Data Network Effects: The Example of Internet Search

Maximilian Schaefer*, Geza Sapi†

December 4, 2019

[Job Market Paper. Do not cite without authors' consent. Updated version: **here**]

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 lacking. In this paper, we show that data as an input into machine learning tasks displays features that favor the hypothesis that data is a source of market power. We study the search result quality for search keywords on Yahoo!. Search result quality improves when more users search a keyword. In addition to this direct network effect caused by more users, we observe an additional externality that is caused by the amount of data that the search engine collects on the users. More data on the users reinforces the direct network effect. We propose to view this reinforcement effect due to additional user-specific data as a data network effect. Our findings are consistent with the consensus that data display diminishing returns to scale for a given prediction task. This feature of data is often regarded as incompatible with the hypothesis that data is a source of market power. Our results rationalize the market power hypothesis through a different mechanism by suggesting that data, in addition to being an input, is also a technology shifter.

*Berlin Scool of Economics, DIW Berlin and Technische Universität Berlin, 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

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 Zürich (2017), the 4th edition of the Industrial Organization in the Digital Economy Workshop (Liège, 2018), the 16th International Industrial Organization Conference (Indianapolis, 2018), the Mannheim Centre for Competition and Innovation 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 Annual Conference of the European Association for Research in Industrial Economics (Barcelona, 2019). All errors are our own.

1 Introduction

The role of data in the success of firms is the subject of much debate. Data is 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, which rely on data as one of their main input for operations, raises concerns that data might constitute a source of market power.[1] The "data feedback loop" hypothesis (Newman, 2014), according to which there is a self-reinforcing cycle between the success and the data amount firms control, is put forward to explain the rise of dominant players in data driven industries.

While the rise of superstar firms in digital industries lends 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. Additionally, the few existing studies suggest that data is an ordinary input, as all find evidence for diminishing returns to scale from additional data (He *et al.*, 2017; Yoganarasimhan, 2019; Claussen *et al.*, 2019). This finding, which is also motivated theoretically (Bajari *et al.*, 2019), appears to be at odds with the narrative of data being a "special" input that conveys a competitive advantage to firms that control of a large amount of data. Rather, sufficient scale appears to be comparatively easy for potential entrants to achieve.

In this paper, we provide empirical evidence that additional data improve the efficiency of the employed technology. According to our results, data do not simply constitute an input into a static production technology because, as they accumulate, they simultaneously shift the efficiency boundary of the technology outwards. While 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 compelling rationale for an inherent competitive advantage from "big data".

[1] 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)

We analyze search traffic data from Yahoo!. We observe users entering keywords in the search bar of the search engine and their subsequent interaction with the search results. The search engine collects information on the interactions of users with the search results, which can be interpreted as feedback regarding the quality. This data is valuable because it allows to improve the quality of search results.

Over time, data on the feedback accumulates 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 dimension the keyword dimension. Second, the more often a user is observed, the more data the search engine collects which is specific to the user. We refer to this dimension as the user dimension of data.

Both dimensions are critical for the predictive performance of the search engine. More data in the keyword dimension allows 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. Search results not receiving any positive feedback from users who searched the keyword in the past will not be suggested to future users searching the same keyword.

The data collected on a user allows the search engine to derive preferences specific to the user. In conjunction with the information that the search engine obtains on other users, data in the user dimension allows 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 it obtained from similar users who previously searched the same keyword.

In this paper, we focus on the interaction between both dimensions of data. 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.

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. Thus, if the search engine has, on average, more data on the searchers providing feedback for a specific keyword, this will lead to a faster quality increase as a function of the data collected on that keyword.

Our result provide a strong rationale for why large amount of data are particularly valuable. An additional data point is not only an incremental input that raises quality along a given technology but simultaneously shifts the efficiency boundary of the technology outwards. Users derive a larger utility from a more efficient technology, because the quality they experience from using the service becomes better. According to our findings, the utility of a user is positively affected by the amount of data collected on the users using the service. This makes the phenomenon we document akin to a direct network effect. Because the novel positive externality that we document is primarily determined by the amount of data collected on users, we call it a 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. Furthermore, our results call for a consideration of database sizes in merger decisions between firms that rely heavily on data for their business models. Our

results suggests that the ability to combine data on users across different services might be extremely valuable.

Our paper is related to a nascent strand of empirical literature investigating the impact of "big data" on firm performance.[2] These studies mostly focus on the channel of economies of scope as a source of market power from data. 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.

The mechanism we propose and investigate is markedly different from the hypothesis of economies of scope according to which algorithms benefit from increased data-variety in addition to increased data-quantity. Instead, our contribution speaks to the idea that artificial intelligence continuously improves with data. This notion is also put forward in the book of 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

[2]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) and Prüfer and Schottmüller (2017) model competition in data-driven markets. Prüfer and Schottmüller (2017) show that a reduction in innovation costs through additional user data (which they call data-driven indirect network effects) leads to monopolized market structures and suggest data sharing as a remedy. Prüfer and Schottmüller (2017) view innovation as a strategic variable. In contrast, we view innovation as a direct consequence of data collection.

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. We benefit from more precise data that allow us to construct a more granular quality measure, which is also defined relative to a more granular unit of observation.

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 from additional data collected on users. Although Yoganarasimhan (2019) mentions potential interaction effects between data collected on the keyword dimension and data collected on the user dimension, she does not further explore them. We propose a method that allows for simultaneously considering both dimensions. In this sense, we also view our paper as a methodological contribution.

In the remainder of the paper, we proceed 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 data network effects. Section 4 presents the results of our empirical analysis, which consists of two parts. The first part presents evidence for data network effects by taking a long run perspective on the data. Using a proxy variable technique, we show that keywords with more data in the user dimension previous to the period our data was sampled achieve a higher observed quality level in our sample. The second part focuses on the quality evolution we observe for keywords during the period our data was sampled. We show that the quality of keywords improves faster with more data in the user dimension. 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 anonymized search logs spanning a period of 32 days from July 1, 2010, through August 1, 2010, inclusive. From hereon, we refer to this time period as the sample period. An observation in our database contains a keyword identifier, a cookie identifier, the time the keyword was entered in the search bar, the ordered list of the top ten organic result URLs and the clicks performed by the user. In total, we have $\sim$ 80 million observations, which comprise $\sim$ 29 million users (identified by the cookie) searching for $\sim$ 67 thousand different keywords.

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.

