Google the Gatekeeper: How Search Components Affect Clicks and Attention

Jeffrey Gleason[1], **Desheng Hu**[2, 1], **Ronald E. Robertson**[3, 1], **Christo Wilson**[1]

[1] Northeastern University
[2] University of Zurich
[3] Stanford University
gleason.je@northeastern.edu, desheng@ifi.uzh.ch, ronalder@stanford.edu, cbw@ccs.neu.edu

Abstract

The contemporary Google Search Engine Results Page (SERP) supplements classic blue hyperlinks with complex components. These components produce tensions between searchers, 3rd-party websites, and Google itself over clicks and attention. In this study, we examine 12 SERP components from two categories: (1) *extracted results* (e.g., `featured-snippets`) and (2) *Google Services* (e.g., `shopping-ads`) to determine their effect on peoples' behavior. We measure behavior with two variables: (1) click-through rate (CTR) to Google's own domains versus 3rd-party domains and (2) time spent on the SERP. We apply causal inference methods to an ecologically valid trace dataset comprising 477,485 SERPs from 1,756 participants. We find that multiple components substantially increase CTR to Google domains, while others decrease CTR and increase time on the SERP. These findings may inform efforts to regulate the design of powerful intermediary platforms like Google.

1 Introduction

The presentation of results on Google Search has evolved significantly from its humble beginnings as a list of ten blue hyperlinks. In 2007, Google introduced "Universal Search," blending results from its Images, Maps, News, and Video vertical search properties into a single *Search Engine Results Page (SERP)*.[1] Five years later, Google added the "Knowledge Graph," which powers the presentation of facts in the main results column and knowledge panels on the right-hand side of the SERP.[2] In 2014, Google added "Featured Snippets," which extract a prominent and readable page description from a website and place it above the website's listing.[3] Collectively, these additions to SERPs beyond simple, organic links are referred to as *components* (Robertson, Lazer, and Wilson 2018).

From a normative perspective, components matter for at least two reasons. First, previous work has shown that specific components can affect peoples' behaviors. For example, Chilton and Teevan (2011) found that direct answers, such as definitions and facts, can reduce traffic from the SERP to 3rd-party (i.e., non-Google) websites. This highlights a tension between stakeholders: users may benefit when search engines directly answer their questions (Diriye et al. 2012; Williams et al. 2016), leading to *good abandonment* (Li, Huffman, and Tokuda 2009) of search activity. Google may also benefit, by bolstering the primacy of their search engine as an authoritative source of information (McMahon, Johnson, and Hecht 2017). However, 3rd-party websites may receive less traffic, leading to a range of problematic outcomes such as loss of revenue and erosion of community. Several studies have highlighted this tension in the context of Wikipedia, which Google Search heavily relies on for answer and knowledge components, but which itself relies on a continuous stream of volunteers to sustain its peer-production community (McMahon, Johnson, and Hecht 2017; Vincent et al. 2019).

The second reason that components matter is concern over self-preferencing, in which a powerful intermediary platform prioritizes its own products and services over comparable offerings from 3rd-parties (Subcommittee on Antitrust, Commercial and Administrative Law 2022; Competition and Markets Authority 2020). In 2020, thirty US state attorneys general filed a lawsuit that cites self-preferencing in SERPs as one form of allegedly anticompetitive conduct in which Google has engaged—"Google throttles consumers from bypassing its general search engine and going directly to their chosen destination."[4] In response to these lawsuits, Google argues that people prefer a SERP full of components, citing as evidence other search engines that emulate Google Search's design.[5]

Given these value tensions—between searchers, Google, and 3rd-party websites—it is critical that we understand the effects of design decisions in Google SERPs on peoples' behaviors. For example, regulators around the world are proposing novel rules on the design of powerful online intermediary platforms like Google (European Commission

[1] https://googleblog.blogspot.com/2007/05/universal-search-best-answer-is-still.html

[2] https://blog.google/products/search/introducing-knowledge-graph-things-not

[3] https://blog.google/products/search/reintroduction-googles-featured-snippets

[4] https://coag.gov/app/uploads/2020/12/Colorado-et-al.-v.-Google-PUBLIC-REDACTED-Complaint.pdf

[5] https://blog.google/outreach-initiatives/public-policy/redesigning-search-would-harm-consumers-and-american-businesses

2022). However, it is difficult to craft these rules thoughtfully or assess their impact without first quantifying the underlying human behaviors at issue. To date, such a comprehensive understanding does not exist, in part because obtaining large-scale, ecologically valid data about peoples' web search and browsing behavior is challenging.

In this study, we aim to address this knowledge gap by studying the causal effect of Google SERP components on peoples' behaviors. We examine two sets of components, corresponding to the two motivations above. First, we use the term *extracted results* to refer to five components—`direct-answers`, `featured-snippets`, `knowledge-panels`, `top-stories`, and `top-image-carousels`—that present extracted information prominently on the SERP in an attempt to directly address peoples' information needs. We analyze how this set of components affects peoples' information seeking behaviors. Second, we use the term *Google Services* to refer to seven components—`videos`, `images`, `ads`, `shopping-ads`, `local-results`, `map-results`, and `scholarly-articles`—that Google either owns or directly monetizes. We analyze whether this set of components increases *click-through rates (CTR)* to Google properties at the expense of 3rd-party websites. Figures 1 and 2 include screenshot examples of all 12 components.

Specifically, our research questions are:

- **RQ1:** What effect do *extracted results* have on behavior, operationalized through (**RQ1a**) CTR and (**RQ1b**) time on the SERP?
- **RQ2:** What effect do *Google Services* have on traffic, measured through (**RQ2a**) organic CTR[6] to 1st-party sites (`google.com` and `youtube.com`) and (**RQ2b**) organic CTR to 3rd-party sites (all other domains)?

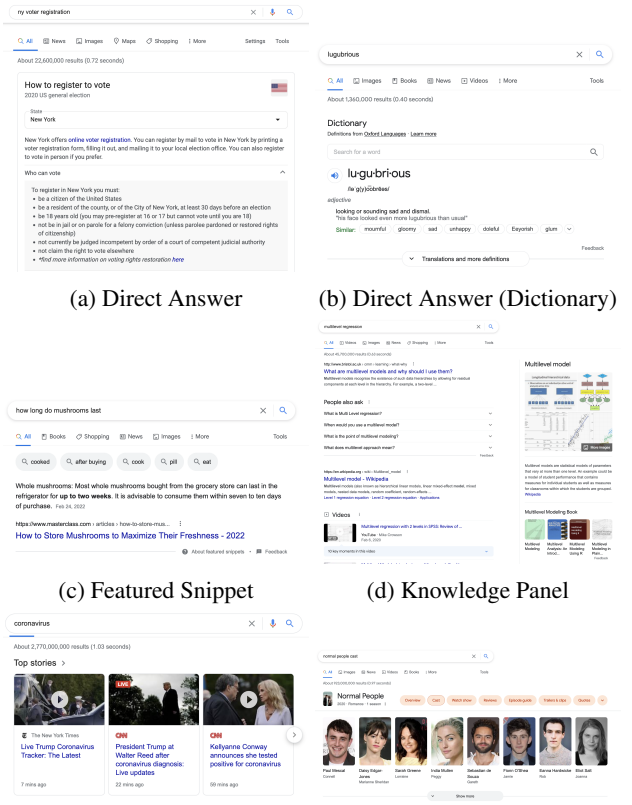
To answer these questions, we apply causal inference methods to a large and ecologically valid trace dataset comprising 477,485 SERPs from 1,756 participants. Our methodological approach exploits variation in SERP composition to find similar pairs of queries that triggered different components. Take `featured-snippets` as an example: if we can find similar pairs of queries—one that triggered a `featured-snippet` and one that did not—and control for remaining contextual differences, we can attribute differences in behavior to the `featured-snippet`.

