

1894

## Discussion Papers

Deutsches Institut für Wirtschaftsforschung

2020

# Learning from Data and Network Effects: The Example of Internet Search

Maximilian Schaefer and Geza Sapi

Opinions expressed in this paper are those of the author(s) and do not necessarily reflect views of the institute.

#### IMPRESSUM

© DIW Berlin, 2020

DIW Berlin  
German Institute for Economic Research  
Mohrenstr. 58  
10117 Berlin

Tel. +49 (30) 897 89-0  
Fax +49 (30) 897 89-200  
<http://www.diw.de>

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:  
<http://www.diw.de/discussionpapers>

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:  
<http://ideas.repec.org/s/diw/diwwpp.html>  
<http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html>

# Learning from Data and Network Effects: The Example of Internet Search

Maximilian Schaefer\*, Geza Sapi†

September 4, 2020

## Abstract

The rise of dominant firms in data driven industries is often credited to their alleged data advantage. Empirical evidence lending support to this conjecture is surprisingly scarce. In this paper we document that data as an input into machine learning tasks display features that support the claim of data being a source of market power. We study how data on keywords improve the search result quality on Yahoo!. Search result quality increases when more users search a keyword. In addition to this direct network effect caused by more users, we observe a novel externality that is caused by the amount of data that the search engine collects on the particular users. More data on the personal search histories of the users reinforce the direct network effect stemming from the number of users searching the same keyword. Our findings imply that a search engine with access to longer user histories may improve the quality of its search results faster than an otherwise equally efficient rival with the same size of user base but access to shorter user histories.

**Keywords:** Competition, Network Effects, Search Engines, Big Data

**JEL Codes:** L12, L41, L81, L86

---

\*Berlin School of Economics, DIW Berlin and Technische Universität Berlin, [mschaefer@diw.de](mailto:mschaefer@diw.de)

†European Commission DG COMP - Chief Economist Team and Düsseldorf Institute for Competition Economics, Heinrich-Heine-Universität Düsseldorf, 40204 Düsseldorf, Germany, [sapi@dice.uni-duesseldorf.de](mailto:sapi@dice.uni-duesseldorf.de)

*The views expressed in this article are solely those of the authors and may not, under any circumstances, be regarded as representing an official position of the European Commission. This is personal research based entirely on publicly available information and not related to any activity of the European Commission.*

The authors thank Tomaso Duso and Hannes Ullrich for their continued advice and support. We are also grateful for the input provided by Andres Hervas Drane, Chiara Farronato, Szabolcs Lorincz, Christian Peukert, Ananya Sen, Jean Tirole, Kevin Tran, Catherine Tucker, Tommaso Valletti, Christoph Wolf and the participants of the Data Workshop at the University of Zurich (2017), the 4th edition of the Industrial Organization in the Digital Economy Workshop (Liège, 2018), the 16th IIOC (Indianapolis, 2018), the MaCCI Annual Conference (Mannheim, 2018), the 10th bi-annual Postal Economic Conference (Toulouse, 2018), the 10th Conference on Digital Economics (Paris, 2018), the 2nd Doctoral Workshop on the Economics of Digitization (Paris, 2018), the 20th Summer Workshop for Young Economists (Mannheim, 2018) and the 46th EARIE Conference (Barcelona, 2019). This paper builds on previous work that circulated under the title “The Effect of Big Data on Recommendation Quality: The Example of Internet Search”, see [Schaefer et al. \(2018\)](#). All remaining errors are our own.

# 1 Introduction

The role of data in the success of firms is the subject of much debate. Data are often mentioned as a crucial input of production, so much so that academics and the press label data as “*the world’s most valuable resource*” (The Economist, 2017). From a competition policy perspective the success of tech companies relying on data as one of their main inputs raises concerns that data might constitute a source of market power.<sup>1</sup> According to the *data feedback loop* hypothesis (Newman, 2014) there is a self-reinforcing cycle between the success of firms and the amount of data they control. This hypothesis is often put forward to explain the rise of dominant players in data driven industries.<sup>2</sup>

While the rise of data-driven superstar firms in digital industries such as Amazon, Google, Facebook, Netflix and Spotify lends some credit to the hypothesis that data might constitute a source of market power, independent empirical evidence systematically examining the impact of data on the performance of firms remains scarce. The handful of existing studies on the topic suggest that data are essentially like any other ordinary input: All of these studies report evidence for diminishing returns to scale from additional data (He *et al.*, 2017; Yoganarasimhan, 2019; Claussen *et al.*, 2019; Bajari *et al.*, 2019). If there were diminishing returns to scale in data, it may be relatively easy for potential entrants to obtain sufficient data to be viable. The finding of these studies therefore cast some doubt on the popular narrative of data being a *special input* and the root cause of competitive advantage.

In this paper we provide empirical evidence that additional data matter and document how data improves the quality of internet search results. Our findings indicate that data do not simply constitute an input into a static production technology: instead, they simultaneously shift the efficiency boundary of the search result matching technology outwards. While

---

<sup>1</sup>Depriving rivals of data was an important pillar of the European antitrust proceedings against Google, resulting in a record fine of 4.3 billion euro (European Commission, 2018, recitals 111, 114, 458, 514, 739, 860(3), 1318, 1348)

<sup>2</sup>The European Commission announced its plan to enforce opening up the data of dominant tech companies to rivals. It argued that *the high degree of market power resulting from the data advantage can enable large players to set the rules on the platform and unilaterally impose conditions for the access of data.* (European Commission, 2020)

our empirical evidence is consistent with a technology that displays diminishing returns to scale, the improvement of efficiency with larger amount of data provides a novel rationale for an inherent competitive advantage from *big data*.

We analyze search traffic data from Yahoo!. We observe users entering keywords in the search bar of the search engine and follow their subsequent interaction with the search results. The search engine collects the logs of users' clicks on the search result page shown in response to these queries. The data obtained are valuable because they allow the search engine to learn from the observed click behaviour and improve the quality of search results.

Over time, data on the users' action accumulate along two different dimensions. First, as more users enter a specific keyword, the amount of data collected on the feedback for that keyword increases. We call this the *keyword dimension* of data accumulation. Generally, data on a specific search keyword are collected across different users. The amount of data collected in the keyword dimension is therefore directly related to the number of individuals searching that keyword. Second, the more often the search engine observes the clicks of individual users, the more information it can collect on the individual preferences and interests of users. We refer to this as the *user dimension* of data accumulation.

Both dimensions are relevant for the predictive performance of the search engine. More data in the keyword dimension allow the search engine to determine common preferences across users. A user being the first to enter a keyword will generally be confronted with results of a lower quality than a user entering the same keyword at a later point in time. This is because, over time, the feedback provided by users allows for determining a general quality ranking of search results. Learning in the keyword dimension can be viewed as a data-enabled direct network effect: Additional users will generally increase the quality experienced by a specific user through learning from data in the keyword dimension.

The data collected on the user dimension allow the search engine to derive preferences specific to the user. In conjunction with the information that the search engine obtains on other users, data on the user dimension allow the search engine to determine user profiles.

Users with similar profiles, revealed through overlapping preferences in the past, will also be more likely to have overlapping preferences in the future. When confronted with the task of finding relevant search results for a specific user searching a keyword, the search engine builds its prediction based on the feedback obtained from similar users who previously searched the same keyword.

In this paper, we highlight the importance of the interaction between the user and keyword dimensions of data for the ability of search engines to learn. Our main result is that the improvement of search result quality through additional data in the keyword dimension is positively affected by the amount of data collected in the user dimension. Keywords that are repeatedly searched by different users improve faster in quality when the search engine has more information, i.e. data, on the users searching that keyword. Learning from additional data in the keyword dimension becomes more efficient with additional data in the user dimension. Hence, search engines with longer search histories on their users learn faster as their user base grows.

Intuitively, the result can be understood in the following way. The value of the feedback that a user provides is determined by the amount of information the search engine has on the user. More information on the user makes the feedback she provides more valuable because the search engine can relate the feedback to a more detailed user profile. More information on users searching a specific keyword allows the search engine to tailor search results to specific user profiles more efficiently. If the search engine has, on average, more data on the searchers that provide feedback on a specific keyword, this will lead to a faster quality increase as a function of the data collected across different users searching that keyword. More data in the user dimension allow the search engine to determine more rapidly what a specific user wishes to see when searching a specific keyword. Without personal information, the search engine cannot learn which type of user prefers which type of search results for a specific keyword. It is only by combining individual data across different users that the search engine can learn which information in the user profiles is relevant to individualize search results for a

keyword. The more information the search engine has on the user profiles, the more likely it is that it will pick up the relevant information necessary to improve the quality of search results for specific user types.

Our results provide evidence that data-driven services benefit from an externality that goes beyond usual direct network effects. Learning in the keyword dimension is comparable to direct network effects: More users generate more data on keywords which benefits other users of the search engine. Additionally, our analysis suggests that these direct network effects are reinforced by data collected in the user-dimension. Augmented by user-specific data, learning in the keyword dimension occurs faster and, thus, more efficiently.

This additional externality is unique to data-driven industries and confer an additional competitive advantage to the incumbent firm. To understand the implications of our results consider two firms A and B with the same number of users: Under conventional direct network effects, both firms should be equally competitive because they benefit from the same size in user base. However, our results suggests that if one firms has more information on its users, it will be more competitive by virtue of the additional externality that we reveal, which can be viewed as an additional *data network effect*.

From a competition policy perspective, our results are consistent with the hypothesis that a firm with a larger database has a competitive advantage and that lack of data constitutes a barrier to entry. Our results therefore call for a consideration of database sizes in merger decisions between firms that rely heavily on data for their business models. Our findings suggest that the ability to combine data on users across different services might be extremely valuable.

Our paper is related to a nascent strand of literature dealing with the impact of *big data* on firm performance. Several contributions approach the topic of scale economies in data from a policy perspective: [Lambrecht and Tucker \(2015\)](#), [Sokol and Comerford \(2015\)](#), and [Tucker \(2019\)](#) argue that the era of digitization poses no special challenge for antitrust authorities and that network effects from data accumulation should be expected to be weak.

Newman (2014) and Grunes and Stucke (2015), on the other hand, argue that data can play an important role for firms in securing competitive advantages over rivals and call for a reorientation of antitrust policy to better account for the role of data as a barrier to entry. Schepp and Wambach (2015) submit that current competition law should be flexible enough to address the new challenges posed by digitization but emphasize the role of data in understanding dynamics in digital marketplaces. Argenton and Prüfer (2012), Prüfer and Schottmüller (2017) and Hagiü and Wright (2020) model competition in data-driven markets. Economies of scale due to nonrivalry of data and associated welfare effects have been analyzed in Jones and Tonetti (2020). Casadesus-Masanell and Hervas-Drane (2015) analyze the implication of consumer privacy for competition.

Bajari *et al.* (2019) provide theoretical support for diminishing returns from data and empirically analyze the impact of data on the predictive performance of Amazon’s retail forecast system. Their findings are consistent with diminishing returns from repeatedly observing a product in the forecast system. Additionally, Bajari *et al.* (2019) investigate to what extent observing additional products in the same product category improves performance of the algorithm and find no noticeable effect. As a result, they conclude that economies of scope from data are weak. In contrast, He *et al.* (2017) find indication for economies of scope in the context of search engine data.

Our findings are related to the notion of economies of scope according to which algorithms benefit from increased data-variety in addition to sheer data-quantity. Our results reveal that combining different sources of data (user-specific data with data gleaned across users) benefits the firm by further increasing the efficiency of the matching technology. However, the mechanism of action we propose is also distinct from economies of scope in that it builds on synergies between user-specific data and the size of the user base. Our insights speak to the notion that artificial intelligence continuously improves with data. This notion is also put forward in Posner and Weyl (2018), who argue that returns from data in the machine learning context follow a different paradigm than in classical statistics. Our findings provide empirical

support for their view, according to which the true value of data in the machine learning context can only be assessed by considering the *overall learning of the system* (Posner and Weyl, 2018, p. 227).

Our research is further related to Chiou and Tucker (2017) who use an exogenous policy change in the data retention policy as an identification strategy to analyze returns from data. Chiou and Tucker (2017) find no indication that reducing the retention time of user specific information affects search result quality. Claussen *et al.* (2019) and Yoganarasimhan (2019) document the important role of personalized data for the predictive performance of algorithms. Both studies find evidence for diminishing returns to additional data collected on users. We add to this literature by exploring the simultaneous interaction between the user and keyword dimensions of data and by proposing a new mechanism whereby data on individual users reinforce the learning from data on keywords. Our paper is also broadly related to Decarolis and Rovigatti (2019) and Decarolis *et al.* (2020) who study competition in the online advertising market for search engines.

The paper proceeds as follows: In Section 2, we present the data and define the quality measure we use for analysis. Section 3 outlines our empirical strategy to assess network effects in data as the search engine learns from more data. Section 4 presents the results of our empirical analysis, which consists of two parts. The first part presents evidence for the search engine learning from data by taking a long run perspective. Using a proxy variable technique, we show that keywords with more data on the user dimension prior to the period our sample was constructed achieve a higher observed quality level in our sample. The second part focuses on the observed quality evolution of search results returned to keywords during our sample period. We show that the quality of search results improves with more data accumulating on the keyword. The quality improves faster when the searches are entered by users on whom the search engine already gathered more data. Section 5 discusses identification using a stylized model of statistical learning. Section 6 concludes.

## 2 The Data

The data we use stem from Yahoo! and contain fully anonymized search logs spanning a period of 32 days from July 1, 2010 till August 1, 2010, inclusive. An observation in our database contains a keyword identifier, a cookie identifier essentially corresponding to the computer on which the search was conducted, the precise time the keyword was entered in the search bar, the ordered list of the top ten organic result URLs and the sequence of clicks performed by the user. In total, we have approximately 80 million observations from 29 million users (identified by the cookie) searching for 67 thousand different keywords. As the search keywords are anonymized, we do not observe the number of words in the query. We however observe when the same combination of keywords was entered as these search terms have the same identifier.

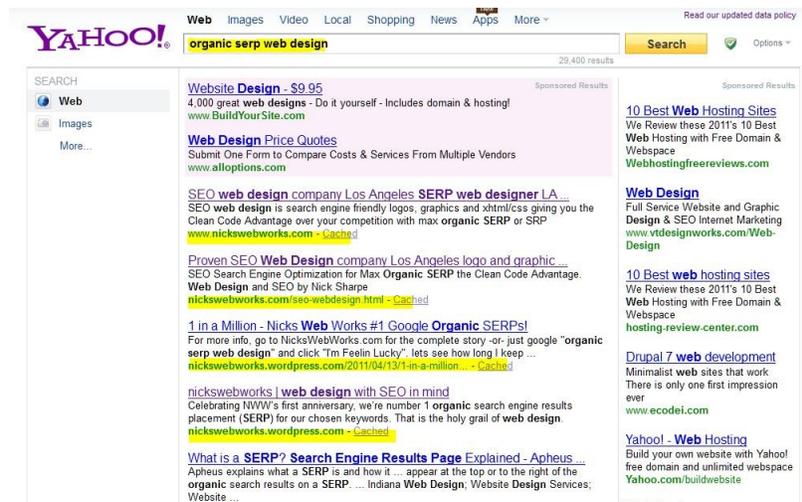


Figure 1: Search Result Layout at Yahoo!, 2011

Figure 1 illustrates the structure of the typical search result page at Yahoo! at the time the data were collected. The search keyword, highlighted in yellow next to the Yahoo! logo, is the sequence of characters the user enters in the search bar in her quest for information. Our analysis focuses on the quality of the organic search results URLs, which are highlighted in yellow in the search result list. Paid advertisements are displayed on the north and east edges of the result list. This general layout of the search result page remained largely intact

up to today, with search engines typically devoting the top and east edges of the result page to ads, and present up to 10 organic search result URLs in the middle area.

## 2.1 Measuring Search Result Quality

For each search session we observe a log that records the actions of the user on the result page. This usually consists of a series of clicks on the retrieved search result links, possibly returning from those URLs to the search result page, until the user clicks on a search result link from where she does not browse back to the result page to continue her search. If the user does not return to the result page to continue her search, the log ends. To measure the quality of search results to a specific keyword, we create a binary variable of search result quality based on the users' actions taking the values *good* (encoded as 1) or *bad* (encoded as 0).

We deem the search result quality as *good* if the last recorded click in a log occurs on the top displayed organic URL. This measure takes into account the fact that users have a natural tendency to perform their first click on the top displayed URL. If a user performs her first click on the top displayed URL and subsequently returns to the result page to choose a URL further down the result list, the search result quality is encoded as *bad*.

Based on this criterion, we calculate the click through rate as the fraction of searches ending on the first URL. The click through rate on the first URL for keyword  $i$  is defined as:

$$ctr_i^{\{1\}} = \frac{\sum_{s_i \in S_i} \mathbb{1}\{lcp_{is} = 1\}}{\sum_{s_i \in S_i} \mathbb{1}\{lcp_{is} \neq 0\}} \quad (1)$$

Where  $\mathbb{1}$  denotes the indicator function and  $lcp_{is}$  the last click position recorded for search  $s_i$ .  $S_i$  is the set of searches considered in the computation of the click through rate for keyword  $i$ . For example, if the click through rate is computed for a particular day, then  $S_i$  is the set of all the searches on that particular day. Finally,  $lcp_{is} = 0$  denotes a final click

on an advertisement URL, which we ignore in the computation of the click through rate. Advertisement URLs are paid content that might be displayed even if the content is not relevant to the user. Because we want to measure the performance of the search engine in its ability to find relevant content for searches where the user was interested in organic content, we decided to ignore clicks on advertisement URLs in the main quality measure. Our results remain qualitatively unaffected if we consider alternative click based quality measures or if we include ads in the analysis (see Appendix A.3 and Appendix A.4).

