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

Experimental evidence

Tobias J. Klein,[1] Madina Kurmangaliyeva,[2] Jens Prüfer[3] and Patricia Prüfer[4]

3 February 2023

Abstract
Do some search engines produce better search results because their algorithm is better, or because they have access to 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), if it has enough data, and that overall differences in the quality of search results are explained by searches for *rare* queries. This is confirmed by results from an experiment, in which we keep the algorithm of the search engine fixed and only vary the amount of data it uses as an input. Because 74% of the traffic in our data come from rare queries, these are the pivotal dimension of competition, where a search engine must perform well to offer users high quality and gain market share. Our results suggest that a small search engine would be able to produce search results that are of similar quality as Google's, if it had access to the data users generated by using Google in the past. We discuss why data sharing may increase innovation here.

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

[1] Tilburg School of Economics and Management, TILEC, Tilburg University, The Netherlands; and C.E.P.R., London, United Kingdom.
[2] ECARES, Université Libre de Bruxelles, Belgium.
[3] University of East Anglia, School of Economics and Centre for Competition Policy, Norwich, United Kingdom; and TILEC, Tilburg University, The Netherlands.
[4] Centerdata, Tilburg University, The Netherlands.

1. Introduction

Search engines are used by billions of users every day. They offer the basic infrastructure for many other industries and are, therefore, 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.[5] There are two potential explanations for this. The first potential explanation is that Google's *algorithm* to generate search results performs better than the one of its competitors'. The second potential explanation is that Google has collected the most *data* from past users about their preferences and characteristics (called *user information* by Argenton and Prüfer, 2012). 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.[6]

Whether the first or the second explanation holds true is highly relevant for policy-making and regulation. If Google just had the better algorithm, there would not be much reason for antitrust policy to intervene (Varian, 2019, Bajari et al., 2019). Other firms could develop an algorithm that is better and enter the market. Google would find it worthwhile to invest in the algorithm to avoid this from happening. Consequently, consumers would benefit from active competition. However, if Google was dominant mainly because it could exploit more data, it 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

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

[6] It has been shown that a search engine with more users improves its quality faster, acquires data on new queries more quickly, and has more other data to make inferences about users' queries (Halevy et al., 2009, He et al., 2017, Schaefer and Sapi, 2020). Notably, in this paper we study the quality of *organic* results of *general* search engines. Feng et al. (2007) use a complementary approach by studying how the relevance of *sponsored* search results can be improved. Ghose et al. (2012) propose an improved ranking system for *product* search engines.

Google would innovate very little. Therefore, Google would have a much smaller incentive to invest in improving its algorithm. In such a situation, which has been theoretically analyzed by Argenton and Prüfer (2012) and Prüfer and Schottmüller (2021), data sharing of user information among competitors would benefit consumers because it would level the playing field between Google and potential competitors and would, therefore, reinstate incentives to innovate for all search engines.

For two reasons, it is difficult to show empirically whether differences in search engine quality are due to different algorithms or to different amounts of user information algorithms have access. The first reason is access to data. Data are usually proprietary and are not public, which inhibits the ability for external review of empirical results through academic peers or public authorities (Persily and Tucker, 2020). The second challenge is that even proprietary data do not stem from an experiment and, hence, are hard to interpret in a causal way (which would be relevant for policy making) (Lewis-Kraus, 2020). 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). Here, we run such an experiment. We collaborated with a small search engine, Cliqz (based in Munich, Germany). They provided us with non-personalized search results for a representative set of queries for German users in April 2020. 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 offers within-search engine comparisons. 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 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.

The results show that differences in the quality of search results exist in the eyes of human assessors. Google's results are evaluated above Bing's, which are evaluated above Cliqz'. We find, however, that overall differences in the quality of search results are explained by searches for *rare* (or tail) search terms, for which a search algorithm can rely on less data that was produced by earlier users searching a similar query. These results are confirmed by a second, independent metric for search engine quality: the overlap between Cliqz' search results with the market leader's, Google. Crucially, this means that we document that the algorithm of a small search engine *can* produce non-personalized results that are of *similar* quality to Google's, measured by human assessors who do not know from which search engine the results came. However, our results also show that this is only the case if the algorithm has access to sufficient user-generated data, which is not the case for rare queries.

The insights are complemented by results from an experiment, in which we keep the algorithm of the search engine fixed and vary the amount of data it uses as input. This offers causal evidence that more user data on rare queries enables search engines to produce better quality. Crucially, 74% of the traffic in our data represents rare queries. Hence, rare queries are the pivotal dimension of competition, where a search engine must perform well to offer users high quality and gain market share.

Taken together, this suggests that Cliqz would be able to produce search results that are of similar quality as Google's, if it would have access to the data users have generated by using Google in the past. Hence, data sharing would put all firms on a level playing field and would thereby create an incentive for all firms to compete against one another by investing in the algorithm. Policymakers are highly interested in this problem and most of them conclude that existing antitrust law regarding search engines and other platform markets needs to be developed further.[7] Indeed, lawmakers are currently in the process of preparing legislation to regulate dominant firms on platform markets.[8] The EU's Digital Markets Act (DMA), just published officially in October 2022, prescribes that large

---

[7] At least 30 top-level advisory reports about competition in online "platform markets" raise such concerns (Beaton-Wells, 2019), including highly regarded ones in the US (Scott Morton et al., 2019), the EU (Cremer et al., 2020), the UK (Furman et al.; 2019), and Germany (Schallbruch et al., 2019).
[8] See European Union Digital Markets Act (2022), United Kingdom CMA (2020), or US Congress (2021).

“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 is generated by end users (Art 6(11)). This paper provides evidence that entrants in the search engine market are not able to compete with Google without data-sharing of query logs. It, therefore, supports these regulatory initiatives, trying to reinstall a level playing field among competitors to some extent. Unlike in other contexts, this remedy does not directly harm the incumbent, as it makes use of the non-rivalry of information: Google will still be able to use the same data. Only the exclusivity of data access would be reduced. Consequently, users would benefit.

In Section 2, we review the literature and explain the theoretical background. Section 3 explains the experimental setup and data from Cliqz. Section 4 reports the results, whereas Section 5 discusses the results, explains why mandatory data sharing may lead to more innovation, and concludes. All additional findings and technical details about the experiment are in the comprehensive appendix.

2. User-generated data, quality, and mandatory data sharing in internet search

Empirical evidence

He et al. (2017) study scale effects in web search 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). As measure of search engine quality, they use 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. Their central finding is a statistically significant improvement in ranking quality of rare queries on the order of 2-3%, with faster improvement at lower levels of data. They interpret this as follows (p.296): “both providers have CTRs on tail queries in the 70% range. Suppose an entrant could achieve 60% “off the shelf.” Then 2-3% represents more like 20-30% of the meaningful range in which we expect competitors to be differentiated and thus appears quite large in this light.”

Schaefer and Sapi (2020) 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 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).

While benefitting a lot from these earlier studies, our paper makes several additional contributions. First, we develop and use various new measures of search engine quality: Our main analysis stems from *human assessment*, where we asked subjects to identify the best and the second-best search result (URL) among five URLs forming one search result set and rate the entire set on a 1-7 Likert scale. Crucially, all result sets were presented in a standardized way not indicating its origin (Cliqz, Google, Bing, or a mixed set). As a second type of search engine quality measure, we assume that Google’s quality is the gold standard and calculate *similarity scores* about the *overlap of Cliqz’ search result sets with Google’s* on the respective queries. As this can be done automatically, we can use all our data and generate a series of robustness checks of the human assessment results. For both types of measures, we also vary the analysis with a focus on the top-1 search result of a result set (mimicking the CTR) with focusing on the top-5 results, which are usually seen at once by search engine users. Our second main contribution is that we cannot only compare the quality of several search engines (between-search engine design) but that we can actually keep Cliqz’s algorithm and the query fixed and vary the amount of historical search data the algorithm has access to (within-search engine design). This generates unambiguously causal results.

Theoretical background

Calvano and Polo (2021) confirm 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, though. 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.[9] 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 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

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

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 cost of innovation --- and regain its quality lead. Foreseeing this situation, rational entrepreneurs and private financiers would not invest into such a smaller firm in the first place.[10] 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.

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).

3. Data and experimental setup

We collaborated with the small search engine Cliqz, which ordered all search queries submitted by German users of their “Human Web” from April 20 to April 26, 2020, by their frequency.[11] They formed five buckets in line with queries’ frequency using the following thresholds: 0.2%, 1%, 5%, and 25%. This means that, for example, the first bucket represents the top 0.2% of search queries by

[10] 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."

[11] 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.

frequency, while the last bucket represents the bottom 75%. Next, 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 which 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 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 conduct. 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.

Then, 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 in 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 remove 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 truncate each result set to top-5 results only. 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 with 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. 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). The appendix contains details on the characteristics of the assessors, and the instructions we gave to the RAs and the clickworkers.

4. Results

Figure 1 shows that the quality of search results differs across search engines. Based on human assessment results, the market leader Google produces the best results, followed by the number two in the market, Bing. The results produced by Cliqz are substantially worse in the eyes of our human assessors. We would like to remind that the assessors were “blind” with respect to the origin.

