a browser extension we developed for Chrome and Firefox.[3] YouGov reached out to a nationally-representative sample of 2,000 people, of which $N = 926$ completed the survey and installed the browser extension. We collected data from these participants' web browsers from August through December 2020. We adjusted all data collected from participants to be representative of the US adult population based on weights provided by YouGov.[4] Specifically, we multiplied counts of participants' online activities by their assigned weight.[5] Additionally, based on self-reports, participants who installed the extension were slightly more likely to have high trust in Google Search and use it daily; we revisit this discrepancy in § 6.3.

To measure participants' web search behaviors on and off Google Search, our browser extension collected two types of passive, observational data from their web browsers: browsing history and *snapshots* of Google SERPs.[6] The browsing history data contains a record of every URL that participants loaded in their browser during our observation window and the timestamp at which each page load occurred. On average our participants loaded 296.6 URLs per day per participant (SD = 49.2). The snapshot data contains the complete HTML of the SERPs that Google Search presented to participants in response to their queries. We collected and parsed 271,062 SERPs in total. On average our participants made 11.6 Google searches per day per participant (SD = 1.7).

We observe that 97% of the searches conducted by our participants on horizontal search engines occurred on Google and Bing (Yahoo and DuckDuckGo were the next two most frequently used). Thus, in the remainder of this study, we exclude activity that occurred on non-Google and non-Bing horizontal search engines.

Parsing SERPs We examine Google SERPs broken down into vertical segments. To facilitate this segmentation (discussed below), we made use of the links that appeared in SERPs and participants' clicks on those links. We used the open source `WebSearcher` package to extract links from SERPs (Robertson and Wilson 2020; Robertson 2023). On average we parsed 16.5 URLs per Google SERP (SD = 9.9), which agrees with prior studies (Robertson, Lazer, and Wilson 2018; Robertson et al. 2018).

[3]This study was IRB approved, see § 7 for details.

[4]https://yougovplatform.zendesk.com/hc/en-gb/articles/360002975617-How-is-the-data-weighted

[5]Underrepresented and overrepresented participants are assigned higher and lower weights, respectively.

[6]Note that the browser extension also collected other data that we exclude from this study.

Click Measurement To identify which, if any, of the links in SERPs were clicked by participants, we examined the URLs that participants loaded immediately after performing a Google Search. Similar to prior work, our high-level approach to click measurement is to compare the exact URLs in a participant's browsing history within (1) thirty seconds and (2) three sequential URLs after performing a search to the exact set of URLs extracted from the SERP (Flaxman, Goel, and Rao 2016; Allen et al. 2020; Guess, Nyhan, and Reifler 2020; Guess et al. 2020).

However, this exact comparison misses clicks on ads and URLs that redirect to a different URL. To address this, we identify ad clicks using the `gclid` URL parameter, which Google uses for conversion tracking and attribution.[7] Further, we identify clicks on redirected URLs by comparing the domains (i.e., not the full URL) in a participant's browsing history within (1) thirty seconds and (2) three sequential URLs to the set of domains extracted from the SERP. To reduce false positives from this approach, we ignore any matches where the domain was included in the three URL visits prior to the search. This exclusion captures instances where a participant was browsing a website immediately before and after a Google search.

Using this approach we identified 103,599 clicks on SERPs and a per-SERP CTR of 38.2% (similar to an estimate of 35% from a recent industry report (Gandhi 2021)). Our approach only detects the first click that participants made on SERPs—a limitation that we discuss in § 6.3.

Filtering Bing Activity Microsoft has a rewards program that offers people monetary incentives to use services like Bing (Microsoft Rewards). One way that people can earn Microsoft rewards is to take quizzes on Bing. Answering a quiz question automatically submits a new query to Bing and appends a query parameter to the URL. This activity accounts for 28.9% of Bing queries and we assume that our sample over-represents people in the rewards program (43.8% of participants who used Bing visited the Microsoft Rewards website a least once). Thus we filter out Bing queries that contain `Rewards`, `Quiz`, or `Gamification_DailySet` in the query segment of the URL. This filter impacted 10.1% of participants.

Additionally, we excluded one Bing user who made over 2000 Bing searches on a single day because we suspect that this participant was using automation to make searches. We did not exclude any Google users.

4.2 Vertical Search

To analyze participants' search behavior on vertical search engines and compare it to their behavior on Google and Bing, we undertook the following steps:

1. identify a mapping of websites to vertical segments (e.g., Shopping, Travel, etc.),
2. identify all searches that participants conducted on websites within the vertical segments, based on their browsing history, and

[7]https://support.google.com/google-ads/answer/9744275

3. divide participants' Google and Bing queries into the same vertical segments.

This process enabled us to examine Google and Bing's market shares within each vertical segment relative to all other independent vertical search engines. We now describe each of these steps.

Mapping Websites to Vertical Segments For this study we use the mapping of websites to vertical segments maintained by FortiGuard. FortiGuard is a vendor of cybersecurity software and their mapping is meant to help companies filter Internet traffic (e.g., to block social media). Vallina et al. found that FortiGuard's mapping had the highest coverage of websites and the most accurate vertical segment labels compared to other vendors' mappings (Vallina et al. 2020). The FortiGuard mapping contains 90 vertical segments, which covered 157,792 (99.7%) of the unique domains loaded by our participants.

Identifying Searches on Websites To identify vertical search engines and participants' queries (if any) on these websites, we used a combination of manual and automated methods. First, we manually examined over 400 websites—a mix of the most popular websites overall and in specific vertical segments, sorted by participants' browsing history—to identify vertical search engines and their respective *search schemas*, e.g., `amazon.com/s?k=QUERY`. This included examining 60 websites that we suspected might have non-keyword-based search functionality (e.g., travel and restaurant reservation products) and 52 websites that required account registration (e.g., social media). In total, these manually checked websites cover 71.1% of all page loads in participants' browsing history.