Our main findings are:

- **RQ1**: `Direct-answers` and `featured-snippets` decrease CTR and increase time on the SERP. We don't find a significant effect of `knowledge-panels` on either outcome.
- **RQ2**: `Local-results` and `images` increase organic CTR to 1st-party domains, while `local-results`, `images` and `shopping-ads` decrease organic CTR to 3rd-party domains.

In the rest of the paper, § 2 contextualizes related work, § 3 introduces our dataset, § 4 presents our approach to causal modeling, § 5 summarizes our results, and § 6 discusses the implications of our findings.

[6]Organic CTR is the rate of clicks on organic links, as opposed to clicks on ads.



(a) Direct Answer (b) Direct Answer (Dictionary)

(c) Featured Snippet (d) Knowledge Panel

(e) Top Stories (f) Top Image Carousel

Figure 1: Examples of extracted results components.

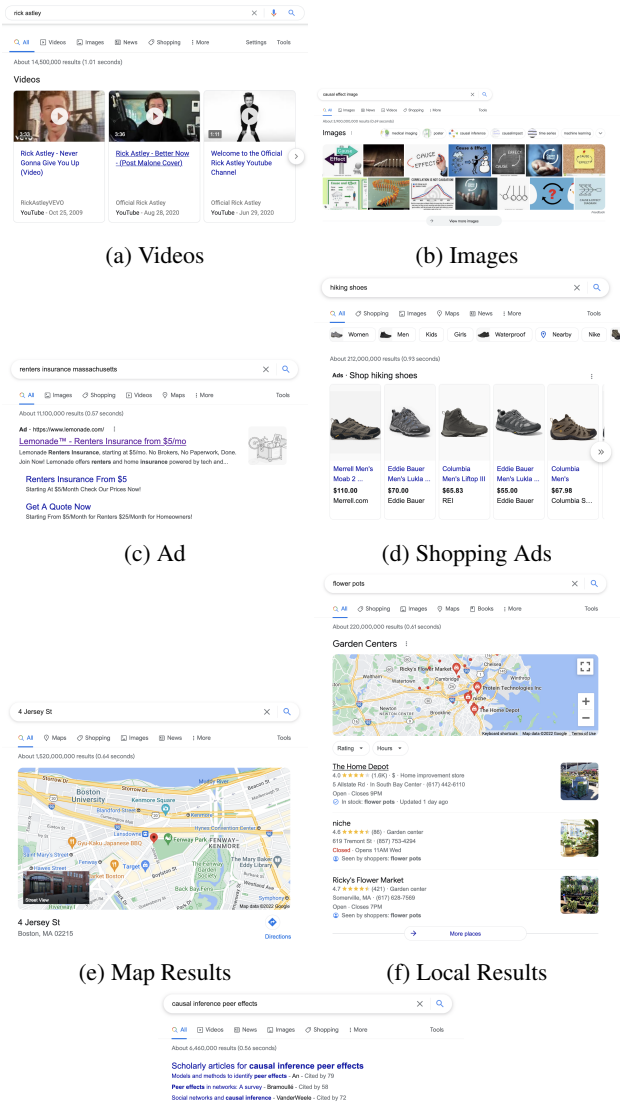
2 Related Work

2.1 Auditing Search Engines

Algorithm audits, especially those focusing on knowledge components, inform **RQ1**. Algorithm auditing is a methodology that systematically scrutinizes algorithmic systems by varying inputs and comparing outputs (Sandvig et al. 2014). Previous audits of search engines have interrogated personalization (Hannak et al. 2013; Kliman-Silver et al. 2015), politics (Robertson et al. 2018; Hu et al. 2019), news (Lurie and Mustafaraj 2018; Trielli and Diakopoulos 2019), and autocomplete (Noble 2018; Robertson et al. 2019).

The most relevant category of audits for this work scrutinizes knowledge components. McMahon, Johnson, and Hecht (2017) examined the effect of knowledge components on traffic to Wikipedia using a controlled study. They found that the "Wikipedia Visitation Rate" increased from 11.1% to 20.5% when Wikipedia-based knowledge components were removed from SERPs.[7] Furthermore, people attributed the source of knowledge components to Google. Lurie and Mustafaraj (2018) found that `knowledge-panels` were

[7]"Knowledge components" here encompass `direct-answers` and `knowledge-panels` in our study.



(a) Videos (b) Images

(c) Ad (d) Shopping Ads

(e) Map Results (f) Local Results

(g) Scholarly Articles

Figure 2: Examples of Google Services components.

insufficient to make definitive credibility assessments, while Lurie and Mulligan (2021) found that 70% of results with misleading information about congressional representatives appeared in `featured-snippets`. Finally, lab studies have found that the information in knowledge components can affect peoples' beliefs. Ludolph et al. (2016) found that comprehensible vaccine information in knowledge panels reduced vaccine skepticism. Epstein et al. (2022) found that featured snippets can potentially shift voting preferences.

2.2 Self-Preferencing

Experimental studies and reports from regulatory agencies on self-preferencing both inform **RQ2**. Luca et al. (2015) conducted a controlled experiment that manipulated listings in `local-results`. They found that people were 40% more likely to click on `local-results` when they contained competitors' listings. Edelman and Lai (2016) evaluated a natural experiment around the launch of Google Flights, in which idiosyncratic differences in search queries affected its presence (our causal identification strategy is very similar, see § 4). The authors found that Google Flights increased the volume of paid clicks by 65% and decreased the volume of organic clicks by 55%. Jeffries and Yin (2020) conducted a scraping experiment of mobile search using a sample of 15,000 trending queries. They parsed SERPs using a spatial approach (Vincent and Hecht 2021) and found that Google devoted 41% of the first page to a combination of its own properties and "direct answers".[8] Methodologically, our paper combines the high-resolution SERP parsing of Jeffries and Yin (2020) with ecologically valid behavioral data like that in Edelman and Lai (2016).

Self-preferencing has received regulatory attention in Europe, the US, and the UK. In its 2017 "Shopping" case, the EC found that Google abused its dominance in general search by favoring its own comparison shopping service (Official Journal of the European Union 2018). In response to this decision, Google was mandated to implement a remedy, and they introduced an auction for placement in `shopping-ads`, which competitors testify might be making the situation worse (Competition and Markets Authority 2020). In a 2020 report, the UK's Competition and Markets Authority (CMA) (Competition and Markets Authority 2020) found that most specialized vertical search engines rely on Google for over 40% of their traffic and spend, on average, 55% of their ad budget on Google. It also cited industry studies finding that Google's Travel results are listed first 98% of the time and that a competitor's CTR decreased by 46.5% when placed below the Flights component.