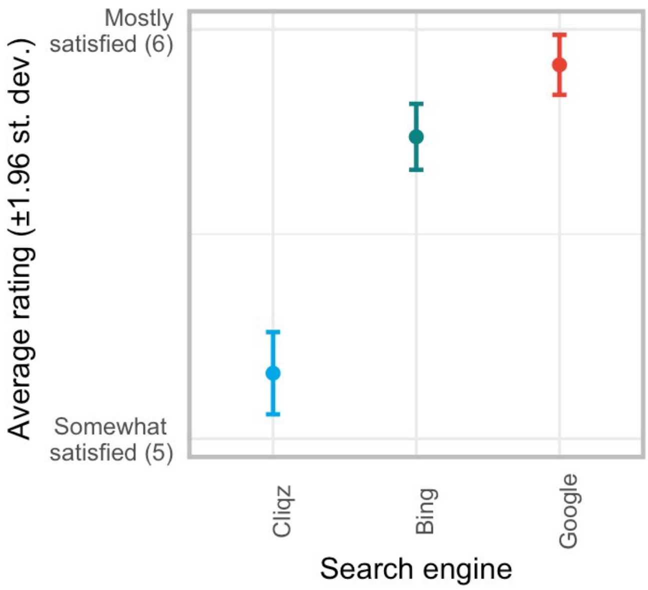


Figure 1: Quality across three search engines.

As pointed out above, there are competing explanations for this. It could either be that Google and Bing simply have better algorithms. Or it could be that they have access to more data. To answer this question, we split the data by the popularity of the query. If there are differences for rare queries but not for popular ones, then this provides a first indication that differences are driven by the amount of data search engines have access to.

We measure popularity by using the defined 5 “buckets” for query frequency. 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) contain 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. Tail queries (in B5) generated 56% of the traffic.

Bucket	% of queries (by design)	av. number searches per week	% of 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%

Table 1: Query buckets

Figure 2 shows that for *popular* queries (B1-3), search results are of comparable quality across search engines. In the light of perceived wisdom (and Figure 1) that Google's quality dominates its competitors' quality, this is already a surprising result: it shows that Cliqz's algorithm is able to 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.

Moreover, it suggests that differences in the overall quality of search results are largely driven by the amount of user-generated data available to search engines for less popular queries. Table 1 shows that, on average, they are searched only 1.2 and 1 times per week, respectively. Apparently, this is not enough to produce sufficient data that can be used to produce high-quality search results. Crucially, buckets 4 and 5 jointly 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.

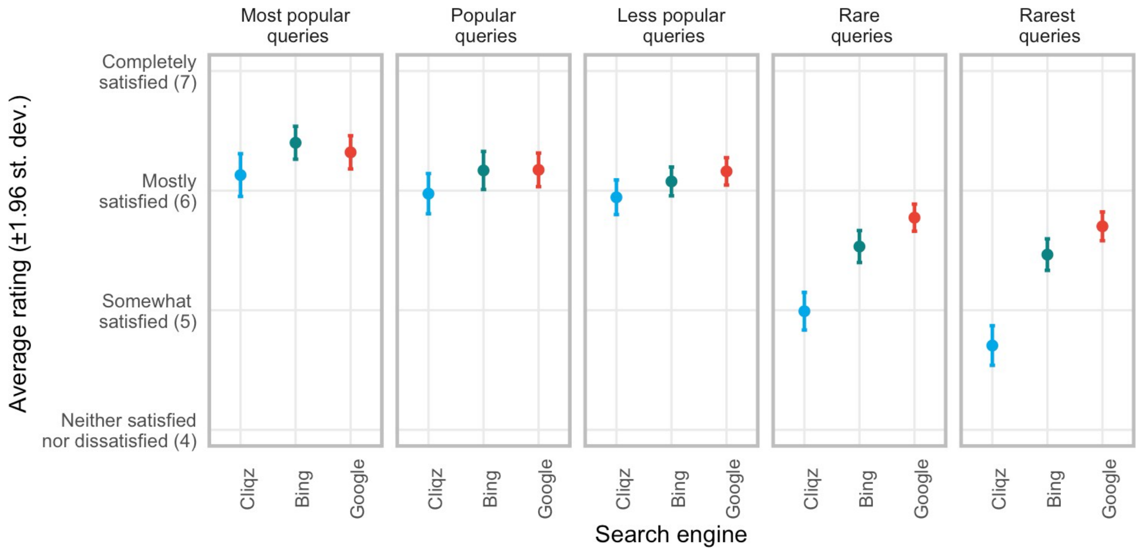


Figure 2: Search result quality across search engines by query popularity (human assessment).

Shown for the 5 buckets in Table 1 (from B1 on the left to B5 on the right).

Next, we report the results from the experiment, where we vary the query log counts. We always use the same algorithm, which yields unambiguous results regarding the impact of more data. Figure 3 shows that, for more popular queries (B1-3), 20% of Cliqz's available data is already enough to produce as much quality as it can (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. This suggests that Cliqz's quality could benefit significantly from access to more search-log data on rare queries.

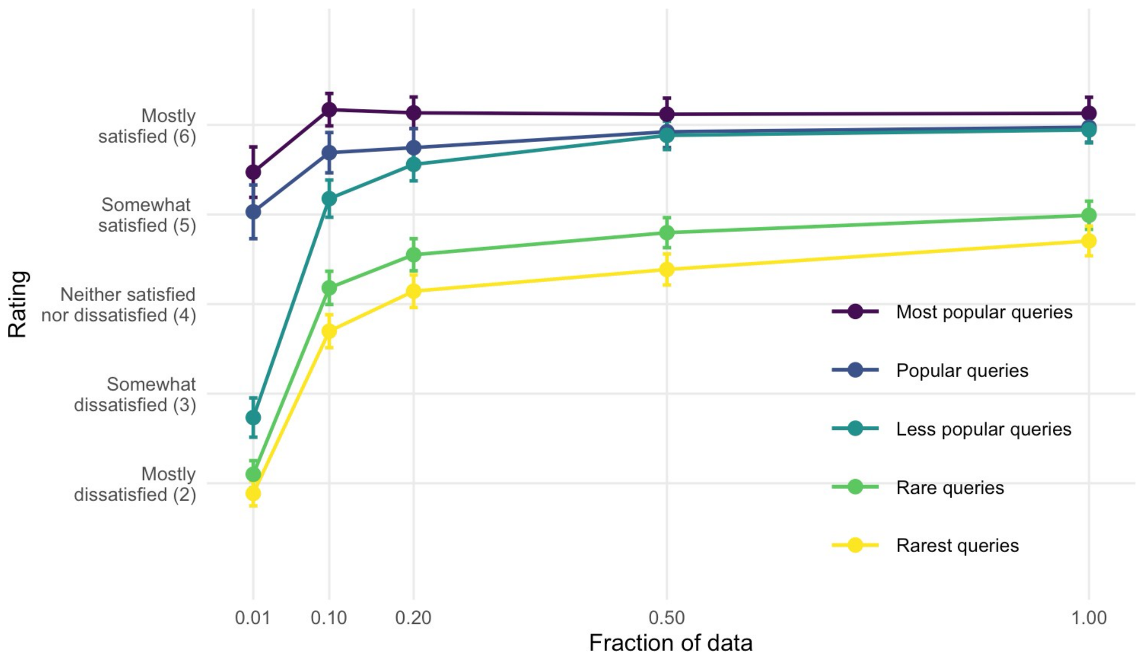
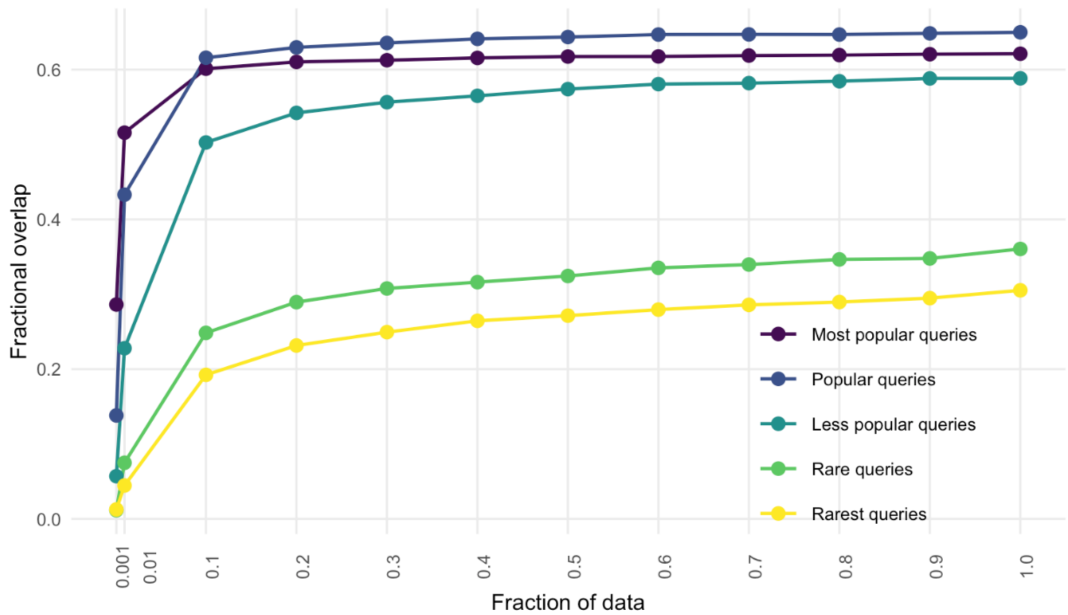


Figure 3: Dependence of search result quality on amount of data (human assessment)

These results are based on between-subject comparisons, pooling 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. Finally, even if there was scope for within-subject learning on what constitutes a good search engine result, because the order of 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 the clickworkers (whose median age was 34 years) were tech-savvy enough to represent an average search engine user.

Finally, we perform a robustness check. In Figure 4, starting from the assumption that Google's search results are the best, we report the overlap between Cliqz's and Google's top-5 results, depending on data available to Cliqz, as measured by similarity scores. This implies that the independent variables in Figures 3 and 4 are the same but that the dependent variables are different (human-assessed quality vs. overlap between Cliqz's and Google's results). Yet, the finding is 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.



.

Figure 4: Overlap between Cliqz' and Google's top-5 results, depending on data available to Cliqz.

