An engine not a camera: Measuring performative power of online search

Celestine Mendler-Dünner[*,§] **Gabriele Carovano**[‡] **Moritz Hardt**[§]

[*]ELLIS Institute Tübingen
[§]Max-Planck Institute for Intelligent Systems, Tübingen and Tübingen AI Center
[‡]Italian Competition Authority

Abstract

The power of digital platforms is at the center of major ongoing policy and regulatory efforts. To advance existing debates, we designed and executed an experiment to measure the performative power of online search providers. Instantiated in our setting, performative power quantifies the ability of a search engine to steer web traffic by rearranging results. To operationalize this definition we developed a browser extension that performs unassuming randomized experiments in the background. These randomized experiments emulate updates to the search algorithm and identify the causal effect of different content arrangements on clicks. Analyzing tens of thousands of clicks, we discuss what our robust quantitative findings say about the power of online search engines, using the Google Shopping antitrust investigation as a case study. More broadly, we envision our work to serve as a blueprint for how the recent definition of performative power can help integrate quantitative insights from online experiments with future investigations into the economic power of digital platforms.

1 Introduction

At the heart of one of Europe's most prominent antitrust case is a seemingly mundane question: How much can a search engine redirect traffic through content positioning? In 2017, the European Commission alleged that Google favored its own comparison shopping service by steering clicks away from search results towards Google's own product comparison service. The technical centerpiece of the case was an ad-hoc data analysis about the position and display biases of Google search results. Google appealed the European Commission's charges, pointing to, among other arguments, methodological errors.[1]

The case is emblematic of a broader problem. Although urgently needed, there is currently no accepted technical framework for answering basic questions about the economic power of digital platforms. Lawyers, economists, and policy makers agree that traditional antitrust tools struggle with multi-sided platforms [1, 2]. Against this backdrop, a recently developed concept from the machine learning literature, called performative power [3], suggests a way to augment existing antitrust enforcement tools and mitigate some of their limitations. Performative power measures how much a platform can causally influence platform users through its algorithmic actions. By directly relating power to a causal effect, it sidesteps the complexities underlying conventional market definitions and offers a promising framework to integrate data and experimental methods with digital market investigations. Although the definition of performative power enjoys appealing theoretical properties, a proof of its practical applicability was still missing.

[1]Case C-48/22 P, Google and Alphabet v Commission (Google Shopping), ECLI:EU:C:2024:14.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1891

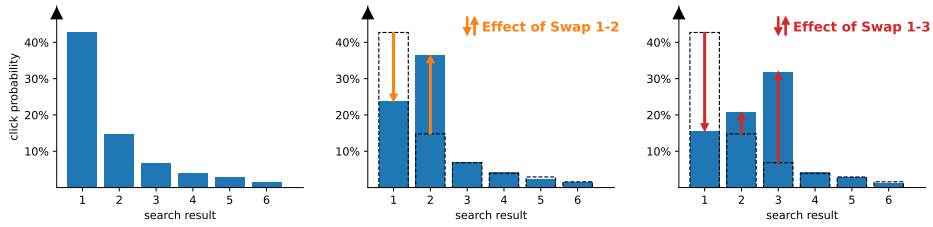


Figure 1: The ability to influence web traffic through content arrangement. Blue bars show average click probability observed for generic search results in position 1 to 6 on Google search under different counterfactual arrangements; default arrangement (left), swapping results 1 and 2 (middle), swapping results 1 and 3 (right). We provide a detailed discussion in Section 5 where we also explore arrangement changes beyond reranking.

Our contributions. We present a first proof of concept showing how to use performative power as an investigative tool in practice. The instantiation of performative power we consider is motivated by the recent Google Shopping antitrust investigation ran by the European Commission against Alphabet Inc. It concerns the ability of a search engine to impact web traffic through decisions of how to arrange content.

Our core contribution is to design and implement an online experiment to establish a lower bound on performative power for the two most widely used search engines, Google Search and Bing, by providing quantitative insights into the causal effect of algorithmic updates on clicks. Our experiment is based on a browser extension, called Powermeter, that emulates updates to the platform's algorithm by modifying how search results are displayed to users. The arrangement to which a user is exposed is chosen at random every time they perform a search. We implement different counterfactual arrangements to inspect the effect of re-ranking and favored positioning (e.g., Ads or Shopping boxes) on clicks, both in isolation and jointly. We discuss several technical steps we implemented to take care of the internal validity of our experimental design.

Using Powermeter we collected data of about 57,000 search queries from more than 80 different subjects, over the period of 5 months. Our experiment is designed to measure the causal effects of arrangement under natural interactions of users with the platform and the queries for any given user follow the distribution of queries under their every-day use of online search. Figure 1 provides a first glimpse into the observed effects. In summary, we find that consistently down-ranking the first element by one position causes an average reduction in clicks of 42% for the respective element on Google search. Down-ranking the same element by two positions yields a reduction of more than 50%. For Bing we find an even larger effect of ranking, although with less tight confidence intervals due to the small number of Bing queries performed by our participants. When combining down-ranking with the addition of Ads or Shopping boxes, the effect of arrangement is even more pronounced, showing a distortion in clicks for the first result of 66% averaged across queries where such elements are naturally present on Google search. Inspecting different subsets of queries we find that the effect of position is larger for queries with a high number of candidate search results. To the best of our knowledge, we are the first to offer independent quantitative experimental insights into display effects on Google search and Bing.

Finally, we outline how to formally relate our quantitative piece of evidence to questions about self-preferencing relevant in the context of the Google Shopping case. Together, we hope our empirical and theoretical results can serve as a first blueprint for what future antitrust investigations of digital platforms' market power based on performative power might look like.

2 Preliminaries and related work

The market power of digital platforms is the subject of a robust debate in policy, legal, and technical circles. See, for example, Newman [4], Crémer et al. [2], Stigler Committee [1], Furman [5], Cabral et al. [6]. Conventional antitrust enforcement tools have been put into question [7] and new concepts of market power have been proposed to deal with the complexities of digital markets [8]. These account for the multi-sided nature of the markets as well as the role of behavioral weaknesses of consumers—albeit with limited success. We refer to a comprehensive literature survey about behavioral aspects in online market by the UK Competition and Market Authority [9].

Performative power. The concept of performative power is inspired by recent developments in performative prediction [10] from the computer science literature. We refer the reader to Hardt and Mendler-Dünner [11] for a recent survey on the topic. A robust insight from performative prediction is that beyond learning patterns in data, the ability to *steer* the data-generating distribution similarly factors into a predictive system's performance. Performative power [3] recognizes that the ability to steer depends on *power*—in terms of reach and scale—of the platform making the predictions. Thus, the core idea behind the new notion of power is to measure the degree to which predictions are performative to obtain an estimate of the power of a platform. Formally, performative power relates the ability of a platform to steer the population of participants, to the causal effect of algorithmic actions.

