How important are user-generated data for search result quality?[1]

Tobias J. Klein,[2] Madina Kurmangaliyeva,[3] Jens Prüfer,[4] and Patricia Prüfer[5]

14 January 2025

Do search engines produce better results because their algorithm is better, or because they can access more data from past searches? We document that the algorithm of a small search engine can produce non-personalized results that are of similar quality to the dominant firm's (Google) for certain types of search queries. Overall differences in the quality of search results are explained by searches for *rare* queries, which constitute 74% of the traffic in our data. We conduct an experiment, in which we keep the algorithm of a small search engine fixed and only vary the amount of data it uses as input. Our results show that giving small search engines access to more data about rare queries improves their quality. This suggests that mandatory data sharing by large search engines is a necessary, yet probably not a sufficient condition to increase competition in the search market.

Keywords: Search engines, user-generated data, search result quality, experiment, rare search queries, data-driven markets.

[1] We are extremely grateful to Marc Al-Hamez and Josep Pujol from Cliqz, who allowed us to run the experiment and answered many technical questions about the way search engines work. We are also grateful for the many useful comments we received when we presented the results of this paper at various occasions, including at the universities of Tilburg, Passau, Stockholm, Paris-Dauphine, East Anglia, Toulouse (15th Digital Economics Conference), and at the virtual Economics of Platforms Seminar by TSE Digital Center, the 2022 SIOE conference in Toronto, and the TILEC Annual Conference. We would also like to thank Gregory Crawford for his discussion of our paper, Seyit Höcük and Pradeep Kumar from Centerdata for helping us to collect the data and preparing them for analysis, and Marco Alberti and Magdalena Kuyterink for excellent research assistance. We acknowledge funding from the German Ministry of Finance (grant no: fe 11/19).
[2] Tilburg University; t.j.klein@tilburguniversity.edu.
[3] Department of History, Trinity College Dublin; madinakur@gmail.com.
[4] Department of Economics and TILEC, Tilburg University; and School of Economics and CCP, University of East Anglia; j.prufer@tilburguniversity.edu.
[5] Centerdata; and Tilburg School of Economics and Management, Tilburg University; patricia.prufer@centerdata.nl.

1. Introduction

Search engines are used by billions of users every day. They are an important part of the infrastructure for many other industries and are of very high economic, political, and social importance (Ducci, 2020). Google is the dominant provider of online search, with a market share above 90% in many national markets.[6] One potential explanation for this is that Google's *algorithm* generates search results that are better than the ones provided by its competitors. Another potential explanation is that Google has access to more relevant data and can therefore produce better search results. Every time a user performs a search on Google, they click on some of the results; this generates data that is useful to produce better search results in the future (*user information*).

Quantifying the contributions of these two explanations is highly relevant for policymaking, regulation, and firms' decision-making. If Google only had a better algorithm, then there would not be much reason for antitrust policy to intervene (Varian, 2019, Bajari et al., 2019). Other firms could develop an algorithm that performs equally well and enter the market.[7] Google would find it worthwhile to invest in the algorithm to avoid this from happening. Or firms would invest into building an even better algorithm. In any of these cases, consumers would benefit from active competition. However, if Google was dominant mainly because it had access to more data, then this would mean that there was no level playing field between Google and existing and potential competitors. It would be much more difficult for entering firms to provide search results that are of similar or better quality than Google's, even if Google would innovate very little. In such a situation, sharing user-generated data among competitors would help level the playing field between Google and (potential) competitors. This, in turn, might benefit consumers by increasing incentives for all search engines to innovate (Argenton and Prüfer, 2012; and Prüfer and Schottmüller, 2021).

[6] https://alphametic.com/global-search-engine-market-share

[7] This is one way to interpret the introduction of large language models that are integrated into search engines (notably in Bing, which has integrated a version of ChatGPT since February 2023).

Users of search engines produce data by entering a search query (a search string users submit when they use the search engine) and selecting one result from the list of search results the engine provides. *Query logs* record how many users select a given result for a given query. These data are useful input for providing search results. For two reasons, it is difficult to study empirically whether differences in search result quality are due to different algorithms or to different amounts of user information algorithms can use as input. The first reason is *access to data*. While search results are public, usage data are usually proprietary and cannot be accessed, which inhibits the ability for external review of empirical results through academic peers or public authorities (Persily and Tucker, 2020). The second reason is that it is hard to identify the *causal effect* of having access to more data related to past searches on search result quality, which is the object of interest for policy making (Lewis-Kraus, 2020). This is hard because the number of past searches for the same or similar queries is not exogenous, but likely correlated with the error term in an estimation equation for current searches. In this paper, we address both challenges.

"[A]n ideal experiment would be to fix the 'query difficulty' and exogenously provide more or less historical data" (He et al., 2017). This paper reports the results of such an experiment. We collaborated with a small search engine, Cliqz (based in Munich, Germany).[8] They provided us with non-personalized search results for a representative set of queries for German users in April 2020. Importantly, they also provided us with a measure of the popularity of the underlying queries. Moreover, Cliqz conducted an experiment on our behalf, where they kept the search algorithm fixed and varied the amount of user-generated data they used to produce search results. This allows us to conduct *within-search engine* comparisons. In particular, it allows us to look at the search results for the same queries, varying only the amount of data that is used as input. We complement the Cliqz data with non-personalized search results from Google and Bing on the same queries in the same period in the same

[8] More details about the business background of Cliqz are in Appendix A.

country. We asked external assessors to assess the quality of the search results on a Likert scale (not mentioning the origin of the results). This offers insights about *between-search engine* comparisons.

We first report the results from an experiment, in which we keep the search engine algorithm fixed and vary only the amount of data it uses as input. This produces causal evidence that more user data on *rare* (or tail) queries enables search engines to produce better quality. We find that, at the margin, more data leads to a substantial increase in search result quality for rare queries and almost no increase for popular queries. Second, we compare quality across search engines and find that the algorithm of a small search engine *can* produce non-personalized results that are of *similar* quality to Google's. We do so by comparing the quality of search results for *popular* queries. By contrast, for *rare* search terms, our results show that Google's quality depreciates only slightly relative to the one of popular queries, whereas Cliqz' quality decreases a lot. Finally, we compare average quality levels across search engines and find that our assessors evaluate Google's results above Bing's, which are evaluated above Cliqz's, producing a ranking of perceived quality that is in line with the ranking of market shares in Germany.[9] Taken together, this paper shows that the overall differences in the quality of search results are explained by searches for rare queries. Crucial for competition policy, rare queries represent 74% of the traffic in our data. Hence, they represent the critical margin for competition.