2.3 Web Search

Two lines of research on web search are relevant to our research questions: good abandonment (**RQ1**) and aggregated search (**RQ2**).

Li, Huffman, and Tokuda (2009) introduced the concept of good abandonment, defining it as "an abandoned query for which the user's information need was successfully addressed by the search results page, with no need to click on a result or refine the query." Diriye et al. (2012) found that participants abandoned 22% of queries and were satisfied with 38% of those abandonments. On mobile search, Williams et al. (2016) found that people credit `direct-answers` as the reason for 56% of good abandonments.

Aggregated search integrates results from multiple specialized search verticals into a single SERP (Arguello et al. 2017). In a study on image, news, and "encyclopedia" verticals, Liu et al. (2015) found that (1) relevant and visually appealing verticals garner more attention, (2) organic results below a vertical receive less attention, and (3) irrelevant verticals increase attention on organic results. Bota, Zhou, and Jose (2016) found that the presence of `knowledge-panels` increased time spent inspecting organic results,

[8] "Direct answers" here encompass `featured-snippets`, `direct-answers`, and `knowledge-panels` in our study.

	Number of SERPs	Number of Users
original	857714	1932
basic cleaning	761797	1900
refined	587373	1890
parsing errors	542488	1774
navigational	477485	1756

Table 1: Summary of SERP dataset filtering steps, including our handling of *refined* queries and *navigational* searches.

while Navalpakkam et al. (2013) found a second *golden triangle*—an attention focal point—in the top left corner of `knowledge-panels`.

3 Data

This section describes (1) how we collected and prepared our dataset, (2) how we operationalized our treatment variable—component type, (3) how we measured our outcome variables—clicks and time on the SERP, and (4) presents a baseline comparison of outcome variables.

3.1 Data Collection and Preparation

From April through December 2020, we recruited participants to install a custom browser extension that we made for Chrome and Firefox. Recruitment was handled in two, parallel waves by the survey company YouGov and the panel aggregator PureSpectrum. Our experimental protocol was approved under Northeastern IRB #20-03-04. See § 7.1 for further discussion of our experimental protocol and approach to protecting participants.

To measure participants' web search behavior, our browser extension collected two types of passive, observational data: browsing history and snapshots of Google Search Engine Result Pages (SERPs). The browsing history data contains the sequence of URLs that participants loaded in their browser. The snapshot data contains the complete HTML of the Google SERPs that participants saw. We use `WebSearcher` to parse the SERPs into ordered lists of components (Robertson and Wilson 2020).

We applied four filtering steps to prepare our Google Search data for analysis. Table 1 lists these steps, along with how much data—in terms of total searches and participants who contributed at least one search—remained after applying each step. First, we performed *basic cleaning*, e.g., removing searches with the `tbm` URL parameter, which indicates that the search was made on a separate vertical like Images or News.

Second, we removed queries that were immediately *refined*, i.e., the next URL in a participant's browsing history is another Google search. Diriye et al. (2012) found that 45% of abandoned (zero-click) searches were triggered by query refinement. Filtering refined queries increases the likelihood that the remaining zero-click searches represent good abandonment instead of bad abandonment (Diriye et al. 2012). Third, we filter out SERPS that show evidence of a *parsing error*. This gives us confidence that we label the treatment correctly on the SERPs we keep.

Finally, we remove *navigational* queries, where a person searches for a specific website, often by name, and then clicks on the result that points to that website. Navigational queries are a specific class of searches with very strong intent, which makes it inappropriate to compare click behavior on these searches to click behavior on other types of searches. Information about how we identify navigational queries can be found in § 7.1 in the Appendix.

3.2 Treatment Variable

We operationalize our treatment variable—component type—as the component in the **top vertical position** on a SERP. For each treatment, the SERPs in the control group have (1) a `general` component (classic blue link) in the top vertical position and (2) no treatment components anywhere else on the SERP. Figure 3a shows an example of the `featured-snippet` treatment condition, while Figure 3b shows a possible control for this treatment. We make one exception to this definition when the treatment variable is a `knowledge-panel`, which appears on the right-hand-side of the SERP. In this case, both the treatment and control conditions have a `general` component (classic blue link) in the top vertical position.

We focus on the top vertical position because 30% of the clicks in our dataset are on the top vertical result and half of the components appear almost exclusively in the top vertical position (see Figure 4).

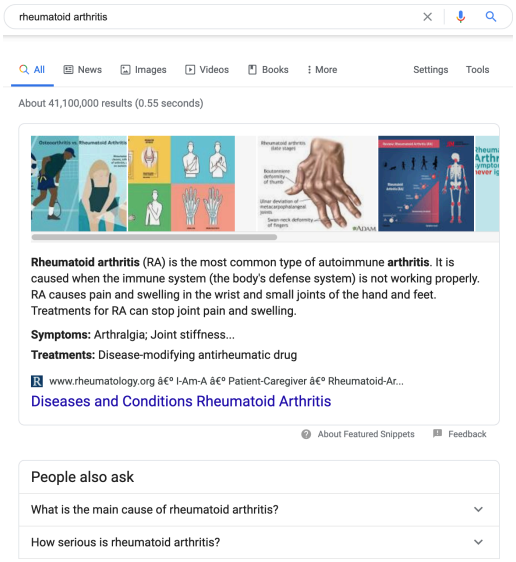
3.3 Outcome Variables

Clicks Click measurement consists of two steps: (1) extract all URLs from a given SERP, and (2) check if the next URL in a participant's browsing history is contained in this set. This means that we measure clicks (1) as binary variables and (2) at the page-level. It is common practice to measure clicks at the page-level in work on good abandonment (Williams et al. 2016; Diriye et al. 2012).

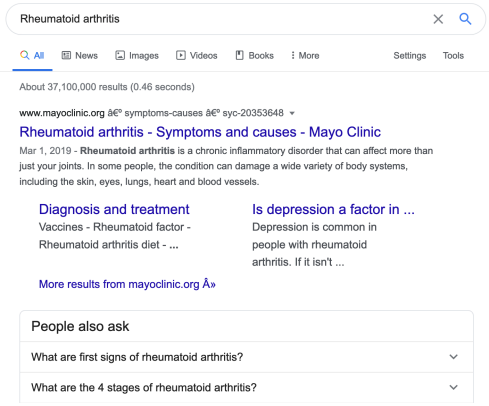
We also measure two characteristics of clicks to answer our research question about self-preferencing (**RQ2**). First, we separate clicks on ads from clicks on organic links using the `gclid` URL parameter.[9] This is important because paid and organic clicks are different forms of traffic. From now on, we refer to clicks on organic links as "organic clicks." Second, we distinguish between clicks to *1st-party* domains (i.e., transitions from a SERP to `google.com` or `youtube.com`) and clicks to *3rd-party* domains (everywhere else). Using these methods, we find that the page-level CTR is 52.5%, 95.7% of clicks are on organic links, and 86.0% of clicks are to 3rd-party domains.