Definition 1 (Performative power [3])**.** Given the algorithmic action a_0 and a set of alternative conducts $\mathcal{A}$, a population $\mathcal{Q}$ and an outcome variable z. Performative power is defined as

$$\mathrm{PP} := \sup_{a \in \mathcal{A}} \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathrm{E} \left\| z_{a_0}(q) - z_a(q) \right\|_1 , \tag{1}$$

where $z_{a_0}(q)$ denotes the outcome for unit $q \in \mathcal{Q}$ under a_0 and $z_a(q)$ denotes the counterfactual outcome, would the platform implement $a \in \mathcal{A}$ instead. Expectations are taken with respect to the randomness in the potential outcome.

Performative power is a measure of influence that predictive systems can have over their participants. It offers a family of definitions that can be instantiated flexibly in a given context. The specific meaning is determined by each instantiation. Performative power can be applied forward-looking to understand whether a platform has the ability to plausibly cause a specific change, as well as in retrospect to measure the effect of an observed conduct. In this work we use performative power to quantify the effect of an algorithmic update a^* central to a recent antitrust investigation against Alphabet Inc. ran by the European Commission.[2] In practical terms, we instantiate $\mathcal{A}$ with a set of conservative and implementable counterfactuals such as to provide a plausible lower bound on the effect of a^*.

The Google Shopping case. In 2017 the European Commission imposed a fine of 2.42 billion EUR on Alphabet Inc. for "abusing its dominance as a search engine by favouring its comparison shopping service.".[2] The General Court dismissed Google's action against the decision in 2021 and the Court of Justice of the European Union upheld the Court's ruling in 2024. It represents a landmark in EU competition law. The conduct under investigation concerned a specific update to the Google search algorithm. The update **a)** demoted rival comparison shopping services among the general search results, often by multiple positions, and, at the same time, **b)** systematically gave prominent placement to Google's own comparison shopping service by triggering visually appealing boxes for shopping queries, reserved for Google's own service. The goal of this work is to provide quantitative insights into the effect of this conduct on web traffic by means of online experiments.

Display effects. Consumer choices on digital platforms are critically mediated by how platforms present content to users. Choice architecture designs [12], presentation bias [13], position bias [14, 15], and trust bias [16] are known to play an important role. There is a rich literature in machine learning aiming to mechanistically understand such biases for debiasing click data [17–24], building better ranking models and auctions [25, 26], and interpreting user feedback in recommender systems [14], to name a few. Unfortunately behavioral aspects often resist a clean mathematical specification. By focusing on measuring a directly observable statistic, performative power circumvents the challenges of modeling behavioral biases for monitoring, auditing and measuring digital economies.

Measuring the effect of algorithmic updates. Several works have been interested in measuring the effects of potential arrangement changes of online platforms. For example, Ursu [27] rely on public data collected under randomized result ordering to investigate the role of positioning on Expedia. Narayanan and Kalyanam [28] investigated position bias in search advertising using a regression discontinuity design. Also focusing on online advertising, Agarwal et al. [29] investigate position bias by experimentally randomizing bids to indirectly influence the ranking. Similarly in information retrieval researchers have studied active interventions in the form of order randomization [30], or relied on harvesting click data collected under multiple historical rankings [31]. In our work we collect experimental data ourselves. We use a browser extension to emulate the algorithmic updates of interest *without* requiring control over the platform's algorithm.

[2]European Commission, AT.39740, *Google Search (Shopping),* 27.06.2017.

https://doi.org/10.52202/079017-1891

Browser extensions have previously been used as a tool for automatically collecting data to audit systems. Robertson et al. [32] audit Google search for polarization on politically-related searches. Gleason et al. [33] collect data via an extension to directly investigate the effect of search result components on clicks. Also the ongoing National Internet Observatory [34] relies on a browser extension to collect web traffic data. While prior works focus on collecting observational data for monitoring systems, we use the extension to conduct online experiments.

3 Performativity in online search

We start by formalizing the causal question under investigation. We model an online search platform as a distribution over *events*. An event is a triplet of a user query Q, content arrangement A and click outcome C. A user query corresponds to a person visiting the search page and entering a search query in the search bar. The query is processed by the platform and results in an arrangement of content on the website. The mapping is typically defined by a proprietary pipeline involving a ranking algorithm that determines the order in which search results are ranked and displayed, including the positioning of components such as Ads or featured elements. Then, mediated by the arrangement, the user query leads to a click outcome C. The categorical random variable C indexes the element clicked over by the user. It is a function of the user query and the arrangement.

3.1 From the causal effect of arrangement to performative power

Assume the platforms were to change their algorithm that determines the content arrangement. We seek to answer the following causal questions: *How much does a change to the arrangement impact clicks of a content element on the search page?*

If clicks were solely determined by stable preferences, then we would see no effect. Performativity is the reason why we see an effect. Display baises, the limited ability to process large amounts of data, and trust in the platform can be a source for performativity. The more performative the arrangement is, the stronger the effect. We use a_0 to refer to the reference arrangement of results on Google search. For a given user query q we define the potential outcome of a click event under the arrangement a as $C_q(a)$. The variable C takes on categorical values, indexing the elements on the page. Let $\{c_1, ...c_K\}$ denote the top K general search results indexed in the order they appear under a_0.

Definition 2 (Performativity gap)**.** Given a counterfactual arrangement a', we define the *performativity gap* at position i with respect to a population of queries $\mathcal{Q}$ as

$$\delta^i(a') = \mathrm{E}\left[1\{C_q(a') = c_i\}\right] - \mathrm{E}\left[1\{C_q(a_0) = c_i\}\right],$$

where expectations are taken over queries $q \in \mathcal{Q}$ and the randomness in the potential outcome.

The performativity gap quantifies how much the click through rate of search item c_i changed, in expectation across queries $\mathcal{Q}$, had the platform deployed arrangement a' instead of arrangement a_0. The following result generalizes Theorem 8 in Hardt et al. [3]:

Theorem 1 (Lowerbound on performative power)**.** *Let* PP *be the performative power of a search platform defined with respect to a set of arrangements $\mathcal{A}$, a population of search queries $\mathcal{Q}$ performed on the platform, and the outcome variable $z_a(q) = 1\{C_q(a) = c_1\}$. Then, performative power is lower bounded by the performativity gap as* $\mathrm{PP} \geq \sup_{a \in \mathcal{A}} \delta^1(a)$.

Note that the instantiation of performative power in Theorem 1 to which we relate the performativity gap measures a platform's ability to steer *outgoing* traffic from its online search website. We will discuss how to relate this notion to a broader discussion of the power of online search in vertically integrated markets in Section 6.

Algorithmic distortion. Often it can be useful to express the performativity gap relative to the base click rate. Thus, we define the *algorithmic distortion factor* as the smallest factor $\beta > 0$ such that

$$\delta^i(a') \leq \beta\, \mathrm{E}\left[1\{C_q(a_0) = c_i\}\right]. \tag{2}$$

This quantity serves as a way to denote the fraction of clicks taken away from content item c_i as a result of the update $a_0 \to a'$. Also, as we will see, it helps to express performative power relative to a base click through rate which offers a more interpretable quantity for investigators.