Our paper contributes to the literature in at least two ways. It is well known that there is a positive relationship between the availability of past user data on search engine quality (He et al., 2017; Schaefer and Sapi, 2023). Our paper is the first to quantify the importance of past user data for an actual entrant in the search engine industry. It shows that an entrant can produce non-personalized search results that are of similar quality as Google's for popular queries and identifies that rare queries constitute the critical challenge for competition in the search-engine industry, where rare queries constitute the majority of all queries. This finding is both of academic interest and highly policy-relevant.[10] Indeed,

[9] https://gs.statcounter.com/search-engine-market-share/all/germany/

[10] At least 30 top-level advisory reports about competition in online "platform markets" raise concerns related to markets where data serves as an input (Beaton-Wells, 2019), including highly regarded ones in the US (Scott

lawmakers are currently in the process of preparing legislation to regulate dominant firms on platform markets.[11] The EU's Digital Markets Act (DMA), which has been enforced since March 2024, prescribes that large "gatekeeper" firms must provide business users with data generated in the context of the use of their services (Art 6(10)). It even has a special clause on search engines, giving third-party providers of online search engines the right to ask gatekeepers for data on search queries that are generated by end users (Art 6(11)). Our results support this provision because they show that small search engines could most likely improve their quality for rare queries—and hence increase the competitiveness of the market—if they had access to more user information.

Our second contribution is that we offer results from an experiment—and one conducted by a previously not studied small search engine. This allows us not only to shed some light on a provider that is usually not in the spotlight but also to obtain clean estimates of the dependence of search result quality on data as an input.

In Section 2, we provide a more detailed discussion of the literature to which we relate and contribute. Section 3 provides details on the setup and the experiment. Section 4 reports the results. Section 5 discusses the results and concludes. This paper is deliberately short. A comprehensive appendix contains many technical details and presents additional findings.

2. Background and related literature

Our paper seeks to contribute to the vast literature studying digital markets. Several papers provide experimental and simulation-based evidence related to the importance of the data for platforms. For example, Sun et al. (2023), Wernerfelt et al. (2022), and Decarolis et al. (2023).[12]

Morton et al., 2019), the EU (Cremer et al., 2020), the UK (Furman et al.; 2019), and Germany (Schallbruch et al., 2019).

[11] See European Union Digital Markets Act (2022), United Kingdom CMA (2020), or US Congress (2021).

[12] In addition, Microsoft argued that "obtaining the large quantity of data necessary to develop an effective [general] search engine (e.g., the information upon which relevancy algorithms can be built and improved) would be a significant barrier to entry". See paragraph 286 of the Google Shopping case: https://ec.europa.eu/competition/antitrust/cases/dec_docs/39740/39740_14996_3.pdf.

Calvano and Polo (2021) conclude in their literature review that digital markets have a strong natural tendency towards concentration or market tipping, which suggests that models of competition *for* the market are more relevant than models for competition *in* the market. Krämer and Schnurr (2022) offer an excellent survey of the literature about economies of scale and scope in data and discuss various policy proposals, with a focus on data-sharing obligations.

Both the academic and the policy discussion about data sharing suffer from unclear definitions. Most of the literature studies situations in which a user knows more about their type or willingness to pay for a service than the provider of the service.[13] Then, the *voluntary* balancing of that information makes markets more efficient (or enables follow-on innovation) but comes at a cost for the individual, including a decrease in privacy, and, hence, the net welfare effects may be positive or negative. By contrast, the search engine market is often referred to as a *data-driven market*, where the interaction between a service provider and a user is administered electronically such that it is possible to store users' choices (e.g. clicking behavior) and characteristics (e.g. location) with very little effort, i.e. virtually for free. Hence, the one provider who interacts with a user already has access to the user's data at the start of the analysis. In such an environment *mandatory* data sharing is needed because one party, the incumbent, has no incentives to share voluntarily.

Prüfer and Schottmüller (2021) define a *market as data-driven if a firm's marginal costs of innovation decrease in the amount of user information*, that is if it is subject to specific feedback effects ("data-driven indirect network effects"). They show in a dynamic model of R&D competition that, in data-driven markets, user information leads to *market tipping (monopolization) and l*ow incentives to innovate both for the dominant firm and for (potential) challengers. The intuition is that the smaller firms, even if they are equipped with superior production technology, face higher marginal costs of innovation because they lack access to enough user information. If a smaller firm were to heavily invest in innovation and roll out its high-quality product, the dominant firm could imitate it quickly --- at a lower

[13] See Bergemann et al. 2020 and the literature cited therein.

cost of innovation --- and regain its quality lead. Foreseeing this situation, rational entrepreneurs and private financiers would not invest in such a smaller firm in the first place.[14] The dominant firm knows about the deterring disincentive to innovate for its would-be competitors and can rest on a lower level of innovative efforts, too.

Some authors argue that mandating the sharing of (anonymized) data on user preferences and characteristics amongst competitors could mitigate market tipping and would have positive net effects on innovation and welfare if data-driven indirect network effects are sufficiently strong (Prüfer and Schottmüller, 2021; Argenton and Prüfer, 2012).

Our paper aims to study how important user-generated data are for search result quality. We study this for non-personalized search results using across-search engine comparisons and an experiment conducted with an actual entrant, Cliqz.[15] This complements the work by He et al. (2017) and Schaefer and Sapi (2023). He et al. (2017) study scale effects in web searches by comparing query logs of several hundred billion searches of two large search engines. They distinguish "popular queries, which do account for a majority of *searches*" from "rarer queries, which account for the majority of *queries*" (p.295, emphasis added). They measure search engine quality using the click-through rate (CTR), i.e., the percentage of clicks on the *top* URL of a search result page. They document a concave relationship between the historical number of queries and the CTR.

Schaefer and Sapi (2023) study with observational data from Yahoo.com whether there are economies of scale in internet search. They also focus on the CTR as a quality measure and show that more data enhances search engine quality and that personal information (for instance, the ability of the search engine to track the browsing behavior of specific users) amplifies the speed of learning. Their findings

[14] This result is reflected by Edelman (2015), who cites the oral testimony of Yelp's CEO before the Senate Judiciary Subcommittee on Antitrust, Competition Policy and Consumer Rights on September 21, 2011, and writes: "Google dulls the incentive to enter affected sectors. Leaders of TripAdvisor and Yelp, among others, report that they would not have started their companies had Google engaged in behaviors that later became commonplace."

[15] See Hagiu and Wright (2023) for delineations and modeling of across-user learning and within-user learning.

are consistent with an incumbent data advantage due to the possession of personal information. A similar result is shown by Bajari et. al (2019) studying Amazon data. They find that the prediction accuracy of their models increases with the time dimension (but with diminishing returns to scale).

Our results are broadly in line with He et al. (2017) and Schaefer and Sapi (2023) but go beyond these two papers. Since our study focuses on the quality of search results of an entrant, we can quantify the actual gap in search result quality between the entrant and the incumbent that is due to rare queries (and thus the availability of user-generated data).

Another interesting complementary paper is Lei et al. (2023), who partnered with a search engine in China and ran a Randomized Control Trial, also with non-personalized search results. In the control condition, they exposed users to the typical product of the smaller search engine, which was trained both by own ("internal") data of the search engine and by "external" data, which they received via an API from the market-leading Chinese search engine. In the treatment condition, they removed access to the API. They also show that less data the algorithm has access to (in the treatment condition) reduces search-engine quality, as measured by the CTR. Confirming our results qualitatively, they conclude (p.4): "API removal leads to a 4.6% decline in average CTR on search suggestions, demonstrating the value of the market leader's data capability."