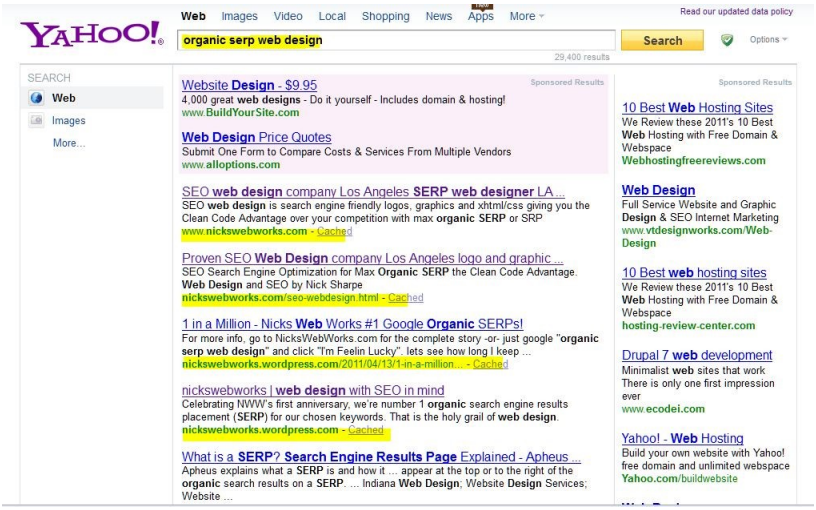


Figure 1: Search Result Layout at Yahoo!, 2011

2.1 Measuring Search Result Quality

For each search performed by a user, we observe a log that records the interaction of the user with the result page. This interaction usually consists of a series of clicks on the URLs proposed on the result page until the user chooses a proposed URL and does not return 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. Our quality measure builds on a dichotomous assessment of search result quality as either " good" (encoded as 1) or " bad" (encoded as 0).

The click-based quality measure employed in the main section encodes search result quality as "good" if the last recorded click in a log occurs on the top displayed organic URL. By considering the last recorded click, we account for 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, which defines the fraction of searches ending on the first URL. The click through rate on the first URL for keyword i is defined as:

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

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. S_i can be arbitrarily chosen. 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 quality measure.

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.[3]

3 Empirical Strategy

3.1 Measuring Data Network Effects

In this subsection, we discuss the variables that we use in order to study data network effects. As mentioned in the introduction, we focus on two different dimensions of data.

The first dimension captures the amount of data that accumulates on a keyword. In the terminology used in machine learning, it captures the amount of training data available to the search engine to solve a given prediction task. The prediction task is to optimize the quality of search results for a given search keyword. The amount of training data is determined by the number of users who searched a specific keyword.

The second dimension captures the amount of data that the search engine collected on the users searching a specific keyword. The hypothesis of data network effects states that the

[3]Our data set also contains 659,000 query-URL pairs with editorial relevance judgments collected from human experts. In Appendix A, we discuss the editorial quality measure in more detail and assess the robustness of our results based on it. In Appendix A.1, we argue that the editorial quality measure has several shortcomings for the sake of our analysis. Nevertheless, as we show in Appendix A.2, the results obtained with the editorial quality measure are in line with the results presented in the main section. The correlation coefficients between the ctr^1 measure and the two editorial quality measures that we consider is 0.45 and 0.55, respectively. We also perform robustness checks on a series of alternative click based quality measures. For instance, we encode the search result quality as "good" based on a broader range of URLs, not just the first. The results for the alternative click based quality measures is found in Appendix A.3. In A.4, we include ads in the analysis.

prediction task can be optimized more efficiently (i.e. requires less training data to achieve a given level of quality) as more data on the user becomes available (i.e. when the search engine collected more data on the users).

The mechanism we have in mind to rationalize network effects is as follows: 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 behavior.

With more data collected in the keyword dimension, the search engine learns which search results are relevant for similar user profiles by deducing which information in the overlapping browsing behavior 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 prediction task can be solved more efficiently 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.[4] 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 will generally be confronted with 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. Our hypothesis is that the amount of data in the user dimension increases efficiency of learning in the keyword dimension.

Whether this effect is present or not is ultimately an empirical question. While the

[4]See Adomavicius and Tuzhilin (2005) for an overview of the different architecture types of recommender systems.

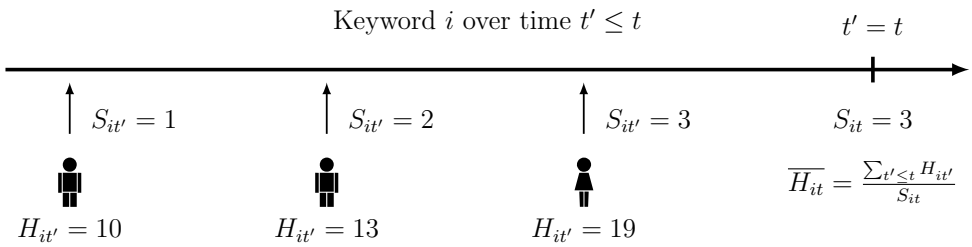


Figure 2: Variable Description

above discussion suggests that the overlap in data on the past behavior might be key for data network effect, we simplify the analysis. Our method focuses on the amount of data collected on users and relies on the intuition that increasing amounts of data on users are likely to reveal more overlap in the preferences of users. To the best of our knowledge, we are the first to systematically study the interaction effect between these two data dimensions.

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.

- $\overline{H_{it}}$: **The average user history** for keyword i. It describes 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. The data network effect describes how much faster quality improves as a function of S_{it} when the average amount of data collected on the users, captured by $\overline{H_{it}}$, increases.[5]

[5]Note that, technically, we only observe the cookie identifiers, which allows us to identify the device from which a search originates. Under the assumption that a device is used by a single user, the cookie identifies an individual. We have no means to determine if a computer is used by multiple individuals. For

3.2 Data Generating Process

The data that we have do not allow us to directly observe S_{it} and $\overline{H_{it}}$ because we only observe a point in time snapshot of the overall search traffic. The variables that we observe constitute the monthly counterpart of S_{it} and $\overline{H_{it}}$ (because we observe one month of data). Intuitively, this monthly counterparts might be informative about S_{it} and $\overline{H_{it}}$ in the sense that 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 subsection 4.1, we use the variables we observe as proxy measures for S_{it} and $\overline{H_{it}}$ and demonstrate that keywords for which the proxies indicate larger values of S_{it} and $\overline{H_{it}}$ are on a higher quality level. The goal of this subsection is to formalize on the relationship between the variables we observe and S_{it} and $\overline{H_{it}}$ to determine conditions under which we minimize the proxy-variable error when substituting for S_{it} and $\overline{H_{it}}$. We will use these conditions to systematically reduce the proxy-variable error in our analysis. The conditions will be stated as a data generating process because they describe how S_{it} and H_{it} evolve over time.

For the variables that we observe in our data, we use the following notation:

- s_i: **The total number of searches** for keyword i in the month of our sample.

- $\overline{h_i}$: **The average user history** for keyword i in the month of our sample.

Both quantities are defined over the entire sample period, i.e. based on all observations we have on a keyword. Because the sample period spans one month of search traffic, this gives both variables the interpretation of monthly quantities. s_i is the number of searches of a keyword during the month of our sample. $\overline{h_i}$ is the average user history during the month of our sample.