3.2 Estimating the performativity gap using an RCT

To estimate the performativity gap for the different arrangements, we rely on a randomized controlled trial (RCT), the gold standard methodology to estimate causal effects [35–37]. As we can not observe a search query simultaneously exposed to different arrangements, the idea of an RCT is to randomly select, for each query $q \in \mathcal{Q}$, the arrangement they are exposed to. We write $Q_a \subset \mathcal{Q}$ for the subset of queries that are exposed to treatment $A = a$. We also refer to these subsets as treatment groups. Comparing the click events across groups allows us to obtain an estimate of the performativity gap as

$$\bar{\delta}^i(a') = \mathrm{CTR}^i(a') - \mathrm{CTR}^i(a_0) \quad \text{with} \quad \mathrm{CTR}^i(a) = \frac{1}{|Q_a|} \sum_{q \in Q_a} 1\{C_q(a) = c_i\}.$$

For $\bar{\delta}^i(a')$ to provide an unbiased estimate of $\delta^i(a')$, we rely on an application of the stable unit treatment value assumption (SUTVA) [38], referred to as isolation assumption by Bottou et al. [39]:

Assumption 1 (Independence across queries)**.** User behavior in response to query q is not affected by the treatment status of other queries, i.e., for all $q \in Q$ we have $C_q(A_q) \perp\!\!\!\perp A_{q'}\ \forall q' \neq q$ where A_q denotes the random variable assigning query q to a treatment group.

This assumption justifies why we can interleave the measurement of different interventions. It requires that the intervention performed on one query does not change individuals' browsing behavior in response to subsequent queries. Crucially, this can only be satisfied, if individual interventions under investigation do not impede user experience in a lasting manner. In the following section we discuss steps we take in our experimental design towards justifying Assumption 1.

More broadly, the key advantage of using an experimental approach to measure the performativity gap is that, while the mechanism mapping user queries to clicks can be arbitrarily complex, this complexity does not affect the experiment. Aspects such as users' preference for clicking links on the left side of the screen [40], the effect of visually appealing elements [41], users' trust in the platform [16], or the relevance gap between search results will naturally enter our measurement.

4 Powermeter: Experiment in the wild

We designed an online experiment to measure the performative power of two popular online search platforms operated by Google and Bing. The experiment is built around a Chrome browser extension that modifies the arrangement of search result pages and records user click statistics in a privacy-preserving fashion. The extension allows us to observe an organic set of user queries and click outcomes under different arrangements without having control over the platforms' algorithm.

4.1 Browser extension

Browser extensions can add functionalities to the web-browser and change how a website is displayed to the user. Powermeter makes use of these functionalities to emulate algorithmic updates by implementing different counterfactual arrangements on Google search and Bing search. We emphasize that Powermeter only hides or reorders, but never modifies or adds any content on the search page.

Technical details. Once activated, the extension triggers the experiment whenever the user enters a search query on either Google search or Bing search. This can be identified by monitoring the url string of the current tab. Before search results are loaded the extension immediately hides the content of the website, inspects the html document, randomly assigns the user to one of the experimental groups and then implements the respective changes to the website before making the page visible. The implementation of the counterfactuals is done by identifying the relevant items to hide or swap by their html class names or ids. We also add custom tags and event listeners to the identified elements that we can fall back on at a later stage. The entire setup of the experiment usually takes around 40 milliseconds. This delay is far below what was found to be noticeable to users [42, 43]. Hiding the html body of the website with the first possible Chrome event is crucial to avoid glitches in case of bad internet connection and make sure the control arrangement is not revealed to the participant. To ensure internal validity of our experiment, we also have to ensure a participant is never reassigned to a new experimental group when reloading a page, navigating between tabs or repeatedly entering the same search query. This is done by storing a hash of user ID and search query together with the assigned experimental group in the browser cache.

 https://doi.org/10.52202/079017-1891

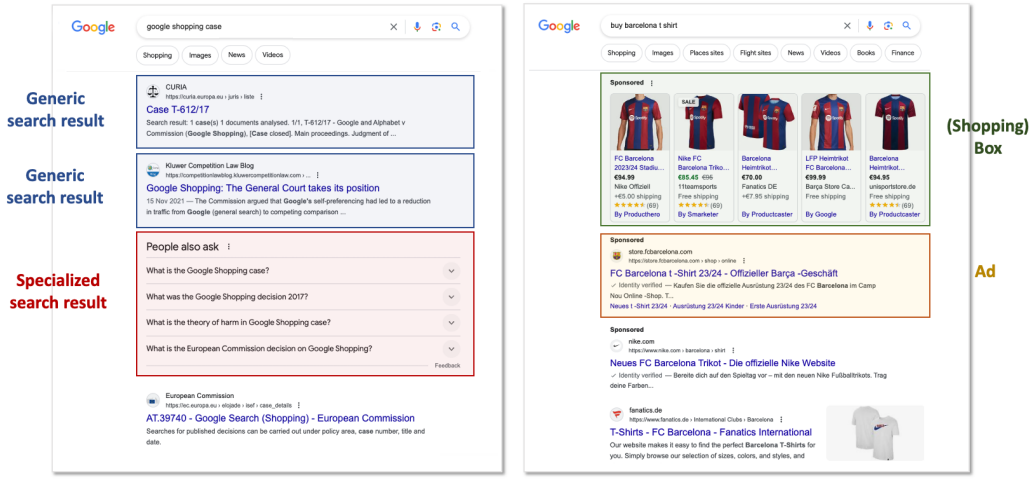


Figure 2: Illustration of different elements on the Google search website.

Table 1: Counterfactual content arrangements implemented by Powermeter as part of the RCT.

	Arrangement	Description
a_0	control	Search results are displayed without any modification.
a_1	swap 1-2	The position of the first and the second generic search result are swapped.
a_2	swap 1-3	The position of the first and the third generic search result are swapped.
a_3	swap 2-3	The position of the second and the third generic search result are swapped.
a_4	hide Ads/Box	Top Ads and Shopping boxes are hidden.
a_5	hide + swap	Combines the latter modification (a_4) with swap 1-2 (a_1).
a_6	hide Box	The shopping boxes are hidden.

Backend and data collection. Every participant is assigned a unique random number that serves as anonymous user ID upon installation of the extension. This user ID persists throughout the experiment. Every time a click event on an element on the search page is registered, the click data is aggregated into a json object and sent to a database server hosted locally at out institution via a post request using the encrypted https protocol. This concerns information about the index of the clicked search result, the click element type, the page index, and the experimental group. In addition, statistics about the website such as the number of search results, the presence of ads and boxes, the number of candidate results, and the position of specialized search results are extracted from the website are recorded. The database server is built using the Microsoft .Net core framework and deployed within a docker container. The database access is rate limited and the Get endpoint of the database is key protected. We use a SQlite database that is mapped to persistent memory.