Time on the SERP We measure time on the SERP (**RQ1b**) as the number of seconds elapsed between a visit to a SERP and the visit to the subsequent URL (regardless of whether the subsequent visit was the result of a click on the SERP). Thus, time on the SERP is a positive, continuous variable with a right-skewed distribution. Following Athey, Mobius, and Pal (2021), we cap time on the SERP at five minutes—longer times suggest a participant has left their

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



(a) `Featured-snippet` treatment condition



(b) `Featured-snippet` control condition

Figure 3: Example of treatment and control conditions for the `featured-snippet` component type from our dataset. The two queries are exactly the same, yet the SERPs differ. We exploit this variation in SERP composition to make causal claims from observational data: does the component (a `featured-snippet` in this example) increase the likelihood of clicking on a link or the expected amount of time spent on the SERP? We choose an example with a query that was searched by eight different participants to protect participants' anonymity.

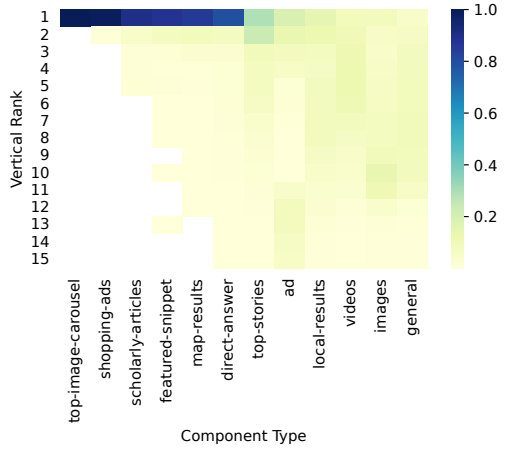


Figure 4: Distributions of components over vertical ranks.

computer. After truncation, the median time on the SERP is 11.7 seconds. 86.7% of the observations have time spent between one second and one minute (inclusive).

3.4 Baseline Comparison

Tables 2 and 3 present baseline difference-in-means between outcome variables in the treatment and control groups for extracted results (**RQ1**) and Google Services (**RQ2**). The `direct-answer` type includes answer boxes, dictionary results, and finance and sports widgets.

We see that four out of five extracted results are associated with lower CTRs and all five are associated with longer times on the SERP. Five out of seven Google Services are associated with higher organic CTRs to 1st-party domains and six out of seven are associated with lower organic CTRs to 3rd-party domains. However, these baseline differences might simply reflect the types of search queries that trigger each component. This motivates us to apply more sophisticated methods to understand the casual relationships between components, CTR, and time on the SERP.

4 Methods

This section describes how we (1) identify confounders that must be controlled, (2) operationalize those confounders, (3) adjust for those confounders to estimate causal effects, and (4) evaluate the sensitivity of our estimates. At a high-level, our estimation approach involves (a) finding similar pairs of queries in different treatment conditions (i.e., *matching*) and then (b) adjusting for the remaining contextual differences with a regression model.

Throughout this section we use the effect of a `featured-snippet` on CTR (**RQ1a**) as a running example to motivate and explain our methodological choices.

4.1 Identifying the Effect

Figure 5 shows a causal diagram that represents our assumptions about the data-generating process for one search. A user has a search intent, from which they formulate a query that they submit to Google Search. In response, Google Search generates a SERP that contains results spanning different topics and component types. Next, the person decides

Component	Sample Size	CTR Difference	Time Difference (s)
featured-snippet	82589	-0.10	11.15
knowledge-panel	60246	-0.12	4.88
direct-answer	26054	-0.37	21.14
top-stories	16904	0.02	4.38
top-image-carousel	10495	-0.05	4.45

Table 2: Baseline differences for extracted results (**RQ1**).

		Organic CTR Difference to	
Component	Sample Size	1st-Parties	3rd-Parties
ad	34999	-0.01	-0.05
shopping-ads	32957	0.04	-0.09
videos	12966	0.15	-0.13
images	9803	0.25	-0.18
local-results	6447	0.22	-0.15
map-results	4036	0.28	-0.18
scholarly-articles	2376	-0.01	0.02

Table 3: Baseline differences for Google Services (**RQ2**).

whether to click on any of the links on the SERP based on their search intent and the displayed results. Finally, the person's interaction with the current SERP influences the generation of the next SERP. In this example, the treatment variable is the presence of a `featured-snippet` at the top of the SERP, and the outcome variable is a click.

We use the backdoor criterion (Glymour, Pearl, and Jewell 2016) to identify the *minimal adjustment set*—the smallest set of confounding variables that block all biasing paths in Figure 5. The minimal adjustment set consists of four confounding variables that are observable in our dataset: (1) the query, (2) topics on the SERP, (3) other components on the SERP and, (4) click behavior on the previous SERP.[10]

4.2 Measuring Confounders

We now describe how we operationalize the four confounders in our causal model.

Query In a recent comparison of text adjustment strategies, Weld et al. (2022) found that transformer-based representations outperformed other text representations. Thus, we use a pre-trained language model based on the BERT architecture (Reimers and Gurevych 2019).[11] This model was fine-tuned to identify semantically similar sentence pairs, which makes it ideal for the *matching* step in our estimation approach (see § 4.3).

Topics We assign topical labels to SERPs based on the domains that appear in links. First, we use the FortiGuard domain classification service to classify domains into 91 categories (Vallina et al. 2020). Second, we label each SERP as a weighted distribution over categories, with weights taken from the empirical click distribution in our dataset to account for decaying attention (Papoutsaki, Laskey, and Huang 2017). For example, our method would assign a SERP containing three links each with a different topic the topical distribution $[0.50, 0.28, 0.22]$. We include the weights of the 20 most common FortiGuard categories in the subsequent regression model, which cover 90.9% of the total topic weight across all SERPs. Figure 6 shows the distribution of these topics over all SERPs.

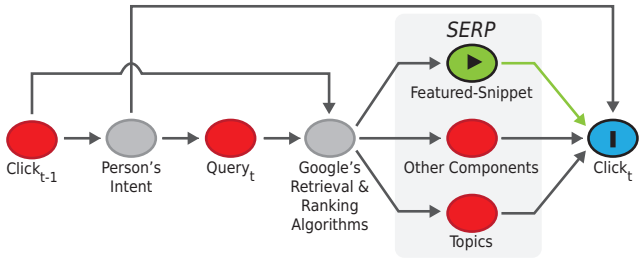


Figure 5: A causal diagram for searching and clicking. In this case, the treatment variable is the presence of a `featured-snippet` at the top of the SERP, and the outcome variable is a click. The green path represents the causal effect of interest, while the grey paths represent biasing paths. Red nodes represent confounding variables and gray nodes represent unobserved variables. The subscript t denotes events that happen within the current search, while events at $t-1$ occurred during a previous search.

Other Components We represent non-treatment components as binary variables in the subsequent regression model, indicating whether or not they were present on the SERP. For example, the `people-also-ask` confounder has the same value (*True*) for both SERPs in Figure 3.

Previous Behavior We represent previous behavior using the participant's click outcome (**RQ1a, RQ2**) and dwell time on the previous search (**RQ1b**). Searches are continuous in time, thus we include the number of seconds since the previous search (on the log scale) and interact it with the outcome on the previous search.

4.3 Estimating the Effect

We now describe our two-step estimation approach, which consists of matching and regression in the matched sample.