We are interested in the relationship between these variables and their unobserved counterparts at time $\underline{t}$, which denotes the beginning of our sample period. Reducing the ap-

the sake of the analysis, we will assume that a cookie represents a single user. The notation we choose for $\overline{H_{it}}$ highlights that the identity of the user is irrelevant to our analysis. All variables are defined with respect to the keyword, which is at the core of our analysis.

proximation error we make from substituting $S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$ by their proxy variables s_i and $\overline{h_i}$ amounts to reducing $\text{Var}[S_{i\underline{t}}|s,h]$ and $\text{Var}[\overline{H_{i\underline{t}}}|s,h]$ for each tuple $\{s_i = s, \overline{h_i} = h\}$ that we condition on. Conditioning on a tuple $\{s_i = s, \overline{h_i} = h\}$ induces a distribution over $S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$. The expectations of these distributions, $\text{E}[S_{i\underline{t}}|s,h]$ and $\text{E}[\overline{H_{i\underline{t}}}|s,h]$, describe the expected values of $S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$ for each tuple $\{s_i = s, \overline{h_i} = h\}$. The variances describe the heterogeneity over $S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$ for each tuple $\{s_i = s, \overline{h_i} = h\}$. The larger the variances, the larger the error we make from substituting $S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$ by s_i and $\overline{h_i}$.

We now introduce the data generating process under which $\text{Var}[S_{i\underline{t}}|s,h]$ and $\text{Var}[\overline{H_{i\underline{t}}}|s,h]$ are minimized.

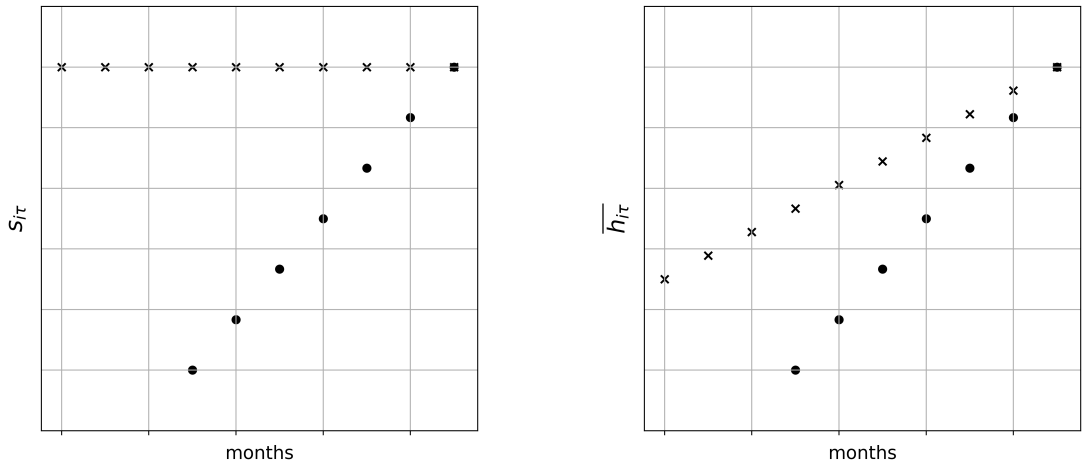
DGP 1: Keywords originate at random time and have constant monthly popularity. i.e. the number of searches for the keywords is constant in every month. I.e. $s_{i\tau} = s_i \quad \forall \tau$, where τ denotes a month.

DGP 2: If two keywords follow **DGP 1** and have the same average user history in month τ, 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 $\text{Var}[S_{i\underline{t}}|s,h]$ and $\text{Var}[\overline{H_{i\underline{t}}}|s,h]$: (i) the number of months a keyword existed before the period of our sample, and, (ii) the monthly variability of $s_{i\tau}$ (for $\text{Var}[S_{i\underline{t}}|s,h]$) and $\overline{h_{i\tau}}$ (for $\text{Var}[\overline{H_{i\underline{t}}}|s,h]$). The conditions formulated in **DGP 1** and **DGP 2** eliminate the second source of heterogeneity.[6]

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

[6]For the sake of brevity, we provide the formal discussion in Appendix B. It should be intuitive, however, that the elimination of one source of heterogeneity should reduce the overall heterogeneity.



(a) Heterogeneity ruled out by **DGP 1**. **(b)** Heterogeneity ruled out by **DGP 2**.

Figure 3: Heterogeneity ruled out by **DGP 1** (left panel) and **DGP 2** (right panel). The last observation in each panel represents the month of the sample. **DGP 1** and **DGP 2** impose that if the realizations of $s_{i\tau}$ and $\overline{h_{i\tau}}$ are the same during our sample period, they must also be the same in each month previous to our sample period. The keyword represented by the crosses follows **DGP 1** (and hence **DGP 2**). The keyword represented by the dots violates **DGP 1** (and hence **DGP 2**). The example illustrates that **DGP 1** imposes constant monthly realizations. **DGP 2** does not impose this stationarity.

$S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$ more accurately through s_i and h_i.[7]

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.

[7]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.

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. Figure 3 summarizes the assumptions imposed by the data generating process.

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 $\text{Var}[S_{i\underline{t}}|s,h]$ and $\text{Var}[\overline{H_{i\underline{t}}}|s,h]$ 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 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**.

To implemented the heuristic approach, we calculate for each keyword the number of searches we observe each day. A query that does not experience popularity changes during the sample period, should have exactly the same number of searches for each day of the sample. In other words, for each day, it should accumulate exactly 1/32% of the total searches we observe over the entire sample period. We call this accumulation rate the "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 $(1/32 + 10)\%$ of its total searches in order to be retained in the sample.

Intuitively, the narrower the tolerance interval, the better the quality of s_i and h_i as proxy variables. The largest tolerance level we consider is 50 percentage points, such that a keyword is allowed accumulate up to $1/32 + 50 = 53.125\%$ 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 13.125% of its total searches in one day without being dropped.[8]

4 Results

We now present the results of our empirical analysis, which consists of two parts. In the first part, we take a long run perspective of the data. The analysis heavily draws on the discussion in subsection 3.2. We build on the insight that the keywords we observe already existed before our 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. Since we cannot directly observe the number of searches and the average user history previous to our sample, we choose a proxy variable approach based on the variables we observe in the sample. We use the heuristic introduced in subsection 3.2 to increase the quality of the proxy variables. We show that when increasing the quality of the proxy variables, the magnitude of the measured network effect monotonically increases.

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

[8]Note that keywords with less than 32 searches can never perfectly fulfill the even accumulation criterion. In this sense, the criterion is "ill-defined" for those keywords. For the keywords for which it is properly defined, 89% of the keywords fulfill the narrowest tolerance level. We interpret this as indication that **DGP 1** seems to be a plausible approximation for the real data generating process of a large fraction of keywords.

of searches and average user history previous to the sample period. By directly analyzing quality differences of keywords, the hypothesis of data network effects can be tested in a more explicit way than in the first part of the analysis. The hypothesis of data network effects implies that keywords with a longer average user history should learn faster. We will document that this is the case. Furthermore, we will emphasize that the results of the second part are consistent with the long run dynamics documented in the first part of the analysis.