Privacy considerations. The information that is stored with every click does not contain any personally identifiable information. While we record the position of the clicked element on the search page, we never store search queries or any information about the websites visited by the user. This is an intentional choice to preserve user privacy, and to demonstrate that valuable insights can be gathered without privacy invasive data collection. The experiment went through an internal approval procedure and the privacy policy can be found on our website.[3]

4.2 Experimental groups

We implement six different counterfactual arrangements, summarized in Table 1, each defining a treatment group. We refer to Figure 2 for the terminology used to refer to individual elements on the general search page. It equivalently applies to both, Google search and Bing. Arrangements $a_1 - a_6$ are designed to emulate conservative variants of the conduct a^* of interest, to inform a plausible lower bound on performative power. The first three arrangements a_1–a_3 concern the reordering of organic search results, leaving the other elements on the website untouched. The arrangements

[3]The project website can be found at: https://powermeter.is.tue.mpg.de/

a_4 and a_6 perform modifications not directly concerning organic search results: Arrangement a_6 hides a specific element, called the Shopping box, appearing either in the right side panel or on top of the search results page. Arrangement a_4 hides the box together with all the Ads. Finally, Arrangement a_5 combines the latter change with a change in search result order. For Bing we only implemented the counterfactual a_1 to ensure statistical power despite data scarcity. A practical reason not to implement larger modifications is also users' sensitivity to the resulting deterioration of quality towards ensuring Assumption 1. The Bing experiment of the European Commission's investigation had to be discontinued after one week for that exact reason.[4] We made sure to avoid a similar failure point. Based on user feedback collected during an initial test round there was no indication that the modifications were even noticeable to users.

4.3 Onboarding

Participants were provided the link to the project website as an entry point. The website contains information about the experiment, the purpose of the study, an onboarding video, as well as the privacy policy of the extension. The extension itself is distributed through the official Chrome webstore and there is a button directly navigating the user to the item in the store. We did not list the extension publicly to ensure participants are informed about the purpose of the study, and protect the integrity of our data. The installation follows the standard procedure of adding a browser extension to Chrome. The user has to give consent to access Google and Bing websites, as well as to use the storage API. The extension remains active until participants remove it from their web-browser, or until the experiment is stopped. The study participants are trusted individuals of different age groups and backgrounds, recruited by reaching out personally or via email. We provide demographic statistics over our pool of participants in Figure 9 in the appendix.

Data preprocessing. For each participant we ignored the clicks collected during the first four days after onboarding, as suggested by Keusch et al. [44] in order to avoid potential confounding due to participation awareness. We also removed clicks where the search elements could not be identified reliably for implementing the RCT to avoid selection bias towards the control group.

5 Empirical results

Using our Powermeter browser extension we collected click data from 85 participants over the course of 5 months, from September 2023 until January 2024. This resulted in $56,971$ click events, and a total of $45,625$ clicks after preprocessing. Out of the clicks 98.9% were registered on Google, and 1.1% on Bing. Figure 8 in the appendix visualizes some aggregate statistics over the clicks collected. We will consider several subsets of these events for which we measure the performativity gap and algorithmic distortion. In the following we discuss the main insights from the collected data. For all plots, we provide bootstrap confidence bounds over 200 resamples.

5.1 Reordering search results

We first inspect the three counterfactual arrangements a_1, a_2, a_3 concerning reranking. Recall that c_i indexes search results in the order in which they appears under a_0. In Figure 3 we visualize the event probabilities $C = c_i$ for each search result $i = 1, 2, ..., 6$ under the control group (blue bars) and compare it to the respective probabilities under the three counterfactuals (orange bars). The figures on the left show the results across Google search queries. Here the counterfactuals correspond to swapping the position of the first two results (left), the first and the third (middle) and the second and the third (right). The right figure shows the results evaluated on Bing search queries when the first two results are swapped. The lower figure visualizes the corresponding performativity gap $\delta^i(a)$, corresponding to the change in clicks to item i caused by the respective arrangement change.

We observe a consistently large effect of arrangement on clicks. Being down ranked by one position on Google search decreases average click through rate of c_1 from 43% to 24%, resulting in $\delta^1(a_1) = -0.19$ and an algorithmic distortion of $\beta = 0.44$. Being down-ranked by 2 positions results in $\delta^1(a_2) = -0.27$ and an average loss of more than 50% of traffic. Note that a similar effect size has been reported in the case decision for the UK market for a two position shift [45, para 460], indicating

[4] See Case T-612/17, *supra* 2, para 399

 https://doi.org/10.52202/079017-1891

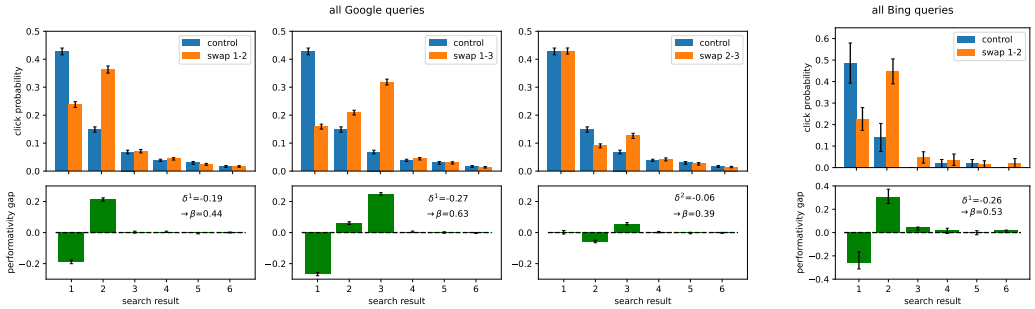


Figure 3: Click through rate and performativity gap for general search results c_1 to c_6 under the counterfactual arrangements a_1, a_2, a_3 for Google and the counterfactual arrangement a_1 for Bing, compared to the control arrangement a_0 (in blue).

that the estimate is robust across different user populations. Finally, by down-ranking the second content element by one position we still observe a significant traffic reduction, corresponding to an algorithmic distortion of $\beta = 0.39$.

An interesting observation is also that for every counterfactual arrangement, the element shown first ends up getting most clicks on average. Implying that all the rankings are close to performatively stable [10] with respect to the non-personalized reranking strategies considered. However, there are several indications of Google's ranking a_0 reflecting relevance of search results better then the other arrangements. Namely, c_1 gets more clicks when displayed first, compared to other results displayed in the same position (c_2 corresponds to first result under a_2 or and c_3 to the first result under a_3). A second indicator is that c_2 benefits from arrangement a_3; under the hypothesis that users consider search results in order this indicates that content item c_3 absorbs less clicks than c_1.

For Bing the position effect seems to be even more pronounced, although confidence intervals are significantly larger in this case. We conjecture that this could be due to the average number of specialized search results and Ads present on the search page being larger on Bing. This results in a larger spacial separation of search results and potentially larger display effects. Statistics about the type of elements present on the the search pages are reported in Figure 8 in the Appendix.