Matching Recent work argues that matching is particularly well-suited to address confounding from text in observational studies because it is interpretable—humans can inspect matches and assess them qualitatively (Roberts, Stewart, and Nielsen 2020; Mozer et al. 2020; Keith, Jensen, and O'Connor 2020). In our application, text—the query—is the most important confounder because it blocks the biasing path through a person's search intent. According to Stuart (2010), matching designs consist of four steps: (1) define a similarity measure, (2) implement a matching method,

[10] A subset of participants in our sample responded to a demographic survey. In an online appendix, we describe these responses and additionally adjust for demographics, which can be interpreted as proxies for a person's intent: https://github.com/jlgleason/google-the-gatekeeper.

[11] Its model card can be found here: https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

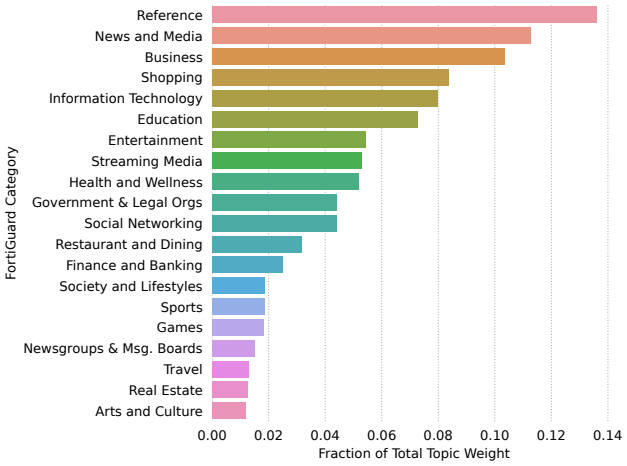


Figure 6: Distribution of topics, one of our confounding variables, over all SERPs.

(3) evaluate the quality of the matches, and (4) analyze the outcome (e.g., with a regression model).

Our *similarity measure* is the cosine similarity between two query vectors (e.g., $\mathbf{A}$ and $\mathbf{B}$): $\frac{\mathbf{A}*\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$. Mozer et al. (2020) found that the cosine similarity produced the best text match quality in a study of five different similarity metrics applied to news articles. There are also strong reasons to prefer the cosine similarity over both propensity score matching (PSM) (Rosenbaum and Rubin 1983) and coarsened exact matching (CEM) (Iacus, King, and Porro 2012). First, matched pairs from PSM cannot be directly assessed because PSM replicates a randomized experiment (King and Nielsen 2019)—matched groups are only similar in expectation. Second, CEM suffers from the curse of dimensionality: there may be vanishingly few exact matches, even if the query string has been coarsened (Mozer et al. 2020; Roberts, Stewart, and Nielsen 2020).

Our *matching method* is 1:1 matching without replacement. After finding matches, our goal during *match evaluation* was to select a cosine similarity threshold above which matched pairs for all components were substantively similar. This is where we took advantage of text matching's interpretability. Specifically, two authors manually examined 30 random matches for each treatment condition and assessed whether the matched control queries could have reasonably elicited the treatment component.[12] Based on this evaluation procedure, we chose a threshold of 0.85. Table 4 shows the mean cosine similarity and the number of matches retained at this threshold for all treatments. Figure 7 shows that matching also improves balance with respect to the other confounders not included in the matching specification.

Finally, one disadvantage of dropping so many matches is that we cannot generalize to the entire treated population (i.e., all SERPs with a `featured-snippet` at the top of the page) (Ho et al. 2007). However, Greifer and Stuart (2021) argue—and we agree—that this trade-off is worth it for treatment effect discovery, i.e., a precise and robust estimate in *some* population.

[12]During this evaluation procedure, we discovered that there were not enough high-quality matches for six additional components that we parsed successfully—`hotels`, `flights`, `jobs`, `translation`, `recipes`, and `twitter-cards`. This demonstrates the importance of directly assessing matched pairs.

	Component	Mean Cosine Similarity	Matches Retained
Extracted Results	featured-snippet	0.92	3330
	knowledge-panel	0.92	1557
	top-stories	0.94	1057
	top-image-carousel	0.96	915
	direct-answer	0.92	345
Google Services	ad	0.95	2223
	shopping-ads	0.93	1090
	local-results	0.90	306
	videos	0.93	300
	images	0.91	170
	map-results	0.90	127
	scholarly-articles	0.91	126

Table 4: Matching summary after filtering matches with cosine similarity below 0.85.

Outcome Analysis The last step in a matching design is to analyze the outcome in the matched sample. We do so using a regression model that adjusts for the remaining differences between covariates in Figure 7. For click outcomes (**RQ1a, RQ2**), we use a logistic regression model with independent Cauchy$(0, 2.5)$ priors on the coefficients, which prevents complete separation (Gelman et al. 2008). For time on the results page (**RQ1b**), we use a generalized linear model with a gamma distribution and log link.

After fitting the outcome model, we use g-computation to estimate the marginal effect, the quantity that would be estimated in a randomized experiment (Snowden, Rose, and Mortimer 2011). This approach uses the regression model to simulate potential outcomes under treatment and control for each search. The marginal effect is then the average difference between the potential outcomes. This is the effect measure that we present in Figures 8 and 9. We compute standard errors using the delta method and cluster at the participant level to account for dependence within participants.

4.4 Sensitivity Analysis

A key threat to validity in observational studies is unmeasured confounding. We evaluate how strong unmeasured confounding would have to be to explain away our estimates using the E-value (VanderWeele and Ding 2017). The E-value measures the minimum strength of association confounders must have with both the treatment and outcome—*above and beyond measured covariates*—to shift the confidence interval to the null (i.e., no effect). The E-value is similar to Cinelli and Hazlett (2020)'s robustness value, but it is measured on the risk ratio scale instead of the partial R^2 scale. Thus, we can use it with our non-linear outcome models (i.e., logistic and gamma regression).

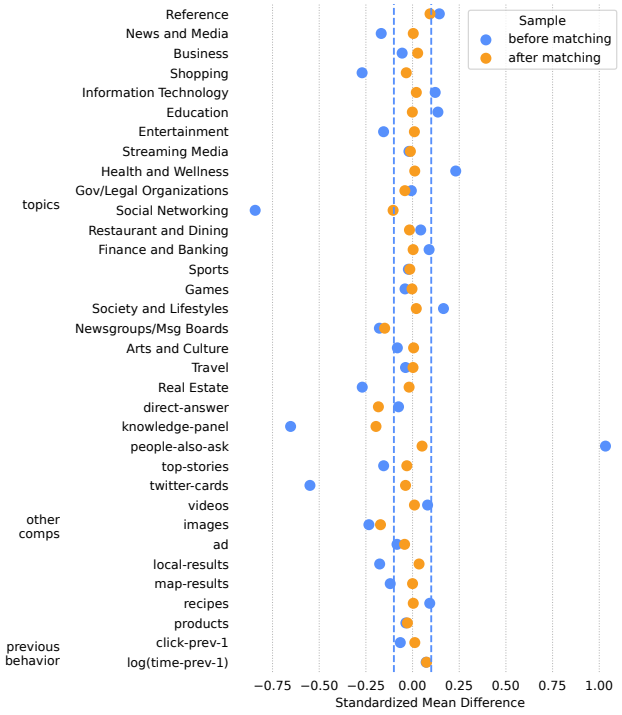


Figure 7: Covariate balance before and after matching measured using standardized mean differences. The y-axis includes all covariates that are not used for matching. The dashed vertical lines represent a standardized mean difference of 0.1 (Greifer 2022).