5. Discussion and Conclusion

Taken together, we find that differences in the quality of non-personalized search results are to a large extent driven by differences in the amount of past user-generated data that search engines have access to. This is not the case, however, for popular queries. For those, we find that a small search engine can in fact produce results of similar quality to Google's. However, for rare queries, which constitute the better part of the traffic (74% in our data), a small search engine does not have access to enough user-generated data to produce results that are equally good as Google's.

These results suggest that the mandatory sharing of user data is an appropriate remedy: it would allow entrants, such as Cliqz, to successfully compete with the incumbent by enabling Cliqz to provide search results that are also of high quality for rare queries. Unlike in other contexts, this remedy should not harm the incumbent because user data are non-rival. Data-sharing only removes the incumbent's exclusivity advantage to access data access of. Hence, data sharing would improve users' welfare.

Moreover, mandatory sharing of user-generated data would not necessarily reduce the incentives of firms to invest in their services and algorithms in the first place: as the proposal is to only share raw, unprocessed data (such as search logs), which is virtually a by-product of regular business operations in digital markets, all investments in algorithms or data analytics of the incumbent firm would be protected.[12]

By contrast, under today's exclusive data access regime, an incumbent firm is protected from viable competition through the gigantic and constantly renewing amounts of user-generated data, which more efficient competitors with lower market shares do not have access to. This relaxes the incumbent's incentives to spend high innovation costs and secures them continuously flowing high profits. Under a mandatory data-sharing regime, incumbents would have to heavily invest in innovation because otherwise, they would risk being driven out of business by more innovative competitors (and risk all profits).[13] This means that mandatory sharing of user-generated data may well lead to more innovation and not less.

From a broader perspective, data as an input is also important in many contexts other than online search (Mayer-Schönberger and Ramge, 2018). The policy proposal by researchers (Argenton/Prüfer,

---

[12] 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.

[13] See Prüfer and Schottmüller (2021) for details.

2012) and the EU's DMA ask to share *anonymized* user information that cannot identify a specific user.[14] This requirement strongly narrows down the scope of data that can be shared, e.g. when a user is logged into a provider's service or otherwise tracked. However, data with personal identifiers are particularly valuable for search engines to predict what information users are looking for (Schaefer and Sapi, 2020). On the internet search market and other platform markets, incumbent firms have enjoyed very high market shares and very high profits for many years. During these years, they could invest in all kinds of complementary infrastructures, such as human resources, data centers, global internet connections, and software and hardware innovations, which smaller firms that may dare to take up competition could not. This means that they are in a very strong position even if data sharing becomes mandatory. In that sense, levelling the playing field by making the sharing of anonymized user information can be seen as a necessary first step, but not a sufficient condition for competitive data-driven markets.

Notably, shortly after our experiment was conducted, Cliqz announced that they would go out of business (https://cliqz.com/announcement.html).

Materials and Methods

Appendix 1-4 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).

[14] Li and Sarkar (2012) develop a clustering method to preserve privacy in shared data. Li and Qin (2017) show how even highly sensitive and personal data, such as health records, can be anonymized.

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.

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.

Ghose A, Ipeirotis PG, Li B (2012) Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science* 31(3):493-520.

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.

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

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

Li XB, Qin J (2017) Anonymizing and Sharing Medical Text Records. *Information Systems Research* 28(2): 332-352. https://doi.org/10.1287/isre.2016.0676

Li XB, Sarkar S (2012) Class-Restricted Clustering and Microperturbation for Data Privacy. *Management Science* 59(4): 796-812.

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).

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 (2020) Learning from Data and Network Effects: The Example of Internet Search. DIW Working Paper.

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.

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

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

How important are user-generated data for search result quality? Experimental evidence

Supplementary materials

This document contains four appendices. Appendix A provides details on the experiment. Appendix B describes how we used human assessment to rate the quality of search results. Appendix C contains further details on the instructions we gave to the assessors. Appendix D contains robustness checks.

A Details on the Cliqz experiment

Overview. We randomly drew 1,000 queries, respectively, in 5 buckets from the population of queries submitted on the Human Web (see below). Then, we conduct the experiment by obtaining search results for each query at 12 levels of available data on past searches. This leaves us with a data set consisting of 60,000 result sets. We augment this data set with 5,000 result sets from Google and Bing, respectively.

Human Web. 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.

Mozilla, as part of another experiment, installed the Cliqz software extension for a 1% random sample of all Firefox downloads in Germany starting in October 2017.[1] This makes the population of the Human Web users somewhat more representative of the general German population than the population of users of the Cliqz browser.

Sample of queries. In online search, very few queries are searched many times, while many others are searched only very rarely. To account for this, we ordered all queries that were submitted on the Human Web between April 20 and April 26, 2020, by their frequency, from the most popular to those that appeared only once within the week.

[1]See https://www.zdnet.com/article/firefox-tests-cliqz-engine-which-slurps-user-browsing-data/ and https://0x65.dev/blog/2019-12-03/human-web-collecting-data-in-a-socially-responsible-manner.html.

Table A.1: Example of query logs

query	clicked URL
google	http://www.google.com
wnmu	http://www.wnmu.edu
ww.vibe.com	http://www.vibe985.com
www.accuweather.com	http://www.accuweather.com
weather	http://asp.usatoday.com
college savings plan	http://www.collegesavings.org
pennsylvania college savings plan	http://www.patreasury.org
pennsylvania college savings plan	http://swz.salary.com

Notes: Taken from Cliqz blog 0x65.dev and AOL query logs dataset.

Then, we formed five buckets using the following thresholds: 0.2%, 1%, 5%, and 25%. This means that, for example, the first bucket represents the top 0.2% of search queries by frequency, while the last bucket represents the last 75%. Next, we randomly drew 1,000 queries from each of the 5 bucket, leaving us with a stratified sample of 5,000 queries.

Index, query logs, and query log counts. Like other search engines, the Cliqz search engine relied on two main input components. The first one is their own index of webpages, which is generated by crawling the web to maintain the up-to-date directory of all webpages.

The second input is the data on user-generated query logs, i.e., actual user queries linked to the URLs they clicked on. Table A.1 provides an example of several query logs. These data are useful because past choices of users might be predictive of future choices. Hence, the search engine may want to put the most clicked result in the past at the top of the new search results.

Query logs are aggregated into query log counts. These say how many times a given URL was clicked by users who searched for a given query. These are the raw data.

Starting from those, Cliqz performed semantic analysis to also use information from its own index of webpages. This allows Cliqz to use the data more efficiently. For example, if someone searches for “Lady Gaga best hits”, the search engine also uses the query log counts from other similar queries such as “Lady Gaga best songs” or “Lady Gaga hits”.[2] This is held fixed in our experiment, in the sense that the algorithm is not re-trained when less data is available.

The experiment. The experiment with the Cliqz search engine was conducted in the evening of April 27, 2020. For each of the 5,000 sampled queries, Cliqz obtained search results at different fractions of the query log counts. Thereby, we simulate the counterfactual search engine results at different availability of user-generated data.

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.

Table A.2 provides an example. The left part shows the query log counts for full data (top) and half of the data (bottom, obtained in the way that was described above). We can see that the query log counts for all but two URL’s are lost when we move from the top to the bottom panel. Search results for a new query, query 4, are generated for the

[2]More details of how Cliqz search engine works are at https://0x65.dev/blog/2019-12-06/building-a-search-engine-from-scratch.html.

full data (top right) and half of the data (bottom right), using the respective query log counts for query 1, 2, and 3. We can see that for half of the data, URL3 and URL4 are not anymore part of the results, as the query log counts for those URL's become zero (bottom left vs. top left). Also, we can see that the order of search results is affected.

Search results. For each query and each level of data, the Cliqz algorithm produces two sets of search results, each consisting of a set of ranked URLs: organic search results and news search results. For example, at the time of the experiment, searches for "Kim Jong-un" – the supreme leader of North Korea – were popular due to rumours of his death. Hence, the Cliqz search engine provided two sets of results: the URL links to the latest news about Kim Jong-un, and general results that are not related to the news, i.e., his wikipedia page. Our analysis focuses only on the organic search results, disregarding news search results.

Data used for analysis. Cliqz provided us with 60,000 search result sets, one for every query at every different fraction of query logs.

We also collected search result sets from Google and Bing for the same 5,000 queries. For this we used the application programming interface (API) of a for-pay service called SerpAPI (see https://serpapi.com/), for http://www.google.com and http://www.bing.com. The API allowed us to specify that we would like to obtain results for users from Germany.

Table A.2: Example of removing 50% of user data and its effect on search results

query	URL	count (100%)
query 1	URL 1	5
query 1	URL 2	1
query 1	URL 3	1
query 2	URL 2	5
query 2	URL 3	1
query 2	URL 4	1
query 3	URL 4	1

algorithm $\Rightarrow$

query	search results
query 4	1. URL2
	2. URL1
	3. URL3
	4. URL4

query	URL	count (50%)
query 1	URL 1	2
query 1	URL 2	0
query 1	URL 3	0
query 2	URL 2	2
query 2	URL 3	0
query 2	URL 4	0
query 3	URL 4	0

algorithm $\Rightarrow$

query	search results
query 4	1. URL1
	2. URL2

Notes: This table illustates how the algorithm generates search results at full data and at half the data. The search results on the right are for a new query. Removing data affects the search results because it affects the query log count.

B Human Assessment

Why not click-through-rate? Other papers, for instance (2) and (3), use the click-through-rate (CTR) as a measure of quality. The CTR is the likelihood that a user clicks on one of the search result of a given set. We do not use this measure for two reasons. First, we would like to make comparisons across search engines and do not have access to data on CTR's from Google and Binck. Second, we create artificial search results in our experiment, which were never shown to actual users. For this reason, there are no data on CTR's. In principle, Cliqz could have shown the results to randomly drawn users and record the CTR, but did not want to do so, because it would lower their user experience.