Second, we built a web crawler that attempted to identify websites that supported search and their associated search schema. We instrumented the Chrome web browser to visit each website in our participants' browsing history and then applied the following two heuristics:

1. The crawler tried to detect support for OpenSearch,[8] which is a web standard that allows websites to programmatically expose their search functionality to web browsers. We used the search URL schema specified in the OpenSearch XML description to validate the effectiveness of our crawler.
2. The crawler tried to locate an HTML `<input>` element where the keyword "search" appeared in (1) the `role` attribute of the `<form>` or (2) the `id`, `name`, `title`, `type`, or `class` properties of the `<input>` tag. If the crawler identified an input element matching these criteria, then it injected a unique query into the detected form, submitted the form, and then attempted to identify the query in the resulting URL. If the crawler found the query in the URL path (e.g., `amazon.com/search/QUERY`) or in the URL parameters (e.g., `amazon.com/s?k=QUERY`) we used this as the search schema for the website.

[8]https://developer.mozilla.org/en-US/docs/Web/OpenSearch

These are the same heuristics used by prior work to investigate search queries (Kats, Silva, and Roturier 2022). Overall, our crawler successfully visited 89.5% of websites that appeared in participants' browsing history and parsed 96.7% of the crawled websites' HTML. The crawler detected that 39.5% of the parsed websites supported search functionality.

We validated the effectiveness of our crawler with both manual and automated checks. First, three authors manually reviewed the top 200 websites that the crawler identified as having search functionality. The authors identified the same search URL parameter as the crawler on 94% of these sites. Second, on websites that had a syntactically-valid OpenSearch XML description, we compared the search URL parameter defined in the XML description file to that identified by our crawler. Our crawler agreed with the XML description file 95% of the time.

Using the search schemas that we isolated for each of these websites, we separated the search and non-search URLs in participants' browsing history. From the 158,272 unique websites that appeared in participants' browsing history, we identified 48,978 (31.0%) vertical search engines. Of the 7,848,032 page loads to these vertical search engines in our dataset, 293,401 (3.7%) corresponded to searches.

When we analyze participants' usage of vertical search engines, we include searches they performed on most Google and Microsoft products. For example, we include searches on GMail in the Web-based Email vertical segment and searches on YouTube in the Streaming Media segment.

Assigning Google Queries to Vertical Segments To divide participants' Google queries into vertical segments—e.g., Shopping queries, Travel queries, etc.—we performed a two stage classification process. First, we identified navigational queries and placed them in their own isolated vertical. Researchers have recognized that navigational queries are a distinct use case from *informational queries* (Broder 2002; Jansen, Booth, and Spink 2008), and regulators have noted that vertical search engines cannot answer navigational queries (Competition & Markets Authority 2020a).

We use Jansen et al.'s rules-based approach to identify navigational queries (Jansen, Booth, and Spink 2008) because it focuses on the content of the query, which allows us to apply it to both Google and Bing searches. We classify a search as navigational if the Jaro-Winkler similarity between a participant's query and the top-level domain of the next URL in their browsing history is ≥ 0.95 (Cohen et al. 2003). Overall, we identified 19,231 (7.1%) navigational queries, which is similar to the 10–21% navigational query rate identified by prior studies (Jansen, Booth, and Spink 2008; Teevan, Liebling, and Ravichandran Geetha 2011).

Second, we classified the remaining 251,831 (92.9%) Google queries into the vertical segments from FortiGuard. If a participant clicked a link on a SERP, then we classified the query into the same vertical segment as the website in the clicked link. The intuition behind this strategy is that a person's intent when searching is revealed through their choice of result, as exemplified by their click. If a participant did not click any links on a SERP, then we treat the query as a weighted distribution over vertical segments. The distribu-

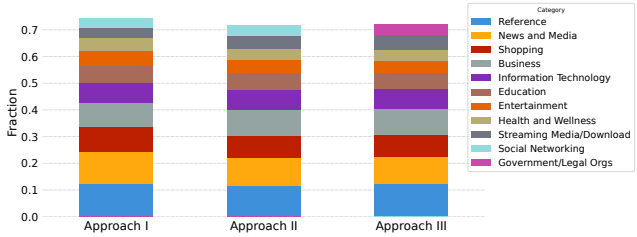

Figure 1: Shares of top ten vertical segments according to three different classification approaches. In each approach, SERPs with clicks are assigned the vertical segment of the clicked URL. SERPs without clicks are assigned a vertical segment based on the most frequently appearing segment (Approach I), segment distribution (II), and weighted segment distribution (III), respectively.

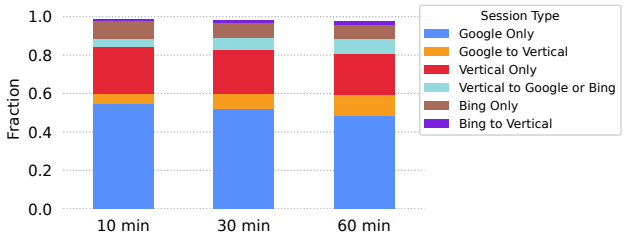

Figure 2: Share of vertical sessions types under three rolling window sizes.

tion for a given query corresponds to the FortiGuard segments of the websites that are linked in the SERP, adjusted by weights that decrease geometrically with rank to account for attenuation in attention (Papoutsaki, Laskey, and Huang 2017).

To assess the sensitivity of our SERP classification method, we evaluated two other SERP to vertical segment classification approaches: most frequent segment on the SERP and unweighted distribution over segments. Figure 1 shows that the top ten vertical segments are extremely similar no matter what classification method is used. Furthermore, the three methods produce overall segment distributions (shown in Figure 3) that are strongly correlated (Pearson $r \geq 0.914$, $p < 0.001$ in all cases), which strongly suggests that our findings are robust to the choice of SERP classification method.

Note that we could not map 24.1% of the Google SERPs in our corpus to vertical segments and we exclude them from all vertical analysis. This issue occurs because some Google SERPs contain results from Google's vertical search engines (e.g., Google Images and Google Videos). It is unclear how to determine the appropriate vertical segment assignment for these searches, as the destinations of the links are not strong indicators of search intent. Given this constraint, our analysis of Google's share of vertical segments should be interpreted as a lower bound.