Click through rates on the top URLs are an intuitive measure for search result quality, whose variations are widely used in the information retrieval literature (Joachims, 2002). Users expect the most relevant search result be displayed at the most prominent position on the search result page. A user clicking on top-displayed URLs and not returning to the search result page means she was satisfied with the information found under that link.

In the Appendix we assess the robustness of our results when using an editorial quality measure. In particular, our data set also contains 659,000 query-URL pairs with editorial relevance judgments Yahoo! collected from human experts. The correlation coefficients between the  $ctr^1$  measure and two editorial quality measures that we consider are 0.45 and 0.55, and hence clearly show positive correlation of moderate magnitude. The editorial quality measure is explained and discussed in depth in Appendix A.1. We present the results of our main analysis when using the editorial quality measure in Appendix A.2.

## 3 Empirical Strategy

### 3.1 Measuring Learning from Data

In this subsection we discuss the variables used to study the search engines ability to learn from data and improve the quality of search results. We focus on two different dimensions of data: First, the *keyword dimension*, which captures the amount of data that accumulates through searches performed on a particular keyword. The number of users determines how

much training data the search engine has for a given keyword. In the *keyword dimension*, learning from data is akin to direct network effects: with more users searching on the same keyword, the search engine has more training data available to solve the task to predict the most relevant search results to display to that particular keyword.

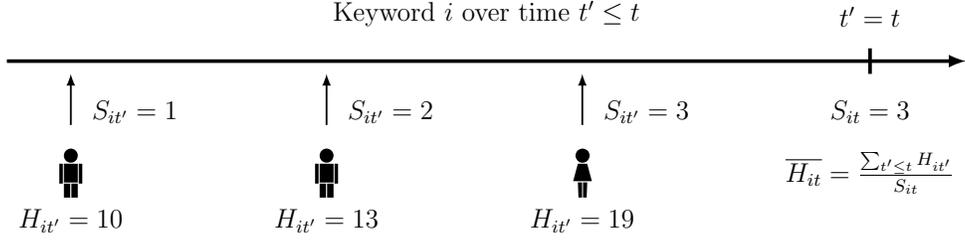
The second dimension captures the amount of data that the search engine collected on the users searching a specific keyword. The value of user-specific data is widely recognized by companies relying heavily on predictive technologies, such as Netflix, Amazon and Google.<sup>3</sup> For internet search, it is common wisdom that data on user search histories are useful because they allow fine-tuning search results for the individual preference of each searcher. In this article we emphasize that the importance of data on the *user dimension* goes beyond the fact that they allow fine tuning search results for a particular user: We show that more data on the *user dimension* - longer observed search histories for users - allow the search engine to learn faster from more data on various keywords. Data in the *user dimension* therefore acts as a catalyst for learning from data in the *keyword dimension*. They can amplify network effects stemming from more users and are a key yet so far underemphasized source of competitive advantage and market power.

Our results are consistent with the following learning mechanism that drive network effects based on data: When a user searches for a particular keyword, the search engine engages in user profiling to establish similarities between the current user and past users who entered the same keyword. Search results that proved relevant to past users with a similar profile are also more likely to be relevant for the current user. Establishing similarities between the current user and past users is facilitated by the amount of data the search engine collects about users. Intuitively, the more data the search engine collects on users, the more likely it is to find overlap in their browsing behaviour

With more data collected in the *keyword dimension*, the search engine learns which search

---

<sup>3</sup>Netflix explains on its website the use of person-related information to suggest movies: “*Personalization is one of the pillars of Netflix because it allows each member to have a different view of our content that adapts to their interests and can help expand their interests over time. It enables us to not have just one Netflix product but hundreds of millions of products: one for each member profile*” (Netflix, 2020).



**Figure 2:** Variable Description

results are relevant for similar user profiles by deducing which information in the overlapping browsing behaviour of user is informative about their preferences for the keyword. Because more data on the users makes it easier to elicit the relevant characteristics to determine relevant search results for specific user profiles, the search engine is able to retrieve more relevant search results when more data on the users are available.

In the machine learning literature, the idea of training algorithms for prediction tasks based on the overlap of past user preferences is known as collaborative filtering.<sup>4</sup> It appears intuitive that this method requires both training data in the *keyword dimension* and data in the *user dimension*. A user being the first to search a keyword may likely be confronted with relatively poor search results because the search engine has no previous training data on the keyword. Similarly, a lack of data on the users does not allow the search engine to tailor results to user profiles. We argue here that these two dimensions of data are not independent from each other: the amount of data in the *user dimension* increases the efficiency of learning in the *keyword dimension*. This in turn has serious implications for economies of scale, and consequently for competition between search engines. To the best of our knowledge, we are the first to systematically study the interaction between the *user* and *keyword dimensions* of data. We now introduce the variables capturing both data dimensions.

- $S_{it}$ : **The cumulative number of searches** for keyword  $i$  at time  $t$ . It describes the size of training sample for keyword  $i$  at time  $t$ .

<sup>4</sup>See [Adomavicius and Tuzhilin \(2005\)](#) for an overview of the different architecture types of recommender systems.

- $\overline{H_{it}}$ : **The average user history** for keyword  $i$ . It captures the average amount of data the search engine had on the users that entered the keyword in the search bar of the search engine until time  $t$ . Denote by  $H_{it'}$  the length of the search history of the user querying keyword  $i$  at time  $t' \leq t$ , then  $\overline{H_{it}} = \frac{\sum_{t' \leq t} H_{it'}}{S_{it}}$ .

Figure 2 illustrates how both variables are computed at a given point in time  $t$ . The average user history is a natural measure for the amount of data the search engine collected on the users who searched for a specific keyword. Throughout the analysis, we track search result quality as a function of  $S_{it}$ , a measure for the amount of training data available for keyword  $i$ . We explore a novel externality in learning from data that provides a new perspective on the value of data. Data on individual users are particularly valuable, because learning from data along the *keyword dimension* becomes faster when the average amount of data collected in the *user dimension*, captured by  $\overline{H_{it}}$ , increases.

### 3.2 Data Generating Process

Before turning to the analysis of search result quality and data volume we pause to discuss the nature of our sample in more detail.

Our dataset is a snapshot of searches observed through a one month period. We do not know how often the search engine had observed the sampled keywords and users prior to the sample period. For example, in theory it is possible that the keywords we observe often are only popular during the month of our sample, but had never been searched upon before. This may be the case for example for news related searches. We do not directly observe the overall amount of data the search engine had collected on keywords and users before our sample started, because we only observe a snapshot of the overall search traffic.

With our data spanning over a one-month period, the variables that we observe constitute the monthly counterpart of the true variables we are interested in, namely  $S_{it}$  and  $\overline{H_{it}}$ . We denote the observed counterparts  $s_i$  and  $\overline{h}_i$  for which we drop the index  $t$  since we focus on a fixed time span of one month. Intuitively, keywords with more searches and longer user

histories during the month of our data are likely to also have experienced more searches and longer user histories previously. In the further analysis we will restrict attention to keywords for which we observe approximately the same number of searches every day. The vast majority of the keywords in the sample fall into this category, as relatively few keywords experienced rapid jumps or drops in the number of searches over time in our sample.

Focusing on keywords with a relatively stable search activity over time is useful: since these keywords see a fairly constant number of searches every day during the period we observe them, it appears reasonable to assume that these keywords had seen a similar search pattern also prior to our sample period. If that is the case, we can compare keywords that received more and less searches in our sample and assume that this also held in the period before the sample started. In other words, the monthly counterparts  $s_i$  and  $\bar{h}_i$  constitute good proxy measures for  $S_{it}$  and  $\overline{H}_{it}$ .

This would be much less likely to hold for queries that experience a rapid boost in searches in the sample: such queries may for example relate to news, where we would not know how often the keyword had been entered prior to our sample period. For example, consider a rapidly booming query observed 50 times, all on a single day and never otherwise in the sample, and another query observed once on each of the 32 days of our sample. While in the sample it appears that the booming query was observed more often (50 times), it is possible and even likely that the search engine has more observations on the query that occurs less often (32 times) in the sample: that query was most likely entered regularly also prior to our sample.<sup>5</sup>

To select the keywords with relatively stable search patterns over the days of our sample, we calculate for each observed keyword the number of searches per day. Take first the limiting case of a query that does not experience popularity changes during the sample period and receives the same number of searches each day. Then this query accumulate each day 100/32 percent of its total observed monthly searches. We will refer to the special case

---

<sup>5</sup>In Appendix B, we discuss in more depth the detailed assumptions under which it is reasonable to focus on constantly searched queries.

with the number of daily searches uniformly distributed as *even accumulation criterion*. We define tolerance levels that determine the maximum percentage point deviation a keyword is allowed to have on one particular day. For instance, a tolerance level of ten percentage points indicates that on each day of the sample, a keyword is not allowed to accumulate more than  $(100/32 + 10)$  percent of its total searches in order to be retained in the sample.

Intuitively, the narrower the tolerance interval, the better the quality of  $s_{it}$  and  $\overline{h_{it}}$  as proxy variables. The largest tolerance level we consider is 50 percentage points, such that a keyword is allowed accumulate up to  $100/32 + 50 \approx 53$  percent of its total searches in one day without being dropped. The narrowest tolerance level is ten percentage points, meaning that a keyword is allowed to accumulate up to 13 percent of its total searches in one day without being dropped.<sup>6</sup>

Importantly, while focusing attention to keywords that see relatively similar number of searches every day allows us to reduce the noise in our data, this is not crucial for our analysis. We obtain the same qualitative effects when using the full sample as well.

In subsection 4.1 we will show that keywords with longer user histories and more searches previous to our sample have higher quality search results.

## 4 Results

Our empirical analysis consists of two parts. In the first part, we take a long run perspective on the data. We take into account that the search engine may have gathered data on the keywords we observe already before the sample period. Depending on the number of searches and the average user history before our sample period, keywords should differ in the quality level we observe in our sample. If the search engine learns faster with additional data in the *user dimension*, keywords with a longer average user history in the past should, *ceteris paribus*, experience a higher quality level.

---

<sup>6</sup>Nearly 90 percent of the keywords display a search pattern over time that falls within the narrowest tolerance level we consider.

We first discuss long term trends, looking at the search result level of queries in our sample as a function of the proxied number of searches and average user history prior to the sample period. Focusing on keywords with a relatively stable number of searches over time helps better isolate how additional data drives learning.

In the second part, we analyze the quality evolution of keywords over the sample period. Building on the insights from the first part, we take into account the unobserved number of searches and average user history prior to the sample period. If the search engine learns faster with additional data in the *user dimension*, keywords with a longer average user history should, *ceteris paribus*, experience a larger increase in quality. We confirm that this is indeed the case. The learning observed within the sample period is consistent with the long run dynamics documented in the first part of the analysis.

## 4.1 Long Run Analysis

In this subsection we study the quality level of queries in our sample as a function of the number of searches and the average user history previous to the sample period. If search result quality increases faster with more data in the *user dimension*, we would expect that search results for keywords with a longer average user history prior to the sample period lead to higher quality search results when we observe them. If we could access the entire search history for each keyword and each user we observe in the sample, we would directly estimate the following equation:

$$ctr_i^1 = f(S_{it}, \overline{H_{it}}) + \epsilon_i \quad (2)$$

Where  $ctr_i^1$  is the click through rate on the first URL for keyword  $i$  based on all the searches we observe.  $S_{it}$  the number of searches previous to the sample period and  $\overline{H_{it}}$  the average user history previous to the sample.  $\epsilon_i$  is the error term. Since we do not observe  $S_{it}$  and  $\overline{H_{it}}$ , we estimate:

$$ctr_i^1 = f(s_i, \bar{h}_i) + \epsilon_i \quad (3)$$

Where  $s_i$  and  $\bar{h}_i$  denote the number of searches and the average user history we observe for keyword  $i$  during our sample period.

Throughout this analysis, we assume that larger values of  $s_i$  and  $\bar{h}_i$  induce larger average values of  $S_{it}$  and  $\bar{H}_{it}$ . Thus, we assume that (i) keywords for which we observe more searches during the sample were on average searched more often before the sample period, and, (ii) keywords with a longer average user history during the sample have, on average, a longer user history previous to the sample period.

Estimation is performed by local linear regression on a grid  $s \times h$ , where  $s = \{0, 0.1, 0.2, \dots, 1\}$  and  $h = \{Q_1, Q_3\}$ . The values of  $s$  denote number of searches in tens of thousand.  $Q_1$  and  $Q_3$  denote the lower and upper quartiles of the distribution of the average user histories.<sup>78</sup>

Figure 3 shows the results of our analysis. The solid line in Figures 3a and 3b correspond to keywords with longer user histories, taken at the upper quartile ( $h = Q_3$ ) of the user history distribution. The dashed line depicts keywords with short user histories, which we set at the lower quartile ( $h = Q_1$ ) of the distribution of user histories. In Figure 3a, estimation was performed using all the keywords in the sample. In Figure 3b keywords that deviate no more than ten percentage points from the even accumulation criterion (100/32 percent of searches per day) were considered. According to the discussion in subsection 3.2, the results in Figure 3b are more informative about the true relationship between the quality of keywords and  $S_{it}$  and  $\bar{H}_{it}$ .<sup>9</sup>

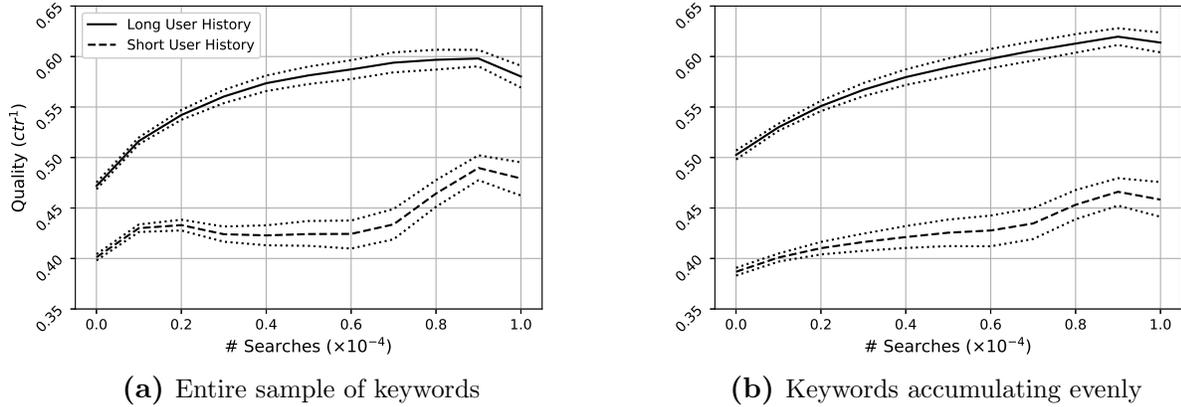
The solid lines in Figure 3 map out the quality level as a function of the number of times the keyword was searched upon,  $s$ , for keywords with a long average user history. The dashed

---

<sup>7</sup>The number of searches we observe is truncated at 10,000. The lower quartile is equal to 2.79, the upper quartile is equal to 4.78. These numbers might appear low but can be explained by the fact that we observe a random sample and that users are observed only three times on average in our sample.

<sup>8</sup>In Appendix C, we present the details on the estimation method.

<sup>9</sup>The reader is referred to Appendix B for an in depth discussion of the relationship between  $s_i$  and  $\bar{h}_i$  and the unobserved  $S_{it}$  and  $\bar{H}_{it}$ .



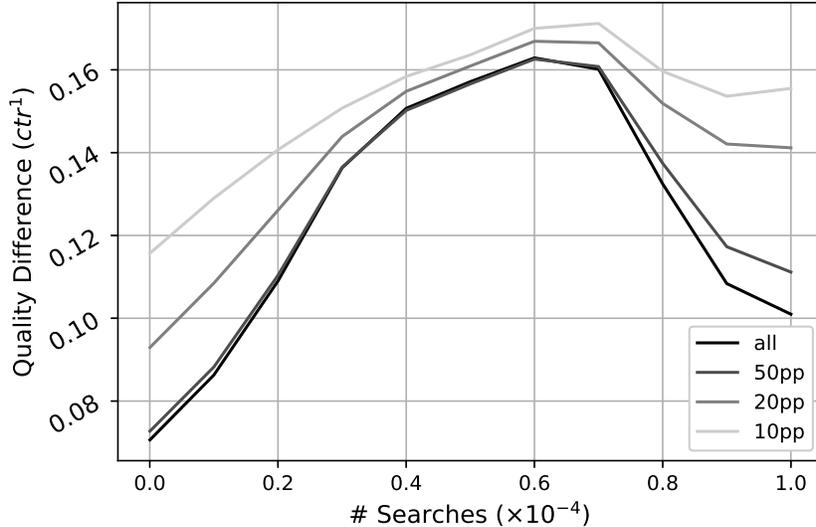
**Figure 3:** Average search result quality by keyword depending on the total number of searches, with 95 % confidence intervals

lines map out the quality level for keywords with a short average user history.<sup>10</sup> Note that when we focus on keywords that accumulate searches more evenly over time in Figure 3b, the observed pattern for keywords with a short average user history normalizes.

Figure 4 shows the difference in the search result quality level between keywords with a long and short average user history for a given number of searches,  $s$ . The black line in Figure 4 shows the difference for the results in Figure 3a. The line with the lightest grayscale shows the difference for the results in Figure 3b, when only those keywords were considered that accumulate searches over time most evenly. The other lines show the differences for intermediate tolerance levels with a lighter grayscale indicating a stricter tolerance level. Figure 4 reveals that the measured impact of the average user history increases as we gradually drop keywords that accumulate searches less evenly over time: Gradually reducing the measurement error in the proxy variables also increases the measured impact of the average user history. This is in line with what we expect to see when the relationship between the average user history and the quality of search results was causal and we increase the quality of the proxy variables.

From Figure 4, we also see that the difference in quality between keywords with a long

<sup>10</sup>The quartiles of the average user history distributions are determined based on all the keywords in the sample, i.e. they do not vary across the different populations of keywords considered.