Sample of queries. Recall that our data set consists of 60,000 result sets for Cliqz (5,000 for 12 levels of data) and 5,000 result sets each for Google and Bing (only full data, as we did not have the opportunity to conduct an experiment with them).

Since human assessment is costly, we use a random sample of the 5,000 sampled queries for evaluation. We restrict attention to queries which are either in German or in English and that are at least 3 characters long. Then, out of 3,918 queries, we sample 500 randomly: 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. After sampling, we remove 7 queries with inappropriate content, resulting in 493 queries for human assessment.

Top-5 results and mixed result sets. Previous studies (30,31) show that search engine users usually look only at the results that appear at the top of the result list. In order to reduce the load on the assessors, we therefore restrict the result sets only to top-5 results.

Additionally, for each sampled query, we construct a "mixed" result set from Google, Bing, and Cliqz (at full data) result sets, using the following algorithm:

1. Assign order: randomly map Google, Bing, and Cliqz result sets to set_1, set_2, set_3;
2. Pop the first element: add the first URL (i.e., result) of set_1 to the mixed result set, and remove that URL from set_1, and also from set_2 and set_3, if those sets also contain that URL;
3. Rotate the order: make set_2 to be the new set_1, set_3 to be the new set_2, and set_1 to be the new set_3;
4. Repeat steps 2 and 3 until the mixed set has 5 elements;
5. Shuffle the mixed set: randomize the order of results within the mixed set.

By randomizing the order with which Google, Bing, and Cliqz result sets contribute to the combined mixed set, we ensure that all search engines get an equal chance to contribute to the mixed set for each query. For example, if all three result sets – Google, Bing, and Cliqz – are distinctly different, the union of the top-5 results will give 15 results in total. However, the mixed set is limited for 5 results only. Hence, those search engines that have been chosen to be the first two to contribute to the mixed set contribute two results each, while the last one will contribute only one. But which search engine is chosen to be the first is random, therefore, the mixed sets on average provide equal opportunities to every search engine. By randomizing the final order of the results in the mixed set, we also remove any residual correlation in the positions of results supplied by the same search engine in the mixed set.

Assessors. In order to measure the quality of these result sets, we asked human assessors to rate their satisfaction with the search results on a seven-level Likert scale for a random sample of queries. 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. One of the research assistants received all the result sets corresponding to queries in German language: and another, 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 2,848 result sets was evaluated on average by four different people (one RA and three clickworkers). Appendix C contains details on the instructions we gave to the RA's and the clickworkers.

In general, individual assessments of the same result set might differ from person to person, which will generate noise. However, since the assessors were unaware about which search engine had generated the results, we expect this noise to be unsystematic and to vanish for the average assessment.

Assessment. For each result set, human assessors was asked to rate the quality of the result set on a scale from 1 to 7, where 7 means "completely satisfied", 4 is "neither satisfied, nor dissatisfied", and 1 means "completely dissatisfied, as if no results." See Table B.3. The assessors were explicitly asked to take the order of the results into consideration when rating the result set.

As an alternative measure of quality, we also asked human assessors to pick the best and second-best results within each result set. The assessors could choose an option "None of the above", in case they find none of the results satisfactory. Although we collected the choices of the best and second-best results for all result sets, we were interested mostly in their choices within mixed result sets. The idea is that the assessors were not aware

Table B.3: Likert scale

value	description
7	completely satisfied
6	mostly satisfied
5	somewhat satisfied
4	neither satisfied, nor dissatisfied
3	somewhat dissatisfied
2	mostly dissatisfied
1	completely dissatisfied, as if no results

Notes: This table shows the Likert scale we used for human assessment by the RA's and the click-workers.

about the fact that they were evaluating a mixed result set. We use this to conduct a robustness check in Appendix D, where we measure which search engine produces the best result by looking at the fraction of times the best rated result from the mixed result set was produced by that search engine.

Presentation. The assessors were shown the result sets simulating a browser experience, where each result showed not only the URL itself, but, in most cases, also the title and the snippet of the page (Figure C.1). The titles and snippets for Google and Bing results were provided directly by the API we used to obtain them (see above). For Cliqz results, we directly copied the title and snippet from Google or Bing results if those results also contained that URL. In this way we recovered titles and snippets for 2,166 out of 3,846 unique URLs in Cliqz results. For the remaining 1,680 URL, we queried those URLs to Google API and scraped titles and snippets provided by Google to those URLs. This helped to find titles and snippets for 1,512 URLs, leaving just 168 URLs without a match. The remaining URLs were mostly web-pages which no longer existed. We kept those 168

Table B.4: Empty result sets out of 493 queries used for human assessment.

search engine	fraction of user data	empty result sets in all queries number	empty result sets in all queries share	empty result sets in rare and rarest queries number	empty result sets in rare and rarest queries share
Cliqz	1%	250	0.51	195	0.40
Cliqz	10%	64	0.13	54	0.11
Cliqz	20%	42	0.09	36	0.07
Cliqz	50%	28	0.06	27	0.05
Cliqz	100%	16	0.03	15	0.03
Google	100%	0	0.00	0	0.00
Bing	100%	0	0.00	0	0.00

Notes: This table shows the number and fraction of empty result sets in all queries (third and fourth column) and the number and fraction of empty result sets in rare and rarest queries (fifth and sixth column). Rare and rarest queries are from bucket 4 and 5, respectively.

URLs in the result list, asking human assessors not to penalize the result simply for the absence of the title and snippet. We discuss the potential influence of the missing titles and snippets on the ratings by human assessors in the robustness checks in Appendix D.

In total, there were 3,944 result sets to be evaluated: i.e., 493 queries times eight result sets per query (mixed, Google, Bing, and Cliqz at five different fractions). However, 400 out of those results sets were empty: i.e., a search engine did not provide any result to the query. Unsurprisingly, empty result sets mostly occurred for rarer queries and at lower fractions of user data (See Table B.4). Moreover, out of 3,544 non-empty result sets, 696 result sets were duplicates, so we did not need to evaluate them again. The duplicate result sets are those that have the same set of URLs in exactly the same order as an already evaluated result set. Overall, there were $2,848$ result sets to be evaluated by the human assessors, net of duplicates and empty sets.

Table B.5: Number of assessments

evaluator	unique	with duplicates	with duplicates and zero sets
click workers	8,544	10,632	11,832
research assistant 1 (DE)	1,544	1,901	2,301
research assistant 2 (EN)	1,304	1,643	2,043
total	11,392	14,176	16,176

Notes: This table shows the number of assessments by type of evaluator. Duplicate result sets have the exact same search results in the same order. Zero result sets are empty.

We decided not to remove empty result sets from our analysis, as it would bias severely our results. We believe that the fact that the Cliqz search engine struggled to provide results at lower fractions of user data or for rarer queries is in itself a sign of deteriorating quality. Hence, even if the actual empty result sets were not evaluated by the assessors to save costs, we restored empty sets for our analysis by imputing the lowest rating of 1 for them. We used the “as if no results” wording for the lowest rating, in order to anchor all the other ratings by the assessors with respect to empty result sets. In robustness checks, we show that our results remain qualitatively similar even if we restrict attention to non-empty result sets only.

In total, we received 11,392 assessments of result sets (without duplicates and empty sets). Then we restored evaluations for duplicate result sets and imputed evaluations for zero sets per each worker, resulting in 16,176 evaluations ready for the analysis. See Table B.5 for more details about the sample size per each type of the evaluator.

Table B.6 provides summary statistics for the evaluations using only unique result sets (without duplicates or empty result sets). We split the answers by the type of the assessor

Table B.6: Summary statistics for ratings

evaluator	lang	median rating	mean rating	shr of 7	shr of 1	shr of no best URL	n obs
clickworker	de	6	5.42	0.33	0.04	0.06	4,632
DE RA	de	6	5.21	0.40	0.11	0.11	1,544
clickworker	en	6	5.35	0.29	0.04	0.05	3,912
EN RA	en	7	6.02	0.53	0.03	0.06	1,304

Notes: This table shows summary statistics on ratings by type of assessor and language. The column headers use the following abbreviations: lang: language of the query, shr: share, shr of 7: share of "completely satisfied" ratings, shr of 1: share of "completely dissatisfied, as if no results at all" ratings, no best URL: the evaluator decided that no result in the result list is satisfactory, n obs: total number of evaluations.

and also by language of the query to facilitate comparison. Overall, the distribution of ratings seem to be broadly in line with each other by different assessors, although clearly there are certain idiosyncrasies. In robustness checks, we discuss relative merits of answers by research assistants relative to answers by clickworkers and show that our result remain qualitatively the same independent of which type of assessors we use.