We also contribute to the existing empirical literature by using various measures of search engine quality: Unlike He et al. (2017), Schaefer and Sapi (2023), or Lei et al. (2023), we use *human assessment* and the *overlap of Cliqz's search result sets with Google's* as measures of quality. Finally, most importantly, He et al. (2017) and Schaefer and Sapi (2023) use observational data, while we use an experimental setting to estimate the relationship between search result quality and the amount of data from past searches the search engine has access to. Lei et al. (2023) also ran an experiment but did so in the institutional context of China (while ours is Europe). Moreover, in their study, the target search engine's algorithm can (and has to) adjust to the consequences of having access to less data in the treatment condition, which changes various parameters. In our experiment, by contrast, we keep

everything but access to data constant, and hence, quality differences are directly related to the availability of data.

Finally, while search results at Google are personalized to some degree[16], a recent study based on German search data produced by Google found that search results are highly *regionalized* but rarely *personalized* (Krafft et al., 2019). This finding increases the relevance of our results, as they are based on non-personalized data collected by a privacy-protective small search engine.

3. Data and experimental setup

We describe the data and experimental setup using the following conceptual framework. A search engine produces a set of results r for a given query q, a given web index I,[17] an algorithm with parameters θ, and data about previous searches in D. That is,[18]

$$r = f\big(q, I, D; \theta(D)\big).$$

Data in D has a *direct effect* on results (more data is useful to predict what users are looking for) as well as an *indirect effect* (because it can be used to re-train/improve the algorithm). For our *between-search engine* comparisons (details below), we evaluate the quality of search-result set r for the same query q by three search engines operating with different indices I,[19] different sets of user data D, and different algorithms θ.[20] For our experiment (the *within-search engine* comparison), we keep the function f, query q, index I, and algorithm θ constant and only vary the amount of data D the algorithm has access to.

---

[16] See https://support.google.com/websearch/answer/12410098.
[17] A web index is a database used by search engines to store and organize information about web pages, enabling quick retrieval of relevant results when users perform searches.
[18] We think of r as being a vector, q as being a text string, I as a more complicated version of a table containing the list of web pages and information about them that can be collected, D as a table containing all the user-generated data the search engine has collected in the past (possibly aggregated in some way), and θ as a vector of parameters that are obtained in a first step for given D, which is why we write $\theta(D)$. f is a vector-valued function.
[19] See https://www.seo.com/basics/how-search-engines-work/
[20] We think of the function f itself as the same across search engines, but this is not essential for the interpretation of the results, as one can think of the same f with a different θ as another function.

We collaborated with the small search engine Cliqz.[21] The mission of Cliqz was to offer an alternative search engine that protects the privacy of its users. This is one of the reasons we focus on non-personalized search results in this paper.

To construct the data we use in this paper, Cliqz ordered all search queries submitted by German users of their "Human Web" from April 20 to April 26, 2020, by their frequency. [22] They formed five buckets in line with queries' frequency using the following thresholds: 0.2%, 1%, 5%, and 25%. Table 1 shows that the distribution of traffic across queries is highly skewed; see also Goel et al. (2010). The second column in Table 1 shows that bucket 1 (B1) contains the 0.2% most popular queries, bucket 2 the next 0.8% most popular queries, etc. B5 contains the 75% least popular queries (tail queries). The third column shows the average number of searches per query per week. The fourth column shows the implied percentage of the traffic that is generated. Rarest queries (in B5) generated 56% of the traffic.

- Table 1 about here –

We randomly drew 1,000 queries from each of the five buckets, leaving us with a stratified sample of 5,000 queries. For each of those 5,000 queries, Cliqz gave us non-personalized search results at different levels of user information. *Search results* consist of a list of URLs (Universal Resource Locator: a website's address) and additional information on each item, like a short preview of the website. The user data used to produce search results include so-called *query log counts*, which summarize how often URLs have been clicked on by past users for a given query. Importantly, the search algorithm detects the similarity between queries, so that it can also use query log counts of similar queries to produce search results, for instance by applying cluster analysis to the query-URL bipartite graph (Liu et al., 2012, Sadikov et al., 2010). We asked Cliqz to keep the search engine's algorithm as it is and only vary

[21] See Appendix A for some background on the Cliqz search engine.
[22] The Human Web is a software integrated in the Cliqz browser or, alternatively, a software extension to Mozilla Firefox. It allows for the anonymous collection of user browsing activity and user-generated query logs. For example, if a user of a Cliqz browser -- or a Firefox browser with installed Cliqz extension -- searches for "ebay auto" using Google, Bing, Cliqz, or any other search engine, the information on the search, the results and choices made by the user were transferred in an anonymized manner to Cliqz. Hence, these search queries represent all searches on any search engine for that subpopulation of users.

the amount of user data used to produce search results on the night between April 26 and April 27, 2020. *This* is the experiment that we conducted. It helps us to provide direct evidence on the dependence of search engine quality on the amount of data an algorithm uses.

Specifically, Cliqz provided results at twelve different levels of data on past searches: 100% (or full data), 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, 10%, 1%, and 0.1%. To obtain respective query log counts, we multiply the query log counts by the assumed fraction of available data and take the floor of that value as the new log counts. For example, if a given query/URL pair has a count of 10 (i.e., people who searched for that query clicked on that URL ten times in the past), then the new count for that query/URL pair would be 5 under 50% of user data availability: 1 under 10%, and 0 under 1% or below. Hence, if the Cliqz search engine would only have 1% of its actual user data, it would completely miss that query/URL pair.

We complemented the Cliqz data with non-personalized search results from Google and Bing on the same 5,000 queries in the same period. For this, we used the application programming interface (API) of a for-pay service for Google and Bing search engines. The API allowed us to specify that we would like to obtain results for users from Germany, to make the results comparable with the search results of Cliqz.

Finally, we asked external assessors to assess the quality of the search results on a 7-point Likert scale for a subset of queries. We restrict attention to queries that are either in German or English and that are at least 3 characters long. Then, we sample 500 queries: we draw 50 queries from buckets 1 and 2 each, 100 queries from buckets 3, and 150 queries from each of buckets 4 and 5. We over-sample rare queries (buckets 4 and 5) to reduce possible noise as we expect that rare queries might be more difficult to assess. We also removed 7 queries with inappropriate content, resulting in 493 queries for human assessment.

For each query, we keep seven result sets to assess: five for Cliqz (at 100%, 50%, 20%, 10%, and 1% of user data), one for Google, and one for Bing. We limit each result set to the top 5 results.[23] Additionally, for each sampled query, we construct a "mixed" top-5 result set from Google, Bing, and Cliqz (at 100% of user data) result sets. We randomize the order in which Google, Bing, and Cliqz contribute to the mixed result set. See the details of the "mixed"' result set construction in the appendix.