**Figure 4:** Impact of longer average user history on quality level. Different tolerance levels for deviation from even accumulation criterion are considered.

and short average user history tends to increase with additional searches on the keywords. This divergence in quality levels is in line with the hypothesis of more efficient learning from additional searches through a longer average user history. The measured divergence will be at the core of Section 5, where we will analyze whether potential sources of confoundedness might rationalize the documented divergence when a longer average user history does not lead to faster learning.

The findings of this subsection can be summarized as follows: First, our data are consistent with diminishing returns to scale in the total number of searches,  $S$ . Second, we find that more data on the individual users amplify network effects: learning from the *keyword dimension* of data ( $S$ ) becomes faster as more data on the *user dimension* ( $H$ ) accumulates. Third, the relationship between search result quality and data in the *user dimension* becomes stronger when we focus on keywords that accumulate searches more evenly over time. Fourth, we observe a divergence in quality between keywords with a long and short average user history as more data in the *keyword dimension* ( $S$ ) becomes available.

These results are consistent with the claim that data may constitute a source of market power. A search engine with more data on the individual search histories of its users will

always outlearn the rival on new keywords, and any keyword that the two search engine had observed equal times previously. The following analysis of the learning in the sample confirms this finding.

## 4.2 Quality Evolution within the Sample Period

We now turn to the analysis of how search result quality changed during the 32 days of our sample period. We again focus on changes in search result quality as a result of more data accumulating in the *keyword* and *user dimensions*, respectively  $s_i$  and  $\bar{h}_i$ . Consistent findings with the results from subsection 4.1 would imply a positive concave relationship between the quality evolution of keywords and  $s_i$ . Furthermore, for the same value of  $s_i$ , we would expect the quality increase to be more pronounced for keywords with a larger average user history,  $\bar{h}_i$ .

Subsection 4.1 reveals a strong relationship between the unobserved  $S_{it}$  and  $\bar{H}_{it}$  and the quality level of a keyword during the sample period. It is also to be expected that  $S_{it}$  and  $\bar{H}_{it}$  will strongly impact the quality evolution we observe for keywords during the sample period. For instance, in the presence of diminishing returns to scale, larger  $S_{it}$  should, *ceteris paribus*, reduce the measured quality evolution of keywords.

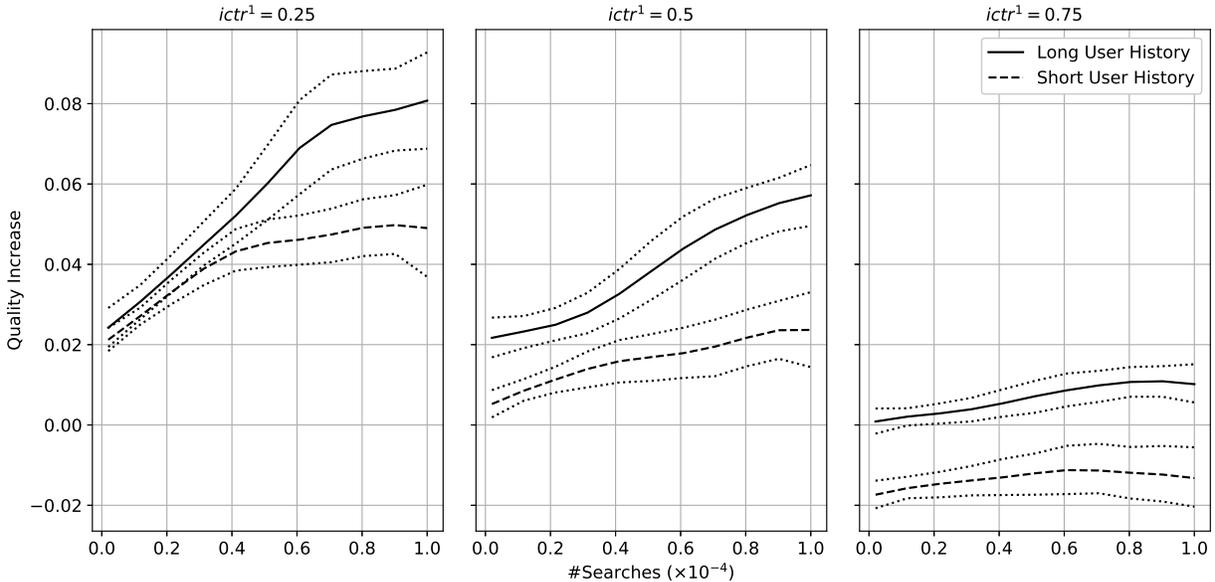
The results from subsection 4.1 suggest that controlling for the quality level that keywords reached should help to control for  $S_{it}$  and  $\bar{H}_{it}$ . We therefore estimate the following equation:

$$\Delta ctr_i^1 = f(s_i, \bar{h}_i, ictr_i^1) + \epsilon_i \tag{4}$$

$\Delta ctr_i^1$  denotes the difference in the click through rate on the first URL between the first and last 100 searches that we observe for a keyword.  $ictr_i^1$  denotes the click through rate on the first URL for the first 100 searches that we observe for a keyword.<sup>11</sup>  $ictr^1$  is an intuitive

---

<sup>11</sup>Note that by construction estimating Equation 4 requires us to focus on keywords with at least 200 searches during the period of our sample. Estimating Equation 4 for all the keywords in our sample is frustrated by regression to the mean. Regression to the mean arises in any kind of analysis where observations are classified based on a noisy measure of the initial outcome. The regression to the mean phenomenon and



**Figure 5:** Average quality increase of keywords with 95% confidence intervals. Each panel stands for a different initial quality level.

measure for the quality a keyword reached at the beginning of our sample period, which helps to account for differences in the unobserved  $S_{it}$  and  $\overline{H_{it}}$ .

Estimation is performed by local linear regression on the grid  $ictr \times s \times h$ .<sup>12</sup> The values of  $ictr$  are given by  $\{0.25, 0.5, 0.75\}$  and the values of  $s$  by  $\{0.02, 0.1, 0.2, \dots, 1\}$ . The values of  $h$  are identical to the ones chosen in subsection 4.1. Figure 5 shows the result of estimating Equation 4 on the grid  $ictr \times s \times h$ . Each panel stands for a different initial quality level. The solid lines within each panel map  $\hat{f}$  as a function of  $s$  for keywords with a long average user history, the dashed lines for keywords with a short average user history.

From the left panel of Figure 5, we can read that keywords with  $ictr^1 = 25\%$ ,  $h = Q_3$  and  $s = 10,000$  experience an average increase of eight percentage points in the click through rate on the first URL. Keywords with  $h = Q_1$  and otherwise identical parameters experience a quality increase of only 4.8 percentage points.

Data on the *user dimension* amplifies learning from data on the *keyword dimension*: Keywords with longer average user history, *ceteris paribus*, display a larger quality increase

its impact on our analysis is discussed in more detail in Appendix D.

<sup>12</sup>Details on the estimation method and bandwidth selection are given in Appendix C.

than keywords with a shorter average user history.<sup>13</sup> Consistent with diminishing returns to scale from additional searches on a keyword, we observe a concave pattern between the measured quality evolution and the number of searches,  $s$ . Furthermore, the average quality increase diminishes *ceteris paribus* with a larger initial quality level, which is also in line with diminishing returns in  $S$ , the number of searches on a keyword.<sup>14</sup>

In summary, four key results emerge from this analysis on how data matters for search result quality. First, search engines learn from data on user response to search results shown in response to the same keyword. Search result quality improves for queries that are searched upon more. Second, the personal dimension of data matters for learning. Data in the *user dimension* enable faster learning from data in the *keyword dimension*. Third, the learning does not fade off fast. We observe quality improvements even at the most searched upon keywords. Fourth, personal data matters even for new queries. This last point is particularly important for the question whether data may constitute an entry barrier. It implies that between two hypothetical search engines, even if they had an equal number of users, the new entrant will be at a disadvantage and increasingly so: the established firm will have observed longer search histories of its users and will be able to provide higher quality search results even to queries that are new to both search engines.

The findings from this subsection match our expectations from the analysis of the long run dynamics. The fact that keywords with a higher quality level improve less in quality is in line with the pattern of diminishing returns to scale from  $S_t$ , which we found in subsection 4.1. Additionally, we find larger quality increase for keywords with a longer average user history. This indicates that data on the *user dimension* allows the search engine to learn faster from

---

<sup>13</sup>For estimation, we removed all keywords that accumulate more than 50% of the total searches during the sample period within a single day. Further narrowing the tolerance level of the even accumulation criterion does not significantly impact the results.

<sup>14</sup>The negative quality evolution measured for keywords with  $ictr^1 = 0.75$  and  $h = Q_1$  (dashed line in lower right panel) might be due to the fact that those keywords experience an increase in the click through rate on URLs positioned below the first URL. Thus, clicks on the URLs below the first one might cannibalize clicks on the first URL. In Appendix A.3, we analyze the quality increase of keywords based on the click through rate of the first three URLs. Based on this quality measure, we find no indication for a negative average quality evolution for keywords that start from a higher initial quality level.

data in the *keyword dimension*. This effect explains the marked quality differences between keywords with a long and short average user history in subsection 4.1.

## 5 Identification

Sections 4.1 and 4.2 showed that data in the *user dimension* enhance learning from data in the *keyword dimension*. We argued that more data may constitute a source of market power, as more data in the *user dimension* may allow an incumbent to learn faster than an equally efficient rival.

In this section we discuss in detail to which extent unobserved heterogeneity can confound our results. To do so, we introduce a functional form assumption that allows us to model faster learning in the *keyword dimension* through additional data in the *user dimension*. The model also allows us to take into account the impact of two potential confounding factors: (i) unobserved heterogeneity across keywords of different types and (ii) unobserved heterogeneity with respect to the age of keywords (i.e. the *time* elapsed since the keyword was first entered into the search engine). If keywords with a longer average user history are either *easier* or *older* this could explain the higher quality levels we observed in subsection 4.1. We show here that it is unlikely that our results are rationalized by confounding factors.

We start by introducing and discussing our functional form assumption for the quality evolution process of keywords. The quality of the search results shown for keyword  $i$  at time  $t$ ,  $(Q_{it})$ , is a function of the number of times the keyword has been searched till that time  $(S_{it})$  and the average amount of user specific data for that keyword  $(\overline{H}_{it})$ :

$$Q_{it}(S_{it}, \overline{H}_i, \mu_i) = 1 - \frac{(1 - \mu_i)}{S_{it}^{\delta(\overline{H}_i)}} = 1 - \frac{(1 - \mu_i)}{(T_i \times s_i)^{\delta(\overline{H}_i)}} \quad (5)$$

The function in Equation 5 is concave in  $S_{it}$  and its image always lies in the interval  $[0, 1]$  if  $S_{it} > 1$ ,  $\delta(\overline{H}_i) > 0$  and  $\mu_i \in [0, 1]$ , which we assume.  $Q_{it}$  converges to 1 as  $S_{it}$  approaches infinity. The function  $\delta(\overline{H}_i)$  determines the speed of convergences as a function

of  $S_{it}$ . Larger values of  $\delta$  imply a faster convergence to 1. More data in the *user dimension* increases the speed of learning from more searches on the same keyword if  $\partial\delta(\overline{H}_i)/\partial\overline{H}_i > 0$ . If in turn  $\partial\delta(\overline{H}_i)/\partial\overline{H}_i = 0$ , data on user history do not impact the speed of learning from more searches on a keyword.<sup>15</sup>

We denote the type of keywords by  $\mu_i$ , which captures the intrinsic difficulty to find relevant search results for the keyword. A larger value of  $\mu_i$  corresponds to a keyword for which it is easier to retrieve relevant search results. Keywords with a larger value of  $\mu_i$  start from a higher quality level and, *ceteris paribus*, remain on a higher quality level. Heterogeneity in  $\mu_i$  is one source of potential confoundedness that we will study in this section. The second source of potential confoundedness stems from the fact that a search engine may have observed some keywords a longer time ago first. We will refer to  $T_i$  as the age of the keyword, namely the time elapsed since the search engine first ever observed that keyword. The decomposition  $S_{it} = T_i \times s_i$  emphasizes that the total number searches at time  $t$  can be written as the product of the number of months a keyword existed until time  $t$ ,  $T_i$ , and the average number of searches per month,  $s_i$ .<sup>16</sup>

Despite its simple parametric form, we believe that the function in Equation 5 provides a realistic approximation of statistical learning for two reasons. First, it captures diminishing returns to scale from additional searches through the concavity of the functional form in  $S$ . Second, the maximum achievable quality level is bounded, which captures the idea that the prediction accuracy of a model can not be increased indefinitely, i.e. that there is an irreducible error term.<sup>17</sup>

Based on the functional form of Equation 5, we can analytically derive the expected

---

<sup>15</sup>In the functional form assumption of Equation 5, we neglect the possibility that  $\overline{H}_{it}$  might be time varying. As discussed in Appendix B,  $\overline{H}_{it}$  should generally increase over time. Neglecting this dynamic in the functional form assumption amounts to assuming that the time varying nature of  $\overline{H}_{it}$  is negligible. We consider this restriction innocuous under the detailed assumptions formulated in the data generating process in Appendix B. It implies that differences in the age of keywords only impact quality through the resulting difference in  $S_{it}$  when keywords have the same average user history in a given period.

<sup>16</sup>The reader should note that this corresponds to the assumption that keywords accumulate evenly over time. We refer to Appendix B for a detailed discussion of this assumption.

<sup>17</sup>See [Bajari et al. \(2019\)](#) for an extensive treatment of the properties of statistical learning.

quality level of keywords conditional on observing  $s_i = s$  and  $\bar{h}_i = h$ . This mimics the scenario we face in subsection 4.1 and allows us to analyze under which assumptions on the distribution of  $\mu$  and  $T$  we can generate data consistent with ours when the average user history does not affect the speed of learning from additional data in the *keyword dimension*.<sup>18</sup>

Consider the conditional expected quality, given a tuple  $\{s_i = s, \bar{h}_i = h\}$ :

$$\mathbb{E}[ctr_i^1|s, h] = \int_0^{\bar{T}} \int_0^1 \left(1 - \frac{1 - \mu_i}{(s \times T_i)^{\delta(h)}}\right) f(\mu_i, T_i|s, h) d\mu_i dT_i \quad (6)$$

Where  $\bar{T}$  describes the maximum number of months a keyword existed before the sample period.<sup>19</sup> Through the conditional density function  $f(\mu_i, T_i|s, h)$ , we can introduce correlation between the unobserved and the observed variables. We consider two scenarios. First, potential confoundedness related to the type only, which we model by assuming  $f(\mu_i, T_i|s, h) = f(\mu_i|s, h) \times f(T_i)$ . Second, potential confoundedness related to the age only, which we model by  $f(\mu_i, T_i|s, h) = f(T_i|s, h) \times f(\mu_i)$ .

We want to understand under which assumptions on  $f(\mu_i|s, h)$  and  $f(T_i|s, h)$ , the conditional expectation in Equation 6 generates the divergence observed in Figure 4 in subsection 4.1 if  $\partial\delta(\bar{H}_i)/\partial\bar{H}_i = 0$ , i.e. when learning in the *keyword dimension* is not strengthened by more data in the *user dimension*. Divergence is defined in the following way: Denote by  $h^l$  a long average user history and by  $h^s$  a short average user history. For the difference in the conditional expectations between  $h^l$  and  $h^s$  given  $s$ , we write  $\mathbb{E}[ictr_i^1|s, \Delta h] = \mathbb{E}[ictr_i^1|s, h^l] - \mathbb{E}[ictr_i^1|s, h^s]$ . We observe a divergent pattern if  $\mathbb{E}[ictr_i^1|s, \Delta h] > 0$  and  $\partial\mathbb{E}[ictr_i^1|s, \Delta h]/\partial s > 0$ , i.e. if the difference in the expected quality between  $h^s$  and  $h^s$  is positive and increases with  $s$ .

Confoundedness is modelled by first order stochastic dominance. For instance, we model the case in which a longer average user history is associated with a higher type by assuming

<sup>18</sup>We focus on subsection 4.1 because the pattern in the data that we observe in subsection 4.2 can be generated under weaker assumptions on confoundedness.

<sup>19</sup>The reader should note that since we assume that  $\bar{H}_{it}$  is not time varying, we can write  $\bar{H}_{it} = \bar{H}_i = \bar{h}_i = h$ .

that  $F(\mu_i|s, h^s) > F(\mu_i|s, h^l)$ , which implies that  $E(\mu_i|s, h^s) < E(\mu_i|s, h^l)$ . We denote the difference in the conditional expectations between  $h^l$  and  $h^s$  for type and age by  $E[\mu_i|s, \Delta h]$  and  $E[T_i|s, \Delta h]$ . Changes in the conditional expectations  $E[\mu_i|s, \Delta h]$  and  $E[T_i|s, \Delta h]$  are assumed to be caused by changes in the degree of first order stochastic dominance. For instance,  $\partial E[\mu_i|s, \Delta h]/\partial s \geq 0$  corresponds to  $\partial [F(\mu_i|s, h^s) - F(\mu_i|s, h^l)]/\partial s \geq 0$ .