C Instructions for human assessors

C.1 Instructions to research assistants

The text below contains the instructions that were given to the research assistants (university students):

Here are the detailed instructions for your RA-task:

1. You are asked to evaluate 1,544 result sets provided by a search engine.
2. Please, click on the following link: https://madinak.shinyapps.io/assessment_app_mag/
3. You will see a field which asks you to put the result list with which you want to start your evaluation. Choose result list #1 and proceed in chronological order. You would see a webpage like in an example below: [Figure C.1 was shown here]
4. Each result set consists of a search query term (highlighted with red rectangle in the picture above) and up to five results (blue rectangle), where each result usually includes a URL link to a website and a short description of that website. For example, the picture above represents results provided by a search engine to someone who was searching for "ptgui". The search engine provided five results. The first result, for example, is a company page https://www.ptgui.com/. The fifth and last result is a webpage which allows to download ptgui software within https://www.giga.de.
5. You are asked to do three things for each result set:
   (a) Evaluate how satisfied you are with the results for a given query overall on a scale from 1 to 7 from a drop down menu (see light green rectangle), where 1 would be equivalent to a situation when the search engine does not give you

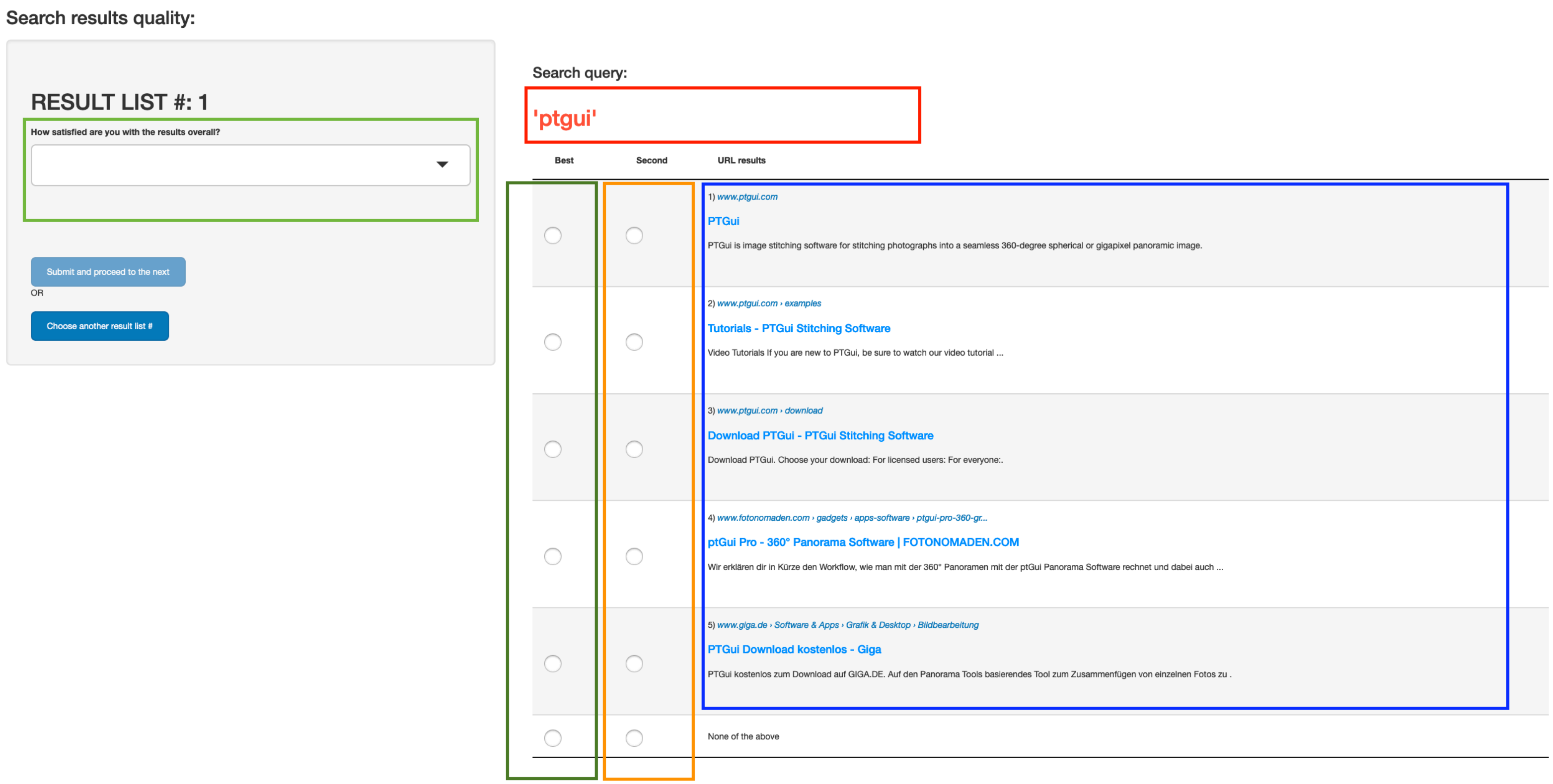


Figure C.1: Example of a web page with search results used for human assessment any results and 7 means that you are extremely satisfied with the result. Please, evaluate the quality of the result set as if you are really searching for the answer. For example, as a search engine user you want the relevant information to appear first in the search results, and less relevant — later. So, please take the order of the results into account when evaluating overall quality.

(b) Among the results provided by the search engine, please choose the result that answers the query the best. You need to click on the radio button in the first column of radio buttons (see the dark green rectangle) in the row that corresponds to the result you have chosen as the best. If you think that none of the results provided by the search engine answer the query well, you can always choose “None of the above” by clicking the radio button in the last row. Please, click on the URL links, if brief descriptions are not enough to give you an idea about each website. Of course, sometimes it is clear without clicking, but sometimes it is not.

(c) You also need to choose the second-best result, by clicking on the radio button in the second column of radio buttons (see the orange rectangle) in the row that corresponds to the result you have chosen as second-best. You can also choose “None of the above”.

6. Note that you cannot simultaneously choose the same result as best and second-best (except for, of course, the “None of the above” option).

7. After you have selected the overall rating of the results, the best, and the second-best result, you should push the submit button which will automatically load the next result list in chronological order. If you made a mistake and you want to return

back to some of the result sets you have already evaluated, you can always manually choose the result set number by clicking “Choose another result list #”.

8. Note that you may see that some queries may be repeating, but result sets are different. This is on purpose.

9. Also, sometimes there will be fewer than five results in a result set.

10. If there is only one result in the result set, please, choose “None of the above” as best and second-best result.

11. I expect that on average you will spend around 1 minute on evaluating one result set. Of course, there will be queries which will be harder to understand, for which you will have to click on every link and explore the results better. As for example, the example result set’s query on “ptgui”. If you have never heard about such software, you would need more time to click on the URL links in the result set, to get acquainted with the concept. However, there will be queries which are straightforward for you (some common knowledge popular queries). So those will not take much time to evaluate. Moreover, as many queries will repeat from time to time, the process should go faster than at the beginning.

12. Also, sometimes some results will not have a brief description under the URL. It will say ”(Description not available)”. This may happen at random. Or this may happen because the web-page no longer exists (the result lists have been collected several months ago). Please, do not penalise such results, this is not the search engine’s fault. Rather try to infer whether it was a valid result or not. You are encouraged to click on those URLs.

13. In general, do not hesitate to click on the links if you want to understand more the context of the query and the results.

14. Note that the result sets are real results by a search engine for a random sample of queries people search on the internet. I filtered out inappropriate content, however, should you still find any inappropriate queries and/or URL results, please skip that result list and let me know the number of the problematic result set, so I would know the reason you skipped.

15. When you want to make a break in your work, please write down the number of the last result list for which you completed the evaluation, and continue with the next one after the break.

16. You have 3 weeks to finish the evaluations, i.e., by July 15. Please let me know when you start evaluations, so we can cross-check for the first few evaluations that the app works as intended.

C.2 Instructions to clickworkers

The text below contains the instructions that were given to people hired through the clickworker.com platform to perform the assessment.

> Please decide how good search results match a search term.
>
> We will show you up to 5 results.
>
> Important:
>
> - If you are not sure how good a result matches the query please follow the link.
> - Please keep in mind that the order of results is also relevant for the quality of results.

- If titles or snippets are missing do not evaluate the results. Only judge the results that are visible.

D Robustness

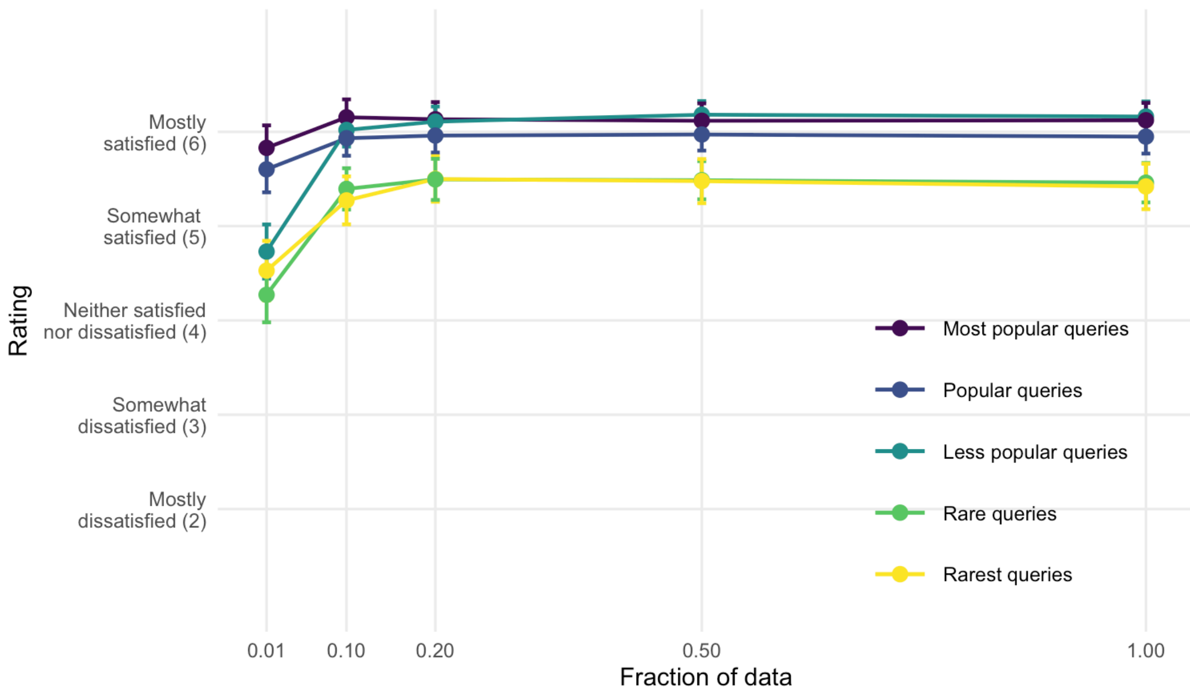
Human assessment of search results shows that if we reduce the amount of user data available to the search engine algorithm, users find that the quality of search results becomes worse, especially for rare queries. In this Appendix, we present additional robustness checks. First, we keep our preferred measure of quality – the average rating on the Likert scale – and show that the main result holds even if we remove empty result sets from the analysis. We also show that the result holds independent of the identity of the assessor. Finally, we show that the result holds if we use alternative measures of quality, whether coming from human assessment or through automated comparison of the overlap with Google results.