Assigning Bing Queries to Vertical Segments We perform the same two-stage classification process to assign Bing queries to vertical segments, with one important caveat: our browser extension did not collect snapshots of Bing SERPs. Therefore, we re-crawled participants' Bing queries on January 9–11, 2023 from a Boston IP address and extracted links from the SERPs using the open source `SearchParser` package.[9] Although the links on individual Bing SERPs likely differ from the ones participants viewed in 2020 (which prevents us from measuring clicks on Bing SERPs), we verified that the aggregate distribution over vertical segments for a fixed sample of 1,000 Google queries was similar between late 2020 and early 2023.[10] Thus, we treat Bing SERPs as weighted distributions over vertical segments and rely on the assumption that this represents aggregate Bing search behavior from late 2020 with high fidelity.

As with Google, we exclude 15.6% of Bing SERPs in our corpus from our analysis of vertical segments because they contain results from Microsoft's vertical search engines that do not clearly map to a vertical segment (primarily Bing Images and Bing Videos).

4.3 Search Sessions

One goal of our study is to assess Google Search and Bing's gatekeeping power by examining where our participants begin and end information seeking tasks. If participants predominantly begin seeking information via Google Search, for example, this grants Google the power to steer participants to subsequent vertical search engines (owned by third-parties or Google itself) where they may refine their queries.

To investigate gatekeeping power we examine participants' propensity to switch between search engines during a single *search session*. Like prior work, we define a search session as searches that occur within a rolling 30 minute window of each other (Jansen et al. 2007; White and Dumais 2009; Hassan et al. 2014). 77% of Google searches and 88% of Bing searches in our dataset have an inter-arrival time under 30 minutes, which further motivates this threshold.

In this study, we examine *vertical search sessions*, which include searches made on Google Search, other Google products (e.g., YouTube, GMail, and Drive), Bing, other Microsoft products (e.g., Bing News, Shopping, and Travel), and/or independent vertical search engines. We represent vertical search sessions as a distribution over verticals in which each search receives equal weight. Specifically, the vertical distribution for a session is $\mathbf{s} = \frac{1}{n} * \sum_{j=1}^{n} \mathbf{c_j}$, where n is the number of searches in a session and $\mathbf{c_j}$ is the vertical distribution for search j in the session. Using this approach we constructed 131,802 vertical search sessions, 82.2% of which include only a single search engine.

To assess the sensitivity of our assignment of search sessions to vertical segments, we repeated the assignment procedure as we varied the rolling window size from 10 to 60 minutes in increments of 10 minutes. Figure 2 shows that the distributions of session types are very similar regardless

[9] https://github.com/jlgleason/SearchParser

[10] Specifically, the Jensen-Shannon distance (JS) between the aggregate 2020 and 2023 distributions was 0.1. As a reference, $JS([0.5, 0.3, 0.2], [0.55, 0.35, 0.1]) = 0.1$.

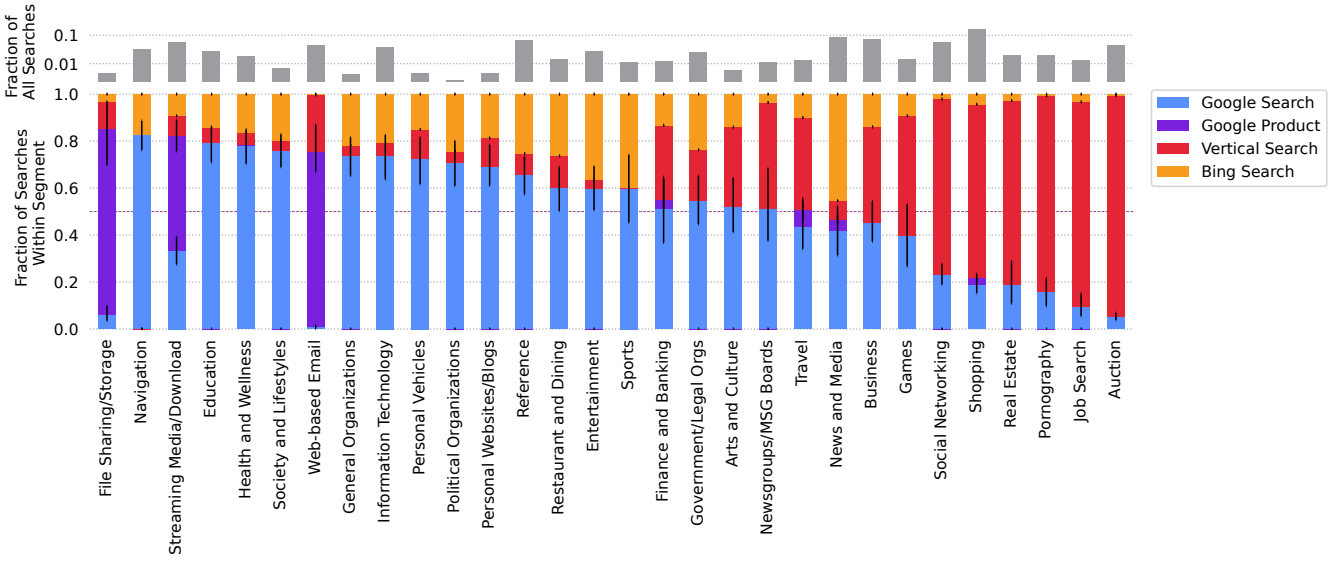

Figure 3: **Google receives $>$ 50% of searches in 21 out of 30 vertical segments.** We examine the fraction of searches in each of the top 30 vertical segments that occurred on Google Search, another Google product, Bing, or an independent vertical search engine, and perform 1000 bootstrap replications to compute 95% confidence intervals. Five of the verticals where Google Search receives $>$ 50% of searches are within the confidence interval. The top inset shows the overall distribution of searches.

of the window size. Specifically, we found that the vertical segment distributions were strongly correlated (Pearson $r \geq 0.966$, $p < 0.001$ in all cases), which validated that our assignment approach is robust.