5.2 Indirect effect of visually appealing elements

Next, we consider the counterfactuals a_4 and a_6 that leave generic search results untouched and hide certain elements on the website. We first inspect the number of clicks these elements absorb if present on the page. We focus on Google search. In Figure 4 we compare the fraction of clicks going to generic search results, Ads, and Boxes for a_0, a_4, a_6. We plot the statistics across the aggregate queries (left panel), the subset of queries where Ads are present on the page under a_0 (middle panel) and the subset of queries where the box is present on the page under a_0 (right panel). We find the addition of boxes absorbs 22.4% of the clicks on average across queries where it is present under a_0 and these clicks are mostly taken away form the generic search results. Similarly, Ads absorb close to 30% of the clicks on average for queries where they are present. However, considering the overall number of clicks the effect is smaller since a large fraction of queries contains neither Ads nor Boxes.

Combined conduct. We now consider the combined conduct of adding the box and down-ranking an element. We constrain our focus on queries where the Shopping box is present under a_0, either in the center column or in the right sidebar. These are 3.2% of all the events. We show the corresponding click probabilities for the three first search results in Figure 5. In both figures the blue bars correspond to the control group a_0, and the red bars correspond to a_1. For these groups boxes are present on the page. In the left figure we investigate the effect of hiding boxes only and the orange bars correspond to arrangement a_4. In the right figure we investigate the effect of also hiding Ads, here the orange bars correspond to a_6. We find that when adding Boxes, all three content items loose a significant fraction of clicks, whereas Ads mostly take away from c_1. In the right figure we additionally show a_5 using the hatched bars (i.e., down-ranking the first element by one position if box is hidden). For c_1 the combined effect of adding Ads and Boxes on the click through rate is almost as large as the effect

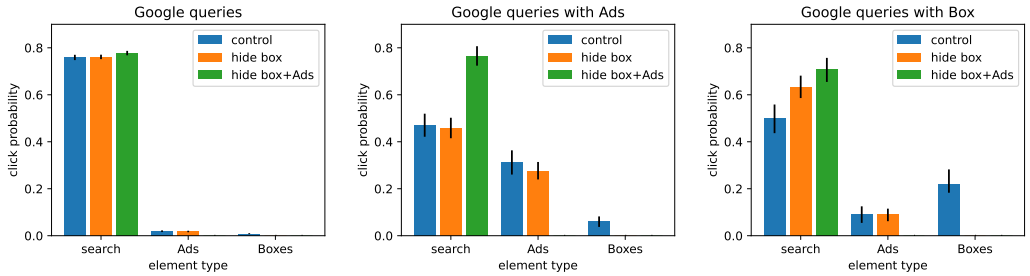
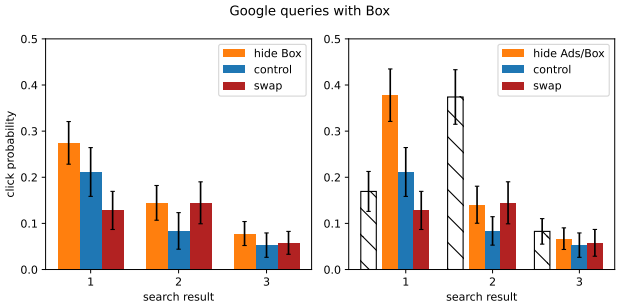


Figure 4: Effect of arrangement on the click distribution across different element types (generic search results, Ads, boxes), visualized for three different subsets of Google queries.



Conduct	Distortion of c_1
add Box	$\beta = 0.23$
add Ads/Box	$\beta = 0.44$
swap+add Box	$\beta = 0.53$
swap+add Ads/Box	$\beta = 0.66$

Figure 5: Effect of hiding box and swapping elements on the click probability of generic search results. Statistics are evaluated for the subset of queries for which box is naturally present. The hashed bar shows the click probability under a_5 when top content is hidden and the first two elements are swapped. The right table reports the empirical measure of algorithmic distortion for different conducts, extracted form the results in the left figure.

of down ranking the same item by one position. What we can see consistently is that combining the conduct of adding visually appealing elements on top of the page, and down-ranking a content item, has a combined effect that is larger than the effect of the two individual modifications alone. For element c_1 the measured distortion is reported in the table. The combined effect leads to a reduction of 25% in clicks and an algorithmic distortion of 66% when considering the effect of Ads/Boxes, and 53% when only considering Boxes (comparing orange and red bar). We believe this to be the first time that quantitative insights into this combined conduct are made public.

5.3 Factors that impact performative power

We perform additional investigations into what factors have a reinforcing effect on the performativity of ranking. To this end, we inspect different subsets of Google queries and measure the performativity gap as well as algorithmic distortion for c_1 under the counterfactual a_1. First, we split the data across two different axes depending on whether Ads or boxes are present, and whether Specialized Search results (SSR) are present in between the first two generic search results. The respective comparisons are visualized in the left and middle panel of Figure 6. We observe that the performativity gap in the presence of Ads and boxes is smaller, and about the same whether special search results are present in between search results. However, if we plot distortion we get a different picture, since the base click probability of content item c_1 across different splits is different for the three cases. We find that while Ads and Boxes have little effect on distortion, the presence of a specialized search result in between the swapped results tends to increase the effect of down-ranking c_1.

For the second investigation we group the queries by the number of candidate search results available on Google. This number was extracted from the website where it appeared as a string on top of the page in the form `‘About 323’000’000 results (0.65 second)’`. The right panel of Figure 6 shows algorithmic distortion for each percentile of the data. We see a clear trend that the performativity gap increases with the number of candidate results. We suspect this to be connected to the smaller relevance gap across results for queries with more potential results, leading to a higher influence of the arrangement on clicks. However, note that findings in this subsection are no causal claims.

https://doi.org/10.52202/079017-1891

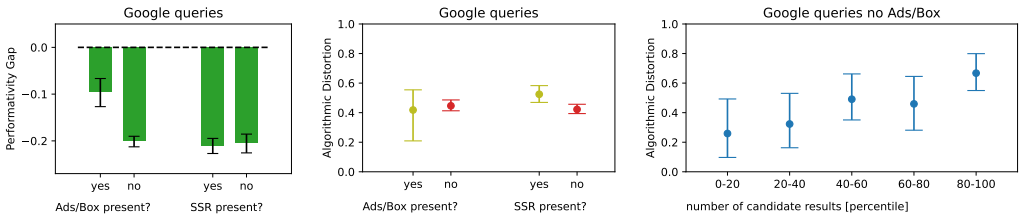


Figure 6: Performativity gap and algorithmic distortion for content item c_1 under the counterfactual arrangement a_1 measure across different subsets of Google search queries.

6 Discussion

We presented a flexible experimental design, based on a browser extension, to investigate the effect of search algorithms on user clicks. The browser extension performs interventions at the level of display to emulate algorithmic updates, without access to the platform algorithm. We implement different counterfactuals relevant for the Google Shopping antitrust investigation, and provide quantitative insights into the causal effect on clicks. Theorem 1 formally relates our quantitative findings to an instantiation of performative power, measuring the platform's ability to steer *outgoing* traffic from search.