D.1 Empty result sets

In Table B.4 of Appendix B, we showed that the incidence of empty result sets was increasing for rarer queries and for lower fractions of user data used to generate the results. In other words, as data available to Cliqz search engine became scarcer, the search engine found it harder to provide search results. At full data, the Cliqz search engine failed to return any results for 3% of queries, while at 1% of user data, it failed for half of the queries.

In our analysis, we assumed that such empty result sets should receive the lowest quality rating of 1 (and we anchored the rating scale by explicitly stating that a rating of one is "as if no results at all"). Here, we assess whether our main results still hold even if we only use the queries that always generated non-empty result sets at all five levels of user-data availability. Indeed, Figure D.2 shows that human assessors give lower ratings to result sets generated at lower fractions of data. The lines in Figure D.2 are now less steep in comparison to Figure 3 of the main analysis, since the average ratings in Figure

Figure D.2: Average ratings as function of query popularity and user-data availability: queries with no empty sets at any fraction of data



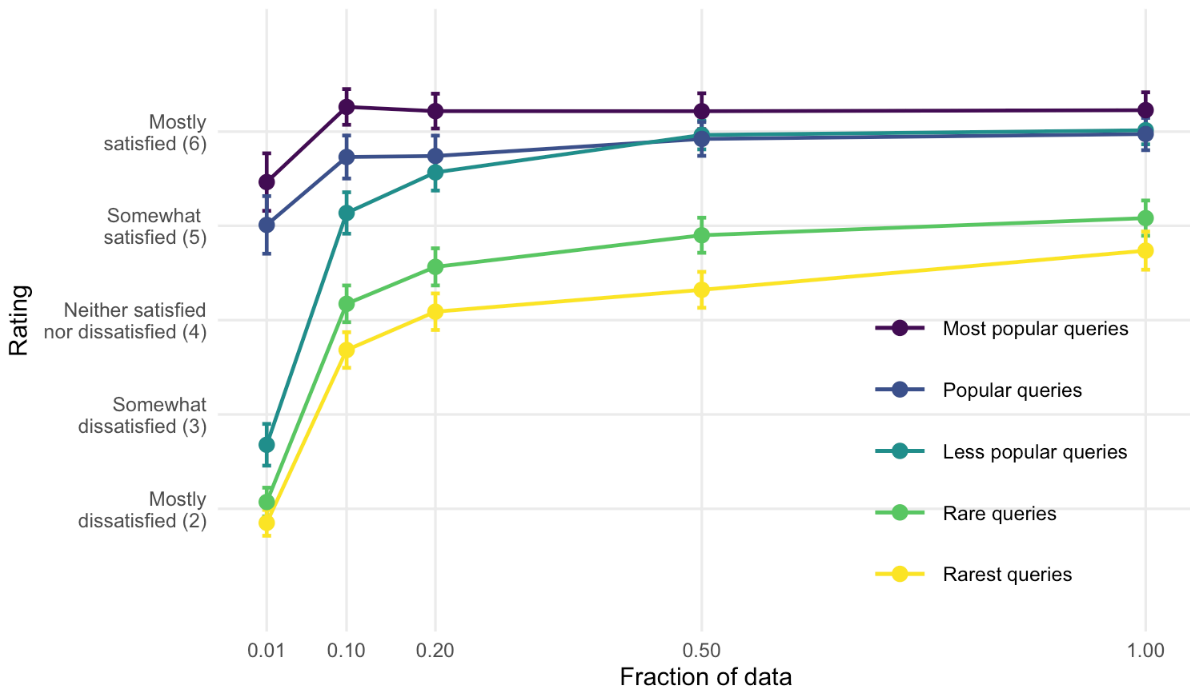
Notes: This figure is based on the sample of queries for which Cliqz was able to generate a non-empty result set at all five levels of data availability: in total, 4,860 assessments for for 243 queries (by popularity: 47 queries in most popular, 44 in popular, 51 in less popular, 57 in rare, and 44 in the rarest bucket).

D.2 are now vastly overestimating quality at lower user data fractions. Nevertheless, it is reassuring to see that the main conclusions hold even in the restricted sample.

D.2 Missing snippets

By trying to construct as natural a web-browser experience for the assessors as possible, we framed each URL result with a corresponding title and a snippet as it is usually represented on the web pages of search engines. However, as we noted earlier, 168 out of 3,846 unique URLs in Cliqz results did not have a matching snippet. It meant that 1,020

Figure D.3: Average ratings as function of query popularity and user-data availability: no missing snippets



Notes: This figure is based on 9,240 assessments for Cliqz result sets for 485 queries at 5 different levels of data availability (see Appendix B for details). Result sets with no missing snippets.

result sets of Cliqz (out of 10,260) were visually distinct since some results had incomplete snippets. Figure D.3 shows that the main result remains unchanged even if we remove result sets that contained missing snippets, suggesting that our results are not driven by slightly different representation of search results across search engine sources.

D.3 Differences across types of assessors

The main analysis pools answers from research assistants and clickworkers. Here, we discuss the implications of this and show that this does not affect our conclusions.

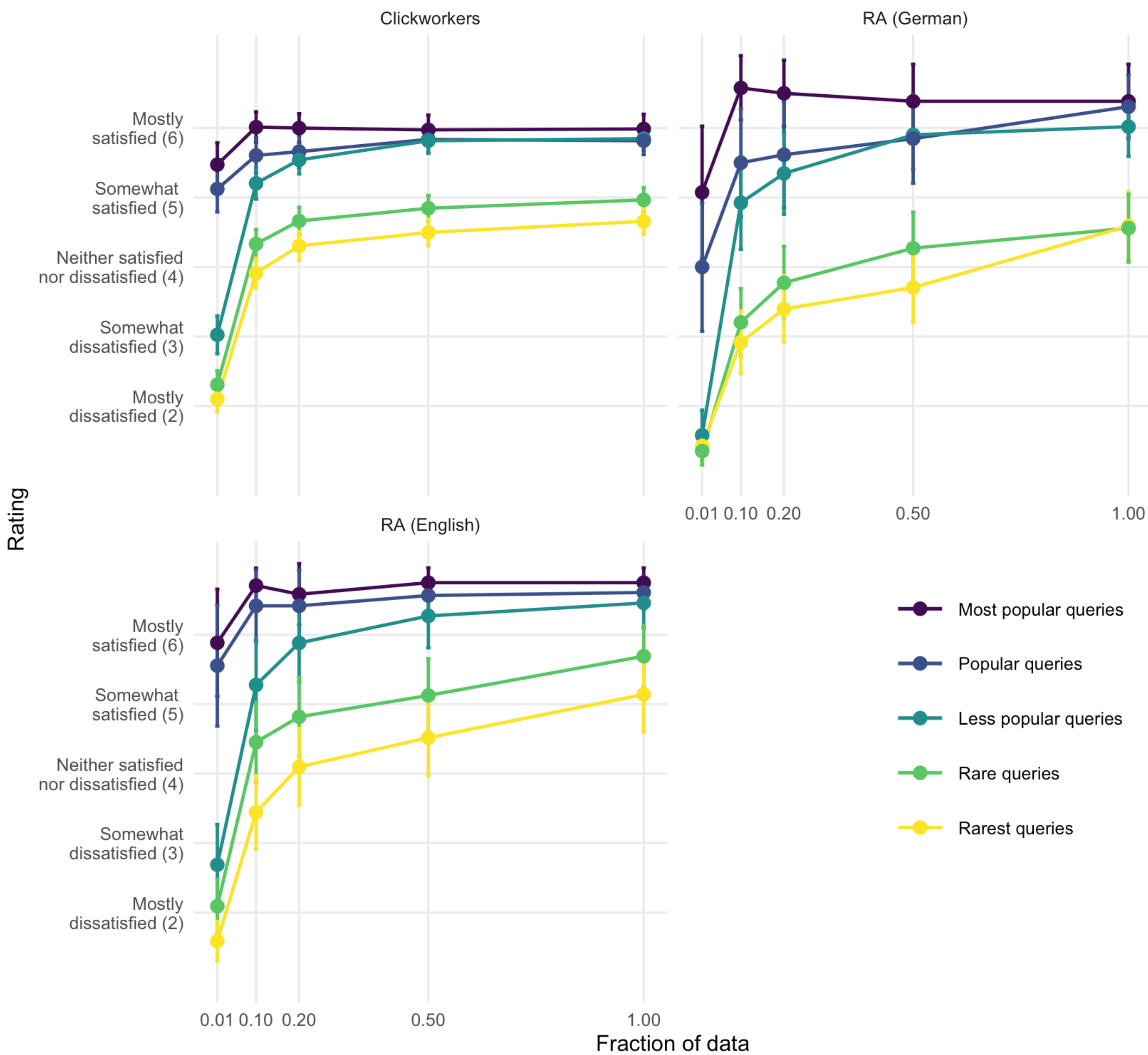
For two main reasons, we expect the assessments by the two research assistants to be more consistent. First, they evaluated more than 1,000 result sets, which provided them with the opportunity to learn what good result sets look like. Second, they first evaluated all the mixed sets (in random order), and only afterwards they were given all the original result sets (also in random order); mixed result sets were more likely to be of good quality.