4.1 Long Run Analysis

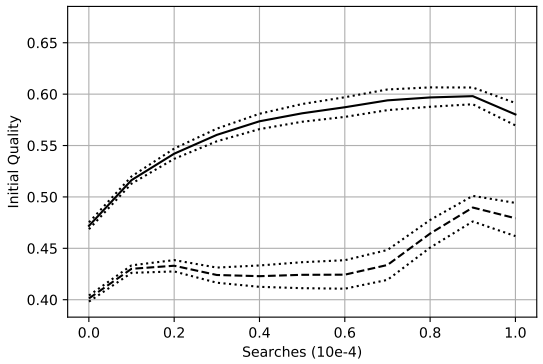
In this subsection, we study the quality level of queries in our sample as a function of the number of searches and the average user history previous to the sample period. With data network effects, we would expect that keywords with a longer average user history previous to the sample period have a higher quality level 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:

$$ctr_i^1 = f(S_{i\underline{t}}, \overline{H_{i\underline{t}}}) + \epsilon_i \tag{2}$$

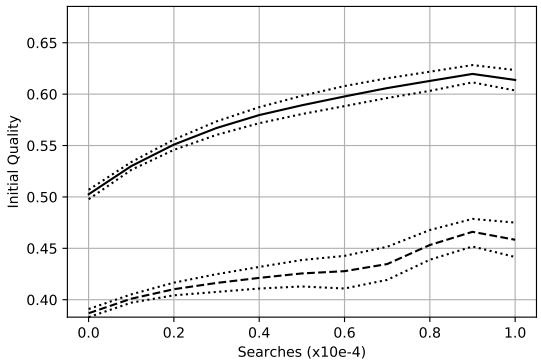
Where ctr_i^1 is the click through rate on the first URL for keyword i based on all the searches we observe. $S_{i\underline{t}}$ the number of searches previous to the sample period and $C_{i\underline{t}}$ the average user history previous to the sample. ϵ_i is the error term. Since we do not observe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

$$ctr_i^1 = f(s_i, \overline{h_i}) + \epsilon_i \tag{3}$$

Where s_i and $\overline{h_i}$ denote the number of searches and the average user history we observe for keyword i during our sample period. In subsection 3.2, we discussed how conditioning on a tuple $\{s_i = s, \overline{h_i} = h\}$ induces a distribution over $S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$. These distributions have conditional expectations $\mathrm{E}[S_{i\underline{t}}|s,h]$ and $\mathrm{E}[\overline{H_{i\underline{t}}}|s,h]$. Throughout this analysis, we assume



(a) Estimation results based on entire sample



(b) Estimation results based on keywords which deviate no more than ten percentage points from the even accumulation criterion

Figure 4: Results of local linear regression. Solid line: $h = Q_3$, Dashed line: $h = Q_1$. Dotted lines: respective 95% confidence intervals.

that larger values of h and s induce larger average values of $S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$. 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 $s = \{0, 0.1, 0.2, .., 1\}$ and $h = \{Q_1, Q_3\}$.[9][10] 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.

Figure 4 shows the results of our analysis. In Figure 4a, estimation was performed using all the keywords in the sample. In Figure 4b, only keywords that deviate no more than ten percentage points from the even accumulation criterion we considered. According to the discussion in subsection 3.2, the results in Figure 4b are less affected by the proxy variable error and should therefore be more informative about the true relationship between the quality of keywords and $S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$.

[9]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.

[10]In Appendix C, we present the details on the estimation method.

The solid lines in Figure 4 map out the quality level as a function of s for keywords with a long average user history. The dashed lines map out the quality level for keywords with a short average user history.[11] Note that once we remove keywords that do not follow **DGP 1** and **DGP 2** in Figure 4b, the observed pattern for keywords with a short average user history normalizes.

Figure 5 shows the difference between the quality level between keywords with a long and short average user history for a given s. The black line in Figure 5 shows the measured network effect for the results in Figure 4a. The line with the lightest grayscale shows the measured network effect for the results in Figure 4b, for which we applied the strictest tolerance level for the even accumulation criterion. The other lines show the measured network effect for intermediate tolerance levels with a lighter grayscale indicating a stricter tolerance level. Figure 5 reveals that the measured impact of network effects gradually increases as we drop an increasing number of keywords which do not follow **DGP 1** and **DGP 2**. Thus, gradually reducing the measurement error in the proxy variables gradually increases the measured impact of network effects. This is in line with what we would expect to see under the presence of a causal relationship when reducing the proxy measurement error.

From Figure 5, we also see that the difference in quality between keywords with a long and short average user history tends to increase with additional searches on the keywords. This divergence in quality levels is reassuring because it 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 discuss potential sources of confoundedness that might rationalize the documented divergence in the absence of network effects.

We conclude this subsection by giving a brief summary for our findings. First, our data are consistent with diminishing returns to scale in S. Second, our data are consistent with

[11] 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.

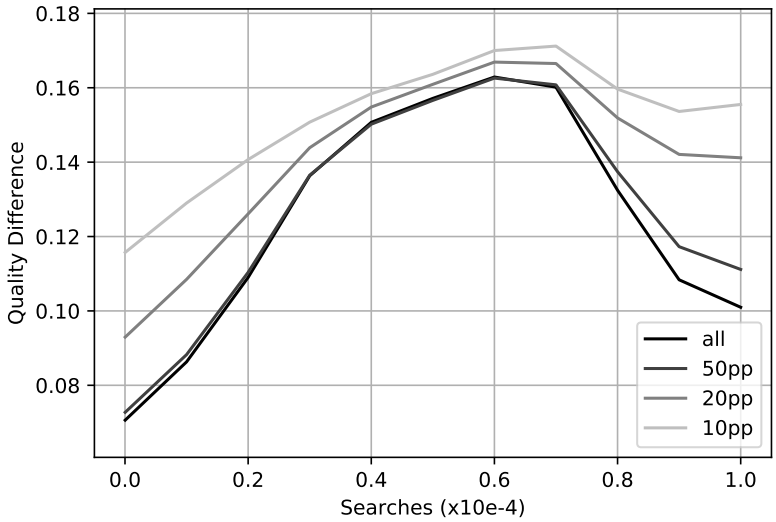


Figure 5: Measured network effect when reducing the tolerance levels for the even accumulation criterion. The legend denotes the tolerated percentage point (pp) deviation from the even accumulation criterion.

positive data network effects, according to which learning from S becomes more efficient through longer $\overline{H}$. Third, when we focus on keywords for which the explanatory variables of interest are measured more accurately, the measured impact of network effects becomes larger. Fourth, we observe a divergence in quality between keywords with a long and short average user history as S increases.

4.2 Variation within the Sample Period

In this subsection, we analyze the quality evolution of keywords as a function of s and h. Consistent findings with the results from subsection 4.1 would imply a positive concave relationship between the quality evolution of keywords and s. Furthermore, for the same value of s, we would expect the quality increase to be more pronounced for keywords with a larger average user history, $\overline{h}$.

Subsection 4.1 reveals a strong relationship between $S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$ and the quality level of a keyword during the sample period. It is also to be expected that $S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$ 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 $S_{i\underline{t}}$ 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 $S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$. We therefore estimate the following equation:

$$\Delta ctr_i^1 = f(s_i, \overline{h_i}, ictr_i^1) + \epsilon_i \tag{4}$$

Δctr_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.[12] $ictr^1$ is an intuitive measure for the quality a keyword reached at the beginning of our sample period, which helps to account for differences in $S_{i\underline{t}}$ and $\overline{H_{i\underline{t}}}$.