In a final step, we describe how our findings could fit with a broader anti-trust investigation potentially concerning effects spanning different markets. Take the Google Shopping case as an example. It is concerned with the ability of Google search to distort *incoming* traffic to a business operating in the market of comparison shopping services by changing where it appears on Google search relative to its competitors. Establishing the causal link between arrangements on search and their effect on incoming traffic to a third party website composes into two steps: a) establish Google's ability to steer outgoing traffic, b) quantify how much of the incoming traffic is mediated by Google search. The first step is a notion of power that experiments like ours operationalize, the second is a number that can readily be obtained from web traffic data. The overall performative power will be the product of the two factors. For putting this together, let's work though the following thought experiment:

> *Suppose,* 80% *of the referrals to the competitor's website come from Google Search.*[5] *Further, assume that* 70% *of the referrals from Google happen while the service is ranked among the top two generic search results. Our estimates suggest that distortion of traffic at the first position can be as large as* 66% *for small arrangement changes. Assuming for the second position the effect is* 20% *smaller, giving a conservative average effect size of* $\beta = 0.8 \cdot 0.66$*. Multiplying the effect size by the fraction of incoming clicks it concerns, we get* $0.8 \cdot 0.7 \cdot \beta \approx 30\%$*. This is the fraction of traffic to the site Google can redirect.*

Turning this number into a conservative lower bound on performative power, it offers an interpretable measure of power for an investigator to judge whether the algorithmic lever of self-preferencing through arrangement changes should be a concern for competition in the down-stream market or not. We can use the same logic to compare search engines, and assess the effectiveness of remedies.

More broadly, we hope our work can serve as a blue print for how performative power can be used to integrate experiments with future digital market investigations, and how tools from computer science, causality, and performative prediction can inform ongoing legal debates related to the power of digital platforms. Our work is situated within a growing scholarship [c.f., 46–48] that takes advantage of the accessibility of digital markets for monitoring and regulating digital platforms. Beyond data, we demonstrate how experimental methods can offer an additional tool for power assessments.

From the perspective of computer science our work offers measurements of performativity in the context of online search, contributing quantitative and empirical support to the study of performativity on digital platforms. As its name suggests a search engine is performative, it acts as an *engine* steering consumption through its ranking algorithm, rather than a *camera* merely picturing candidate results—we borrow this analogy from MacKenzie [49].

[5] In 2022 up to 82% of incoming traffic to comparison shopping services in the European Economic Area was mediated by Google search, as reported in the Google Shopping case decision (Section 7.2.4.1, Table 24).

Acknowledgements

We are particularly grateful to everyone who installed the extension and participated in our study; without you such a project would not have been feasible. Further, the authors would like to thank Jonathan Williams for the design of the project website, Alejandro Posada for producing the on-boarding video and helping with the logo, Telintor Ntounis for assistance in setting up the server infrastructure, and Ana-Andreea Stoica, André Cruz and Jiduan Wu for feedback on the manuscript. We would also like to thank two anonymous reviewers at NeurIPS who provided valuable feedback related to the presentation and framing of our work. Celestine Mendler-Dünner acknowledges the financial support of the Hector Foundation. Opinions expressed in the paper do not necessarily reflect the opinions of the AGCM.

References

[1] Stigler Committee. Final report: Stigler committee on digital platforms, September 2019.

[2] Jacques Crémer, Yves-Alexandre de Montjoye, and Heike Schweitzer. *Competition Policy for the digital era: Final report*. Publications Office of the European Union, 2019.

[3] Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative Power. In *Advances in Neural Information Processing Systems*, pages 22969–22981, 2022.

[4] Nathan Newman. Search, antitrust, and the economics of the control of user data. *Yale Journal on Regulation*, 31:401, 2014.

[5] Jason Furman. Unlocking digital competition, report of the digital competition expert panel, 2019. `https://doi.org/10.17639/wjcs-jc14`.

[6] L Cabral, J Haucap, G Parker, G Petropoulos, T Valletti, and M Van Alstyne. The EU digital markets act. Technical Report KJ-02-21-116-EN-N (online), 2021.

[7] Chad Syverson. Macroeconomics and market power: Context, implications, and open questions. *Journal of Economic Perspectives*, 33(3):23–43, August 2019.

[8] OECD. The evolving concept of market power in the digital economy, 2022. OECD Competition Policy Roundtable Background Note.

[9] CMA. Online search: Consumer and firm behavior - a review of the existing literature, 2017.

[10] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609, 2020.

[11] Moritz Hardt and Celestine Mendler-Dünner. Performative prediction: Past and future. *ArXiv preprint arXiv:2310.16608*, 2023.

[12] R.H. Thaler and C.R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008.

[13] Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data. In *International Conference on World Wide Web*, page 1011–1018, 2010.

[14] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, page 154–161, 2005.

[15] Zhiwei Guan and Edward Cutrell. An eye tracking study of the effect of target rank on web search. In *SIGCHI Conference on Human Factors in Computing Systems*, page 417–420. Association for Computing Machinery, 2007.

[16] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. In Google We Trust: Users' Decisions on Rank, Position, and Relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823, 2007.

 https://doi.org/10.52202/079017-1891

[17] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *International Conference on Web Search and Data Mining*, page 87–94, 2008.

[18] Aleksandr Chuklin and Ilya Markov and. *Click Models for Web Search*. Springer Cham, 2015.

[19] Guipeng Xv, Si Chen, Chen Lin, Wanxian Guan, Xingyuan Bu, Xubin Li, Hongbo Deng, Jian Xu, and Bo Zheng. Visual encoding and debiasing for ctr prediction. In *ACM International Conference on Information & Knowledge Management*, page 4615–4619, 2022.

[20] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Trans. Inf. Syst.*, 41(3), 2023.

[21] Aman Agarwal, Kenta Takatsu, Ivan Zaitsev, and Thorsten Joachims. A general framework for counterfactual learning-to-rank. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, page 5–14, 2019.

[22] Zhen Qin, Suming J. Chen, Donald Metzler, Yongwoo Noh, Jingzheng Qin, and Xuanhui Wang. Attribute-based propensity for unbiased learning in recommender systems: Algorithm and case studies. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2359–2367, 2020.

[23] Mouxiang Chen, Chenghao Liu, Zemin Liu, and Jianling Sun. Scalar is not enough: Vectorization-based unbiased learning to rank. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[24] Yunan Zhang, Le Yan, Zhen Qin, Honglei Zhuang, Jiaming Shen, Xuanhui Wang, Michael Bendersky, and Marc Najork. Towards disentangling relevance and bias in unbiased learning to rank. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 5618–5627, 2023.

[25] Gagan Aggarwal, Jon Feldman, and S. Muthukrishnan. Bidding to the top: Vcg and equilibria of position-based auctions. In *International Conference on Approximation and Online Algorithms*, page 15–28, 2006.

[26] Susan Athey and Glenn Ellison. Position Auctions with Consumer Search. *The Quarterly Journal of Economics*, 126(3):1213–1270, 2011.

[27] Raluca M Ursu. The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, 37(4):530–552, 2018.

[28] Sridhar Narayanan and Kirthi Kalyanam. Position effects in search advertising and their moderators: A regression discontinuity approach. *Marketing Science*, 34(3):388–407, 2015.