A clickworker, on the other hand, did not see enough result sets to develop more experience in the given task, so the scope for their learning was limited. Thus, the ratings by each individual clickworker is expected to be noisier than the ratings by the research assistants. We believe that it is valuable to use the ratings that were provided by the clickworkers, as they represent a broader population, with the age ranging from 18 to up to 90 years old and a median age of 34 years. Moreover, the sheer number of evaluations is helpful to reduce noise. Therefore, the ratings by clickworkers might be more representative of a general German population than the ratings by the research assistants.

Finally, since the order of result sets were completely random, in expectation, learning or the absence of it should have impacted all buckets and all search engines results equally. The noise should make it difficult to find any difference at all. If despite all the noise, we still observe that human assessors rate certain types of result sets systematically higher than the other, it must be due to the fact that they are of higher quality.

To assess this, we show the results by type of assessor. Figure D.4 shows that qualitatively the results do not change if we use answers of one group of assessors or the other. In other words, the main result holds independent of the type of the assessor.

Figure D.4: Average ratings as function of query popularity and user-data availability: by type of assessor



Notes: This figure shows the average ratings of Cliqz result sets separately for each type of human assessors. This figure is based on 10,260 assessments for Cliqz result sets for 493 queries at 5 different levels of data availability (see Appendix B for details).

We also conducted a regression analysis in order to control for assessor fixed effects and thus take into account only variation of ratings within the answers of any given assessor. We also account for query fixed effects.

We use the answers on the Likert scale as the dependent variable and estimate the changes in user satisfaction for 24 groups of result sets (5 buckets at 5 different fractions minus one baseline group, which is the group of most popular queries at full data). We fit the linear model

$$y_{iqf} = \alpha_i + \sum_{b=1}^{5} \sum_{f=1}^{5} \beta_{b,f} I\{q \in b, f\} + \delta_q + \varepsilon_{iqf}, \tag{D.1}$$

where y_{iqf} is the rating assessor i ($i \in \{1, \ldots, 565\}$, i.e., 563 clickworkers plus two research assistants) provided for query q ($q \in \{1, \ldots, 493\}$) when fraction f of the data was used. α_i is an assessor fixed effect. $\beta_{b,f}$ are bucket-specific effects of the fractions of data used. Technically, each query q is in bucket b; we use this to construct indicators $I\{q \in b, f\}$ for bucket-fraction combinations that we use as regressors. δ_q is a query fixed effect and ε_{iqf} is the error term. We normalize $\beta_{b,f}$ to be zero for the most popular queries at full data. Given this, the parameters $\beta_{b,f}$ are the difference in the ratings between the group of result sets in bucket b at fraction f and the baseline group of result sets (most popular queries at full data). Reported standard errors are clustered at the assessor level.

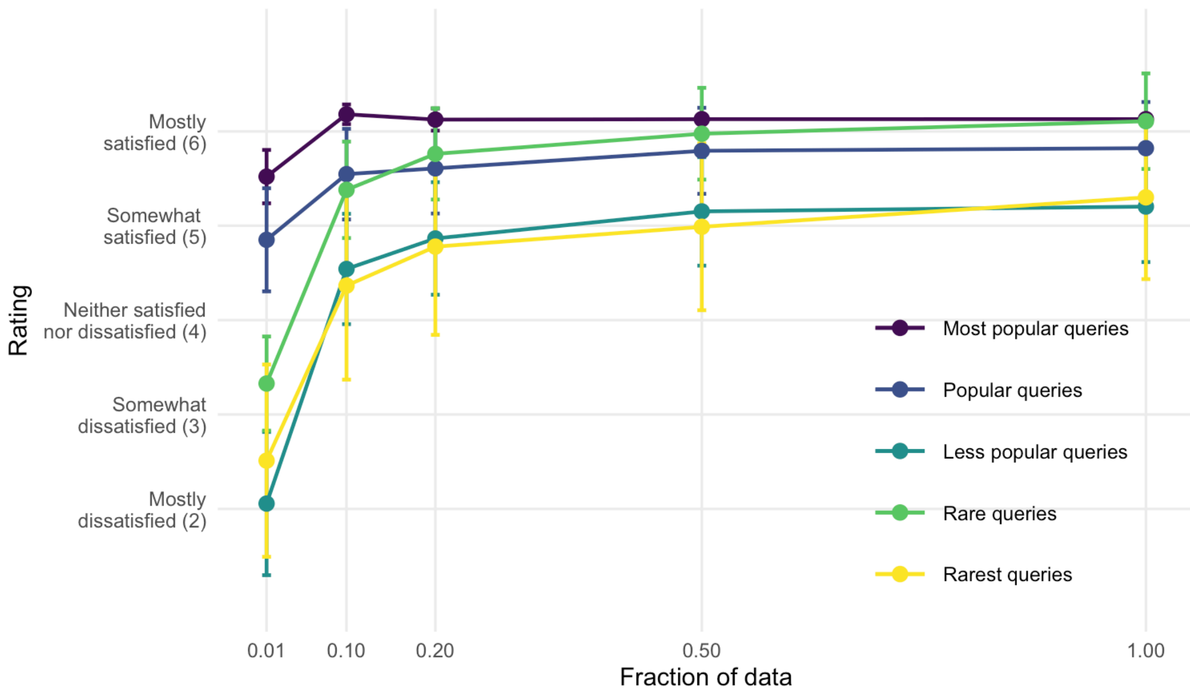
Table D.7 provides the results. For instance, the coefficient estimate -0.52 for "popular; fraction = 0.2" means that the average rating was 0.52 less for popular queries with 20 percent of the data, as compared to the average rating for the most popular queries under full data. Based on the regression results, Figure D.5 again plots the predicted average ratings and shows that the lines have similar shape to the ones in Figure 3 in the main text.

Table D.7: Average ratings as function of query popularity and user-data availability

bucket × fraction	est.	s.e.	t-stat	p-val
most popular; fraction = 0.5	-0.00	0.02	-0.01	0.99
most popular; fraction = 0.2	-0.00	0.06	-0.08	0.94
most popular; fraction = 0.1	0.05	0.05	0.97	0.33
most popular; fraction = 0.01	-0.61	0.14	-4.22	0.00
popular; fraction = 1.0	-0.31	0.25	-1.23	0.22
popular; fraction = 0.5	-0.34	0.23	-1.44	0.15
popular; fraction = 0.2	-0.52	0.24	-2.14	0.03
popular; fraction = 0.1	-0.58	0.24	-2.38	0.02
popular; fraction = 0.01	-1.28	0.28	-4.59	0.00
less popular; fraction = 1.0	-0.93	0.30	-3.09	0.00
less popular; fraction = 0.5	-0.98	0.29	-3.33	0.00
less popular; fraction = 0.2	-1.26	0.30	-4.16	0.00
less popular; fraction = 0.1	-1.59	0.30	-5.32	0.00
less popular; fraction = 0.01	-4.07	0.39	-10.52	0.00
rare; fraction = 1.0	-0.02	0.26	-0.08	0.93
rare; fraction = 0.5	-0.15	0.25	-0.62	0.53
rare; fraction = 0.2	-0.37	0.25	-1.49	0.14
rare; fraction = 0.1	-0.75	0.26	-2.87	0.00
rare; fraction = 0.01	-2.80	0.25	-11.01	0.00
rarest; fraction = 1.0	-0.83	0.44	-1.87	0.06
rarest; fraction = 0.5	-1.14	0.45	-2.53	0.01
rarest; fraction = 0.2	-1.35	0.48	-2.83	0.00
rarest; fraction = 0.1	-1.76	0.51	-3.47	0.00
rarest; fraction = 0.01	-3.62	0.52	-6.96	0.00

Notes: This table reports results from a regression of ratings on bucket times fraction of available data indicators. Based on 10,260 assessments for Cliqz result sets for 493 queries at different levels of data availability (see Appendix B for details). Standard errors are clustered at the assessor level.

Figure D.5: Regression results: Average ratings as function of query popularity and user-data availability



Notes: This figure shows the predicted ratings for Cliqz result sets at different fractions of data. Based on estimating model (D.1), which controls for across-assessor and across-query variation in ratings using fixed effects.

D.4 Alternative measures of quality

Our main result, as depicted in Figure 3 in the main text, is based on the average ratings for Cliqz result sets grouped by the query's popularity (i.e., the search frequency buckets) at different levels of user-data availability (i.e., at different fractions of query logs). One may be concerned that a Likert scale is a categorical variable and not a cardinal one and that our results are solely based on human assessment. Here, we show that our results are robust to using three alternative measures of quality, including one that is not based on human assessment.

The first measure is the share of mostly or completely satisfied ratings, i.e., the share of results sets rated 6 at least on the Likert scale. The advantage of using this measure is that we do not have to impose cardinality. Figure D.6 shows the result. They are qualitatively the same.