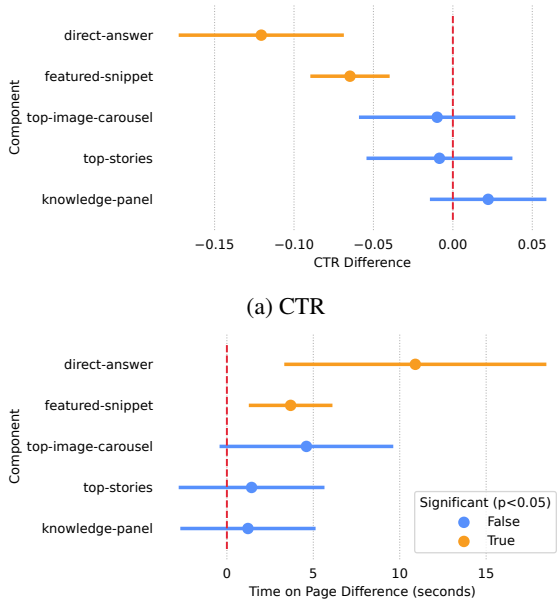
We ground the E-value in our measured confounders using McGowan and Greevy Jr (2020)'s *Observed Covariate E-value*. This approach removes a set of observed confounders (e.g., the topic proportions) from the outcome model and measures how much the effect estimate changes on the E-value scale. This benchmarks hypothetical unmeasured confounding against the observed confounders.

5 Results

5.1 Effects of Extracted Results (RQ1)

Figure 8a shows the effects of extracted results on CTR. `Direct-answers` and `featured-snippets` both have significant negative effects, decreasing CTR by -12.1 percentage points (pp) (95% CI -17.1 to -7.0pp) and -6.5pp (95% CI -8.9 to -4.1pp), respectively. However, both effects are smaller than the corresponding naive estimates in Table 2. The effects of `knowledge-panels`, `top-stories`, and `top-image-carousels` are not distinguishable from zero. The naive estimate in Table 2 would have led us to incorrectly conclude that `knowledge-panels` have a large negative effect on CTR.

Figure 8b shows the effects of extracted results on time on the SERP. `Direct-answers` and `featured-snippets` both significantly increase time on the SERP by 10.9s (95% CI 3.4–18.4s) and 3.7s (95% CI 1.4–6.0s), respectively. However, both effects are smaller than the corresponding naive estimates in Table 2. The point estimate for `top-image-carousels` (4.6s) is also positive, but its confidence interval overlaps with zero. The effects of the other two components are not distinguishable from zero.



(a) CTR

(b) Time on the SERP

Figure 8: Effects of extracted results on CTR and Time on the SERP.

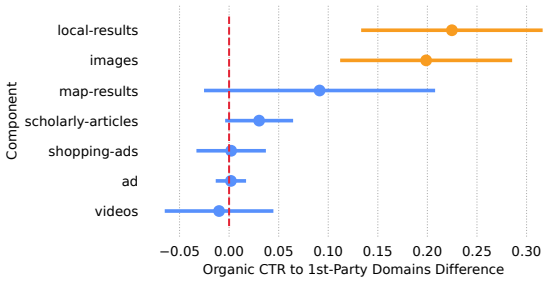
5.2 Effects of Google Services (RQ2)

Figure 9a shows the effect of Google Services on organic CTR to 1st-party domains. `Local-results` and `images` both substantially increase CTR to Google properties by a staggering 22.5pp (95% CI 13.5–31.4pp) and 19.9pp (95% CI 11.4–28.4pp), respectively. The point estimate for `map-results` (9.1pp) is also positive, but its confidence interval overlaps with zero. The effect of `videos` is not distinguishable from zero, which contrasts with the naive estimate in Table 3. The reason for this change is that we adjust for the presence of `youtube.com` links when we control for the streaming media topic. Finally, we are confident that `ads` have no effect on organic CTR to 1st-party domains, which validates our methodological approach because Google does not advertise its own services on Google Search.

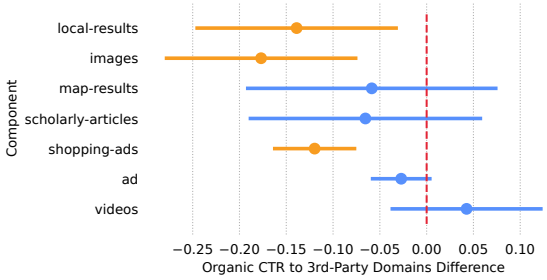
Figure 9b shows the effect of Google Services on organic CTR to 3rd-party domains. `Local-results`, `images`, and `shopping-ads` all have significant negative effects, decreasing CTR to 3rd-party domains by -13.9pp (95% CI -24.5 to -3.3pp), -17.7 pp (95% CI -27.8 to -7.6pp), and -12.0pp (95% CI -16.2 to -7.7pp), respectively. The effects of `map-results`, `scholarly-articles`, and `videos` are not distinguishable from zero.

5.3 Sensitivity Analysis

Figure 10 shows the results of sensitivity analyses using the E-value. In order to tip any of the significant effects from



(a) Organic CTR to 1st-Party Domains



(b) Organic CTR to 3rd-Party Domains

Figure 9: Effects of Google Services on organic CTR to 1st and 3rd-party domains.

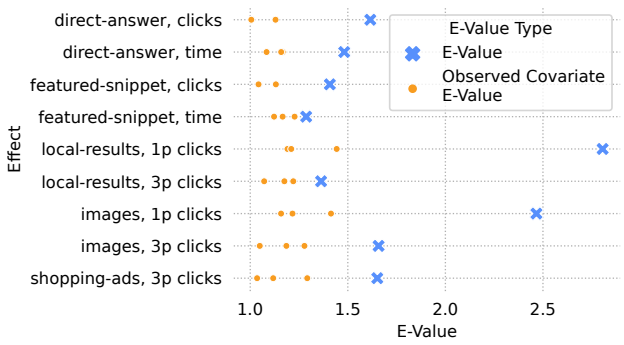


Figure 10: E-value sensitivity analysis. For each treatment-outcome pair, the E-value represents the minimum amount of unmeasured confounding needed to tip the confidence interval. Observed covariate E-values demonstrate how much the confidence bound closer to zero changes on the E-value scale if we drop one group of covariates.

Figures 8 and 9, we require unmeasured confounding that is stronger than all three groups of measured covariates. That being said, the two effects that are most susceptible to unmeasured confounding are the effect of `featured-snippets` on time spent on the SERP and the effect of `local-results` on organic clicks to 3rd-parties.

6 Discussion

6.1 Extracted Results

We find that both `direct-answers` and `featured-snippets` decrease CTR and increase time on the SERP, which is broadly consistent with prior work (Chilton and Teevan 2011; McMahon, Johnson, and Hecht 2017; Epstein et al. 2022). As expected, `direct-answers` have a larger effect than `featured-snippets` on both outcomes. Interestingly, we do not find a significant effect of `knowledge-panels` on either outcome.