The assessment design (more details in the appendix) ensures that the assessors did not know which search engine generated a set of search results. Hence, all brand effects are excluded. We hired two research assistants (RA's) at Tilburg University and 587 people in Germany (37% women, median age 34) through the clickworker.com platform to perform the assessment (clickworkers). One of the research assistants received all the result sets corresponding to queries in German; the other to queries in English (each assessor was proficient in the relevant language). 563 clickworkers provided evaluations, on average for fifteen result sets. In total, each of the result sets was evaluated on average by four different people (one RA and three clickworkers). When evaluating the result sets, the assessors were able to click on the respective links.

The appendix contains further details, including details on the characteristics of the assessors, and the instructions we gave to the RAs and the clickworkers.

4. Results

4.1 The experiment: within-search engine comparisons

In the experiment, we start with a set of queries q, vary the query log counts in D_{Cliqz}, and keep I_{Cliqz} and θ_{Cliqz} constant. This allows us to obtain unambiguous results regarding the impact of more data.[24] These results are conservative in two ways. First, in reality, the parameters θ_{Cliqz} would likely change

[23] The first five organic search results combined typically receive about 80% of the clicks at Google (https://www.smartinsights.com/search-engine-optimisation-seo/seo-analytics/comparison-of-google-clickthrough-rates-by-position/).

[24] Rare queries may also differ from popular ones for other reasons. By conducting an experiment and holding the set of queries fixed, we control for these differences. For that reason we do not have to control for any other query characteristics, e.g., length or number of words or measures of the complexity of the query, in our analysis.

when the algorithm was trained with less data and would produce worse search results. Second, we drop data proportionally (see Section 3). In reality, less data would also be more noisy.[25] For these two reasons, we would expect that the patterns we will document would be stronger in reality and in that sense, our results are conservative.

Figure 1 shows that, for more popular queries (B1-3), search result quality increases with access to more data, but only until about 20% of Cliqz's available data, which is already enough to produce the highest quality it can produce (as the curves start to flatten there). This supports statistical learning theory, which suggests diminishing returns to dataset size in terms of predictive performance (He et al., 2017, Varian, 2019). The remaining differences may come from differences in the algorithm. Alternatively, they may be a consequence of complementary investments and organizational practices generating productivity gains from the use of data (Bresnahan et al. 2002, Bloom et al., 2012). Crucially, however, for rare queries (B4-5) the quality of search results is at a much lower level and the curves are still increasing when using 100% of data available to Cliqz. That is, even if the marginal increases in quality are decreasing, they are positive for all levels of data we could observe. This suggests that Cliqz's quality could benefit from access to more search-log data on rare queries. However, whether Cliqz could eventually produce search results comparable to Google's remains an open question that our experiment cannot answer.

- Figure 1 about here -

These results are based on between-subject comparisons. We pool assessments by the RAs and the clickworkers. In the appendix, we show that these results are robust when we limit the sample only to one type of assessor, or when we measure only within-subject variation in the ratings. Even if there was scope for within-subject learning on what constitutes a good search engine result because the order of

[25] To see this, suppose we start with a large number of query log counts, say 1000 and 300. By the law of large numbers, the relative magnitudes are close to the ones for even more data. But the law of large numbers would fail to hold when we drop for instance 99% of the data. We instead do as if it still holds (and there is no noise from sampling error), by changing the query log counts to 10 and 3, respectively, under 1% of the original data.

result sets shown to the assessors was completely random, in expectation all buckets and all search engines should be affected equally. At the same time, we believe that both RAs and clickworkers were tech-savvy enough to represent an average search engine user.

- Figure 2 about here -

Next, we perform a robustness check. In Figure 2, we report how the overlap between Cliqz' and Google's top-5 results, which is an alternative measure of search engine quality, depends on data available to Cliqz. The overlap is measured as the percentage of times the top-5 search results by Cliqz and Google are the same (in the sense that the set of results is the same without comparing the ordering). This implies that the independent variables in Figures 1 and 2 on the horizontal axis are the same but that *the outcome variables are different* (human-assessed quality vs. overlap between Cliqz's and Google's results). Therefore, the magnitudes are not directly comparable. Yet, the qualitative findings are the same: for popular queries (B1-3), about 20% of the data available to Cliqz's algorithm is sufficient to reach a level beyond which access to more data has minimal or no effects. For rare queries (B4-5), however, there is no quality saturation. More data makes Cliqz's algorithm better both in the eyes of human assessors and brings its results closer to Google's, as measured by machine-calculated similarity scores.

Notably, in Figure 2 even for bucket 1, with most popular queries, the overlap of Cliqz with Google levels off at about 65%. As the human assessment has shown, this is not a sign of the diminished quality of Cliqz's results as compared to Google's: users like the results for the most popular queries of both search engines equally much (see Figure 2). However, there are various ways to answer certain queries, which is why only about 3 of the top-5 links produced by Cliqz (60%) are the same as those produced by Google.

4.2 Between-search engine comparisons

Figures 1 and 2 have shown that more data on previous searches is an important ingredient in increasing search result quality, especially for rare queries. If such data are a driving factor of search result quality,

as compared to other influence factors such as algorithmic quality, we should see differential access to data on previous searches reflected in the search result quality *across* search engines. By definition, a "small" search engine, which has only fewer users than a "large" search engine with a high market share, should therefore be at a disadvantage in producing high-quality results, especially for rare queries.

- Figure 3 about here –

We have also assessed the search results of Google and Bing. This was done for the same queries and the assessors did not know which search engine produced them. Figure 3 shows search result quality for each of the three search engines by popularity of the query. For popular queries (B1-3), search results are of comparable quality across search engines. This suggests that Cliqz's algorithm can provide similar quality to Google's for search queries where it can access substantial amounts of user information (see also Banko and Brill, 2001). For less popular queries, however, Cliqz's results are significantly worse than Google's and Bing's.[26] These results reflect the differential role that data on previous searches play for popular vs. rare queries: the quality of search engines with a higher market share and, hence, more data on both popular and rare queries decreases only a bit when moving from most popular to rarest queries; by contrast, a small search engine's quality depreciates significantly when comparing its most popular to its rarest queries.

This underlines that indeed, differences in the overall quality of search results are largely driven by differences in the quality of results for less popular queries. Table 1 shows that, on average, B4 and B5 are searched only 1.2 and 1 times per week, respectively. It appears that this is not enough to produce sufficient data that can be used to produce high-quality search results. Crucially, buckets 4 and 5 jointly

[26] Referring to our theoretical framework above, differences could also stem from different web indices I --- Cliqz had indeed invested in an own web index. However, the quality of a web index is largely a function of money invested, just as the quality of an algorithm. Economically they are the same in our production-function framework. Therefore, here we lump effects of the index I and the algorithm θ on results r together and distinguish both from data about previous searches D.

produce 74% of the traffic in our data: for a search engine to offer users satisfactory results, it is pivotal to perform well on those rare and rarest queries.

While this is a plausible explanation, it could in principle also be true that differences in the quality of search results for rare queries are driven by differences in the algorithm. To see this, let Google's algorithm have parameters θ_{Google}, and use data D_{Google} and index I_{Google}, while Cliqz's algorithm has parameters θ_{Cliqz}, and use data D_{Cliqz} and index I_{Cliqz}. Then, it could in principle be true that both search engines produce results that are of similar quality for popular queries and different quality for rare queries solely because of differences between θ_{Google} and θ_{Cliqz}. This would be the case when Google's algorithm was particularly good at producing search results for rare queries. However, the within-search engine experiment we reported on in Section 4.1 shows that data matters as an input and that at least some of the differences are related to the availability of data.