[29] Ashish Agarwal, Kartik Hosanagar, and Michael D. Smith. Location, location, location: An analysis of profitability of position in online advertising markets. *Journal of Marketing Research*, 48(6):1057–1073, 2011.

[30] Filip Radlinski and Thorsten Joachims. Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. In *National Conference on Artificial Intelligence*, page 1406–1412, 2006.

[31] Zhichong Fang, Aman Agarwal, and Thorsten Joachims. Intervention harvesting for context-dependent examination-bias estimation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 825–834, 2019.

[32] Ronald E. Robertson, David Lazer, and Christo Wilson. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *International Web Conference*, 2018.

[33] Jeffrey Gleason, Desheng Hu, Ronald E. Robertson, and Christo Wilson. Google the gatekeeper: How search components affect clicks and attention. *AAAI Conference on Web and Social Media*, 17(1):245–256, 2023.

[34] NSF. The National Internet Observatory, 2021. https://nationalinternetobservatory.org.

[35] R.A. Fisher. *The design of experiments*. 1935.

[36] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1): 140–181, 2009.

[37] G.W. Imbens and D.B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[38] Donald B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[39] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

[40] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. Visual positions of links and clicks on wikipedia. In *International Conference Companion on World Wide Web*, page 27–28, 2016.

[41] Erik Fubel, Niclas Michael Groll, Patrick Gundlach, Qiwei Han, and Maximilian Kaiser. Beyond rankings: Exploring the impact of serp features on organic click-through rates. *Arxiv preprint arxiv:2306.01785*, 2023.

[42] Jake D Brutlag, Hilary Hutchinson, and Maria Stone. User preference and search engine latency. *JSM Proceedings, Qualtiy and Productivity Research Section*, 2008.

[43] Ioannis Arapakis, Xiao Bai, and B. Barla Cambazoglu. Impact of response latency on user behavior in web search. In *International ACM SIGIR Conference on Research & Development in Information Retrieval*, page 103–112, 2014.

[44] Florian Keusch, Ruben Bach, and Alexandru Cernat. Reactivity in measuring sensitive online behavior. *Internet Research*, 2022. ISSN 1066-2243.

[45] European Commission. Google shopping commission decision, 2017.

[46] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends in Human–Computer Interaction*, 14(4):272–344, 2021.

[47] Jack Bandy. Problematic machine behavior: A systematic literature review of algorithm audits. *ACM Human-Computer Interaction*, 5, 2021.

[48] Abhisek Dash, Abhijnan Chakraborty, Saptarshi Ghosh, Animesh Mukherjee, Jens Frankenreiter, Stefan Bechtold, and Krishna P. Gummadi. Antitrust, amazon, and algorithmic auditing. *Arxiv preprint arxiv:2403.18623*, 2024.

[49] Donald MacKenzie. *An engine, not a camera: How financial models shape markets*. MIT Press, 2008.

[50] Elizabeth A. Stuart, Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011.

 https://doi.org/10.52202/079017-1891

A Limitations and Broader Impact

We develop a flexible methodology, to provide insights into the power of digital platforms. We hope our framework can support future digital market investigations, complement and address some of the limitations of current antitrust tools. This could help make firms accountable for steering behavior on digital platforms, and support cases of anti-trust, consumer protection and abuse of dominance in digital markets. Compared to traditional tools, our approach offers a more straights forward way to integrate experimental insights with regulatory questions and requires fewer assumptions on the market dynamics. Furthermore, by providing a quantitative measurement, the methodology also allows to compare platforms and assess the effectiveness of potential remedies. That being said, instantiating the definition in the right way is still at the discretion of the investigator and requires substantial domain knowledge. Further, fitting the approach within existing legal frameworks is an open question, we hope to further make this concrete in future work.

On a technical side, our design ensures that for any given users, the observed clicks follow a natural distribution. However, our participants form a convenience sample of online search users. Depending on the target of the investigation this might not be sufficient to argue for external validity of the quantitative insights. We provide statistics about the users to support such a judgement in Figure 9. Further, we propose to link this data with collected clicks which can potentially help to adjust estimates using propensity score reweighting Stuart et al. [50]. More rigorously arguing about the external validity of our results in specific contexts is left for future work. However, we expect the qualitative take-aways of our work to generalize beyond our study, and hope they can inform future modeling and problem statements around performativity in online search.

Lastly, we want to reiterate that our results for Bing should be taken with caution. While we designed our experiment to take most out of the available data, it is still a small sample of ~ 600 search queries. Nevertheless, we decided to share the results with the reader.

B Additional technical results

B.1 Causal model

To support the arguments about composability of performative power in the discussion, consider the simplified diagram of how a user navigates to a website illustrated in Figure 7. The random variable U denotes a user request and T indicates which website is being visited in response. The user either navigates to the website via Google search (gray box), or they navigate to the website on some alternative way. This can be by entering the url directly, or using a different search service. Naturally, the arrangement A only impacts the outcome T if the user uses Google search, otherwise A does not have any influence on the outcome. For a search query, the user query leads to an arrangement of content shown on the website, and the arrangement mediates the click.

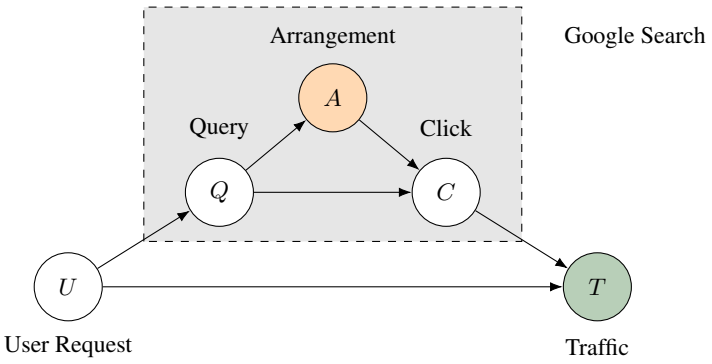


Figure 7: Causal graph of online search users. A web request leads to the visit of a website, partially mediated by Google search.

B.2 Proof of Theorem 1

We instantiate performative power with respect to a set of action $\mathcal{A}$, a population of search queries $\mathcal{Q}$, and the outcome variable $z_a(q) = 1[C_q(a) = c_i]$. Since the outcome is a scalar the L_1 norm reduces to an absolute value and we get

$$\mathrm{PP} = \sup_{a \in \mathcal{A}} \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathrm{E} \left| z_{a_0}(q) - z_a(q) \right| \tag{3}$$

Using the definition of the performativity gap and the definition of the L_1 norm the proof is a direct consequence of

$$\begin{aligned}
\mathrm{PP} &= \sup_{a \in \mathcal{A}} \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathrm{E} \left| 1[C_q(a_0) = c_i] - 1[C_q(a) = c_i] \right| && (4) \\
&\geq \sup_{a \in \mathcal{A}} \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \left| \Pr[C_q(a_0) = c_i] - \Pr[C_q(a) = c_i] \right| && (5) \\
&\geq \sup_{a \in \mathcal{A}} \left| \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \Pr[C_q(a_0) = c_i] - \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \Pr[C_q(a) = c_i] \right| && (6) \\
&= \sup_{a \in \mathcal{A}} \delta^i(a) && (7)
\end{aligned}$$

where the performativity gap is defined with respect to the set of queries $\mathcal{Q}$. In words, Theorem 1 formalizes the idea that the average effect of an arrangement change on an individual query $q \in \mathcal{Q}$ can be bounded by the average aggregate statistics across queries.