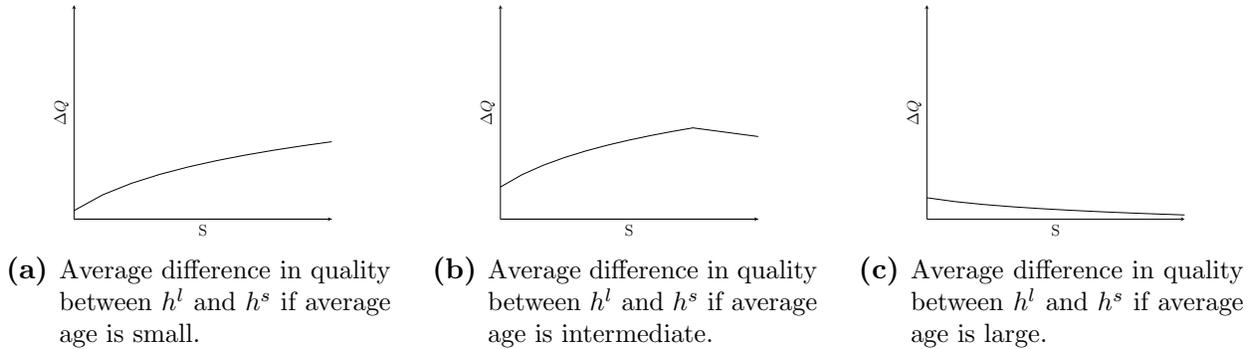
We now state our main result, which we prove in Appendix E:

**Proposition 1** *First, consider the case  $f(\mu_i, T_i|s, h) = f(\mu_i|s, h) \times f(T_i)$ , i.e. confoundedness in types. If  $\partial\delta(\overline{H}_i)/\partial\overline{H}_i = 0$ ,  $E[ictr_i^1|s, \Delta h] > 0 \quad \forall s$  and  $\partial E[ictr_i^1|s, \Delta h]/\partial s > 0 \quad \forall s$  can only occur if  $E[\mu_i|s, \Delta h] > 0 \quad \forall s$  and  $\partial E[\mu_i|s, \Delta h]/\partial s > 0 \quad \forall s$ . Second, Consider the case  $f(\mu_i, T_i|s, h) = f(T_i|s, h) \times f(\mu_i)$ , i.e. confoundedness in age. Further assume that  $\partial f(\mu_i, T_i|s, h)/\partial s > 0$ , i.e. more popular keywords are on average older. If  $\partial\delta(\overline{H}_i)/\partial\overline{H}_i = 0$ ,  $E[ictr_i^1|s, \Delta h] > 0 \quad \forall s$  and  $\partial E[ictr_i^1|s, \Delta h]/\partial s > 0 \quad \forall s$  can only occur if  $E[T_i|s, \Delta h] > 0 \quad \forall s$  and  $\partial E[T_i|s, \Delta h]/\partial s > 0 \quad \forall s$ .*

Proposition 1 states that, if  $\partial\delta(\overline{H}_i)/\partial\overline{H}_i = 0$ , a divergent pattern in  $E[ictr_i^1|s, \Delta h]$  can only occur if the confoundedness exactly replicates this pattern. To understand what this means, consider the case of the unobserved type of keywords. Divergence in  $E[ictr_i^1|s, \Delta h]$  cannot simply be explained by an average difference in type between keywords with  $h = h^l$  and  $h = h^s$ . Instead, what is required is that the average difference in type between  $h = h^l$  and  $h = h^s$  increases with  $s$ . Similarly for age, any positive age difference between  $h = h^l$  and  $h = h^s$  would have to increase with  $s$  in order to rationalize the data in the absence of network effects. Proposition 1 imposes restrictive condition on the confoundedness to generate the data we observe: The positive correlation between types or age and the average user history needs to increase monotonically with  $s$ .<sup>20</sup>

---

<sup>20</sup>Note that for age confoundedness in Proposition 1, we introduce the assumption that keywords with a larger monthly search quantity need to be on average older. We consider this a weak additional restriction. An increasing age gap between keywords with a long and short average user history as a function of  $s$  with a simultaneous decrease of the average age with  $s$  seems artificial. Furthermore, it should also be noted that potential confoundedness between the age and average user history seems at odds with the results in subsection 4.2. If keywords with a longer average user history are older, on average, we would expect



**Figure 6:** Evolution of quality differences when data in the *user dimension* boost learning from data in the *keyword dimension*. We assume no simultaneous confoundedness.

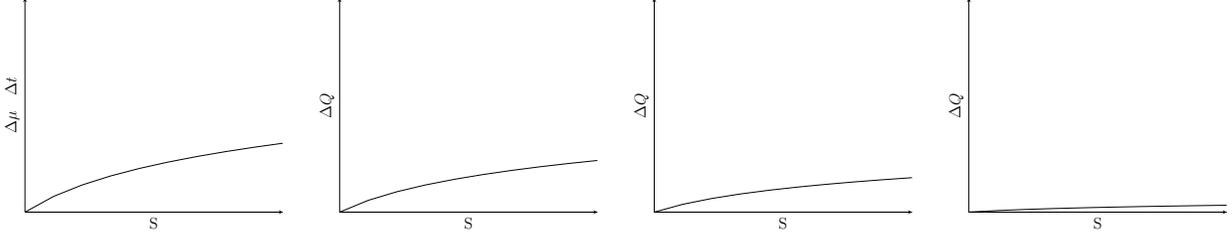
By contrast, if data obtained on the *user dimension* boost learning in the *keyword dimension*, the pattern observed in our data arises naturally. Figure 6 schematically describes the three main patterns we would observe in this case if the data were generated by the model introduced in this section, absent confoundedness. The pattern we observe depends on the average age of the keywords. For a young average age, we would observe a divergent pattern as the one presented in Figure 6a. As the average age increases, initially the pattern becomes similar to that shown in Figure 6b before reaching a state similar to that shown in Figure 6c.<sup>21</sup>

The main reason for the three different patterns is that all keywords eventually converge to the same quality limit. This is what explains the transition from the pattern shown in Figure 6a to the pattern shown in Figure 6b. After an initial phase of divergence in quality between  $h^l$  and  $h^s$ , quality starts to converge after a certain threshold,  $S^*$ , is reached. Because this threshold is first reached by keywords with a large number of searches per month, we observe the pattern in Figure 6b for an intermediate average age of keywords. As the average age of keywords increases, the location of the kink shifts to the left, which eventually leads

---

them to experience a smaller quality increase during our sample than keywords with a shorter average user history. The same cannot be said about potential confoundedness with respect to the type, which is generally compatible with the results found in subsection 4.2. If keywords with a longer average user history are of a higher type, this implies a younger age than compared to keywords with a short average user history if both have the same initial quality level. This difference in age implies that keywords with a longer average user history are on the steeper part of the learning curve as compared to keywords with shorter average user history.

<sup>21</sup>We provide a formal discussion of the change in the pattern in Appendix E.



(a) Average difference in type or age between  $h^l$  and  $h^s$ . (b) Average difference in quality between  $h^l$  and  $h^s$  if average age is small. (c) Average difference in quality between  $h^l$  and  $h^s$  if average age is intermediate. (d) Average difference in quality between  $h^l$  and  $h^s$  if average age is large.

**Figure 7:** Evolution of quality differences induced by unobserved heterogeneity as average age of keywords increases.

to a pattern similar to 6c. The average age that marks the transition between the different phases mainly depends on the difference in the speed of convergences between  $h^l$  and  $h^s$ .

The functional form in Equation 5 implies that, when data on the *user dimension* do not impact learning from data on the *keyword dimension*, initial quality differences continuously diminish. As a consequence, we would not observe persistent differences in the data only due to unobserved heterogeneity. Observed quality differences between keywords with a long and short average user history vanish as the average age of keywords increases. Figure 7 illustrates how the quality differences between keywords with a long and short average user history, which are caused by confoundedness, vanish as the average age of keywords increases.<sup>22</sup>

The analysis of this section highlights that strong assumptions on the unobserved heterogeneity are required to generate patterns consistent with our data if the average user history does not impact the speed of learning from more searches on the same keyword. In contrast, when data in the *user dimension* boost learning in the *keyword dimension*, our model generates patterns consistent with our data if the average age of keywords is in a certain range. The results of this section suggest that simultaneous confoundedness of age and type

<sup>22</sup>The pattern shown in Figure 7 is just one example of many potential patterns generated through heterogeneity. The main point of Figure 7 is to illustrate that any initial differences in quality generated through unobserved heterogeneity vanish monotonically, not to suggest a specific shape of the observed divergence.

would require assumptions similar to the one formulated for each factor of confoundedness in isolation, i.e. confoundedness between the unobservables and the average user history needs to be reinforced with  $S$  in order to generate the patterns observed in the data.

## 6 Conclusion

In this paper we propose a mechanism that rationalizes the hypothesis of data as a source of market power. We find evidence that more comprehensive data about the users' personal search histories triggers faster learning from data on searches upon the same keyword. Our results suggest that a search engine with longer user search histories is able to provide higher quality search results to the same keyword observed previously equal times.

We view the mechanism that we propose as a network effect from additional data: More data increase the efficiency of the technology, thereby causing positive externalities similar to those created by more users in a network. The search engine improves as it observes more users entering the same keyword. This is akin to the classical network effect, whereby more users improve the search result quality. In addition, we document another effect of learning from data: more data on the users' personal search histories lead to faster learning from data obtained on searches on the same keyword.

Our results provide a novel insight in the discussion on the necessary scale for a search engine to operate effectively in a competitive environment: data may provide first movers an advantage that can become difficult for rivals to overcome. This is because more data on user search histories allow the search engine to learn faster on observed click behaviour for the same keyword. Search engines with access to longer user search histories need less users to reach a given search result quality for the same keyword.

Our findings rationalize why data on users might be particularly valuable for firms operating data-driven technologies. Because data is not simply an input but also a technology shifter, our results suggest that data have the potential to confer a significant competitive

advantage and to act as a barrier to entry. Our results call for awareness from antitrust policy regarding any potentially anti-competitive behaviour of firms seeking to deepen knowledge about their existing customer base. For instance, our results suggest that the benefits from locking in customers might be substantial. Similarly, merging databases across different services with a large overlap in the user base might grant firms a data advantage that is difficult to overcome for competitors. In this context, it is important to understand how market demand reacts to quality differences. This is an empirical question that remains to be answered.

Our research also reveals that there may be a strong relationship between privacy and competition policy, since the privacy interest of users may be weighed off against the externalities potentially generated through personalized information. These externalities may be positive when they enable better search results. But they may also accumulate into an advantage in data that helps a firm cement its market power. Our findings call for a careful interaction between privacy and competition policy that preserves the benefit of competition but may even allow for some data sharing to create a level playing field and tap positive externalities. In light of our findings, the right to data portability, which enables users of IT services to easily carry their personal data to other service providers, implemented in Article 20 of the EU General Data Protection Regulation, is a step in a right direction.

However, our findings also suggest that in order to substantially benefit entrants, the right of data portability would have to be exerted by a sufficient number of users switching to the new entrant. A lack of coordination and the under appreciation of the externalities involved in carrying private data to the potential entrant might result in only a few customers making use of data portability and, consequently, in sub-optimal levels of switching. A systematic theoretical and empirical assessment of user switching in light of externalities from data across users would be a valuable contribution to future research.

## References

- Adomavicius, G. and Tuzhilin, A. (2005) Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, *IEEE Transactions on Knowledge & Data Engineering*, **17**, 734–749.
- Argenton, C. and Prüfer, J. (2012) Search Engine Competition with Network Externalities, *Journal of Competition Law and Economics*, **8**, 73–105.
- Bajari, P., Chernozhukov, V., Hortaçsu, A. and Suzuki, J. (2019) The Impact of Big Data on Firm Performance: An Empirical Investigation, in *AEA Papers and Proceedings*, vol. 109, pp. 33–37.
- Barnett, A. G., van der Pols, J. C. and Dobson, A. J. (2004) Regression to the Mean: What It Is and How to Deal With It, *International Journal of Epidemiology*, **34**, 215–220.
- Casadesus-Masanell, R. and Hervas-Drane, A. (2015) Competing with privacy, *Management Science*, **61**, 229–246.
- Chiou, L. and Tucker, C. (2017) Search Engines and Data Retention: Implications for Privacy and Antitrust, *NBER Working Paper Series*, 23815.
- Chuklin, A., Serdyukov, P. and De Rijke, M. (2013) Click Model-Based Information Retrieval Metrics, in *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval*, ACM, pp. 493–502.
- Claussen, J., Peukert, C. and Sen, A. (2019) The Editor vs. the Algorithm: Economic Returns to Data and Externalities in Online News, *SSRN*, <https://ssrn.com/abstract=3399947> (November 3, 2019).
- Decarolis, F., Goldmanis, M. and Penta, A. (2020) Marketing agencies and collusive bidding in online ad auctions, *Management Science*.
- Decarolis, F. and Rovigatti, G. (2019) From Mad Men to Maths Men: Concentration and Buyer Power in Online Advertising, *CEPR Discussion Paper No. DP13897*.
- European Commission (2018) Commission Decision AT.40099 – Google Android, [https://ec.europa.eu/competition/antitrust/cases/dec\\_docs/40099/40099\\_9993\\_3.pdf](https://ec.europa.eu/competition/antitrust/cases/dec_docs/40099/40099_9993_3.pdf) (July 18, 2018).
- European Commission (2020) Communication from the commission to the european parliament, the council and social committee and the committee of the regions: A european strategy for data, Tech. rep., European Commission.
- Grunes, A. P. and Stucke, M. E. (2015) No Mistake About It: The Important Role of Antitrust in the Era of Big Data, *SSRN*, <https://ssrn.com/abstract=2600051> (April 28, 2015).
- Hagiu, A. and Wright, J. (2020) Data-enabled learning, network effects and competitive

- advantage, Tech. rep., Working Paper.
- He, D., Kannan, A., Liu, T., McAfee, R., Qin, T. and J.M., R. (2017) Scale Effects in Web Search, In: R. Devanur N., Lu P. (eds) Web and Internet Economics. WINE 2017. Lecture Notes in Computer Science, vol 10660. Springer, Cham.
- Joachims, T. (2002) Optimizing Search Engines Using Clickthrough Data, in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 133–142.
- Jones, C. I. and Tonetti, C. (2020) Nonrivalry and the Economics of Data, *American Economic Review*, **110** (9), 2819–58.
- Lambrecht, A. and Tucker, C. E. (2015) Can Big Data protect a Firm from Competition?, *SSRN*, <https://ssrn.com/abstract=2705530> (December 18, 2015).
- Netflix (2020) Netflix research - personalization & search: Helping members discover content they’ll love.
- Newman, N. (2014) Search, Antitrust, and the Economics of the Control of User Data, *Yale J. on Reg.*, **31**, 401.
- Posner, E. A. and Weyl, G. (2018) *Radical Markets, Uprooting Capitalism and Democracy For a Just Society*, Princeton University Press.
- Prüfer, J. and Schottmüller, C. (2017) Competing with Big Data, *SSRN*, <https://ssrn.com/abstract=2918726> (February 16, 2017).
- Schaefer, M., Sapi, G. and Lorincz, S. (2018) The effect of big data on recommendation quality: The example of internet search, *DIW Berlin Discussion Paper No. 1730*.
- Schepp, N.-P. and Wambach, A. (2015) On Big Data and its Relevance for Market Power Assessment, *Journal of European Competition Law & Practice*, **7**, 120–124.
- Sokol, D. D. and Comerford, R. (2015) Antitrust and Regulating Big Data, *Geo. Mason L. Rev.*, **23**, 1129.
- The Economist (2017) The world’s most valuable resource is no longer oil, but data, <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data> (May 6, 2017).
- Tucker, C. (2019) Digital Data, Platforms and the Usual [Antitrust] Suspects: Network Effects, Switching Costs, Essential Facility, *Review of Industrial Organization*, **54**, 683–694.
- Yoganarasimhan, H. (2019) Search Personalization Using Machine Learning, *Management Science*.

# A Robustness Analysis with Alternative Quality Measures

This Appendix provides robustness checks for the results presented in sections 4.1 and 4.2 using alternative quality measures. Appendix A.1 provides an introduction and discussion of the editorial quality measures before presenting the corresponding results in Appendix A.2. In Appendix A.3, we present results based on alternative click-based quality measures.

## A.1 Editorial Quality Measures – Introduction

Our data set contains 659,000 query-URL pairs with editorial relevance judgments collected from human experts. The editorial quality judgments assess the relevance of a URL for a specific query by a categorical grade ranging from zero (*not at all relevant*) to four (*highly relevant*). By aggregating the editorial quality judgments for multiple URLs displayed on the same search result page, it is possible to obtain an overall *grade* for the quality of the result page.

Editorial quality measures are often used in the information retrieval (IR) literature to assess the quality of different algorithms in an offline environment and to compare the quality of various algorithms in an experimental setting (Chuklin *et al.*, 2013). Editorial quality measures are criticised for often struggling to capture user preferences in an online setting, which led IR researchers to develop relevance metrics based on user click behaviour (Chuklin *et al.*, 2013). We nevertheless repeat the analysis of repeat the analysis of sections 4.1 and 4.2 based on editorial quality measures.

A commonly used quality measure in the information retrieval (IR) literature is the so-called *discounted cumulative gain* (DCG). The informational *gain* of a specific URL for a specific topic is directly assessed by the relevance grade that ranges from zero to four. The position of the URL on the result page determines by how much this informational *gain* is *discounted*. For example, assume that a specific URL is rated with a relevance grade of four

relative to the searched topic. Furthermore, assume that this URL is shown on the second position of the corresponding result page. Then, we say that the *discounted gain* (DG) of this URL is given by:

$$DG = \frac{2^{rel_j} - 1}{\log_2(j + 1)} = \frac{2^4 - 1}{\log_2(2 + 1)}, \quad (7)$$

Where  $j$  stands for the position and  $rel$  for the relevance grade of the URL. The numerator captures the informational *gain* that the user obtains by being provided this with this URL. The denominator discounts for the fact that the URL is displayed in the second position: The user had to *scan* through the search result page to be provided with this URL. Note that by applying the logarithm of base two to the denominator, the gain of a document displayed on the first position is not discounted.