Finally, we bring both parts of the analysis together by comparing the overall average quality of search results across search engines. Figure 4 shows that the average quality of search results differs: Google produces the best results, followed by the number two in the market, Bing. The results produced by Cliqz are substantially worse. This ordering corresponds to the ordering of the market shares of the three search engines.

- Figure 4 about here -

If we consider that search engine users' consumption choices, on top of ordinal quality considerations, are influenced both by legitimate factors (given that nominal usage prices are zero across all search engines, there is little reason to use another but the highest-quality provider) and illegitimate factors (Google has been accused of various abuses of its monopoly/dominant position),[27] it is conceivable that

[27] For instance, Google has been accused to place its own products above competing sellers' if a user searches for products (Google Shopping case https://ec.europa.eu/competition/antitrust/cases/dec_docs/39740/39740_14996_3.pdf). Recently, a federal judge ruled that Google violated US antitrust law by maintaining a monopoly in the search and advertisement markets, especially by concluding exclusive deals with Apple and other operating system producers to install Google as the default search engine (https://www.theverge.com/2024/8/5/24155520/judge-rules-on-us-doj-v-google-antitrust-search-suit).

moderate average quality differences between Google and Bing in Figure 4 are translated into drastic differences in market shares. Notably, the analysis presented here is orthogonal to these legal cases and to Google's conduct in markets. We just showed that more data on rare queries makes search engines better and that the different access levels to such data across Google, Bing, and Cliqz in the market are consistent with our findings.

5. Discussion and conclusion

Data is an important input for many services that are widely consumed (Mayer-Schönberger and Ramge, 2018). In this paper, we study how important data about previous searches is as an input in the context of online search. We start from the empirical observation that Google has market shares above 90% in many national markets, including the US and Europe. To quantify the relevance of more data about previous searches, compared to all other influence factors (including algorithmic quality), for today's search engine quality, we conduct an experiment with the small search engine Cliqz, in which we *only* vary the amount of data Cliqz uses to produce search results. We show by two completely different outcome measures that data access has a significant effect on search result quality. This effect levels off for popular search terms, but a key result of this paper is that the quality returns to data are monotonically increasing for rare queries at the level of data Cliqz had during the time of our experiment.

Our within-search engine results are reflected in our between-search engine comparisons, which show that Cliqz' algorithm can produce results of similar quality than Google's and Bing's for popular search terms, that is, for queries where Cliqz has access to enough data. For rare queries, however, where Cliqz' algorithm can only draw on data from 1 previous search per week, search result quality decreases substantially, whereas the quality of Google and, to some extent, Bing decrease only a bit. Our final

analysis brings all parts together and shows that the average quality ranking across search engines reproduces the ranking found in the market. [28]

Taken together, we show that search result quality depends on the amount of data a search engine has access to and that using less data has a particularly big effect on search-result quality for rare queries. This suggests in turn that having access to more data would enable small search engines such as Cliqz to produce results for rare queries that are of better quality. This finding is highly policy-relevant, as rare queries constitute the better part of the traffic (74% in our data). For that reason, entrants may have a hard time competing with incumbents.[29] A proposed remedy that might help level the playing field could be to require large search engines to share data about previous searches with smaller competitors. Unlike in other contexts, this remedy would not directly harm the incumbent because user data are non-rival. Data-sharing only removes the incumbent's *exclusivity advantage* to access data.[30] Notably, saving raw data on previous searches is a by-product of operating a search engine that is independent of the fixed cost of operations, which can be huge. Moreover, big tech firms are already engaged in all kinds of *voluntary* data sharing, which suggests that even the costs to share large amounts of data are not prohibitive.[31] To the extent this also holds on the search engine market, data sharing can be a useful policy, too, specifically because data is non-rival.

---

[28] Notably, shortly after our experiment was conducted, the small search engine we worked with (Cliqz) announced that it would go out of business (https://cliqz.com/announcement.html).

[29] One may wonder how, then, Google could have reached market leadership against Yahoo, the leading search engine in the 1990s. This development can be understood through the lens of the theory of data-driven markets (Prüfer and Schottmüller, 2021): in the 1990's, Yahoo categorized websites by hand. Google's original page rank algorithm, as described in Page et al. (1998), let the rank of a website be determined only by the structure of hyperlinks on the world wide web, not by user information. Only since 2001 has Google made use of the data created by logging users' clicking behavior in *search logs* (Zuboff 2016). This innovation transformed the search -engine industry into a data-driven market, leading to Google's market leadership as of 2003.

[30] See Graef and Prüfer (2021) for a fully-fledged proposal how to implement mandatory data sharing, including a governance structure who should be responsible for which tasks, and which is in line with EU privacy-protection, intellectual property, consumer rights, and competition law. See Krämer and Schnurr (2022) for a discussion of several ways of data sharing, including their pros and cons.

[31] Recently, Google, Facebook, Microsoft, Twitter, and other firms showed that sharing of user data is technically and organizationally possible at a large and automatic scale. They had announced a new standards initiative called the Data Transfer Project, designed as a new way to move data between platforms. See https://www.theverge.com/2018/7/20/17589246/data-transfer-projectgoogle-facebook-microsoft-twitter.

Consequently, subjecting a dominant search engine, which has already sunk set-up costs, to a data-sharing obligation would not disrupt its cost structure too much. On the other side, access to the shared data would improve the quality of competitors and newcomers, as shown in this paper, thereby making the search engine market more competitive. Consequently, these competitors would know they stand a chance in the market and have an incentive to innovate, either by improving their algorithms or by offering complementary services. This, in turn, would constitute stronger incentives to innovate than without mandatory data sharing for the dominant search engine if it wants to stay in business (Prüfer and Schottmüller, 2021). Consumers would benefit from both changes.[32]

Importantly, mandatory data sharing is *not* a silver bullet to reinstate leveled competition in the search-engine market. Google, the dominant market leader, has not only 90% market share or more on most national markets,[33] it also has dedicated hardware and data center solutions as compared to smaller competitors, a very strong brand name, and pays hardware manufacturers, such as Apple, to install it as the default search engine on millions of end-user devices.[34] However, the positive and monotonic role that more data on past searches plays for rare queries, which constitute over 70% of all queries and hence strongly affect users' expectations about search-engine quality, will draw them towards the search engine with the highest expected quality (see Figure 4). Therefore, it seems wise to provide small search engines with more data about rare queries independent of all other dimensions of competition. Another way of putting this is by saying that data sharing is a *necessary* condition for competition on the search engine market, but not a *sufficient* condition.