C Additional details on the study

C.1 Aggregate click statistics

In Figure 8 we show aggregate statistics over the clicks collected. In Figure 9 we provide aggregate statistics over the user base. The latter information was collected through the onboarding form.

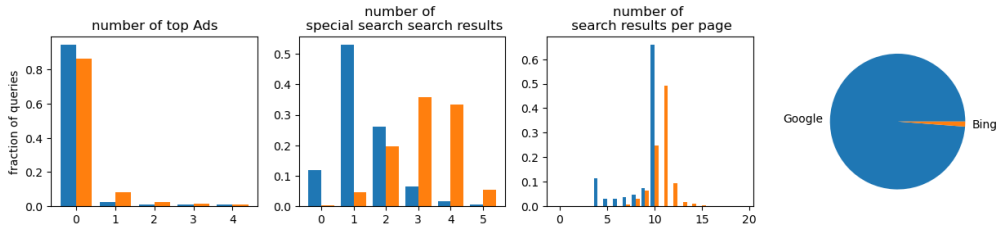


Figure 8: Aggregate statistics over clicks and search result pages collected during our experiment. The blue bars show the statistics for Google and the orange bars show the statistics for Bing. Numbers are aggregated based on original search pages, before any modifications are performed.

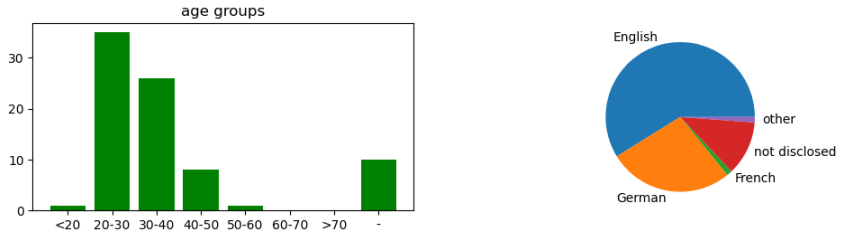


Figure 9: Aggregate user statistics collected from the 85 participants with the onboarding form. Age distirbution (left) and language in which they consume online search (right). That's all the data we have about the demographics of our participants.

https://doi.org/10.52202/079017-1891

C.2 Project website

The project website provides the entry point to the experiment. Users where provided with details about the experiment, instructions for how to participate, and information about data usage. See Figure 10 for a screenshot.

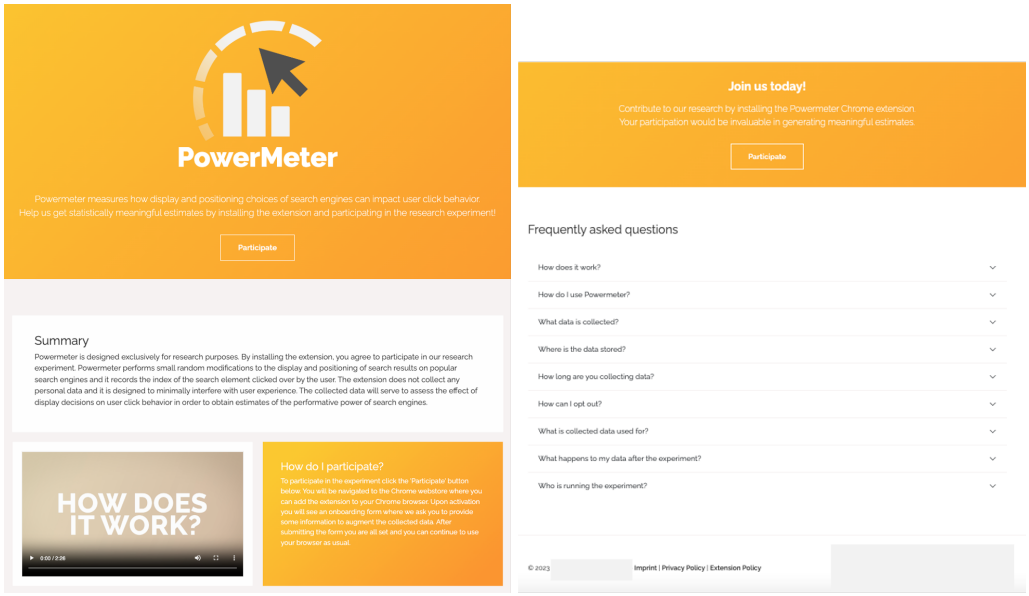


Figure 10: Project website. URL and institution names are removed for the sake of anonymity.

C.3 Onboarding form

Upon installation of the extension the user is navigated to the onboarding form, as illustrated in Figure 11. Providing the information is not mandatory and answers are binned to only provide coarse grained information and no personally identifiable information. The main purpose of the information serves debugging different languages and website versions.

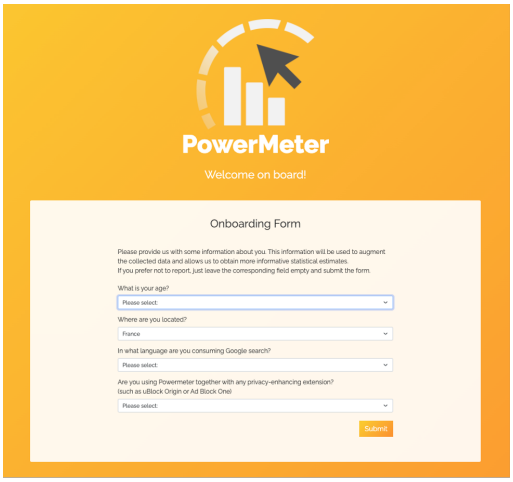


Figure 11: Onboarding form.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-1891

Justification: Assumptions are stated in Section 3, and the proofs can be fund in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The experimental protocol is described in detail, including all the technical steps we took to ensure validity of the experimental design. As such it can be reproduced as a general framework.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Our data was collected under the "need to know" principle, and not intended for publication. The data is also very specific to the question under investigation, providing limited additional insights. Regarding code. The extension is publicly available in the Chrome store, and the code can fully be inspected. The link can be found on the project website.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: We do not report results on model training

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide error bars in all figures, using bootstrapped sampling, as described.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

https://doi.org/10.52202/079017-1891

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Experimental results concern data evaluation, no compute-intensive model training of inference is performed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To clarify the question on privacy. The project is purposefully designed not to collect personal data. The privacy policy of our extension went through internal legal review. It is linked on our website.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section A for a broader impact statement.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release and model or dataset.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

https://doi.org/10.52202/079017-1891

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provide a screenshot of the website that provides the entry point to the experiment. It contains information about the experiment and participation instructions. In addition, the extension policy describes the experiment in full detail, see project website.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: IRB approval was waved for this type of study.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1891