Estimation is performed by local linear regression on the grid $ictr \times s \times h$.[13] The values of $ictr$ are given by $\{0.25, 0.5, 0.75\}$ and the values of s by $\{0.02, 0.1, 0.2, ..1\}$. The values of h are identical to the ones chosen is subsection 4.1. Figure 6 shows the result of estimating Equation 4 on the grid $ictr \times s \times h$. Each column 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 6, we can read that keywords with $ictr^1 = 25\%$, $h = Q_3$ and $s = 10,000$ experience an average increase of ~ 8 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.

Consistent with the hypothesis of data network effects, keywords with longer average user history, *ceteris paribus*, display a larger quality increase than keywords with a shorter

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

[13] Details on the estimation method and bandwidth selection are given in Appendix C.

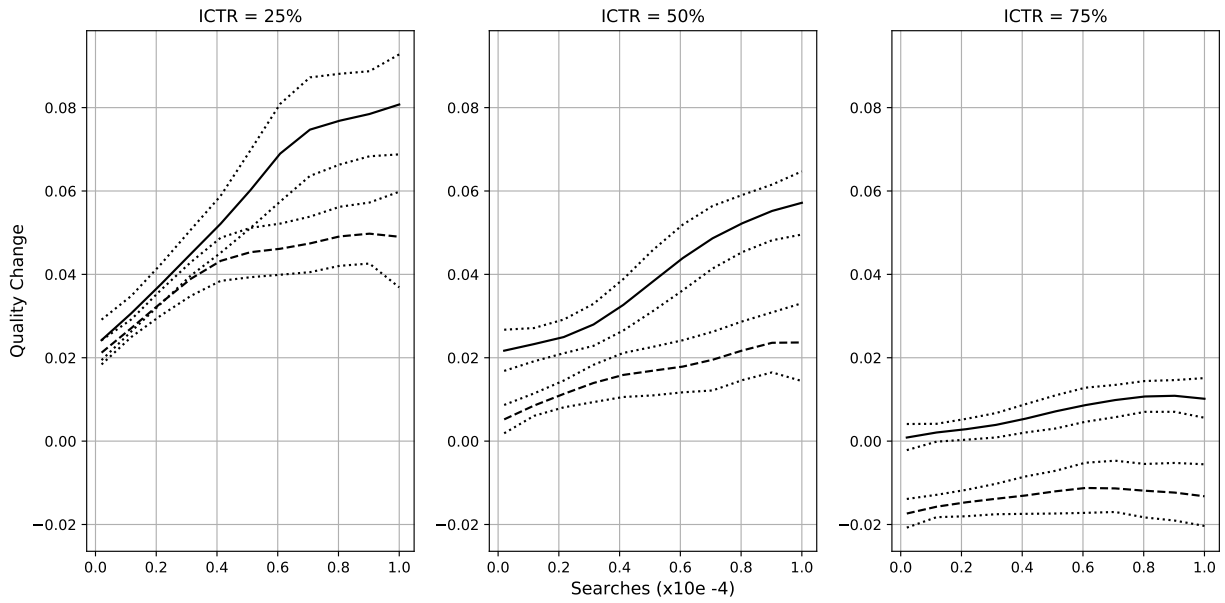


Figure 6: Results for estimation of $\hat{f}$ on the grid $ictr \times s \times h$. Each column stands for a different value of $ictr$. The solid lines map $\hat{f}$ as a function of s for $h = Q_3$. The dashed lines map $\hat{f}$ as a function of s for $h = Q_1$. Dotted lines: respective 95% confidence intervals.

average user history.[14] Consistent with diminishing returns to scale from additional searches on a keyword, we observe a concave pattern between the measured quality evolution and s. Furthermore, the average quality increase, ceteris paribus, diminishes with a larger initial quality level, which is also in line with diminishing returns in S[15]

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, the larger quality increase found for keywords with a longer average user

[14]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. We suspect that controlling for the initial quality level attenuates the impact of keywords that deviate form **DGP 1** and **DGP 2** in the estimation.

[15]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" on 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.

history is consistent with the hypothesis that data network effects explain the marked quality differences between keywords with a long and short average user history in subsection 4.1.

5 Identification

The results in subsections 4.1 and 4.2 are internally consistent and provide strong support for the hypothesis of data network effects. In this Section, we study to which extend unobserved heterogeneity can confound our results. To do so, we introduce a functional form assumption that will allow us to model network effects and the impact of two potential confounding factors: (i) unobserved type heterogeneity across keywords and (ii) unobserved heterogeneity with respect to the age of keywords. Both confounding factors have the potential to rationalize our data in the absence of network effects. Intuitively, 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. Throughout the discussion, we will assume that **DGP 1** and **DGP 2** hold.

We start by introducing and discussing our functional form assumption for the quality evolution process of keywords:

$$Q_{it}(S_{it}, \overline{H_i}, \mu_i) = 1 - \frac{(1-\mu_i)}{S_{it}^{\delta(\overline{H_i})}} = 1 - \frac{(1-\mu_i)}{(\mathrm{T}_i \times s_i)^{\delta(\overline{H_i})}} \tag{5}$$

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 δ imply a faster convergences to 1. Network effects are present if $\partial\delta(\overline{H_i})/\partial\overline{H_i} > 0$ and absent if $\partial\delta(\overline{H_i})/\partial\overline{H_i} = 0$.[16]

[16] In the functional form assumption of Equation 5, we neglect the possibility that $\overline{H_{i\underline{t}}}$ might be time varying. As discussed in subsection 3.2, we believe that $\overline{H_{i\underline{t}}}$ generally increases over time. Neglecting this dynamic in the functional form assumption amounts to assuming that the time varying nature of $\overline{H_{it}}$ is negligible for the quality of search results. We consider this restriction innocuous under the assumptions formulated in the data generating process. It implies that differences in the age of keywords only impact

μ_i is the type of a keyword and captures the intrinsic difficulty fo find relevant search results for the keyword. A larger value of μ_i models a keyword for which it is easier to find relevant search results. keywords with a larger value of μ_i start from a higher quality level and, *ceteris paribus*, always remain on a higher quality level. Heterogeneity in μ_i is one source of potential confoundedness that we will study in this Section. The second source of potential coundoudedness is introduced by T_i, which captures the age of the keyword. The decomposition $S_{it} = \mathrm{T}_i \times s_i$ emphasizes that the total number searches at time t can 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.

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. The interested reader is referred to Bajari *et al.* (2019) for a more formal treatment of the properties of statistical learning.