---

[32] This is the case as long as the part of the additional fixed cost associated with operating multiple search engines that is passed on to consumers would not be too high. That cost could in principle be passed on indirectly in the form of more advertising that consumers are exposed to, even if using the search functionality r saving raw data on previous searches emains free, as it is now. At the same time, competition between search engines that results from leveling the playing field may discipline firms when it comes to passing on such fixed costs to consumers.
[33] See https://gs.statcounter.com/search-engine-market-share.
[34] In a survey we conducted with a representative sample of the Dutch population (in the so-called LISS panel), among 538 individuals who were aware of other search engines than Google the three most frequently mentioned reasons to use Google were: 1. Google provides high quality search results. 2. Google is a well-known search engine. 3. Google is the standard search engine on the device I most often use.

Finally, as a litmus test for the relative impact of data versus algorithmic quality, one could take the recent surge of large language models (LLMs), notably ChatGPT, since November 2022. ChatGPT's integration into Bing since February 2023, following a U$10bn investment of Microsoft in ChatGPT's owner OpenAI could be regarded as a very large investment in Bing's algorithmic quality. Several commentators expected that this would reshuffle market shares in the search-engine market.[35] Nevertheless, empirically global market shares have been virtually fixed over the past two years.[36] Even the huge investment in Bing's algorithm has not done a scratch to Google's 90%+ market share.

Online Appendix and Replication Package

Appendix A-E provide details and robustness checks.

A replication package with data and code is available at https://www.dropbox.com/s/k4u02vh9dbn38vg/submission_data_code.zip?dl=0 (this will be made public, e.g. as a Github repository, at a later point).

[35] See https://prufer.net/2023/02/08/chatgpt-vs-bard-in-search-engines-good-to-have-a-theory/ for some background and references about those discussions and a theory-led perspective on the likely impact of LLMs on the search engine market, as of February 2023.
[36] https://gs.statcounter.com/search-engine-market-share > set time window May 2022-September 2024.

References

Argenton C, Prüfer J (2012) Search engine competition with network externalities. *J Comp Law Econ* 8(1): 73-105.

Bajari P, Chernozhukov V, Hortaçsu A, Suzuki J (2019) The impact of big data on firm performance: An empirical investigation. *AER P&P* 109: 33-37.

Banko M, Brill E (2001) Scaling to very very large corpora for natural language disambiguation. *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*: 26-33. https://doi.org/10.3115/1073012.1073017

Beaton-Wells C (2019) Ten Things To Know About The ACCC's Digital Platforms Inquiry. *Comp Pol Int*.

Bergemann D, Bonatti A, Gan T (2022) The economics of social data. *RAND J Econ* 53(2): 263-296.

Bloom N, Sadun R, Van Reenen J (2012) Americans do IT better: US multinationals and the productivity miracle. *Amer Econ Rev* 102(1): 167-201.

Bresnahan TF, Brynjolfsson E, Hitt LM (2002) Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *Quarterly J Econ* 117(1): 339-76.

Calvano, E, Polo, M (2021) Market power, competition and innovation in digital markets: a survey. *Inf Econ Policy* 54: 100853.

Crémer J, Montjoye YA, Schweitzer H (2020) Competition Policy for the Digital Era. Report for European Commission, DG Competition, 2020, https://ec.europa.eu/competition/publications/reports/kd0419345enn.pdf.

Decarolis, F, Rovigatti, G, Rovigatti, M and Shakhgildyan, K. (2020) Artificial Intelligence & Data Obfuscation: Algorithmic Competition in Digital Ad Auctions. CEPR Discussion Paper No. 18009.

Ducci F (2020) *Natural Monopolies in Digital Platform Markets* (Cambridge University Press, Cambridge, UK).

Edelman B (2015) Does Google leverage market power through tying and bundling? *J Comp Law Econ* 11(2): 365–400.

European Union (2022) Digital Markets Act. https://eur-lex.europa.eu/eli/reg/2022/1925/oj

Feng J, Bhargava HK, Pennock DM (2007) Implementing Sponsored Search in Web Search Engines: Computational Evaluation of Alternative Mechanisms. *INFORMS Journal on Computing* 19(1): 137-148.

Furman J, Coyle D, Fletcher A, Marsden P, McAule D (2019) Report of the Digital Competition Expert Panel: Unlocking digital competition. Report for UK Treasury.

Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ (2010) Predicting consumer behavior with Web search. *PNAS* 107(41): 17486–17490.

Graef I, Prüfer J (2021) Governance of Data Sharing: a Law & Economics Proposal. *Res Pol* 50: 104330.

Hagiu A, Wright J (2023) Data-enabled learning, network effects, and competitive advantage. *RAND J Econ* 54(4): 638-667.

Halevy A, Norvig P, Pereira F (2009) The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems:* 8-12.

He D, Kannan A, McAfee RP, Liu TY, Qin T, Rao JM (2017) Scale Effects in Web Search. *International Conference on Web and Internet Economics - WINE 2017*. https://doi.org/10.1007/978-3-319-71924-5_21.

Krämer J, Schnurr J (2022) Big Data and Digital Markets Contestability: Theory of Harm and Data Access Remedies. *J Comp Law Econ* 18(2): 255–322. https://doi.org/10.1093/joclec/nhab015

Lei X, Chen Y, Sen A (2024) The Value of External Data Capabilities in the Search Market: Evidence from a Field Experiment. Available at SSRN: https://ssrn.com/abstract=4452804.

Lewis-Kraus G (2022) How Harmful Is Social Media? *The New Yorker*, 3 June 2022. Available at https://www.newyorker.com/culture/annals-of-inquiry/we-know-less-about-social-media-than-we-think

Liu X, Song Y, Liu S, Wang H (2012) Automatic taxonomy construction from keywords. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 1433–1441.

Mayer-Schönberger V, Ramge T (2018) *Reinventing Capitalism in the Age of Big Data* (Basic Books, London).

Page L, Brin S, Motwani R, Winograd T (1998): The Pagerank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford University.

Persily N, Tucker JA (2020) Social Media and Democracy: The State of the Field, Prospects for Reform (Cambridge University Press)

Prüfer J, Schottmüller C (2021) Competing with big data. *J Ind Econ* 69: 967-1008.

Sadikov E, Madhavan J, Wang L, Halevy A (2010) Clustering query refinements by user intent. *Proceedings of the 19th International Conference on World Wide Web*: 841–850.

Schaefer M, Sapi G (2023) Complementarities in learning from data: Insights from general search. *Information Economics and Policy*, 65: 101063.

Schallbruch M, Schweitzer H, Wambach A (2019) Ein neuer Wettbewerbsrahmen für die Digitalwirtschaft. Report for German Economics Ministry.

Scott Morton F, Bouvier P, Ezrachi A, Jullien B, Katz R, Kimmelman G, Melamed D, Morgenstern J (2019) Committee for the Study of Digital Platforms: Market Structure and Antitrust Subcommittee. Report. Stigler Center for the Study of the Economy and the State. University of Chicago Booth School of Business. Available at https://www.chicagobooth.edu/-/media/research/stigler/pdfs/market-structure-report.pdf.

Sun T, Yuan Z, Li C, Zhang K, and Xu J. (2023) The Value of Personal Data in Internet Commerce: A High-Stakes Field Experiment on Data Regulation Policy. *Management Science*.