Our findings highlight the tensions that knowledge components produce between different stakeholders. Users may benefit when they can answer their questions marginally faster using information contained directly on the SERP (Williams et al. 2016). However, the effect of good abandonment on clicks is consequential for 3rd-parties, who may lose revenue and visibility. This impact is especially significant for volunteer-driven non-profits (e.g., Wikipedia) and websites whose value is intrinsically built on having an engaged user base (e.g., StackOverflow) (Vincent et al. 2019). Historically, Google has also misappropriated content from small 3rd-parties, e.g., lyrics site Genius and Celebrity Net Worth, for use in knowledge components (Jeffries and Yin 2020). One interesting development here is the Wikimedia Foundation's release of a commercial product for companies that use high volumes of Wikipedia content.[13] Google and the Internet Archive are its first customers.

The effects we observe increase the importance of work on the credibility and quality of content in knowledge components (Lurie and Mustafaraj 2018; Lurie and Mulligan 2021). Future work should heed Mustafaraj, Lurie, and Devine (2020)'s argument for "voter-centered audits" (i.e., user-centered audits) that focus on the types of information needs for which searchers use these components. For example: what important information needs (e.g., political, health, financial, and/or legal) do knowledge components attempt to satisfy? Lab studies have also investigated the effects of `featured-snippets` and `direct-answers` on peoples' beliefs (Ludolph et al. 2016; Epstein et al. 2022). Future work could extend these studies to ecological settings and assess whether behavioral effects are ephemeral.

Of course, the elephant in the room is the introduction of chatbot-based search on Google and Bing, which is nothing if not a knowledge component on steroids. Chatbot-based search exacerbates concerns about abandonment, misappropriation, and information quality (Robertson 2023) that were originally discussed in the context of knowledge components. Future studies could adapt the methods in this paper to study users' interactions with chatbot-based search.

6.2 Google Services

We find that `local-results` and `images` substantially increase organic CTR to 1st-party domains, while `local-results`, `images` and `shopping-ads` significantly decrease organic CTR to 3rd-party domains. Interestingly, `shopping-ads` have little effect on organic CTR to 1st-party domains. This means that `shopping-ads` cause

[13] https://wikimediafoundation.org/news/2022/06/21/wikimedia-enterprise-announces-google-and-internet-archive-first-customers

people to click on ads, but not to navigate to Google's shopping-specific vertical search engine.

This combination of effects quantifies the power of search engine operators who also own services that compete for attention on their platform: owner-operators can simultaneously boost compatriots and beleaguer competitors. Our results complement the CMA's macro-level findings that most specialized search providers rely on Google for over 40% of their traffic. The 2020 US House report also documented that negative effects on organic traffic force competitors to substitute paid clicks for organic clicks (Subcommittee on Antitrust, Commercial and Administrative Law 2022). Our study did not evaluate this substitution directly, but we do measure organic CTR to acknowledge the difference between organic and paid traffic for 3rd-parties.

Existing remedies for self-preferencing—e.g., the EC's "Shopping" decision and the auction mechanism Google implemented in response—have been criticized for failing to restore competition (Jeffries 2020). Fortunately, new attempts exist: the EC recently passed the Digital Markets Act (European Commission 2022), which explicitly forbids "gatekeepers" from treating their products and services more favorably than similar products or services offered by 3rd-parties. In the US, the American Online Innovation and Choice Act (Klobuchar 2022) would make it illegal for covered platforms to preference their own products or services. Our results could be used to support cases brought under those acts.

6.3 Limitations

Our main limitations revolve around generalization. First, we acknowledged in § 4.3 that dropping matches prevents us from generalizing to the entire treated population (e.g., all SERPs with a `featured-snippet` at the top of the page) within our study. Second, we cannot claim to know how these effects would transfer to mobile search, although previous work implies that they would be stronger (Williams et al. 2016; Jeffries and Yin 2020). Third, these results apply to a snapshot in time—Google constantly experiments with new and redesigned components. Thus, effects for specific components might not generalize temporally.

We also reiterate that our data did not have enough high-quality matches to estimate the effects of specific Google Services (i.e., `hotels` and `flights`) that have received attention in previous work (Edelman and Lai 2016; Competition and Markets Authority 2020). Two factors explain the lack of high-quality matches: (1) `hotel` and `flight` components appeared infrequently in our data, and (2) travel-related searches almost invariably produced SERPs with `hotel`, `flight`, or `ads` components at the top of the page. Thus, the lack of quality matches, in and of itself, is evidence of the extent of self-preferencing in travel-related search.

7 Appendix

7.1 Navigational Searches

Teevan, Liebling, and Geetha (2011) distinguish between two types of navigational queries: general and personal. We classify a query q as general navigation using its click entropy: $CE(q) = -\sum_{d \in D(q)} p(d|q) * \log p(d|q)$, where $D(q)$ is the collection of domains clicked for query q, and $p(d|q)$ is the proportion of clicks on domain d among all clicks for query q. Queries with $CE(q) \leq 1$ are classified as general navigation. We classify repeated queries that a participant uses to navigate to the same domain as personal navigation. Finally, we classify queries as navigational if the start or end of the query matches the top-level domain of the next URL (Jansen, Booth, and Spink 2008). Overall, we label 9.0% of searches as navigational queries, which is similar to the 10–21% navigational query rate identified in previous studies of users on other search engines (Teevan, Liebling, and Geetha 2011; Jansen, Booth, and Spink 2008).

Ethical Statement

In accordance with our IRB-approved experimental protocol, we obtained informed consent from participants before collecting any data, participants were compensated for installing our extension, and we informed participants that they could uninstall our extension at any time. Our extension automatically uninstalled itself at the end of the data collection period. Given that the focus of this study is on the behavior of search engines, not users, we do not anticipate any adverse impact on participants. We protect participants' privacy by not sharing any of their individualized data. After publication we will release the aggregated data and source code to reproduce the figures in the paper.

Acknowledgements

We thank Brendan Nyhan, Jason Reifler, Annie Y. Chen, David Lazer, Olga Vitek, Nick Beauchamp, and the anonymous reviewers for their comments. 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

Arguello, J.; et al. 2017. Aggregated search. *Foundations and Trends in Information Retrieval*, 10(5).

Athey, S.; Mobius, M.; and Pal, J. 2021. The impact of aggregators on internet news consumption. Technical report, National Bureau of Economic Research.

Bota, H.; Zhou, K.; and Jose, J. M. 2016. Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proceedings of the 2016 acm on conference on human information interaction and retrieval*.

Chilton, L. B.; and Teevan, J. 2011. Addressing people's information needs directly in a web search result page. In *Proceedings of the 20th international conference on World wide web*.

Cinelli, C.; and Hazlett, C. 2020. Making sense of sensitivity: Extending omitted variable bias. *Journal of the*

Royal Statistical Society: Series B (Statistical Methodology), 82(1).

Competition and Markets Authority. 2020. Appendix P: specialised search. https://assets.publishing.service.gov.uk/media/5fe496018fa8f56af2a85fea/Appendix_P_-_specialised_search_v.8_WEB.pdf. Accessed: 2023-04-18.