To assess the quality of the entire result page, one can add up the discounted gain of all documents displayed on the first results page. Assume for convenience that all ten documents on the first result page are assigned a relevance judgment, then the discounted cumulative gain is given by:

$$DCG^p = \sum_{j=1}^{p=10} \frac{2^{rel_j} - 1}{\log_2(j + 1)}, \quad (8)$$

Two criteria determine the value of the DCG: (I) the general relevance of the documents available on the result page and (II) the ranking of the documents. (I) simply captures the idea that providing documents with relevant content is generally desirable (i.e. a lot of documents rated four are better than a lot of documents rated with a grade of one). (II) captures the idea that, given a specific set of documents with a given relevance, it is desirable to display the most relevant documents at the top of the result page (the ordering 4,3,2 is

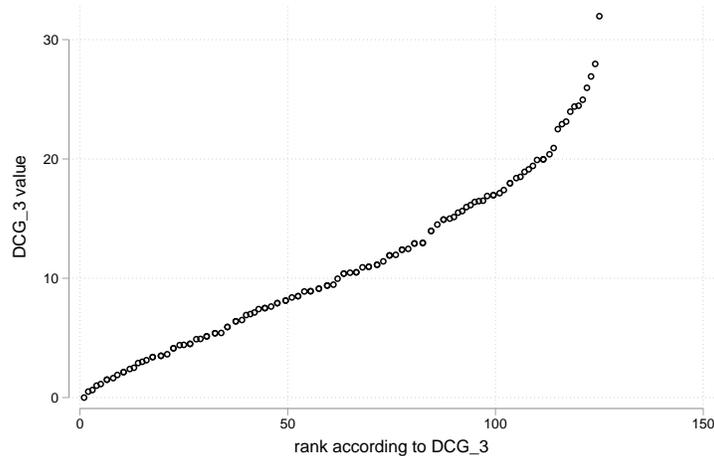
better than the ordering 2,3,4). The DCG captures both dimensions.

Obviously, in order to be able to compute the DCG for the entire result page, we need relevance judgments for all the URLs displayed on the page. This is only rarely the case in our dataset. Another shortcoming of the DCG measure is that the measure only allows a meaningful comparison between result pages if both pages are exactly the same number of consecutively graded URLs starting from the URL displayed on the top. It is, for example, not possible to directly compare a search result page where the first two URLs are graded with a result page where the first three URLs are graded. It is also not possible to directly compare a result page where the first three URLs are rated with a result page where only the first and third URLs are rated but the grade for the second is missing. Therefore, comparing result pages based on a DCG measure with a certain depth  $p$  requires that all compared pages have URLs consecutively rated until  $p$ .

Consequently, the IR literature deals extensively with the imputation of relevance grades for URLs with missing grades. Usually, imputed grades are assigned based on click through rates (CTR) for the URL with the missing grade that take into consideration existing relevance grades of nearby URLs. Repeating such an exercise for the present dataset would be extremely burdensome and costly. It is for this reason that we decided to take another approach and to repeat the analysis based on two DCG-measures with different depth  $p$ :  $DCG^1$  and  $DCG^3$ . The  $DCG_1$  measure can be calculated for approximately 90 percent of the searches in the data set. The  $DCG_3$  measure can only be computed for roughly one-third of the searches.

While the  $DCG^1$  measure allows us to compute a grade for most of the searches in our data, it only takes into account the first URL. As a consequence, a results page can only be assigned 5 possible grades. While the  $DCG^3$  measure allows for 125 different possible grades, it can only be computed for a fraction of the observed searches.

Each combination of relevance judgments gives rise to a particular DCG-value. These DCG-values can be used to establish an ordinal ranking of the relevance judgment combina-



**Figure 8:** y-axis:  $DCG^3$  value, x-axis: Ordinal ranking of relevance judgment-combinations according to their  $DCG^3$  value

tions from 1 (lowest DCG-value) to 5 (highest DCG-value) in the case of the  $DCG^1$  measure and from 1 to 125 in the case of the  $DCG^3$  measure. Figure 8 depicts the DCG-values (y-axis) against their ordinal ranking (x-axis) for the  $DCG^3$  measure.

Figure 8 illustrates the convex nature of the DCG-measure: the difference in DCG-values between the relevance judgment combination (4,4,4) and (4,4,3) is larger than the difference in DCG-values between the relevance judgment combination (0,0,1) and (0,0,0). In other words, incremental improvements of relevance judgment combinations lead to higher DCG increases as we move along the ordinal ranking of URL-combinations. This is due to the fact that the relevance judgments enter the DCG formula in the exponent.

This property is mechanical rather than informative about the true added quality gain for customers. In the above example it is debatable, whether the improvement from (0,0,0) to (0,0,1) is more or less valuable to the consumer than the improvement from (4,4,3) to (4,4,4). Classical economic thinking would suggest that the former is more valuable than the latter, if relevance judgments could be interpreted directly as “information-units”.

If we use the ordinal ranking of the URL-combinations each incremental improvement is valued the same. An improvement for a given keyword would then be larger, the more “steps” it improved on the ordinal ranking. This choice seems the most sensible to us. It is

for this reason that we opt for the ordinal rank dictated by the DCG measure rather than for the DCG measure itself to perform our analysis. For the remainder of the analysis, we refer to the editorial quality measures as  $rank^1$  and  $rank^3$ , respectively. The correlation coefficient between  $rank^1$  and  $ctr^1$  is 0.55. The correlation coefficient between  $rank^3$  and  $ctr^1$  is 0.49.

## A.2 Editorial Quality Measures – Results

Figure 9 shows the results of estimating Equation 3 on the grid  $s \times h$  with  $rank^1$  as the dependent variable. In Figure 9a, estimation is performed for all keywords in the sample. In Figure 9b estimation is performed for the subset of keywords that do not deviate by more than ten percentage point from the even accumulation criterion. Figure 10 shows the corresponding analysis with  $rank^3$  as the dependent variable.

For both editorial quality measures, only minor changes to the observed pattern are caused by dropping keywords that see less steady search patterns over time. The changes are larger when we perform the same exercise using the click-based quality measures and removal keywords that see more rapid boosts in searches on some days.

Nevertheless, Figures 11 and 12 reveal that the effect of removing keywords do not accumulate evenly increases the measured impact of the network effect for both editorial quality measures in a similar way as does the click-based quality measure. Furthermore, the divergent pattern in the quality levels between keywords with a long and short average user history that we observe for the click-based quality measures is also present when using the editorial quality measures.

Figures 13 and 14 show the results of estimating Equation 4 on the grid  $ictr \times h \times s$  for  $rank^1$  and  $rank^3$ , respectively. For both editorial quality measures, we normalized the quality to lie in the interval  $[0, 1]$  when analyzing the quality evolution of keywords.

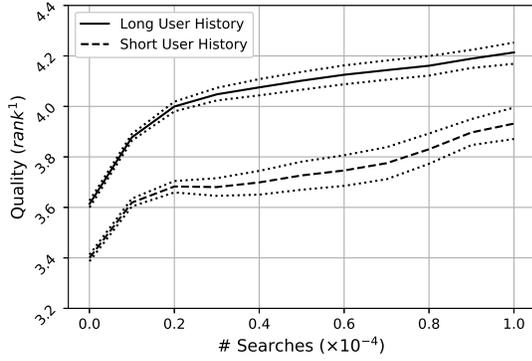
In both Figure 13 and Figure 14, we observe that the measured quality increase generally declines with the initial quality level. Furthermore, the quality increase of keywords with

a long average user history is always weakly larger than the quality increase measured for keywords with a short average user history.

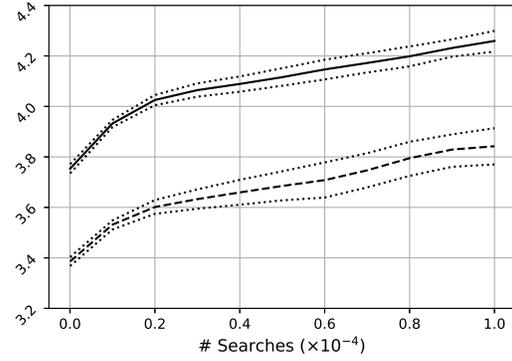
For the  $rank^1$  measure, we observe no difference between long and short keywords for the lowest initial quality level. The large confidence intervals reflect the fact that only few keywords start from this quality level in our sample. For the second highest initial quality level, the observed pattern roughly corresponds to our expectations of differential learning speeds between keywords with a long and short average user history. A general problem that we see with the  $rank^1$  measure is its very crude scale, which only allows for five different quality grades. If differential learning between keywords with a long and short average cookie length occurs on a more granular level within the sample period, the  $rank^1$  measure is likely to poorly capture differential learning.

For the  $rank^3$  measure, the difference in learning between keywords with a long and short average user history is most pronounced for the lowest initial quality level. For larger initial quality levels, the difference is not pronounced. The large confidence intervals that we observe for all initial levels reflect the loss of observations associated with using the  $rank^3$  quality measure.

Altogether, the results obtained with the editorial quality measures are in line with the results obtained using the click-based quality measures. The divergence of the quality levels observed in Figures 11 and 12, and the differential learning speed observed in Figures 13 and 14 for lower initial quality levels are in line with the hypothesis that additional data in the *user dimension* boosts learning from data in the *keyword dimension*.

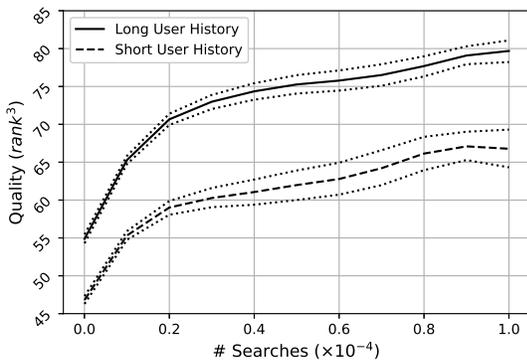


(a) Entire sample of keywords

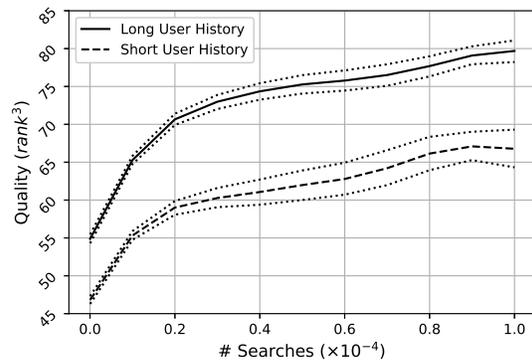


(b) Keywords accumulating evenly

**Figure 9:** Average quality of keywords with 95 % confidence intervals

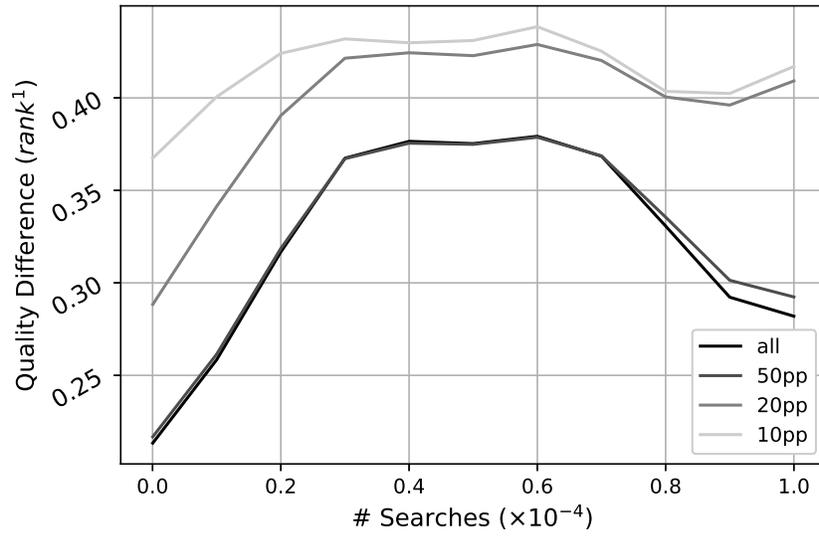


(a) Entire sample of keywords

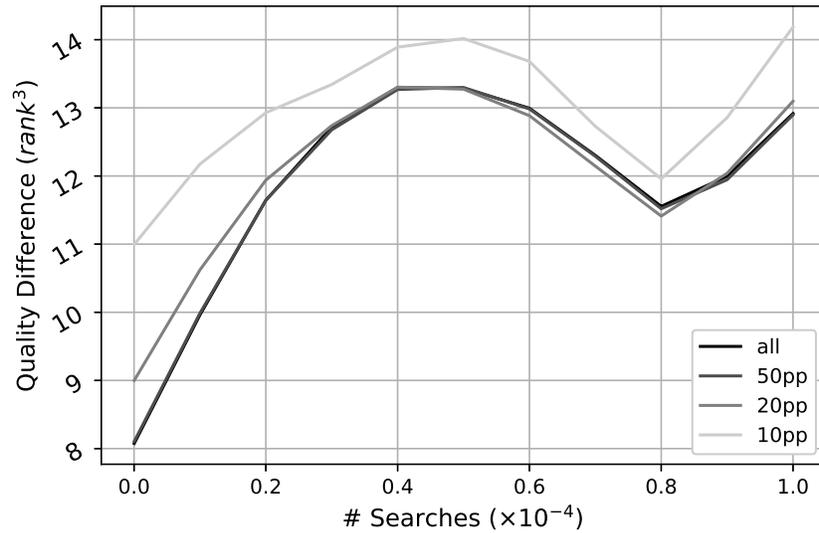


(b) Keywords accumulating evenly

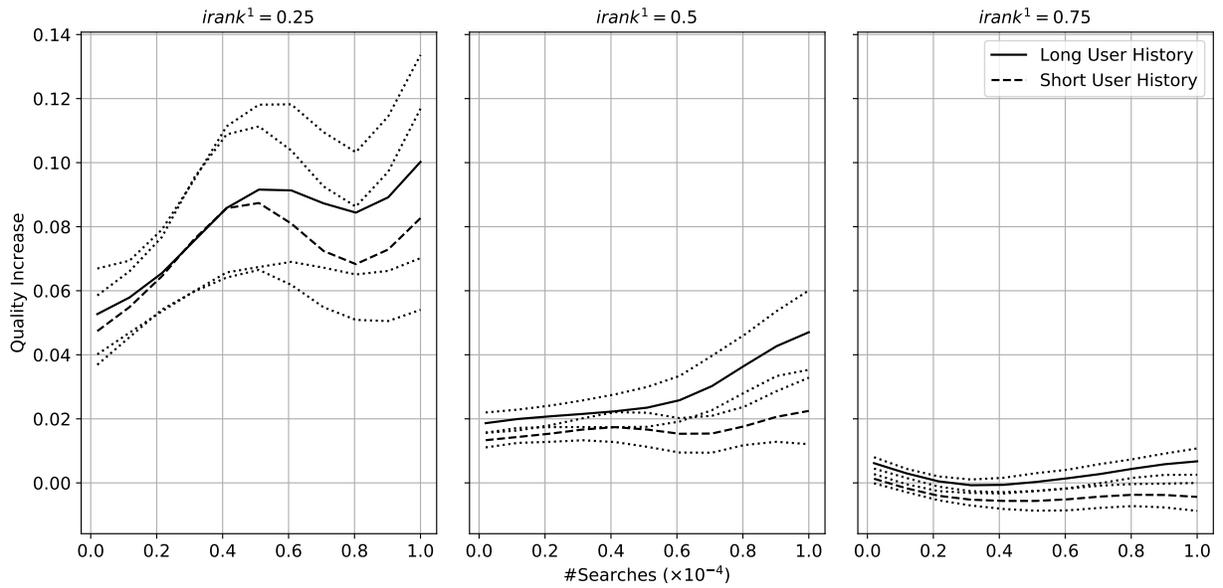
**Figure 10:** Average quality of keywords with 95 % confidence intervals



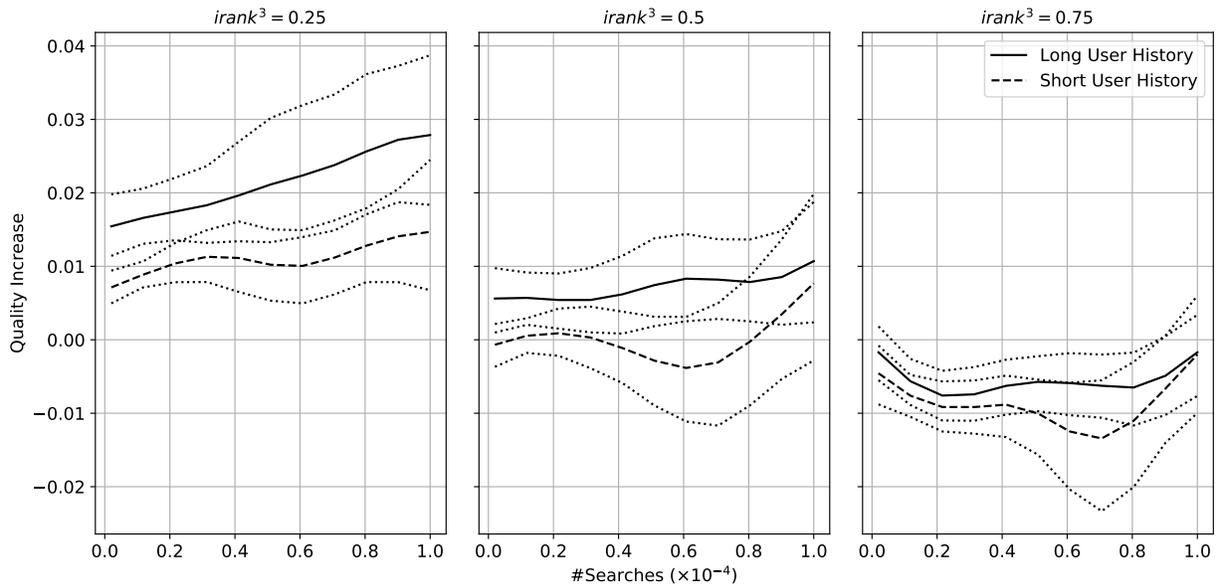
**Figure 11:** Impact of longer average user history on quality level. Different tolerance levels for deviation from even accumulation criterion are considered.