Based on the conditions formulated in **DGP 1** and **DGP 2** as well as the functional form of Equation 5, we can analytically derive the expected quality level of keywords conditional on observing s and h. This mimics the scenario we are confronted with in the analysis performed in subsection 4.1 and allows us to analyze under which assumptions on the distribution of μ

quality through the resulting difference in S_{it} when keywords have the same average user history in a given period.To understand what this implies under the condition formulated in **DGP 2**, consider two queries i and j for which we observe a given number of searches $s_i = s_j = s$ and a given average user history $\overline{h_i} = \overline{h_j} = h$ during the period of ou sample. Both keywords only differ in the number of months they existed previous to the period of our sample, which keyword i being the older keyword. By **DGP 2**, both keywords had exactly the same value h_τ in each period previous to the sample period in which they both existed. Hence, the difference between $\overline{H_{i\underline{t}}}$ and $\overline{H_{j\underline{t}}}$ is only due to the fact that both keywords are different in age. Under the assumption of an increasing trend in $\overline{H_t}$, it is easy to show that this implies that $\overline{H_{i\underline{t}}} < \overline{H_{j\underline{t}}}$. A time varying user length in Equation 5 would therefore imply that an older keyword benefits less from network effects than a younger keyword that is otherwise exactly comparable. Formulating the quality evolution as in Equation 5 can be understood as imposing the restriction that for the keywords in the above example, differences in age only impact quality through differences in S_{it}.

and T we can generate data consistent with ours in the absence of network effects.[17]

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

$$\mathrm{E}[ctr_i^1|s,h] = \int_0^{\bar{\mathrm{T}}} \int_0^1 \big(1 - \frac{1-\mu_i}{(s \times \mathrm{T}_i)^{\delta(h)}}\big) f(\mu_i, \mathrm{T}_i|s,h) d\mu_i d\mathrm{T}_i \tag{6}$$

Where $\bar{\mathrm{T}}$ describes the maximum number of months a keyword existed before the sample period. Through the conditional density function $f(\mu_i, \mathrm{T}_i|s,h)$, 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 $f(\mu_i, \mathrm{T}_i|s,h) = f(\mu_i|s,h) \times f(\mathrm{T}_i)$. Second, potential confoundedness related to the expected age only, which we model by $f(\mu_i, \mathrm{T}_i|s,h) = f(\mathrm{T}_i|s,h) \times f(\mu_i)$.

We want to understand under which assumptions on $f(\mu_i|s,h)$ and $f(\mathrm{T}_i|s,h)$, the conditional expectation in Equation 6 generates the divergence observed in Figure 5 in subsection 4.1 in the absence of network effects. 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 $\mathrm{E}[ictr_i^1|s,\Delta h] = \mathrm{E}[ictr_i^1|s,h^l] - \mathrm{E}[ictr_i^1|s,h^s]$. We observe a divergent pattern if $\mathrm{E}[ictr_i^1|s,\Delta h] > 0$ and $\partial \mathrm{E}[ictr_i^1|s,\Delta h]/\partial s > 0$, i.e. if the difference in the expected quality between h^s and h^s is positive and increases with s.

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

[17] 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.

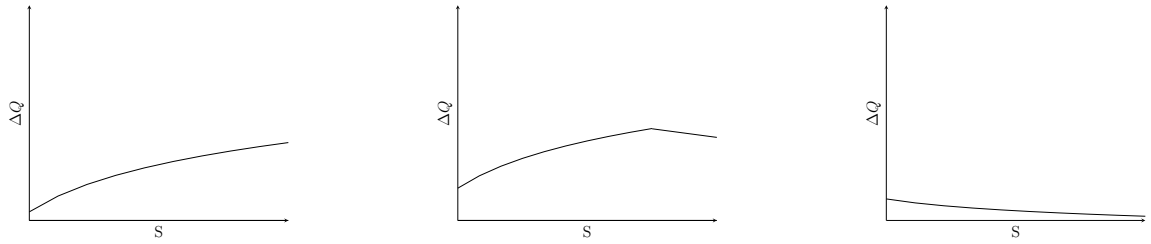
We now state our main result, which we prove in Appendix E:

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

Proposition 1 states that a divergent pattern in $\mathrm{E}[ictr_i^1|s,\Delta h]$ can only occur if the counfoundedness exactly replicates this pattern. To understand what this means, consider the case of the unobserved type of keywords. Divergence in $\mathrm{E}[ictr_i^1|s,\Delta h]$ 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.[18]

By contrast, if we allow for network effects, the pattern observed in our data arises naturally. Figure 7 schematically describes the three main patterns we would observe with

[18]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 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.



(a) Average difference in quality between h^l and h^s if average age is small.

(b) Average difference in quality between h^l and h^s if average age is intermediate.

(c) Average difference in quality between h^l and h^s if average age is large.

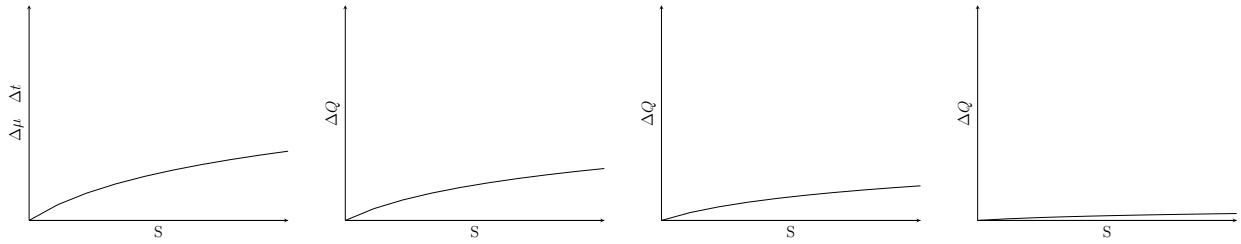
Figure 7: Evolution of quality differences induced by network effects as average age of keywords increases. We assume no simultaneous confoundedness.

network effects if the data would be generated by the model introduced in this Section and no confoundedness is assumed. 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 7a. As the average age increases, initially the pattern becomes similar to that shown in Figure 7b before reaching a state similar to that shown in Figure 7c.[19]

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 7a to the pattern shown in Figure 7b. After an initial phase of divergence in quality between h^l and h^s, quality starts to converge after a certain threshold, S^*_{it}, is reached. Because this threshold is first reached by keywords with a large number of searches per month, we observe the pattern in Figure 7b 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 to a pattern similar to 7c. 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, in the absence of network effects, initial quality differences continuously diminish. As a consequence, any divergent pattern in the data that is due to unobserved heterogeneity is at no point "self sustaining". Observed quality differences between keywords with a long and short average user history vanish as

[19] We provide a brief formal discussion of the pattern generated in the presence of data network effects 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 8: Evolution of quality differences induced by unobserved heterogeneity as average age of keywords increases. We assume absence of network effects.

the average age of keywords increases. Figure 8 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.[20]

The analysis of this Section highlights that strong assumptions on the unobserved heterogeneity are required to generate patterns consistent with our data in the absence of network effects. In contrast, in the presence of network effects, our model generates patterns consistent with our data if the average age of keywords is in a certain range. Although we do not explicitly study it, The results of this Section suggest that simultaneous confoundedness of age and type would require assumptions similar to the one formulated for each factor of confoundedness in isolation, i.e. confoudedness between the unobservables and the average user history needs to be reinforced with S in order to generate the patterns observed in the data.

[20]The pattern shown in Figure 8 is just one example of many potential patterns generated through heterogeneity. The main point of Figure 8 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.

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 improves the efficiency of the employed technology by reducing the number of user interactions required to achieve a given quality for a prediction task. We derive conditions under which our results might be confounded through correlation between unobserved factors and the variables that we use in the analysis. We show that restrictive assumptions about the correlation patterns are needed in order to rationalize our results in the absence of the proposed mechanism.