5 Analysis

In this section we present the results of our analysis by examining market share and gatekeeping power.

5.1 Market Share

Figure 3 presents the fraction of participants' searches within thirty vertical segments stratified by where they occurred: on Google Search, on a non-Search Google product, on Bing, or on an independent vertical search engine. We compute 95% confidence intervals using the percentile bootstrap over participants with 1,000 replications. Figure 3 focuses on the thirty most popular verticals in our dataset, which collectively account for 94.1% of participants' vertical searches. The volume of searches in each vertical segment is shown in the upper portion of Figure 3. We sort the vertical segments along the x-axis based on Google's share of search volume, computed as the sum of Google Search and Google product search volume within each segment.

Figure 3 shows that Google products receive greater than 50% of search volume in 21 vertical segments, although six are within the confidence interval. Google products receive greater than 50% of search volume in many informational segments, such as Health and Wellness, Entertainment, and Reference, despite major websites in these segments having their own search functionality. Searches on YouTube, Google Drive, and GMail account for Google's share in the Streaming Media, File Sharing/Storage, and Web-based Email segments, respectively.[11] Google receives over 80% of participants' navigational queries and no vertical search engines appear in this segment.

In segments where Google products receive less than 50% of search traffic, eBay receives the most queries in the Auction segment, Zillow in Real Estate, Indeed in Job Search, Twitter and Facebook in Social Networking, and Amazon in Shopping. Out of the top 30 segments, Bing matches Google's share in only one segment: News and Media. We hypothesize that this may be driven by links to breaking news that Microsoft includes on MSN, the Bing homepage, and in Windows (Parmar 2021).

5.2 Gatekeeping Power

We investigated Google and Bing's role as gatekeeper by constructing and analyzing vertical search sessions. Our goal is to capture participants' proclivity for switching between Google and Microsoft-owned search engines and vertical search engines, as well as understand whether participants start their information seeking tasks on Google and Microsoft products or on vertical search engines.

Figure 4 presents the fraction of vertical search sessions in each vertical segment that included only searches on Google products, only searches on Bing, only searches on vertical search engines, or sessions that include searches on two of the three, broken down based on where the first search in each session was initiated. To make the figure legible, we omit sessions that included searches from all three (0.7% of

[11] Our results in the Web-based Email and File Sharing/Storage verticals should be interpreted with caution: because Hotmail, Outlook, and OneDrive do not encode queries in their URLs, we are unable to measure searches on these services.

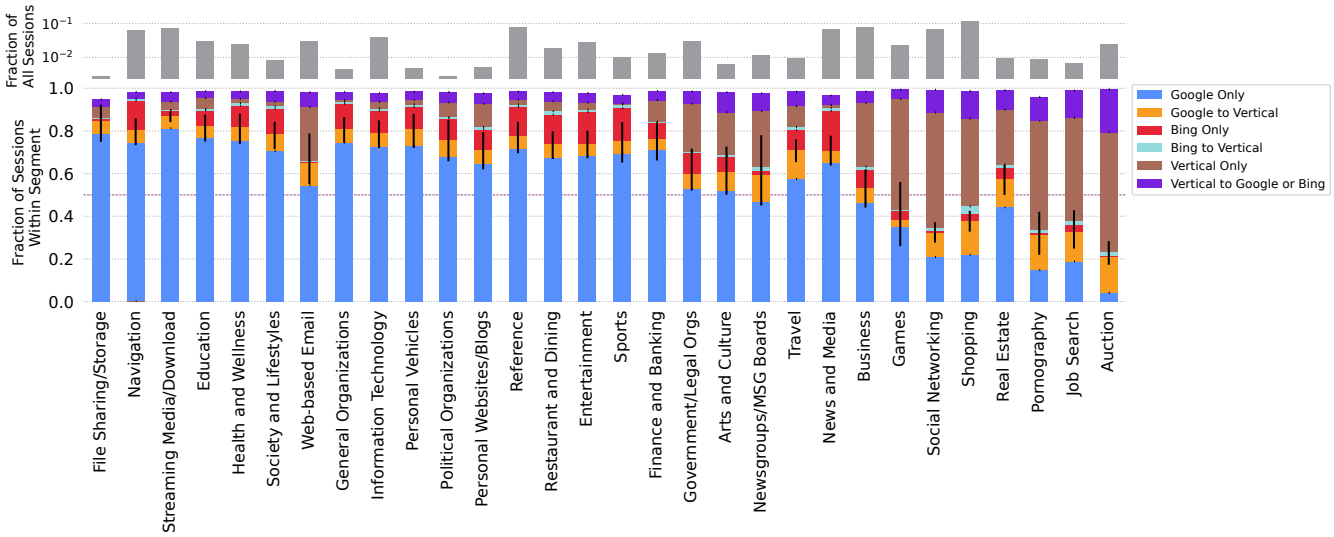

Figure 4: $>$ **50% of participants' search sessions begin on a Google product (or solely involve a Google product) in 24 out of 30 vertical segments.** We examine the fraction of search sessions in the top 30 vertical segments including: only searches on Google products, only on Bing, only on vertical search engines, or searches on two of the three. In the latter case, we divide the search sessions based on where the initial search in each session occurred. We compute 95% confidence intervals for the sum of 'Google Only' and 'Google to Vertical' categories using 1000 bootstrap replications. The top inset shows participants' session distribution.

total sessions) and sessions that only include Google products and Bing (1.0% of total sessions). We compute 95% confidence intervals using the percentile bootstrap over participants with 1,000 replications. Although Figure 4 contains six session types, we only computed a confidence interval for the most consequential: the sum of 'Google Only' and 'Google to Vertical' sessions. Figure 4 retains the same sort order of segments along the x-axis as Figure 3. Moreover, the thirty categories in Figure 4 collectively account for 92.4% of participants' vertical sessions.