**Figure 12:** Impact of longer average user history on quality level. Different tolerance levels for deviation from even accumulation criterion are considered.



**Figure 13:** Average quality increase of keywords with 95% confidence intervals. Each panel stands for a different initial quality level.



**Figure 14:** Average quality increase of keywords with 95% confidence intervals. Each panel stands for a different initial quality level.

### A.3 Alternative Click Based Quality Measures – Results

In this Appendix, we present the results obtained with two alternative click-based quality measures. For the first, search result quality is encoded as “good” if the last recorded click occurs on one of the first three organic URLs. For the second, search result quality is encoded as “good” if the last recorded click occurs on one of the first ten organic URLs. We denote the first alternative click based measure by  $ctr^3$  and the second by  $ctr^{all}$ . For  $ctr^{all}$ , search result is only considered “bad” if the searcher ends the search on the second result page or leaves the search engine without finding a URL she considers relevant.

**Table 1:** Correlation click based vs. editorial measures

	$ctr^1$	$ctr^3$	$ctr^{all}$
$rank^1$	0.55	0.51	0.41
$rank^3$	0.49	0.45	0.34

From Table 1, it can be seen that the correlation between the click based quality measures and each editorial quality measure decreases as the criterion for a “good” search result quality is defined more broadly. Furthermore, each click based quality measure correlates more highly with the editorial quality measure of depth one.

Figure 15 shows the results of estimating Equation 3 on the grid  $s \times h$  with  $ctr^3$  as the dependent variable. The results in Figure 15a are based on all the keywords in the sample. Figure 15b shows the results for the population of keywords that deviate no more than ten percentage points from the even accumulation criterion. Figure 16 repeats the same exercise for the  $ctr^{all}$  quality measure. Figures 17 and 18 shows the consequence of continuously narrowing the tolerance interval for the even accumulation criterion on the measured network effects for the  $ctr^3$  and  $ctr^{all}$  quality measure, respectively. All Results are qualitatively identical to the results obtained based on the  $ctr^1$  quality measure.

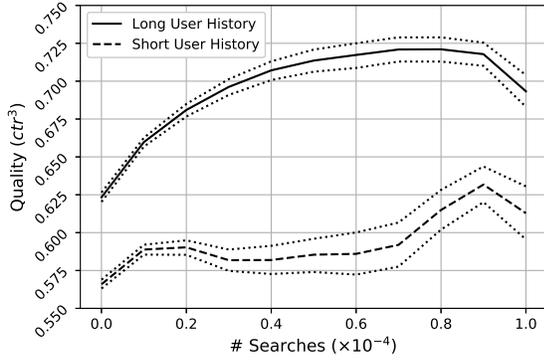
Figures 19 and 20 show the results of estimating Equation 4 on the grid  $ictr \times h \times s$  for  $ctr^3$  and  $ctr^{all}$ , respectively. The results are qualitatively identical to the ones obtained

based on the  $ctr^1$  quality measure.

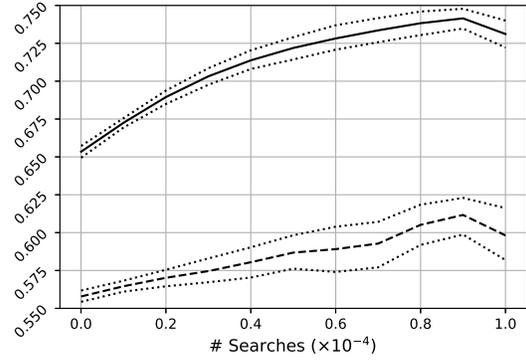
It is noteworthy that, in general, the measured network effect between keywords with a long and short average user history tends to decrease as we broaden the set of URLs that determine a “good” search result quality. We find this result intuitive: It requires less personalized knowledge to get three out of ten or ten out of ten results right as it requires to get one out of ten results. A more broadly defined quality measure therefore de-emphasizes the value of personalized information.

Another noteworthy difference to the results in the main section is that the quality evolution during the sample period is now also positive for keywords with a high initial quality level and a short average user history. This indicates that the decrease in the click through rate on the first URL, which we found in Section 4.2, corresponds to an increase in the click through rate for URLs further down the results list, i.e. that URLs further down the result list cannibalize clicks on the first URL. Again, it seems intuitive to see this phenomenon occur for keywords with a short average user history, where the search engine lacks the information to target individual preferences.

Altogether, the results based on the alternative click-based quality measure prove to be robust to the results presented in the main section of the paper. Minor differences to the results in the main section appear to be sensible.

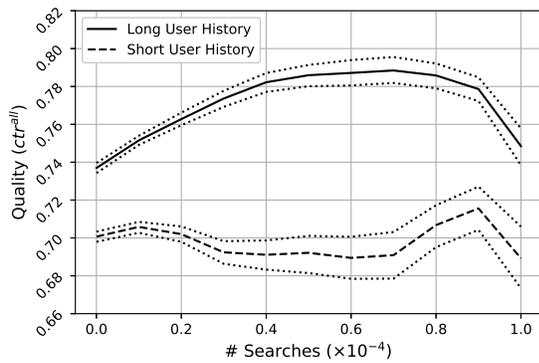


(a) Entire sample of keywords

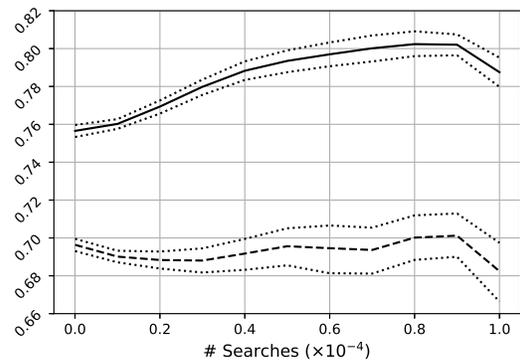


(b) Keywords accumulating evenly

**Figure 15:** Average quality of keywords with 95 % confidence intervals

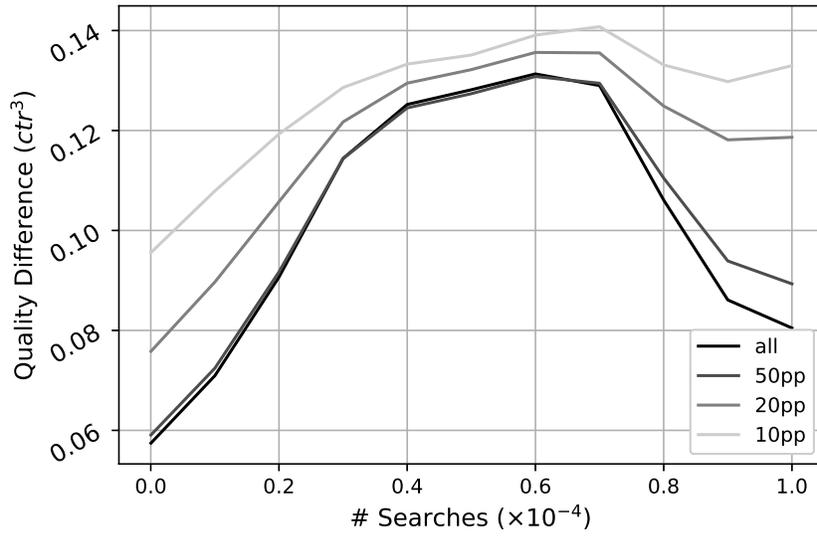


(a) Entire sample of keywords

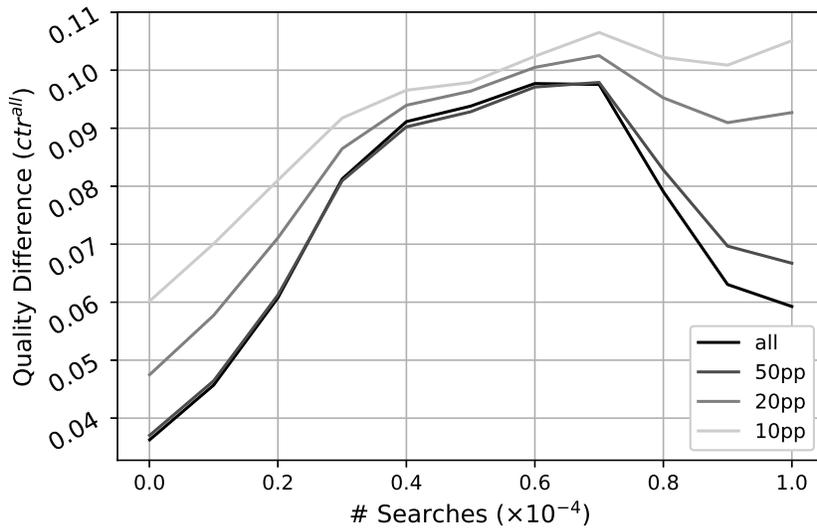


(b) Keywords accumulating evenly

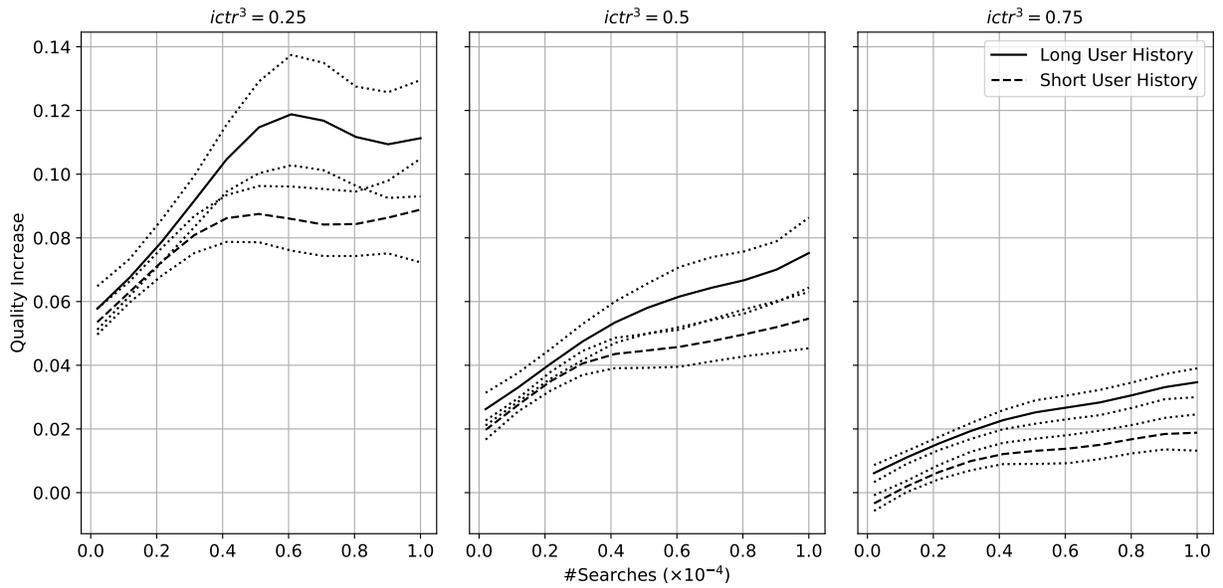
**Figure 16:** Average quality of keywords with 95 % confidence intervals



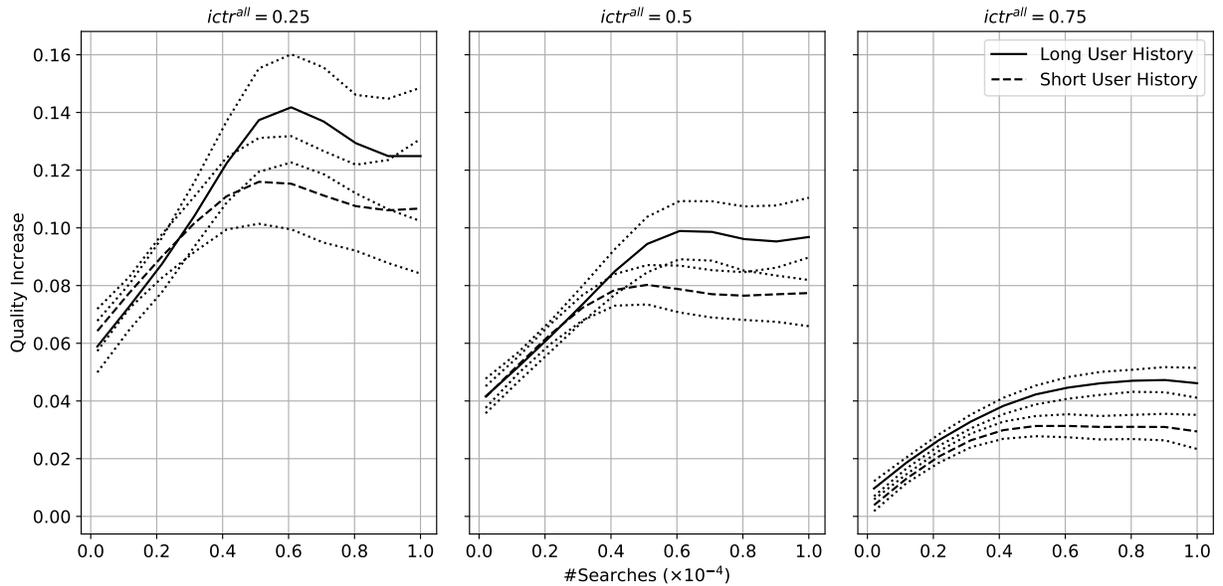
**Figure 17:** Impact of longer average user history on quality level. Different tolerance levels for deviation from even accumulation criterion are considered.



**Figure 18:** Impact of longer average user history on quality level. Different tolerance levels for deviation from even accumulation criterion are considered.



**Figure 19:** Average quality increase of keywords with 95% confidence intervals. Each panel stands for a different initial quality level.



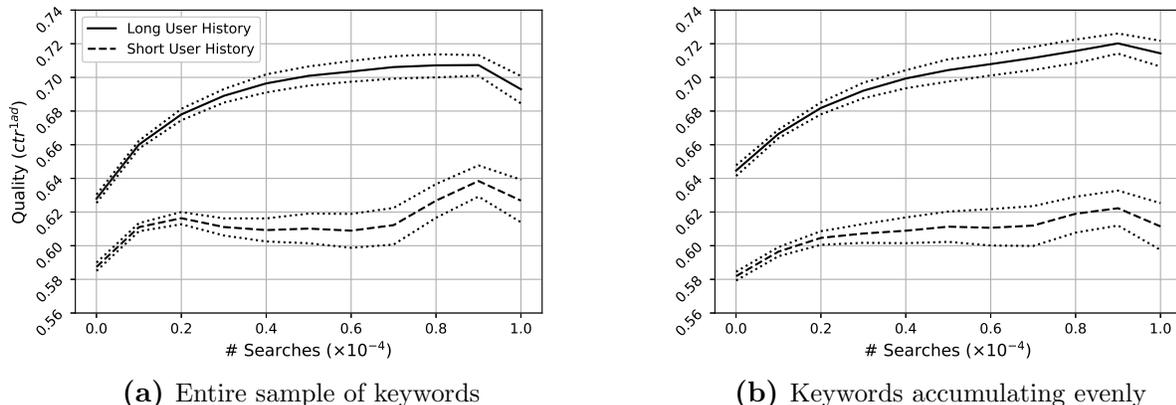
**Figure 20:** Average quality increase of keywords with 95% confidence intervals. Each panel stands for a different initial quality level.

## A.4 Counting clicks on advertisement URLs – Results

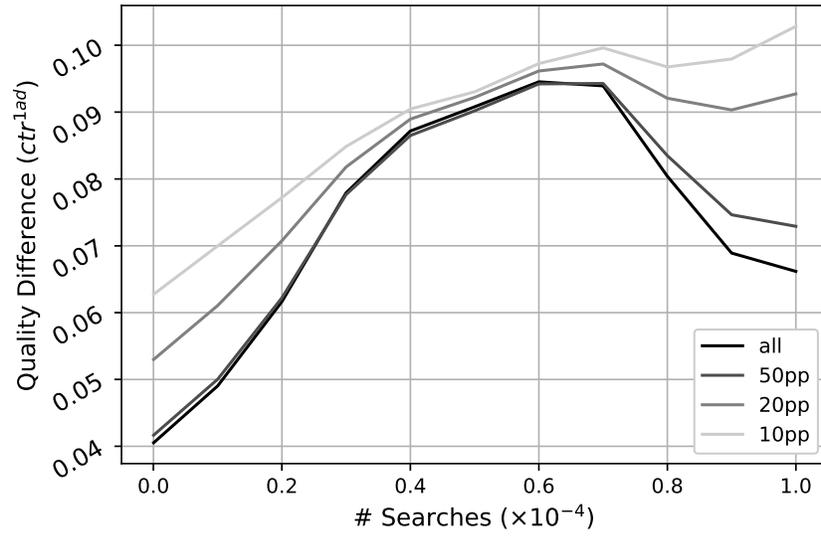
This Appendix gives the results for the robustness checks when search result quality is encoded as *good* if the last click occurs on the first URL or an Advertisement URL. The click through rate is now computed as follows:

$$ctr_i^{\{1ad\}} = \frac{\sum_{s_i \in S_i} \mathbb{1}\{lcp_{is} \in \{0, 1\}\}}{\sum_{s_i \in S_i} \mathbb{1}\{lcp_{is} = \Omega\}} \quad (9)$$

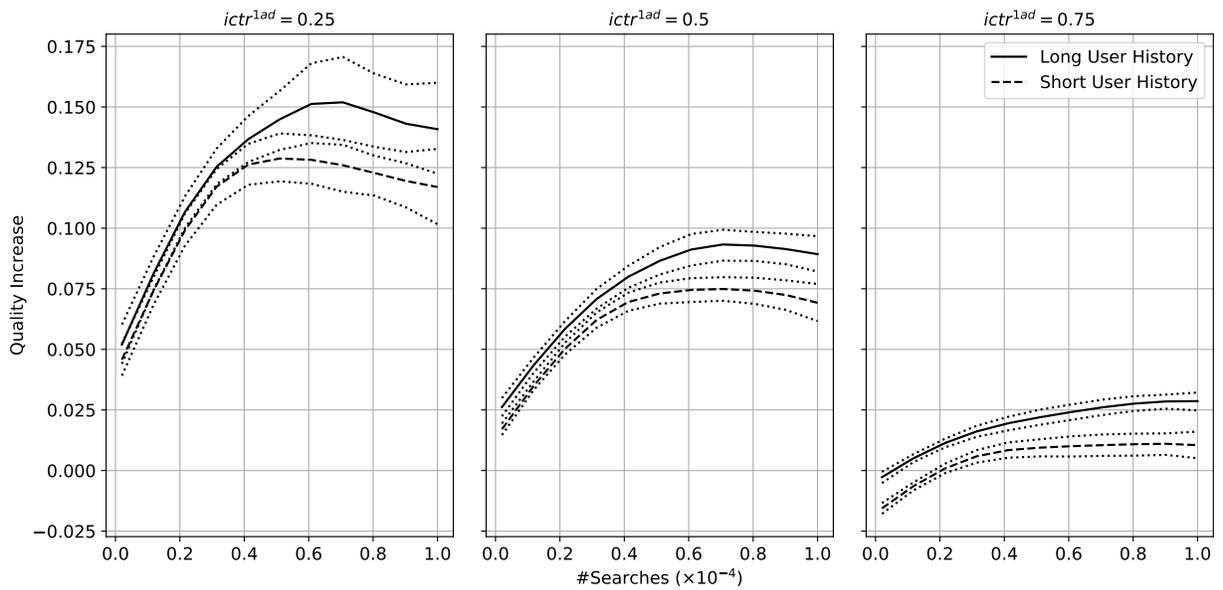
In comparison to the other click based quality measures used in the paper, the denominator now sums up all the searches in  $S$ . The numerator now considers clicks on advertisement URLs as “good” search result quality. For completeness, it should be noted that the data description offers no clear guidance on the exact nature of the clicks that we characterize as “ads”. These clicks could be clicks on ads but also clicks on spelling suggestions or reformulations of the original keyword the user submitted. The results in Figures 21, 22 and 23 show that our conclusions remain unaffected by whether or not, we count these clicks or not.