Diriye, A.; White, R.; Buscher, G.; and Dumais, S. 2012. Leaving so soon? Understanding and predicting web search abandonment rationales. In *Proceedings of the 21st ACM international conference on Information and knowledge management*.

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

Epstein, R.; Lee, V.; Mohr, R.; and Zankich, V. R. 2022. The Answer Bot Effect (ABE): A powerful new form of influence made possible by intelligent personal assistants and search engines. *PloS one*, 17(6).

European Commission. 2022. The Digital Markets Act: ensuring fair and open digital markets. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en. Accessed: 2023-04-18.

Gelman, A.; Jakulin, A.; Pittau, M. G.; and Su, Y.-S. 2008. A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics*, 2(4).

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Greifer, N. 2022. Assessing Balance. https://kosukeimai.github.io/MatchIt/articles/assessing-balance.html. Accessed: 2023-04-18.

Greifer, N.; and Stuart, E. A. 2021. Choosing the Estimand When Matching or Weighting in Observational Studies. *arXiv preprint arXiv:2106.10577*.

Hannak, A.; Sapiezynski, P.; Molavi Kakhki, A.; Krishnamurthy, B.; Lazer, D.; Mislove, A.; and Wilson, C. 2013. Measuring personalization of web search. In *Proceedings of the 22nd international conference on World Wide Web*.

Ho, D. E.; Imai, K.; King, G.; and Stuart, E. A. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3).

Hu, D.; Jiang, S.; Robertson, R. E.; and Wilson, C. 2019. Auditing the Partisanship of Google Search Snippets. In *Proceedings of the Web Conference*.

Iacus, S. M.; King, G.; and Porro, G. 2012. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1).

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).

Jeffries, A. 2020. How to Stop Google Self-Preferencing? Europe May Not Be the Model. https://themarkup.org/google-the-giant/2020/10/15/big-tech-antitrust-google-nondiscrimination-enforcement. Accessed: 2023-04-18.

Jeffries, A.; and Yin, L. 2020. Google's Top Search Result? Surprise! It's Google. https://themarkup.org/google-the-giant/2020/07/28/google-search-results-prioritize-google-products-over-competitors. Accessed: 2023-04-18.

Keith, K. A.; Jensen, D.; and O'Connor, B. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. *arXiv preprint arXiv:2005.00649*.

King, G.; and Nielsen, R. 2019. Why propensity scores should not be used for matching. *Political Analysis*, 27(4).

Kliman-Silver, C.; Hannak, A.; Lazer, D.; Wilson, C.; and Mislove, A. 2015. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of the Internet Measurement Conference*.

Klobuchar, A. 2022. American Online Innovation and Choice Act. https://www.congress.gov/117/bills/s2992/BILLS-117s2992rs.pdf. Accessed: 2023-04-18.

Li, J.; Huffman, S.; and Tokuda, A. 2009. Good abandonment in mobile and PC internet search. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*.

Liu, Z.; Liu, Y.; Zhou, K.; Zhang, M.; and Ma, S. 2015. Influence of vertical result in web search examination. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval*.

Luca, M.; Wu, T.; Couvidat, S.; Frank, D.; Seltzer, W.; et al. 2015. *Does Google Content Degrade Google Search?: Experimental Evidence*. Harvard Business School Boston, MA, USA.

Ludolph, R.; Allam, A.; Schulz, P. J.; et al. 2016. Manipulating Google's knowledge graph box to counter biased information processing during an online search on vaccination: application of a technological debiasing strategy. *Journal of medical Internet research*, 18(6).

Lurie, E.; and Mulligan, D. K. 2021. Searching for Representation: A sociotechnical audit of googling for members of US Congress. *arXiv preprint arXiv:2109.07012*.

Lurie, E.; and Mustafaraj, E. 2018. Investigating the Effects of Google's Search Engine Result Page in Evaluating the Credibility of Online News Sources. In *Proceedings of the 10th ACM Conference on Web Science*.

McGowan, L. D.; and Greevy Jr, R. A. 2020. Contextualizing E-values for Interpretable Sensitivity to Unmeasured Confounding Analyses. *arXiv preprint arXiv:2011.07030*.

McMahon, C.; Johnson, I.; and Hecht, B. 2017. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Mozer, R.; Miratrix, L.; Kaufman, A. R.; and Anastasopoulos, L. J. 2020. Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis*, 28(4).

Mustafaraj, E.; Lurie, E.; and Devine, C. 2020. The case for voter-centered audits of search engines during political elections. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Navalpakkam, V.; Jentzsch, L.; Sayres, R.; Ravi, S.; Ahmed, A.; and Smola, A. 2013. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proceedings of the 22nd international conference on World Wide Web*.

Noble, S. U. 2018. *Algorithms of oppression*. New York University Press.

Official Journal of the European Union. 2018. Summary of Commission decision: Google Search (Shopping). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018XC0112(01)&from=EN. Accessed: 2023-04-18.

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*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Roberts, M. E.; Stewart, B. M.; and Nielsen, R. A. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4).

Robertson, K. 2023. Publishers Worry A.I. Chatbots Will Cut Readership. https://www.nytimes.com/2023/03/30/business/media/publishers-chatbots-search-engines.html. Accessed: 2023-04-18.

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).

Robertson, R. E.; Jiang, S.; Lazer, D.; and Wilson, C. 2019. Auditing autocomplete: Suggestion networks and recursive algorithm interrogation. In *Proceedings of the 10th ACM Conference on Web Science*.

Robertson, R. E.; Lazer, D.; and Wilson, C. 2018. Auditing the personalization and composition of politically-related search engine results pages. In *Proceedings of the 2018 World Wide Web Conference*.

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

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1).

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22.

Snowden, J. M.; Rose, S.; and Mortimer, K. M. 2011. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American journal of epidemiology*, 173(7).

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1).

Subcommittee on Antitrust, Commercial and Administrative Law. 2022. Investigation of Competition in Digital Markets. https://www.govinfo.gov/content/pkg/CPRT-117HPRT47832/pdf/CPRT-117HPRT47832.pdf. Accessed: 2023-04-18.

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

Trielli, D.; and Diakopoulos, N. 2019. Search as news curator: The role of Google in shaping attention to news information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*.

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*.

VanderWeele, T. J.; and Ding, P. 2017. Sensitivity analysis in observational research: introducing the E-value. *Annals of internal medicine*, 167(4).

Vincent, N.; and Hecht, B. 2021. A Deeper Investigation of the Importance of Wikipedia Links to Search Engine Results. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1).

Vincent, N.; Johnson, I.; Sheehan, P.; and Hecht, B. 2019. Measuring the importance of user-generated content to search engines. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13.

Weld, G.; West, P.; Glenski, M.; Arbour, D.; Rossi, R. A.; and Althoff, T. 2022. Adjusting for confounders with text: Challenges and an empirical evaluation framework for causal inference. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16.

Williams, K.; Kiseleva, J.; Crook, A. C.; Zitouni, I.; Awadallah, A. H.; and Khabsa, M. 2016. Detecting good abandonment in mobile search. In *Proceedings of the 25th International Conference on World Wide Web*.