We view the mechanism that we propose as a network effect from additional data, because additional data increase the efficiency of the technology, thereby causing positive externalities similar to those created by more users in a network. Ultimately, our findings suggest that both classical network effects and data network effects are at work. The number of searches on a keyword captures the number of users interacting with the keyword. The positive impact of additional users for the quality of a keyword can be viewed as a classical direct network effect. The data network effect is the additional externality caused by the data collected on each user interacting with a keyword, which reduces the number of users necessary to reach a given quality.

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 an barrier to entry. Our results call for awareness from antitrust policy regarding any potentially anti-competitive behavior of firms seeking to deepening 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 a strong tension between privacy and competition policy. Our results suggest that the privacy interest of users should be weighed against the positive externalities potentially generated through personalized information. Our findings caution against overly strict privacy rules that might hinder the implementation of mechanisms that increase competition, like a data sharing mechanism. 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 suggest that in order to substantially benefit entrants, the right of data portability would have to be exerted by a majority of the 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.

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).

European Commission (2018) Commission Decision AT.40099 – Google Android, https://ec.europa.eu/competition/antitrust/cases/dec_docs/40099/40099_9993_3.pdf (July 18, 2018).

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).

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.

Lambrecht, A. and Tucker, C. E. (2015) Can Big Data protect a Firm from Competition?, *SSRN*, https://ssrn.com/abstract=2705530 (December 18, 2015).

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).

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.

From the information retrieval (IR) literature, it appears that editorial quality measures are most often used to assess the quality of different algorithms in an offline environment when the aim is to compare the quality of algorithms in an experimental setting (Chuklin *et al.*, 2013). It is also known that editorial quality measures often struggle to capture user preferences in an online setting, which led IR researchers to develop relevance metrics based on user click behavior (Chuklin *et al.*, 2013).

Fundamentally, it is questionable to what extent grades that are assigned through experts can capture user-specific preferences. The click behavior of users in the online context might lead the algorithm to treat webpages as relevant for particular user profiles even though these webpages might appear irrelevant to experts. This cautions against the interpretation of editorial quality measure as a "gold standard" to test the hypothesis that we are investigating.

Nevertheless, it seems useful to repeat the analysis of the main section based on editorial quality measures. To the extent that there is some overlap in the judgment of experts and the average user, we expect to draw similar conclusions from both analyses. It is beyond the scope of this Appendix to discuss all the subtleties and potential pitfalls of editorial quality measures. Our approach is to provide an as brief as possible description of the editorial quality measure and to repeat the analysis of sections 4.1 and 4.2 based upon it.

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:

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

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:

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

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

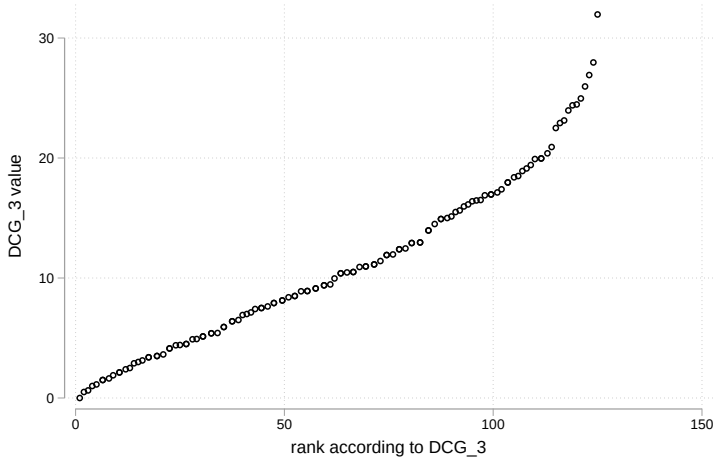


Figure 9: y-axis: DCG^3 value, x-axis: Ordinal ranking of relevance judgment-combinations according to their DCG^3 value

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-

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 9 depicts the DCG-values (y-axis) against their ordinal ranking (x-axis) for the DCG^3 measure.

Figure 9 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 10 shows the results of estimating Equation 3 on the grid $s \times h$ with $rank^1$ as the dependent variable. In Figure 10a, estimation is performed for all keywords in the sample. In Figure 10b estimation is performed for the subset of keywords that do not deviate by more than ten percentage point from the even accumulation criterion. Figure 11 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 do not follow the assumptions formulated in **DGP 1** and **DGP 2**. This is in contrast to the changes we observe when we perform the same exercise using the click-based quality measures, where the removal of keywords that deviate from **DGP 1** and **DGP 2** causes large changes in the observed pattern.

Nevertheless, Figures 12 and 13 reveal that the effect of removing keywords that do

not follow **DGP 1** and **DGP 2** 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.

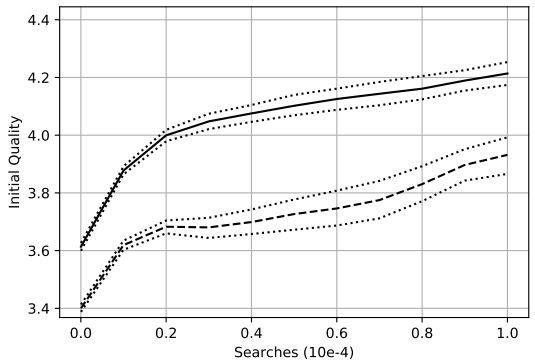
Figures 14 and 15 show the results of estimating Equation 4 on the grid $ictr \times h \times s$ for $rank^1$ and $rank^3$, respectively. For both editorial quality measures, we normalized the quality to lie in the interval $[0, 1]$.

In both Figure 14 and Figure 15, 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.

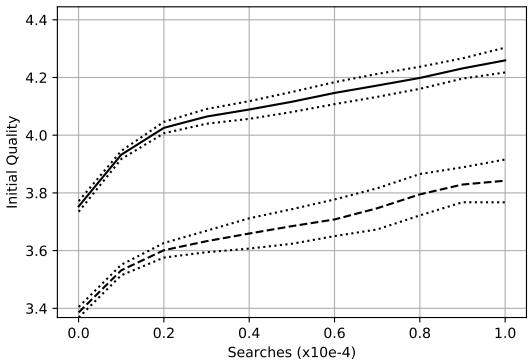
For the $rank^1$ measure, we observe no difference between long and short keywords for the lowest initial quality level. The large confidence intervals reflect the fact that only few keywords start from this quality level in our sample. For the second highest initial quality level, the observed pattern 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 dramatic 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 12 and 13, and the differential learning speed observed in Figures 14 and 15 for lower initial quality levels are in line with the formulated hypothesis of network effects.

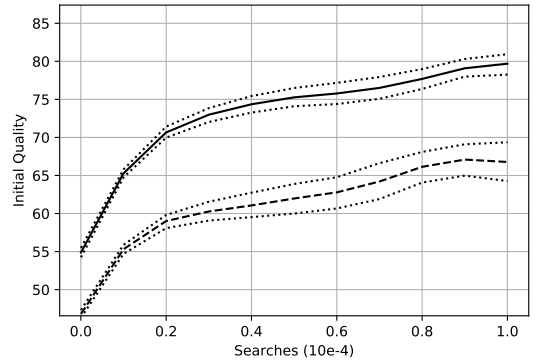


(a) Results of local linear regression for all keywords in the sample.