UK Competition and Markets Authority (2020) Online platforms and digital advertising. https://assets.publishing.service.gov.uk/media/5fa557668fa8f5788db46efc/Final_report_Digital_ALT_TEXT.pdf.

US Congress (2021) American Innovation and Choice Online Act, Section 2. https://www.congress.gov/bill/117th-congress/house-bill/3816/text?r=43&s=1

Wernerfelt N, Tuchman A, Shapiro B, and Moakler R (2022). Estimating the value of offsite data to advertisers on Meta. University of Chicago, Becker Friedman Institute for Economics Working Paper 114.

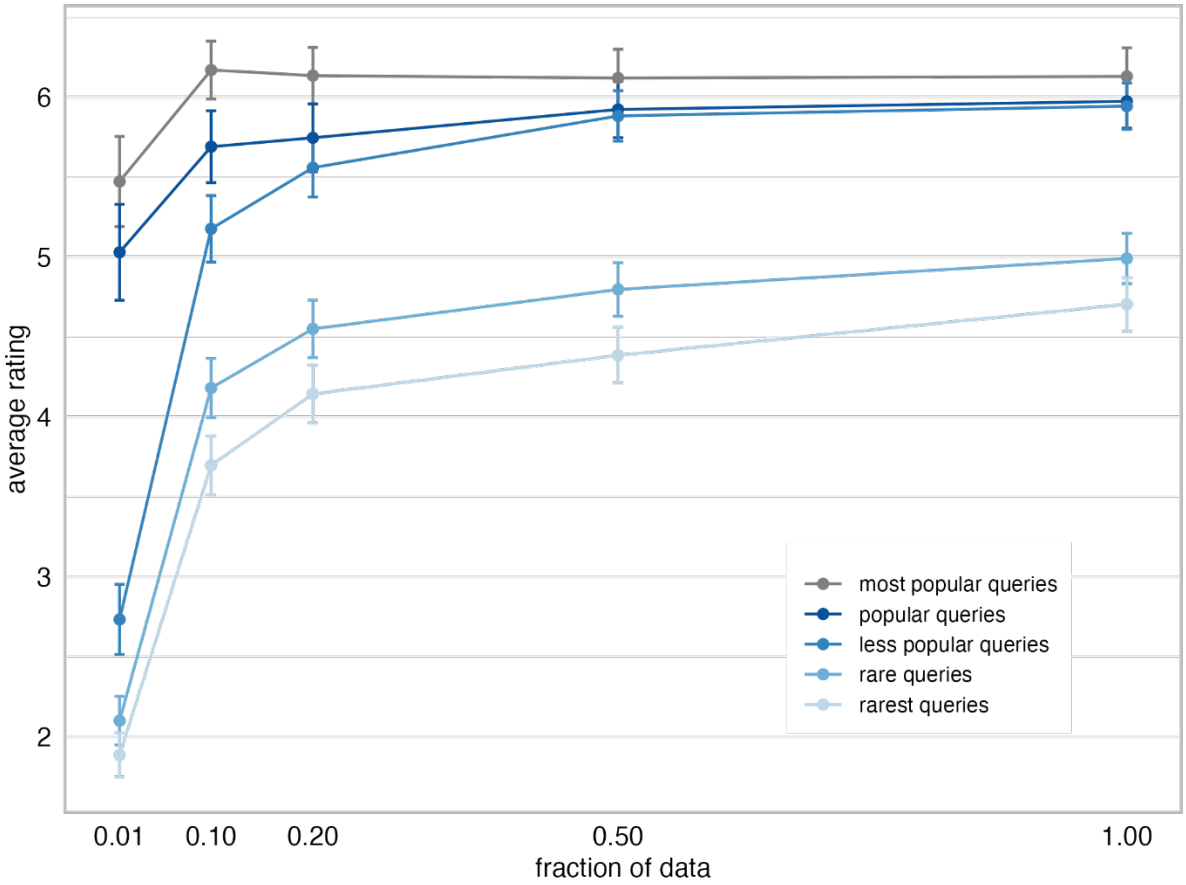
Varian H 2019) *Artificial Intelligence, Economics, and Industrial Organization* (University of Chicago Press, Chicago, IL).

Tables and Figures

bucket	% queries	searches per week	% traffic
B1	0.2%	72.1	11%
B2	0.8%	9.8	6%
B3	4%	3.2	10%
B4	20%	1.2	18%
B5	75%	1.0	56%

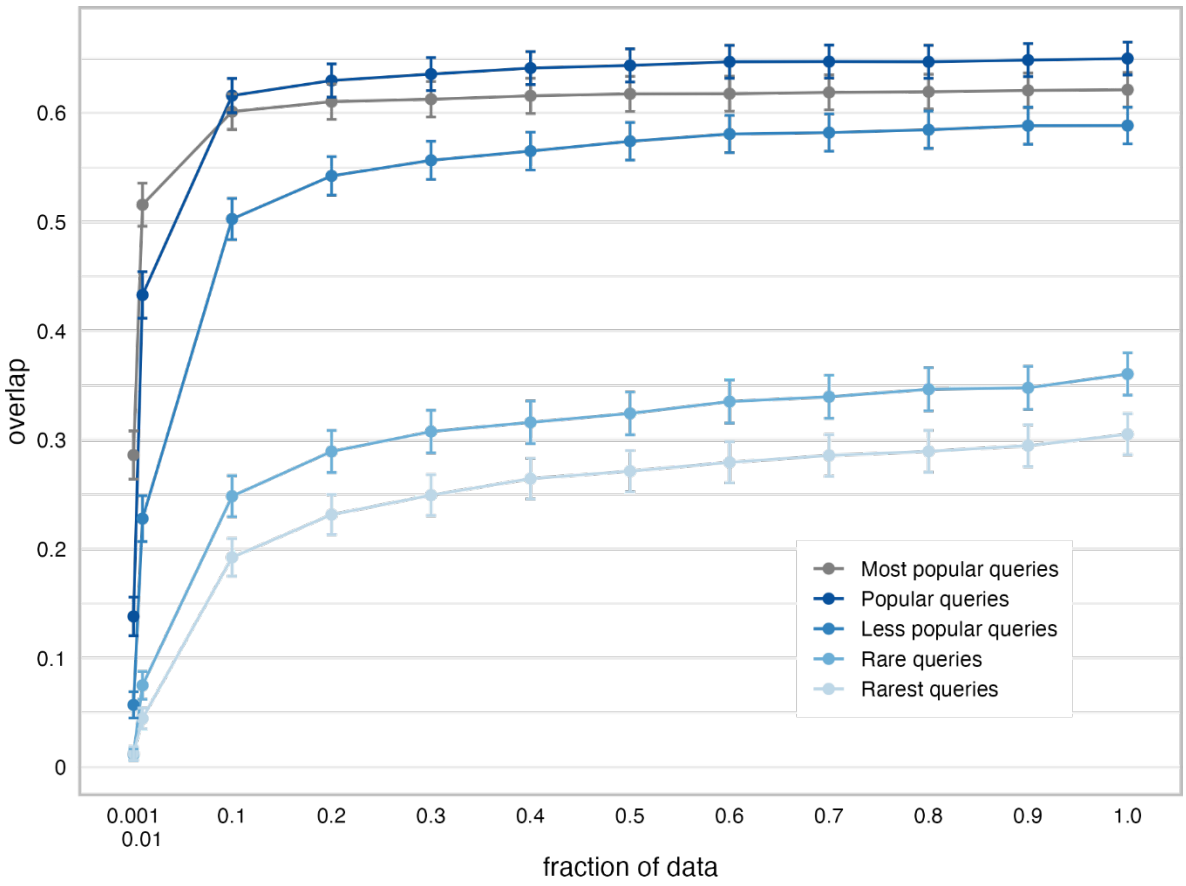
Notes: This table shows statistics for the 5 buckets that we use to classify our queries. B1 contains the most popular queries and B5 the least popular queries. The second column shows the respective percentages of queries. They add up to 100% by construction. The third column shows the average number of searches per week for each query in a bucket. The last column shows the percentage of the total traffic associated with each of the queries. It is calculated from the information in the second and third columns.