Overall, 16.2% of vertical sessions in Figure 4 include switching, versus 1.0% of vertical sessions when we only consider vertical sessions with Google products and Bing. These observations make intuitive sense: as horizontal search engines, Google Search and Bing are obviously substitutable, so there is relatively little incentive for people to switch between them at short time scales. In contrast, vertical search engines are much less substitutable. For example, Google Search can help you find a specific item for sale or a particular retailer, but the best way to see a retailers' full inventory in a legible format is to search on their own website. Furthermore, some information is not public on the web and is thus inaccessible from Google Search, such as private posts on social media websites that can only be surfaced by using their native search interfaces.

The results in Figure 4 demonstrate that Google has significant gatekeeping power. In 24 of the top 30 segments, sessions begin on Google products (or only include searches on Google products) at least 50% of the time.

We observe that in many verticals, Google's proportion of the market grows substantially when we shift from the granularity of individual searches (Figure 3) to search sessions (Figure 4). For example, in the Real Estate vertical, Google Search receives roughly 20% of individual searches, but 57.3% of sessions begin on a Google product. We make similar observations for the Finance and Banking and Travel verticals. Bing's share of the News and Media vertical shrinks when we change the unit of analysis from individual searches to search sessions because long sequences of news queries get collapsed into single sessions.

These results highlight how our conception of market share, and thus market power, in online search may shift depending on how we account for user behavior—in this case, the tendency for people to conduct multiple searches in rapid succession within task-oriented sessions. Further, these findings demonstrate the extent to which vertical search engines are reliant on Google products for traffic.

6 Discussion

We conclude by discussing the implications of our findings and the limitations of our study.

6.1 Search Market Dynamics

Defining the boundaries of relevant search markets lies at the core of some of the most consequential antitrust litigation against Google, and has the potential to shape the future of the digital economy. Our study offers new methods and empirically-derived insights in the market(s) for online search, beyond the existing focus on aggregate usage measures and high-level arguments about corporate revenue (Harkrider 2020). We provide a potential basis for more

granular, vertical segmentation of markets for online search and a better understanding of Google's gatekeeping position.

Our analysis supports the concerns of market participants, regulators, legal scholars, and journalists who posit that Google may have the power to leverage its dominance in the horizontal search market to also dominate specific vertical segments by, for example, affording preferential placement to Google's vertical search products in SERPs (Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary 2020; European Commission a; Competition & Markets Authority 2020b; Khan 2019; Jeffries and Yin 2020; Gleason et al. 2023). Over 80% of our participants' navigational queries were to Google Search, which gives Google an immediate advantage over all vertical competitors since they cannot answer these queries. In 24 out of 30 vertical segments, participants' search sessions began with a Google product greater than 50% of the time (and only 13.9% of search sessions included transitions between a Google product and a non-Google vertical search engine across all verticals). This gatekeeper position grants Google enormous power to collect data about the preferences of Internet users, as well as steer participants towards vertical search engines of Google's choosing.

Among the nine vertical segments where Google commanded less than 50% of search volume, many are dominated by other major tech platforms (e.g., Amazon, ebay, Indeed, Facebook, Twitter, and Zillow). These are segments that Google has entered or attempted to enter in the past (e.g., through the introduction of Google+, Google Shopping, and Google for Jobs). The Shopping vertical has been the focus of previous regulatory actions against Google (European Commission a) and travel companies are advocating for similar regulatory intervention in Europe (Lomas 2020). Our analysis supports regulatory scrutiny of Google's actions in these verticals. For example, we observe that Google receives 51% ($+15\%$, -13%) of searches in the Travel vertical among our participants, but Google could potentially tip the scales unambiguously in their own favor by prioritizing Google Flights and Google Hotels in SERPs while demoting other travel companies.

One of the vertical segments that Google does not dominate—Pornography—may have been intentionally ceded by the company. Google Search filters pornography from search results by default unless users disable the Safe Search feature, and Google prohibits pornography on other platforms they own (e.g., YouTube and the Play Store).

Understanding why people exhibit entrenched online search behaviors requires further study. Google claims that people prefer their search engine because it offers the highest-quality results (Walker 2020b). It is also plausible, however, that bundling of search engines with hardware and software (European Commission b), as well as default effects (Lyons 2020), may ossify peoples' search behavior.

Although generative artificial intelligence (AI) models are rapidly altering the affordances of search engines, it is unclear whether they will have lasting impacts on search market shares. As of May 2023, Google Search and Bing both integrate chat-style AI based on large language models (Peters 2023). While Bing was first to adopt this technology, initial speculation that this would cause Bing to take market share from Google Search does not appear to be coming true (Dotan 2023). This could be because defaults (which favor Google Search) are sticky, users' initial excitement for these technologies has waned, or simply that overcoming entrenched human behaviors is hard for upstarts. More broadly, online search startups are failing (Pierce 2023) and business relationships between Google Search and dominant firms are not changing (Roth 2023).

6.2 Assessing Monopoly Power

In this study we focus on market definitions and shares, which are preliminary questions in establishing whether a company holds monopoly power. When assessing whether a company's market power amounts to monopoly power, courts conduct holistic assessments that may, for example, include considerations of market entry barriers. This holistic assessment remains beyond the scope of this study. Moreover, a finding of monopoly power only opens the door to antitrust enforcement; it is necessary, but not sufficient to support an antitrust claim. Standing antitrust doctrine requires specific anticompetitive behavior and harm in the form of consumer welfare losses to support a claim (Reiter v. Sonotone Corp.; Ohio v. American Express Co.; Areeda and Hovencamp 2020).

A more granular, vertical segmentation of markets for online search does not suggest that the effects of Google's monopoly power are restricted to the particular vertical in which it holds a significant share. Google may leverage its monopoly power in one vertical or its gatekeeping power over navigational queries to exert power in vertical segments in which it holds a comparatively small share. Additionally, our study only examines one side of a multi-sided platform, and does not address Google's potential to leverage its position in the online advertising market (U.S. District Court Southern District of New York 2020).