**Figure 21:** Average quality of keywords with 95 % confidence intervals



**Figure 22:** Impact of longer average user history on quality level. Different tolerance levels for deviation from even accumulation criterion are considered.



**Figure 23:** Average quality increase of keywords with 95% confidence intervals. Each panel stands for a different initial quality level.

## B Formal Discussion of Data Generating Process

In subsection 4.1, we use the variables we observe,  $s_i$  and  $\bar{h}_i$ , as proxy measures for  $S_{it}$  and  $\bar{H}_{it}$  and demonstrate that keywords for which the proxies indicate larger values of  $S_{it}$  and  $\bar{H}_{it}$  are on a higher quality level. In this Appendix, we are interested in the formal relationship between  $s_i$  and  $\bar{h}_i$  and their unobserved counterparts at time  $t$ , which denotes the beginning of our sample period.

Reducing the approximation error we make from substituting  $S_{it}$  and  $\bar{H}_{it}$  by their proxy variables  $s_i$  and  $\bar{h}_i$  amounts to reducing  $\text{Var}[S_{it}|s, h]$  and  $\text{Var}[\bar{H}_{it}|s, h]$  for each tuple  $\{s_i = s, \bar{h}_i = h\}$  that we condition on. Conditioning on a tuple  $\{s_i = s, \bar{h}_i = h\}$  induces a distribution over  $S_{it}$  and  $\bar{H}_{it}$ . The expectations of these distributions,  $E[S_{it}|s, h]$  and  $E[\bar{H}_{it}|s, h]$ , describe the expected values of  $S_{it}$  and  $\bar{H}_{it}$  for each tuple  $\{s_i = s, \bar{h}_i = h\}$ . The variances describe the heterogeneity over  $S_{it}$  and  $\bar{H}_{it}$  for each tuple  $\{s_i = s, \bar{h}_i = h\}$ . The larger the variances, the larger the error we make from substituting  $S_{it}$  and  $\bar{H}_{it}$  by  $s_i$  and  $\bar{h}_i$ .

We now introduce a generating process under which  $\text{Var}[S_{it}|s, h]$  and  $\text{Var}[\bar{H}_{it}|s, h]$  are minimized.

**DGP 1:** Keywords originate at random time and have constant monthly popularity. i.e. the number of searches for the keywords is constant in every month. I.e.  $s_{i\tau} = s_i \quad \forall \tau$ , where  $\tau$  denotes a month.

**DGP 2:** If two keywords follow **DGP 1** and have the same average user history in month  $\tau$ , they also have the same average user history in month  $\tau + 1$ .

To understand why **DGP 1** and **DGP 2** reduce the variance, it is crucial to note that there are two sources of heterogeneity that (potentially) influence  $\text{Var}[S_{it}|s, h]$  and  $\text{Var}[\bar{H}_{it}|s, h]$ : (i) the number of months a keyword existed before the period of our sample, and, (ii) the monthly variability of  $s_{i\tau}$  (for  $\text{Var}[S_{it}|s, h]$ ) and  $\bar{h}_{i\tau}$  (for  $\text{Var}[\bar{H}_{it}|s, h]$ ). The conditions formulated in **DGP 1** and **DGP 2** eliminate the second source of heterogeneity.

By **DGP 1**, all keywords for which we observe a given number of searches during the sample period had the same monthly popularity in all previous periods. By **DGP 2**, all keyword for which we observe a given average user history during our sample also had the same average user history in all previous periods. Therefore, if we can determine a population of keywords for which **DGP 1** and **DGP 2** is a plausible approximation, we can approximate  $S_{it}$  and  $\overline{H_{it}}$  more accurately through  $s_i$  and  $h_i$ .<sup>23</sup>

**DGP 2** can be understood in the following way: keywords differ in the type of user who search the keyword. One dimension of the user type is the usage intensity of the search engine, which determines  $\overline{H_{it}}$ . **DGP 2** states that if two keywords have the same average user type in one month, they will also have the same average user type in the next month. Since neither experience a change in popularity (note that we assume that **DGP 1** holds), this is a sensible assumption to make since no change in popularity indicates that the user type did not change. By contrast if **DGP 1** does not hold, it is also likely that the user type changed and therefore that **DGP 2** does not hold. For the sake of this analysis, we assume that if **DGP 1** is violated than **DGP 2** cannot hold.

Note that **DGP 2** states a weaker assumption than **DGP 1** because it does not require that the average user history is constant. While a constant monthly popularity seems a realistic approximation for the evolution of the total search quantity of a keyword that reached its steady state, a similar assumption for the average user history appears unrealistic. This is because the search engine continuously collects data on users, which suggest that the average user history should have tendency to increase.

Given the above reasoning, it is sufficient to develop a method to detect keywords for which **DGP 1** is a poor approximation. By removing these keywords, we can focus on a population of keywords for which  $\text{Var}[S_{it}|s, h]$  and  $\text{Var}[\overline{H_{it}}|s, h]$  is minimal. By **DGP 1**, we need to drop keywords for which the monthly popularity that we observe is a unreliable measure for the monthly popularity in previous periods. Because we have no information on

---

<sup>23</sup>It should be noted that, to simplify analysis, we abstract from the fact that  $s_i$  and  $\overline{h_i}$  are subject to sampling error.

the previous periods, we rely on the simple heuristic that keywords that experience popularity changes during our sample period do likely not follow **DGP 1**.

The heuristic relies entirely on the intuition that popularity changes in the sample are indicative of long term patterns in popularity: A keyword that increases in popularity during each day in our sample is likely on a long run upward trend. Similarly, a keyword on a downward trend in our sample is likely also on a long run downward trend. Both patterns, upward or downward trends in our sample, could also be indicative of an oscillating long run popularity, with phases of high and low popularity. The key point is that all mentioned scenarios are inconsistent with **DGP 1**.

Next, we provide formal arguments why **DGP 1** and **DGP 2** reduce the variances. We discuss each assumption separately.

## B.1 DGP 1

Throughout, we assume that keywords originate at random time in the past. Denote by  $T_i$ , the number of months a keywords existed previous to the sample period. We denote the average monthly popularity of a keyword by  $s_i$ . Note that, by construction, we have  $S_{it} = T_i \times s_i$ . The total number of searches previous to the month our data were sampled is equal to the number of months a keywords existed previous to the sample period times the average monthly popularity.

Therefore, for the population of keywords that do not follow **DGP 1**, we have that  $\text{Var}[S_{it}|s, h] = \text{Var}[s_i \times T_i|s, h]$ . If we make the assumption of conditional independence between  $s_i$  and  $T_i$ , the conditional variance of the product  $s_i \times T_i$  is given by  $\sigma_s^2 \sigma_T^2 + \sigma_s^2 \mu_T^2 + \mu_s^2 \sigma_T^2$ . Where  $\sigma$  denote conditional variances and  $\mu$  conditional expectations. For a population of keywords that follows **DGP 1**, we have  $\sigma_s^2 = 0$  and, consequently, the variance reduces to  $\mu_s^2 \sigma_T^2$ , which is obviously smaller.

The assumption that we must make above, in addition to conditional independence, is that  $\mu_s$  and  $\sigma_T^2$  do not change when dropping keywords that do not follow **DGP 1**. This

amounts to assuming that keywords that do not follow **DGP 1** have  $s_i$  that randomly deviate from  $\mu_s$ . Furthermore, they must be drawn from the same distribution of  $T_i$ .

## B.2 DGP 2

For **DGP 2**, first note that we can write  $\overline{H_{it}} = \sum_1^{T_i} w_\tau \times h_\tau$ , where  $\tau$  denote months previous to the sample period and  $h_\tau$  is the average user history in month  $\tau$ .  $w_\tau$  is the monthly weight given by  $s_\tau/S_{it}$ , where  $s_\tau$  is the number of searches in month  $\tau$ . We have  $\sum_1^{T_i} w_\tau = 1$ . Thus, the average user history previous to the sample period is the weighted monthly average user history, where the weights are the ratio of the number of searches in a given month in the total number of searches previous to sample period.

If we would assume that  $h_\tau = h \quad \forall \tau$ , then  $\overline{H_{it}} = h$  and by **DGP 2**,  $\text{Var}[\overline{H_{it}}|s, h] = 0$ . While we find it realistic to assume that the monthly popularity of keywords  $s_\tau$  stays roughly constant, as stated in **DGP 1**, we do not believe it is a realistic assumption for the average user history. Instead, it is more realistic to assume that the average user history increases over time as users continuously interact with the search engine. This is why **DGP 2** is formulated more broadly.

**DGP 2** states that for a given  $h$  there is a unique sequence of monthly average user histories previous to the sample period:  $\{h_1, \dots, h_\tau, h_{\tau+1}, \dots, h_{T_i}\}$ . The only source of heterogeneity between keywords is the length of the sequence determined by  $T_i$ .

Thus, if **DGP 2** holds, for two keywords  $i$  and  $j$  with  $T_i \neq T_j$ , we have that  $h_{\tau i} = h_{\tau j}$  for all  $\tau \leq \min\{T_i, T_{ij}\}$ . If **DGP 2** would not hold, in addition to the heterogeneity in the length of the sequences, we would in general have that  $h_{\tau i} \neq h_{\tau j}$  for  $\tau \leq \min\{T_i, T_{ij}\}$ . It is logical that this additional heterogeneity can only increase  $\text{Var}[\overline{H_{it}}|s, h] = 0$ .

## C Details on Kernel Estimation

Throughout the paper, estimation is performed by means of local linear regression. The employed kernel for the weighting of the observations is a radial Gaussian kernel. The bandwidth of the kernel determines the standard deviation of the Gaussian distribution used for weighting, which is identical for each dimension. The off-diagonal elements of the covariance matrix are set to zero. For estimation, all explanatory variables are normalized to lie in the interval between zero and one. Because all explanatory variables only take positive values, this transformation amounts to dividing each explanatory variable by its maximum value.

For the analysis of subsection 4.1 and all the associated robustness checks in appendices A.2 and A.3, the bandwidth is set to 0.1, irrespective of the employed quality measure and the considered subsample resulting from dropping keywords that do not follow **DGP 1** and **DGP 2**. For the analysis of Section 4.2 and all associated robustness checks, in appendices A.2 and A.3, the bandwidth is set to 0.15.

Table 2 reports the optimal bandwidths determined by leave one out cross validation for each quality measure/subsample combination considered in the analysis of subsection 4.1 and the corresponding robustness checks in appendices A.2 and A.3. Table 3 reports the optimal bandwidths determined by leave one out cross validation for each quality measure for the analysis in subsection 4.2 and the corresponding robustness checks in appendices A.2 and A.3.

Throughout the analysis, we trade reduced variance for increased bias, which means that we selected a bandwidth larger than the optimal bandwidth determined by leave one out cross validation. As shown in Tables 2 and 3, the bandwidth chosen for the analysis in subsection 4.1 is further away from the optimal bandwidth determined by cross validation than the bandwidth we chose for the analysis in subsection 4.2.

Figures 24 and 25 show the results for the analysis performed in subsection 4.1 when employing a bandwidth of 0.05, which is close the optimal bandwidth determined by the

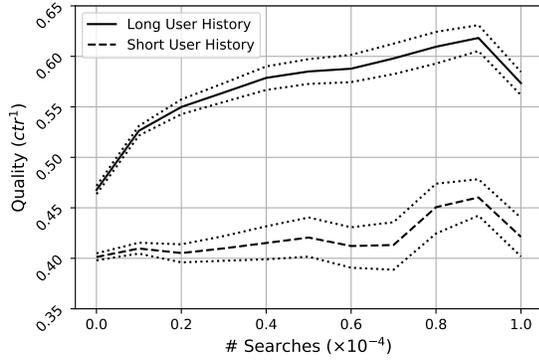
cross validation procedure. The irregular patterns observed in Figures 24a and 24b indicate that the selected optimal bandwidth tends to over-fit the data. Figures 24 and 25 also demonstrate that the results in subsection 4.1 are not driven by the selection of a wider bandwidth, which is reassuring.

**Table 2:** Optimal Bandwidths for Analysis of Section 4.1

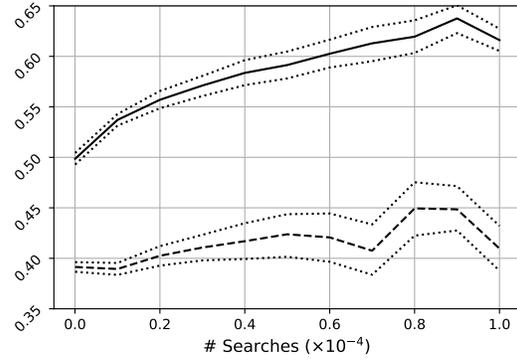
	all	dev:50pp	dev:20pp	dev:10pp
$ctr^1$	0.03	0.03	0.04	0.04
$ctr^3$	0.03	0.03	0.04	0.05
$ctr^{all}$	0.03	0.03	0.04	0.05
$rank^1$	0.04	0.04	0.06	0.1
$rank^2$	0.04	0.04	0.04	0.05
$rank^3$	0.04	0.06	0.1	0.08

**Table 3:** Optimal Bandwidths for Analysis of Section 4.2

	dev:50 pp
$ctr^1$	0.11
$ctr^3$	0.10
$ctr^{all}$	0.09
$rank^1$	0.06
$rank^2$	0.10
$rank^3$	0.14

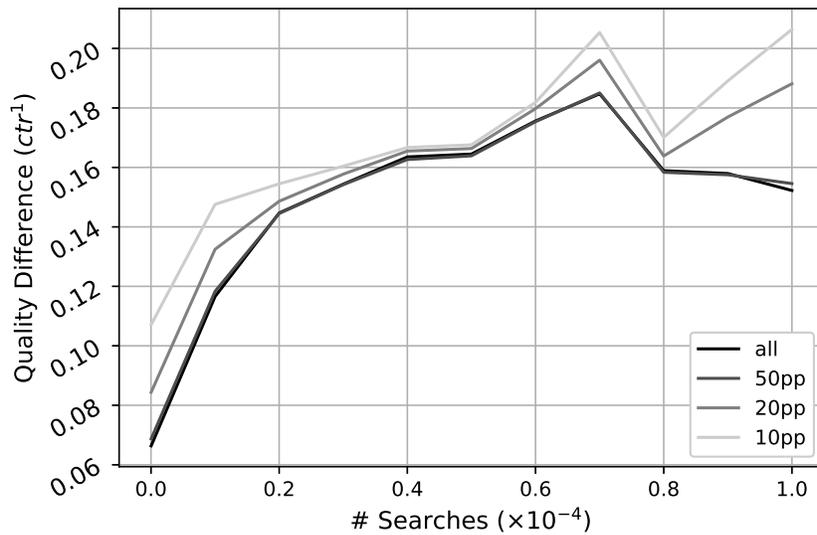


(a) Entire sample of keywords



(b) Keywords accumulating evenly

**Figure 24:** Average quality of keywords with 95% confidence intervals



**Figure 25:** Impact of longer average user history on quality level. Different tolerance levels for deviation from even accumulation criterion are considered.

## D Regression to the Mean (RTM)

The regression to the mean (RTM) problem arises in any analysis in which observations are classified based on an initial outcome (i.e based on the dependent variable). Intuitively, the problem arises because subjects are “erroneously” allocated to a category based on a single (or few) observation(s), which is not representative for the average of that subject. Because over time the outcome associated with a subject tend to revert to the average value commonly observed for that subject, studies that rely on a classification based on an initial outcome are prone to be affected by the RTM effect.