Table 1: Query buckets



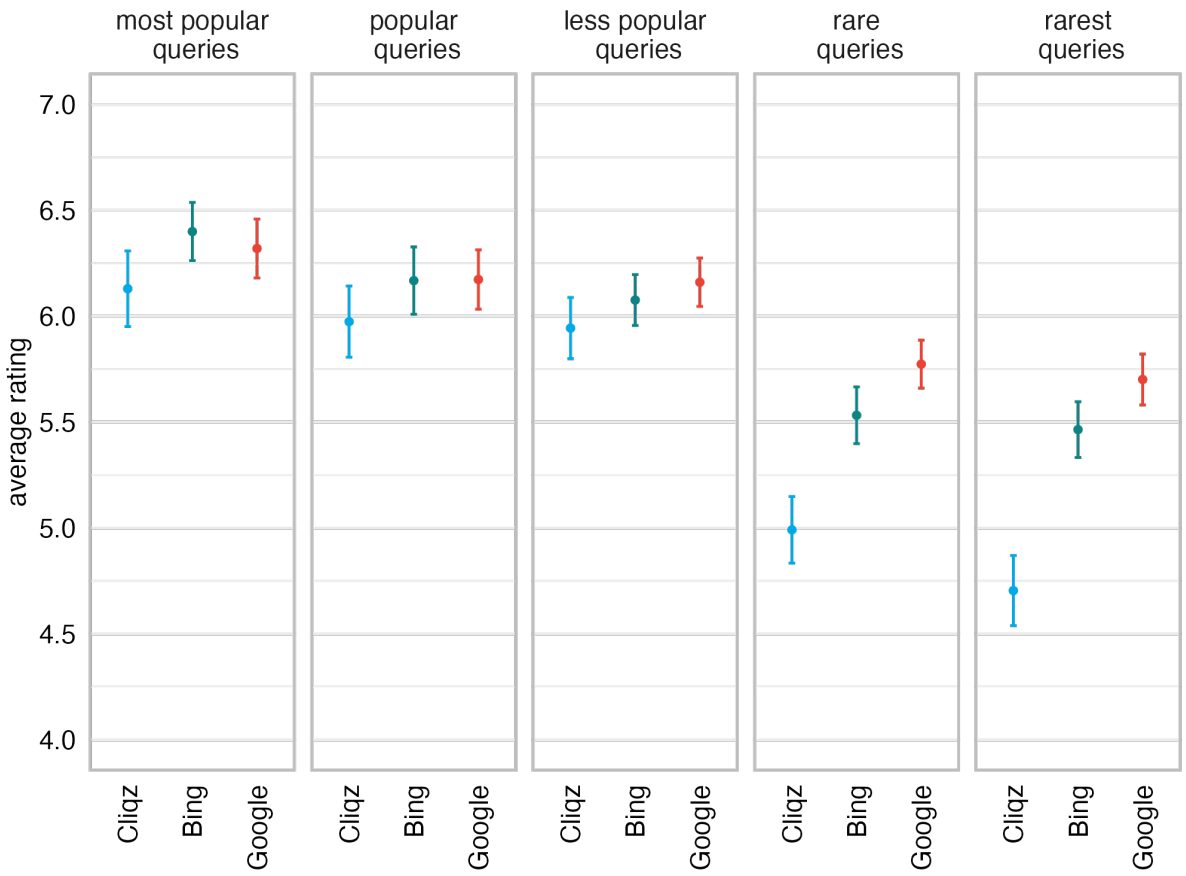
Notes: This figure shows the average quality of the Cliqz results by the amount of data that was used in the experiment (horizontal axis) and by bucket. Whiskers indicate 95 percent confidence intervals. From top to bottom, we report the results for bucket B1 through B5. See Table 1 for details. For each bucket, we held the set of queries fixed and only varied the amount of data that was used to produce the search results. We then used human assessment to evaluate the quality of the search results.

Figure 1: Dependence of search result quality on amount of data



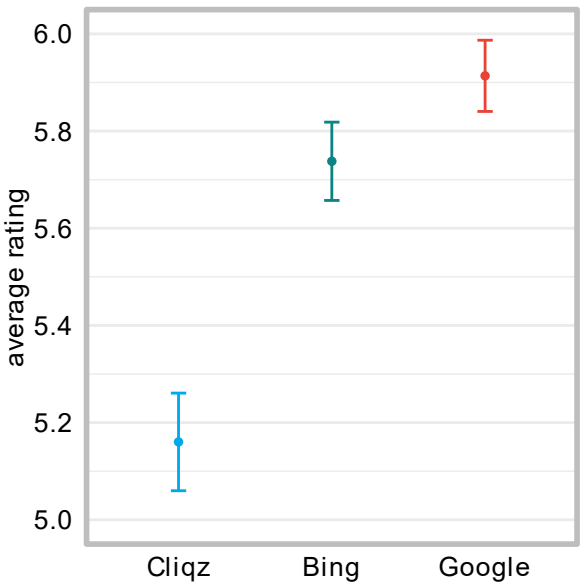
Notes: This figure shows the results from a robustness check in which we re-did the analysis underlying Figure 1 but replaced human assessment with an alternative measure of search result quality. The alternative measure is the average percentage overlap of the respective top-5 search results by Cliqz (in the sense that the set of results is the same without comparing the ordering), for a given fraction of the data, and the results obtained from Google. Whiskers indicate 95 percent confidence intervals.

Figure 2: Measuring search result quality by the overlap with Google's results and repeating the exercise in Figure 1, otherwise



Notes: This figure shows the average quality of the results by Cliqz, Bing, and Google, respectively, by bucket. From left to right, we report the results for bucket B1 through B5. See Table 1 for details. Whiskers indicate 95 percent confidence intervals. We obtained search results for the same representative sample of queries from each of the three search engines. We then used human assessment to evaluate the quality of the search results.

Figure 3: Quality of search results by popularity of the query



Notes: This figure shows the average quality of the results by Cliqz, Bing, and Google, respectively. Whiskers indicate 95 percent confidence intervals.

Figure 4: Average quality of search results