Scholars have criticized the reliance on market shares as an indicator of market power and, more generally, the definition of relevant markets, due to the inherent challenges and uncertainties associated with that practice (Hovenkamp 2022). Instead, they have suggested to rely on direct evidence of harm. Our study takes no position on this issue, but acknowledges that regulators, enforcers, and courts continue to consider market shares and require market definitions (Ohio v. American Express Co.).

6.3 Limitations

Our study has several limitations. Our data is constrained to a sample of online activity from US-based individuals on the desktop platform and our participants may not be perfectly representative. Participants who agreed to install our browser extension expressed slightly higher usage and trust in Google Search than participants who did not agree to install our extension. That said, many of our participants engaged with the Microsoft Rewards Program, which incentivizes users to search on Bing. Publicly available estimates indicate that Google Search has greater than 90% horizontal market share in mobile and non-US markets (e.g., the UK and Europe) (StatCounter d). This suggests that our results

should be treated as a lower-bound on Google's dominance across horizontal and vertical market segments.

To respect participants' privacy, our browser extension did not collect data from incognito browsing windows. It is unclear how often people used incognito mode during our data collection period, although we note that we did capture a significant amount of potentially sensitive browsing activity (e.g., searching for and viewing pornography). This suggests that at least some participants did not use—or did not consistently use—this functionality.

Our analysis also depends on specific data operationalization choices that could impact our conclusions. For example, we only consider the first click on a SERP when assigning it a category; taking additional clicks into account could alter the distribution of query volume across segments.

Finally, there may be false negatives in our detection of participants' searches on independent vertical search engines. However, we manually validated that we correctly identified the search schemas on websites that account for 71.1% of all page loads in participants' browsing history, so the impact of potential false negatives is constrained. A related issue is that Hotmail, Outlook, and OneDrive do not include queries in their URLs, which prevents us from tabulating searches on these services in the Web-based Email and File Sharing segments, respectively.

7 Broader Perspective

This study was approved by the Northeastern IRB under protocol #20-03-04. All participants consented to data collection (see § 8.2) and were compensated. The total amount we paid to YouGov to administer our survey and compensate participants was $78,000. Participants were free to leave our study at any time. Our browser extension used TLS to protect data in transit and uninstalled itself at the end of the study period. Participant data was stored on a siloed server that was only accessible to personnel approved by the IRB.

We do not foresee any negative societal impacts of this study or risks to study participants. The nature of the data we collected from participants precludes deidentification. Thus, in accordance with our protocol, we only present aggregated results in this manuscript and we will not be making identifiable data from this study publicly available.

Acknowledgements

The collection of data used in this study was funded in part by the Anti-Defamation League, the Russell Sage Foundation, and the Democracy Fund. This research was supported in part by NSF grant IIS-1910064. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

References

Allen, J.; Howland, B.; Mobius, M.; Rothschild, D.; and Watts, D. J. 2020. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14).

Alphabet Inc. 2022. 2022 Annual Report (Form 10-K).

American Innovation and Choice Online Act. 2021. H.R. 3816, 117th Congress.

Areeda, P. E.; and Hovencamp, H. 2020. *Antitrust law: An analysis of antitrust principles and their application*. Wolters Kluwer, fifth edition.

Barbaro, M.; and Jr., T. Z. 2006. A Face Is Exposed for AOL Searcher No. 4417749. The New York Times.

Broder, A. 2002. A taxonomy of web search. In *ACM Sigir forum*, volume 36, 3–10. ACM New York, NY, USA.

Bruni, R.; and Bianchi, G. 2020. Website categorization: A formal approach and robustness analysis in the case of e-commerce detection. *Expert Systems with Applications*, 142: 113001.

Buber, E.; and Diri, B. 2019. Web page classification using rnn. *Procedia Computer Science*, 154: 62–72.

Camastra, F.; Ciaramella, A.; Placitelli, A.; and Staiano, A. 2015. Machine learning-based web documents categorization by semantic graphs. *Advances in Neural Networks: Computational and Theoretical Issues*, 75–82.

Cohen, W. W.; Ravikumar, P.; Fienberg, S. E.; et al. 2003. A Comparison of String Distance Metrics for Name-Matching Tasks. In *IIWeb*, volume 3, 73–78.

Competition & Markets Authority. 2020a. Appendix P: specialized search.

Competition & Markets Authority. 2020b. Online Platforms and Digital Advertising.

Digital Markets Act 2022. 2022. Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act).

Do, N. Q.; Selamat, A.; Krejcar, O.; Yokoi, T.; and Fujita, H. 2021. Phishing webpage classification via deep learning-based algorithms: an empirical study. *Applied Sciences*, 11(19): 9210.

Dotan, T. 2023. Microsoft Struggles to Gain on Google Despite Its Head Start in AI Search. The Wall Street Journal.

Downey, D.; Dumais, S.; Liebling, D.; and Horvitz, E. 2008. Understanding the relationship between searchers' queries and information goals. In *Proceedings of the ACM conference on Information and knowledge management*, 449–458.

Downey, D.; Dumais, S. T.; and Horvitz, E. 2007. Models of Searching and Browsing: Languages, Studies, and Application. In *IJCAI*, volume 7, 2740–2747.

Edelman, B.; and Lai, Z. 2016. Design of search engine services: Channel interdependence in search engine results. *Journal of Marketing Research*, 53(6): 881–900.

European Commission. 2017a. CASE AT.39740 Google Search (Shopping).

European Commission. 2019b. Summary of Google Android Commission decision.

Federal Trade Commission. 2012. The FTC Report on Google's Business Practices.

Federal Trade Commission. 2023. Guide to Antitrust Laws: Mergers. Federal Trade Commission.

Flaxman, S.; Goel, S.; and Rao, J. M. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80(S1): 298–320.

Gandhi, M. 2021. 2021 CTR Research Study.