In the context of the analysis in subsection 4.2, there is the concern that the initial quality of keywords is a bad measure for the true initial quality of the keywords. If a keyword is assigned to the low initial quality group “by chance”, because it happened to experience a large idiosyncratic deviation from its true quality in the beginning of the sample period, it will naturally revert to its true average quality subsequently. This might lead to the erroneous conclusion that keywords with a low initial quality display a high quality increase, whereas in reality they simply revert to their natural average. Intuitively, the problem is more pronounced if we rely on a few observations to assess the quality; i.e. if we rely on an imprecise measurement of the true initial quality.

Figure 26 shows the result of the analysis of subsection 4.2 when we compute the initial quality of the keywords based on the searches a keyword experiences during the first day of the sample. In one day, a keyword with 200 searches experiences approximately six searches on average (200/32). Note that in subsection 4.2, we use the first 100 searches to determine the initial quality. As shown in Figure 26, the estimated quality increase is especially “irregular” for keywords with a low total search quantity: Keywords with a low initial quality experience a quality increase of roughly 10 percentage points.

To assess to what degree the results are driven by regression to the mean, we apply the correction formula discussed in [Barnett \*et al.\* \(2004\)](#). This formula is derived under the assumption of normally distributed stationary data and when observations are classified based

on thresholds. For example, the formula can be applied in treatment analysis when interested in the effect for individuals who are below a certain income threshold. The underlying assumption is that, absent treatment, there is no significant income trend and that income is normally distributed for the studied population. To the best of our knowledge, no formula exists for our specific application.

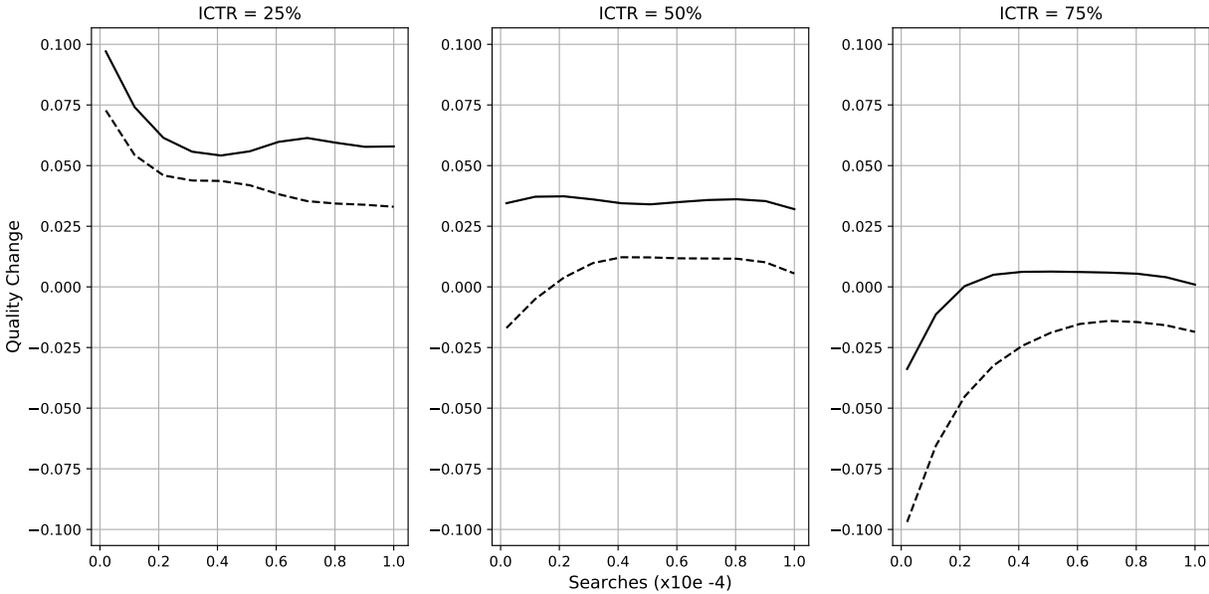
Our analysis does not rely on thresholding and our data are highly non-normally distributed and non-stationary. Our kernel estimation approach defines weights for observations close to the kernel centroid. Observations are assigned less weight the further away they are from the kernel centroid. To apply the formula, we ignore the weighting and estimate the average RTM effect as if all observations are equally weighted. However, we only include observations for which the average user history and the total number of searches are within one standard deviation of the kernel centroid.

More precisely, imagine we want to approximate the RTM effect for the estimate at the kernel centroid defined by an initial quality level of  $ctr^1 = 25\%$   $s = 2000$  and  $h = Q_1$ . Imagine that in the estimation we specified a standard deviation of 0.1 for the radial Gaussian kernel. To approximate the RTM, we apply the formula in [Barnett \*et al.\* \(2004\)](#) to all observations for which the Euclidean distance to the kernel centroid in the space spanned by  $s$  and  $h$  is less than or equal to 0.1. By doing so, the RTM is calculated for the observations that had most weight during the estimation.

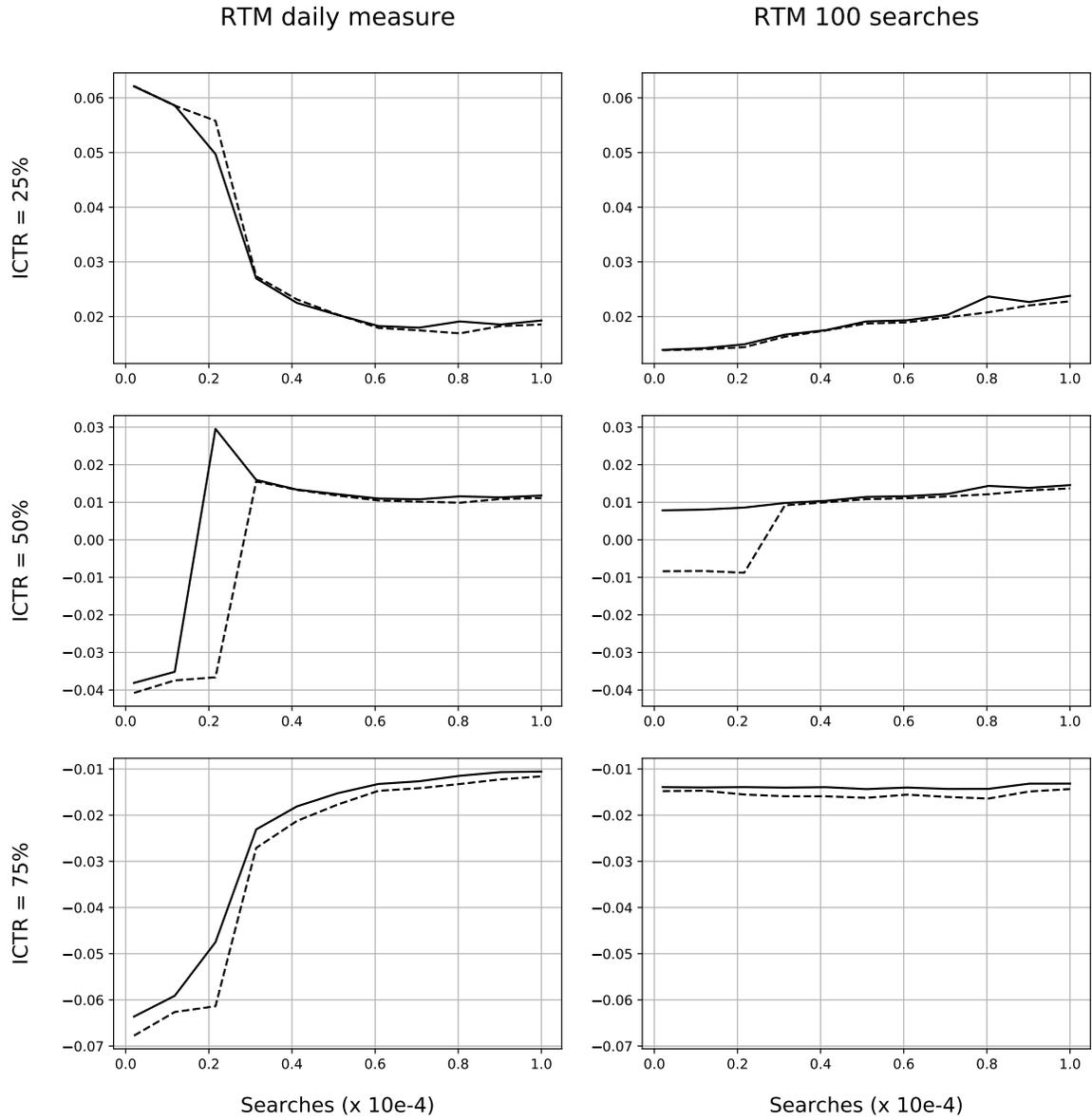
The left column of [Figure 27](#) displays the approximated RTM effect for the results presented in [Figure 26](#). As shown, the RTM effect seems to explain much of the observed quality changes observed for small search quantities in [Figure 26](#). The right column of [Figure 27](#) displays the approximated RTM for the analysis presented in [subsection 4.2](#). As can be seen, the regression to the mean effect appears to be much less pronounced when we use 100 observations to estimate the initial quality.

At around 4,000 searches the estimated RTM effect in the left column of [27](#) has roughly the same magnitude as the RTM effect in the right column of [Figure 27](#). This is the range in

which the number of searches in one day corresponds to roughly 100 searches. Our heuristic to assess the RTM effect seems to deliver sensible results. When both methods to assess the initial quality rely on the same number of searches, the estimated RTM effect is similar. It is also worthwhile to note that the estimated effect seems mostly only marginally different between keywords with a long and short average user history. This lessens the concern that the observed differences in the quality increase between keywords with a long and short average user history observed subsection 4.2 are driven by a differential impact of the RTM effect.



**Figure 26:** Analogue results to Figure 5 of subsection 4.2 if analysis is performed based on quality differences between first and last day and the initial quality is assessed based on the click through rate on the first day.



**Figure 27:** Left column: Estimated Regression to the mean effect when analysis is based on difference in the click through rate between first and last day. Right column: Estimated Regression to the mean effect for results in 5 of subsection 4.2

# E Model Discussion

## E.1 Proof of Proposition 1

### E.1.1 Confoundedness in types

We start by proving the first case considered in Proposition 1. The first case deals with confoundedness in types under simultaneous unconfoundedness in age, which is modelled by  $f(\mu_i, T_i|s, h) = f(\mu_i|s, h) \times f(T_i)$ . The strategy of the proof is to rule out a convergent pattern in the type heterogeneity as data generating process for our data. We remind the reader that changes in expected values are assumed to be induced by first order stochastic dominance.

**Proof.** Consider two distinct values for the average user history  $h$ ,  $h^l$  and  $h^s$ , with  $h^l > h^s$ . Absent network effects, the difference between conditional initial qualities given  $s_i = s$  and  $T_i = T$  is given by  $E[\mu|s, \Delta h]/(s \times T)^{\delta(h)}$ , because  $T_i \perp h, s$ . Assume that  $E[\mu|\bar{s}, \Delta h] < 0$  for some  $s = \bar{s}$ , i.e. queries with a longer average user history are on average of a lower type for  $s = \bar{s}$ . Since this implies that  $E[\mu|\bar{s}, \Delta h]/(s \times T)^{\delta(h)} < 0 \quad \forall T$ , the integral of the latter expression over  $T$ , which yields  $E[ctr_i^1|s, \Delta h]$  must be negative. Hence, absent network effects, it is not possible that the model generates a pattern consistent with our data if  $E[\mu|s, \Delta h] < 0$  for some  $s$ . Now assume that  $E[\mu|s, \Delta h] \geq 0 \quad \forall s$ , and consider a convergent pattern, i.e.  $\frac{\partial E}{\partial s}[\mu|s, \Delta h] \leq 0 \quad \forall s$ . Any  $\bar{s}$  and  $\underline{s}$ , such that  $\bar{s} > \underline{s}$ , implies that  $E[\mu|\underline{s}, \Delta h]/(\underline{s} \times T)^{\delta(h)} > E[\mu|\bar{s}, \Delta h]/(\bar{s} \times T)^{\delta(h)} \quad \forall T$ . Consequently, we must have that  $E[ctr_i^1|\Delta h, \underline{s}] > E[ctr_i^1|\Delta h, \bar{s}]$  for any  $\bar{s}$  and  $\underline{s}$  such that  $\bar{s} > \underline{s}$ . ■

The proof rules out that a convergent pattern in the types could generate a divergent pattern in the initial quality. To see that a divergent pattern in types can in principle generate the data, note that a divergent pattern in types allows the possibility that  $E[\mu|\underline{s}, \Delta h]/(\underline{s} \times T)^{\delta(h)} < E[\mu|\bar{s}, \Delta h]/(\bar{s} \times T)^{\delta(h)}$  for  $T < T^*$ . I.e. if keywords are not too old, for example if all  $T$  in the age distribution are below  $T^*$ , the divergent pattern in types can be conserved.

Note that for larger  $T$  the inequality reverses, which suggest that the divergent pattern is not stable. Furthermore, the expression  $E[\mu|s, \Delta h]/(s \times T)^{\delta(c)}$  illustrates that differences in average type vanish for larger  $T$ , hence, if the age distribution puts more weight on older queries,  $E[ctr_i^1|\Delta h, s]$  becomes smaller.

### E.1.2 Confoundedness in age

The proof for age confoundedness is slightly more complex. In the case of type confoundedness, the proof is facilitated by the property that differences in quality caused by differences in the type of keywords vanish at the same speed, independently of the type of keywords. The absolute quality difference between two keywords  $i$  and  $j$  with  $s_i = s_j = s$  and  $T_i = T_j = T$  caused by differences in types is given by  $|\mu_j - \mu_i|/(s \times T)^{\delta(h)}$ . By contrast, the absolute quality difference between two keywords  $i$  and  $j$  with  $s_i = s_j = s$  and  $\mu_i = \mu_j = \mu$  caused by differences in age is given by  $((1 - \mu) \times |T_j^{\delta(h)} - T_i^{\delta(h)}|)/(s \times T_i \times T_j)^{\delta(h)}$ . The last expression reveals that the same age difference leads to a smaller quality difference if the corresponding product of the ages is larger.

**Proof.** Given  $\bar{s}$  and  $\mu$ , the difference in the expected initial quality between  $h^l$  and  $h^s$  can be written as  $\int (1 - \frac{1-\mu}{\bar{s} \times T}) f(T|\bar{s}, h^l) dT - \int (1 - \frac{1-\mu}{\bar{s} \times T}) f(T|\bar{s}, h^s) dT$ . The latter expression is integrated over  $f(\mu)$  to obtain  $E[ctr_i^1|\Delta h, \bar{s}]$ . If  $F(T|\bar{s}, h^s)$  FODs  $F(T|\bar{s}, h^l)$ , which implies  $E[T|\bar{s}, \Delta h] \leq 0$ , because  $1 - \frac{1-\mu}{\bar{s} \times T}$  is a strictly increasing function in  $T$ , it follows that  $\int (1 - \frac{1-\mu}{\bar{s} \times T}) f(t|\bar{s}, h^l) dT - \int (1 - \frac{1-\mu}{\bar{s} \times T}) f(t|\bar{s}, h^s) dT \leq 0 \quad \forall \mu$  and therefore  $E[ctr_i^1|\Delta h, \bar{s}] \leq 0$ . Hence, absent network Effects, the model cannot generate a pattern consistent with our data if  $E[T|\bar{s}, \Delta h] \leq 0$  for some  $s = \bar{s}$ . Now assume that  $E[t_{i0}|s, \Delta h] \geq 0 \quad \forall s$ , and consider a convergent pattern, i.e.  $\frac{\partial E}{\partial s}[t_{i0}|s, \Delta h] \leq 0 \quad \forall s$ . From  $\int (1 - \frac{1-\mu}{s \times T}) f(T|s, h^l) dT - \int (1 - \frac{1-\mu}{s \times T}) f(T|s, h^s) dT$ , it is easy to see that for larger  $s$ , the difference between the integrals reduces if  $f(T|s, h^l)$  and  $f(T|s, h^s)$  do not depend on  $s$ . Furthermore, reduction in the FOD of  $F(T|s, h^l)$  over  $F(T|s, h^s)$  also reduces the difference between the integrals for each  $s$ . Finally, if we replace  $f(T|s, h^l)$  and  $f(T|s, h^s)$  by  $f'(T|s, h^l)$  and  $f'(T|s, h^s)$ , such that

each  $F'$  FOD  $F$  and  $F'(T|s, h^l) - F'(T|s, h^s) \leq F(T|s, h^l) - F(T|s, h^s) \quad \forall s$ , the difference between the integrals also reduces. Thus, if an increase in  $s$  is accompanied by a decrease in the FOD of  $F(T|s, h^l)$  over  $F(T|s, h^s)$ , which implies  $\frac{\partial E}{\partial s}[t_{i0}|s, \Delta h] \leq 0 \quad \forall s$ , a divergent pattern cannot be generated, unless  $F(T|s, h^l)$  and/or  $F(T|s, h^s)$  increase with  $s$ , which we rule out. ■

## E.2 Discussion of the Model under Network Effects and no Heterogeneity

It is easy to show that the difference in quality between any two keywords  $i$  and  $j$  with  $\overline{H}_i > \overline{H}_j$  and  $S_{it} = S_{ij} = S$  and  $\mu_i = \mu_j = \mu$  is positive  $\forall S, \mu$ . Furthermore it is a concave function in  $S$  with an unique maximum  $S^*$ , which is increasing for  $S < S^*$  and decreasing for  $S > S^*$ . Keywords with a larger monthly popularity  $s$  reach the region  $s \times T = S > S^*$  earlier as a function of  $T$ . Thus, when integrating over  $T$ , keywords with a larger monthly popularity,  $s$ , will be assigned more weight over the domain  $S > S^*$ . This is true  $\forall \mu$ . Therefore  $E[ctr_i^1|s, \Delta h]$  first decreases for larger  $s$ .