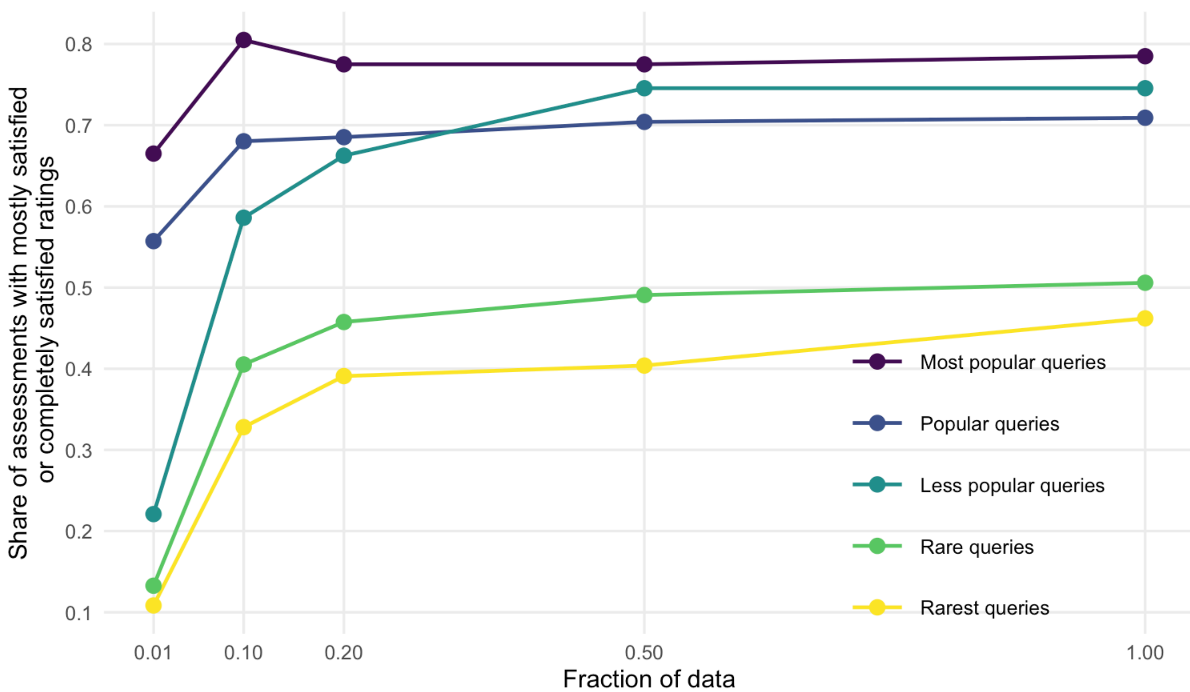
The second measure is the position of the best rated result from the mixed result sets. For all three search engines and all levels of available data (for Cliqz) we determine whether the best rated result for each of the 493 queries is presented as the top result, or in position 2 or 3, 4 to 10, 11 to 18, or not in the top 18. Again, for this, we do not treat the ratings as cardinal.

Figure D.7 shows the result. It confirms that the quality of search results depends on the amount of data that is used to obtain them (for Cliqz). By this measure, overall Google produces the best search results, closely followed by Bing, ahead of Cliqz.

Taken together, these two robustness checks suggests that assuming cardinality and looking at average ratings is appropriate for our purposes.

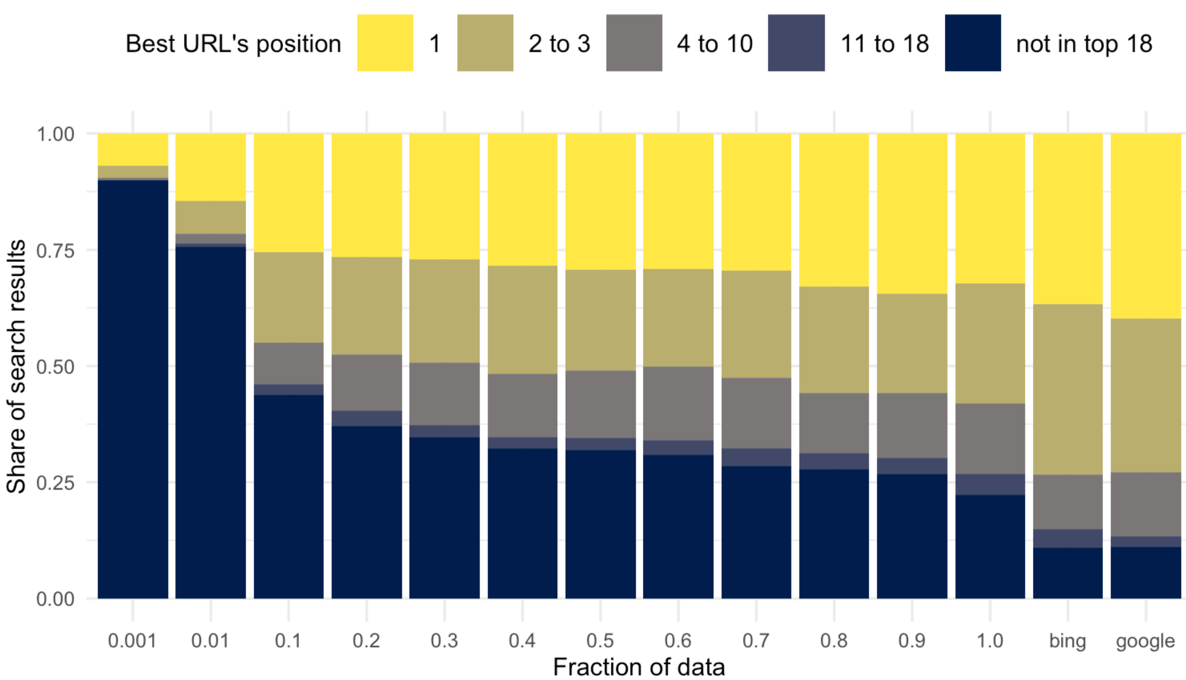
As a third alternative measure, we step away from human assessment and use the Google results as the yardstick. Specifically, our third alternative measure is to what extent Cliqz produces the same top, top 3 and top 5 results as Google. Under this

Figure D.6: Share mostly or completely satisfied assessments



Notes: This figure shows the share of assessments that were mostly satisfied (rating of 6 on the Likert scale) or completely satisfied (rating of 7), by popularity of the query (bucket) and fraction of data that was used to produce the search results. Based on 10,260 assessments for a random sample of 493 queries and the corresponding 2,465 Cliqz result sets (see Appendix B for details).

Figure D.7: Position of best rated result

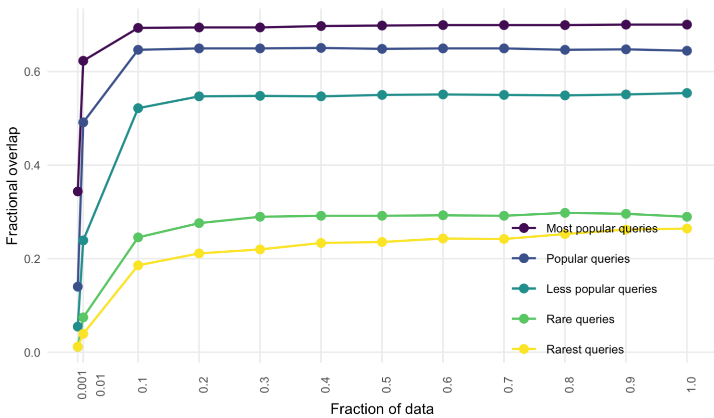


Notes: This figure shows how the position of the overall best result differs across search engines and depends on the amount of data that was used to obtain the search results for Cliqz. The overall best result was determined using the 493 mixed result sets for the 493 sampled queries (see Appendix B).

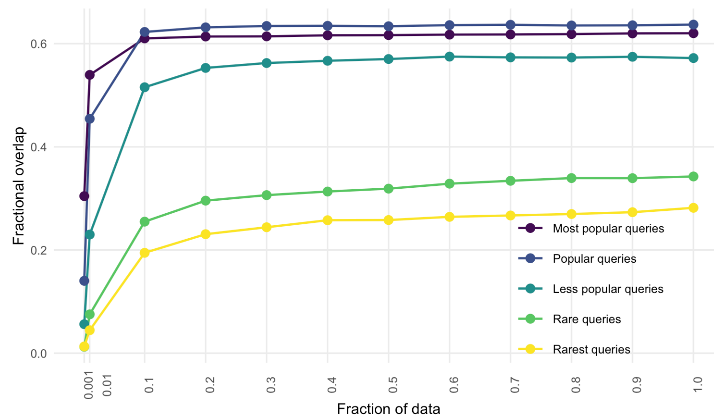
measures, the top x results are considered to be the same, when the respective elements are the same. The ordering is not taken into account. Figure D.7 shows for all 3 versions of this alternative measure that we obtain similar results as the ones in Figure 3 in the main text.

Figure D.8: Overlap with Google results

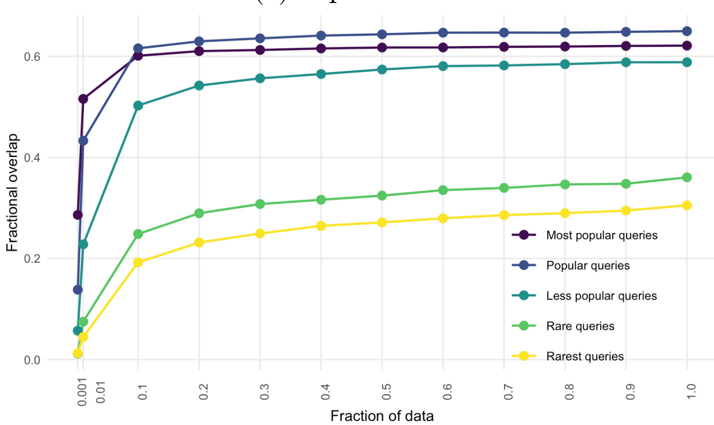
(a) top result



(b) top 3 results



(c) top 5 results



Notes: Fraction of Cliqz results for 493 queries at 5 different levels of data availability that are the same as Google results. Here, “same” means (a) the same top result, (b) the same top 3 results, (c) the same top 5 results. (b) and (c) are in the sense of an unordered set comparison, meaning that the set of results is the same and that the ordering does not matter.