Gleason, J.; Hu, D.; Robertson, R. E.; and Wilson, C. 2023. Google the Gatekeeper: How Search Components Affect Clicks and Attention. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Golub, K.; and Ardö, A. 2005. Importance of HTML structural elements and metadata in automated subject classification. In *Research and Advanced Technology for Digital Libraries: 9th European Conference*, 368–378. Springer.

Guess, A. M.; Nyhan, B.; O'Keeffe, Z.; and Reifler, J. 2020. The sources and correlates of exposure to vaccine-related (mis) information online. *Vaccine*, 38(49): 7799–7805.

Guess, A. M.; Nyhan, B.; and Reifler, J. 2020. Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, 4(5): 472–480.

Haggin, P.; and Dapena, K. 2019. Google's Ad Dominance Explained in Three Charts. *Competition Policy International*.

Harkrider, J. D. 2020. Break Up Denny's. *Competition Policy International*.

Hassan, A.; White, R. W.; Dumais, S. T.; and Wang, Y.-M. 2014. Struggling or exploring? Disambiguating long search sessions. In *Proceedings of the 7th ACM international conference on Web search and data mining*, 53–62.

Heidhues, P.; Bonatti, A.; Celis, L. E.; Crawford, G. S.; Dinielli, D.; Luca, M.; Salz, T.; Schnitzer, M.; Scott Morton, F. M.; Seim, K.; Sinkinson, M.; and Zhou, J. 2021. More Competitive Search Through Regulation. Tobin Center for Economic Policy at Yale.

Hodžić, A.; Kevrić, J.; and Karadag, A. 2016. Comparison of machine learning techniques in phishing website classification. In *International Conference on Economic and Social Studies (ICESoS'16)*, 249–256.

Hovenkamp, H. 2022. Digital Cluster Markets. *Columbia Business Law Review*, 2022(1).

Huang, J.; and Efthimiadis, E. N. 2009. Analyzing and Evaluating Query Reformulation Strategies in Web Search Logs. In *Proceedings of the Conference on Information and Knowledge Management*.

Jansen, B. J.; Booth, D. L.; and Spink, A. 2008. Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management*, 44(3): 1251–1266.

Jansen, B. J.; Spink, A.; Blakely, C.; and Koshman, S. 2007. Defining a session on Web search engines. *Journal of the American Society for Information Science and Technology*, 58(6): 862–871.

Jeffries, A.; and Yin, L. 2020. Google's Top Search Result? Surprise! It's Google. The Markup.

Joseph Johnson. 2021. Share of search queries handled by leading search engine providers in the United States as of April 2021. Statista.

Kats, D.; Silva, D. L.; and Roturier, J. 2022. Who Knows I Like Jelly Beans? An Investigation Into Search Privacy. *Proceedings on Privacy Enhancing Technologies*, 2022(2).

Khan, L. M. 2019. The Separation Of Platforms And Commerce. *Columbia Law Review*, 119(4).

Kim, H.; and Luca, M. 2019. Product quality and entering through tying: Experimental evidence. *Management Science*, 65(2): 596–603.

Kwon, O.-W.; and Lee, J.-H. 2003. Text categorization based on k-nearest neighbor approach for web site classification. *Information Processing & Management*, 39(1): 25–44.

L. Ceci. 2021. Leading mapping apps in the United States in 2021, by downloads. Statista.

Lomas, N. 2020. Travel startups cry foul over what Google's doing with their data. TechCrunch.

López-Sánchez, D.; Arrieta, A. G.; and Corchado, J. M. 2019. Visual content-based web page categorization with deep transfer learning and metric learning. *Neurocomputing*, 338: 418–431.

López-Sánchez, D.; Corchado, J. M.; and Arrieta, A. G. 2017. A CBR system for image-based webpage classification: case representation with convolutional neural networks. In *The Thirtieth International Flairs Conference*.

Lyons, K. 2020. Mozilla and Google renew Firefox search agreement. The Verge.

Merriam-Webster. 2023. "google" verb. Merriam-Webster.

Microsoft Rewards. 2023. Get on board with Microsoft rewards.

Mladenic, D. 1998. Turning Yahoo into Automatic Web-Page Classifier. In *13th European Conference on Artificial Intelligence Young Researcher Paper, 1998*.

Ohio v. American Express Co. 2018. In *S. Ct.*, volume 138, 2274. Supreme Court.

Papoutsaki, A.; Laskey, J.; and Huang, J. 2017. Searchgazer: Webcam eye tracking for remote studies of web search. In *Proceedings of the Conference on Conference Human Information Interaction and Retrieval*.

Parmar, M. 2021. Windows 10 is now nagging users with Microsoft Bing alerts.

Peters, J. 2023. Google wants you to forget the 10 blue links. The Verge.

Pierce, D. 2023. Neeva, the would-be Google competitor, is shutting down its search engine. The Verge.

Qi, X.; and Davison, B. D. 2009. Web page classification: Features and algorithms. *ACM computing surveys (CSUR)*, 41(2): 1–31.

Reiter v. Sonotone Corp. 1979. In *S. Ct.*, volume 442, 330. Supreme Court.

Robertson, R.; Lazer, D.; and Wilson, C. 2018. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *Proceedings of the 27th International Web Conference*. Lyon, France.

Robertson, R. E. 2023. WebSearcher. Github.

Robertson, R. E.; Green, J.; Ruck, D. J.; Ognyanova, K.; Wilson, C.; and Lazer, D. 2023. Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature*.

Robertson, R. E.; Jiang, S.; Joseph, K.; Friedland, L.; Lazer, D.; and Wilson, C. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 148:1–148:22.

Robertson, R. E.; and Wilson, C. 2020. WebSearcher: Tools for Auditing Web Search. In *Proceedings of Computation + Journalism Symposium*.

Roth, E. 2023. Sorry Bing, Samsung's sticking with Google as its default mobile search engine. The Verge.

S. Dixon. 2021. Most popular online video properties in the United States in December 2021, by reach. Statista.

Shih, L. K.; and Karger, D. R. 2004. Using urls and table layout for web classification tasks. In *Proceedings of the 13th international conference on World Wide Web*, 193–202.

Silverstein, C.; Marais, H.; Henzinger, M.; and Moricz, M. 1999. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum*, 33(1): 6–12.

Srinivasan, D. 2019. Why Google Dominates Advertising Markets. *Stanford Technology Law Review*, 24(55).

StatCounter. 2022a. Browser Market Share Worldwide. StatCounter GlobalStats.

StatCounter. 2022b. Desktop Search Engine Market Share United States of America. StatCounter GlobalStats.

StatCounter. 2022c. Mobile Operating System Market Share Worldwide. StatCounter GlobalStats.

StatCounter. 2022d. Search Engine Market Share Worldwide. StatCounter GlobalStats.

Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary. 2020. Investigation of Competition in Digital Markets.

Sun, S.; Liu, F.; Liu, J.; Dou, Y.; and Yu, H. 2014. Web classification using deep belief networks. In *2014 IEEE 17th International Conference on Computational Science and Engineering*, 768–773. IEEE.

Teevan, J.; Adar, E.; Jones, R.; and Potts, M. A. S. 2007. Information Re-Retrieval: Repeat Queries in Yahoo's Logs. In *Proceedings of the SIGIR Conference*.

Teevan, J.; Liebling, D. J.; and Ravichandran Geetha, G. 2011. Understanding and predicting personal navigation. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 85–94.

U.S. District Court for the District of Columbia. 2020a. Colorado and Plaintiff States v. Google LLC.

U.S. District Court for the District of Columbia. 2020b. Defendant Google LLC's Answer and Affirmative Defense to Plaintiffs' Complaint.

U.S. District Court for the District of Columbia. 2020c. U.S. and Plaintiff States v. Google LLC.

U.S. District Court Southern District of New York. 2020. Texas and Plaintiff States v. Google LLC.

Utard, H.; and Fürnkranz, J. 2006. Link-local features for hypertext classification. In *Semantics, Web and Mining: Joint International Workshops, EWMF 2005 and KDO 2005*, 51–64. Springer.

Vallina, P.; Le Pochat, V.; Feal, Á.; Paraschiv, M.; Gamba, J.; Burke, T.; Hohlfeld, O.; Tapiador, J.; and Vallina-Rodriguez, N. 2020. Mis-shapes, mistakes, misfits: An analysis of domain classification services. In *Proceedings of the ACM Internet Measurement Conference*, 598–618.

Walker, K. 2020a. A deeply flawed lawsuit that would do nothing to help consumers.

Walker, K. 2020b. A deeply flawed lawsuit that would do nothing to help consumers. Google Corporate Blog.

White, R. W.; and Dumais, S. T. 2009. Characterizing and predicting search engine switching behavior. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 87–96.

Zhang, D.; and Lee, W. S. 2004. Web taxonomy integration using support vector machines. In *Proceedings of the 13th international conference on World Wide Web*, 472–481.

8 Appendix

8.1 Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see § 4.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see § 4.1.
   (e) Did you describe the limitations of your work? Yes, see § 6.3.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see § 7.
   (g) Did you discuss any potential misuse of your work? NA
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see § 7.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? NA

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see § 7.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see § 7.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, see § 7 and § 8.2.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, see § 7.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, see § 7. Note that YouGov handled individual participant compensation and we are not privy to per participant wages.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, see § 7.

8.2 Browser Extension Informed Consent

Welcome to the study!

This extension implements a user study being conducted by researchers at Northeastern University, Dartmouth, Princeton, and University of Exeter. If you choose to participate, this browser extension will confidentially collect four types of data from your browser.

1. Metadata for web browsing (e.g., URL visited with time of visit), exposure to embedded URLs on websites (e.g., YouTube videos), and interactions with websites (e.g., clicks and video viewing time). This data is collected until the study is completed.

2. Copies of the HTML seen on specific sites: Google Search, Google News, YouTube, Facebook Newsfeed, and Twitter Feed. We remove all identifying information before it leaves the browser. This confidential data is collected until the study is completed.

3. Browsing history, Google and YouTube account histories (e.g., searches, comments, clicks), and online advertising preferences (Google, Bluekai, Facebook). This data is initially collected for the year prior to the installation of our browser extension, and we then check these sources once every two weeks to collect updates until the study is completed.

4. Snapshots of selected URLs from your browser. For each URL, the extension saves a copy of the HTML that renders, effectively capturing what you would have seen had you visited that website yourself. Once per week we conduct searches on Google Search, Google News, YouTube, and Twitter, and collect the current frontpage of Google News, YouTube, and Twitter. These web page visits will occur in the background and will not affect the normal functioning of your browser. There is a theoretical risk of "profile pollution" – that this extension will impact your online profiles, i.e., "pollute" them with actions that you did not

take. To mitigate this risk, the extension will only visit content that is benign and will only execute searches for general terms. Our previous work has found that historical information of this kind has minimal impact on online services.

Additionally, if you choose to participate, you will be asked to take a survey in which we ask you several questions about your demographics, web usage, and media preferences. These data, as well as those mentioned above, will be used to analyze the correlations between your online behavior and your interest profiles.

After the study is complete on December 31, 2020, the extension will uninstall itself. All data collected will be kept strictly confidential and used for research purposes only. We will not share your responses with anyone who is not involved in this research.

You must be at least 18 years old to take part in this study. The decision to participate in this research project is voluntary. You do not have to participant and you can refuse to participate. Even if you begin our experiment, you can stop at any time. You may request that we delete all data collected from your web browser at any time.

We have minimized the risks. We are collecting basic demographic information, information about your internet habits, and copies of web pages that you visit. To the greatest extent possible, information that identifies you will be removed from all collected web data.

Your role in this study is confidential. However, because of the nature of electronic systems, it is possible, though unlikely, that respondents could be identified by some electronic record associated with the response. Neither the researchers nor anyone involved with this study will be collecting those data. Any reports or publications based on this research will use only aggregate data and will not identify you or any individual as being affiliated with this project.