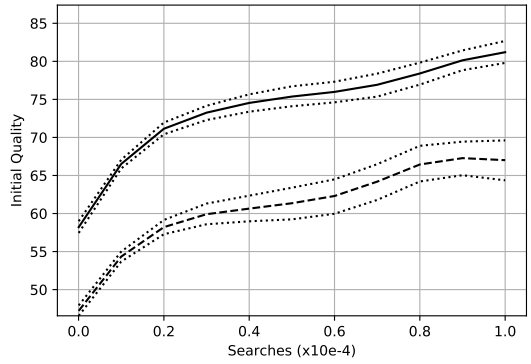


(b) Results of local linear regression for keywords that deviate no more than ten percentage points from the even accumulation criterion.

Figure 10: Estimated average $rank^1$ on grid $s \times h$. Solid line: $h = Q_3$, Dashed line: $h = Q_1$. Dotted lines: respective 95% confidence intervals.



(a) Results of local linear regression for all keywords in the sample



(b) Results of local linear regression for keywords that deviate no more than ten percentage points from the even accumulation criterion

Figure 11: Estimated average $rank^3$ on grid $s \times h$. Solid line: $h = Q_3$, Dashed line: $h = Q_1$. Dotted lines: respective 95% confidence intervals.

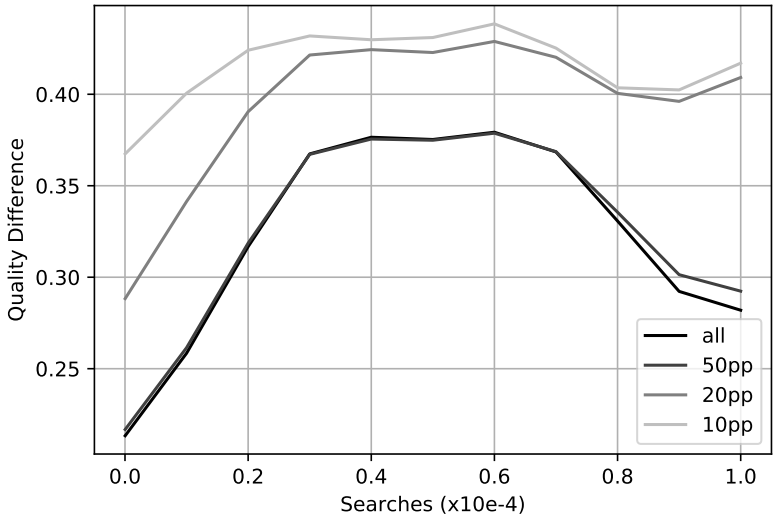


Figure 12: Measured network effect when reducing the tolerance levels for the even accumulation criterion. The legend denotes the tolerated percentage point (pp) deviation from the even accumulation criterion. Dependent variable: $rank^1$.

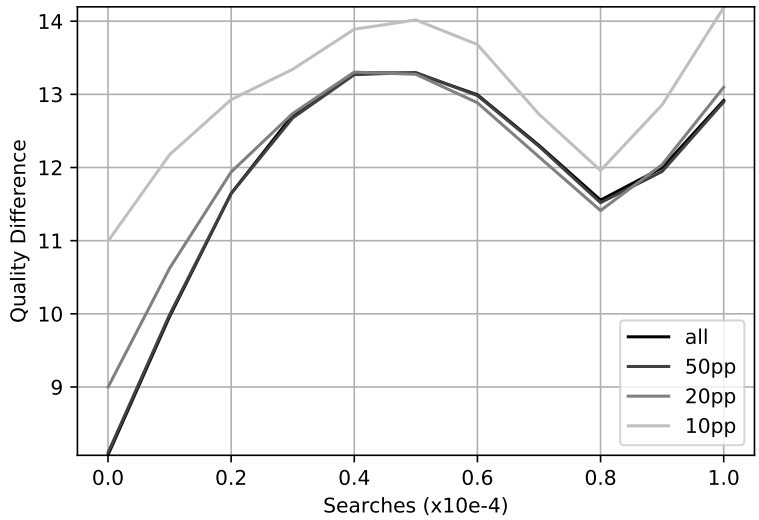


Figure 13: Measured network effect when reducing the tolerance levels for the even accumulation criterion. The legend denotes the tolerated percentage point (pp) deviation from the even accumulation criterion. Dependent variable: $rank^3$.

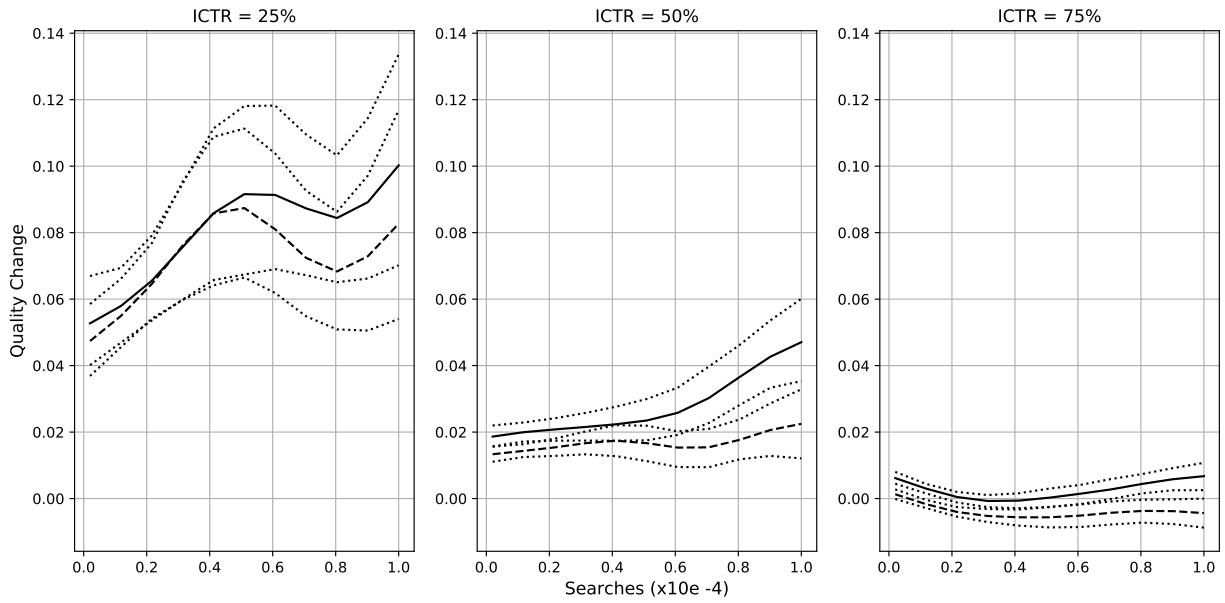


Figure 14: Results for estimation of $\hat{f}$ on the grid $ictr \times s \times h$. Each column stands for a different value of $ictr$. The solid lines map $\hat{f}$ as a function of s for $h = Q_3$. The dashed lines map $\hat{f}$ as a function of s for $h = Q_1$. Dotted lines: respective 95% confidence intervals. Dependent variable: $rank^1$.

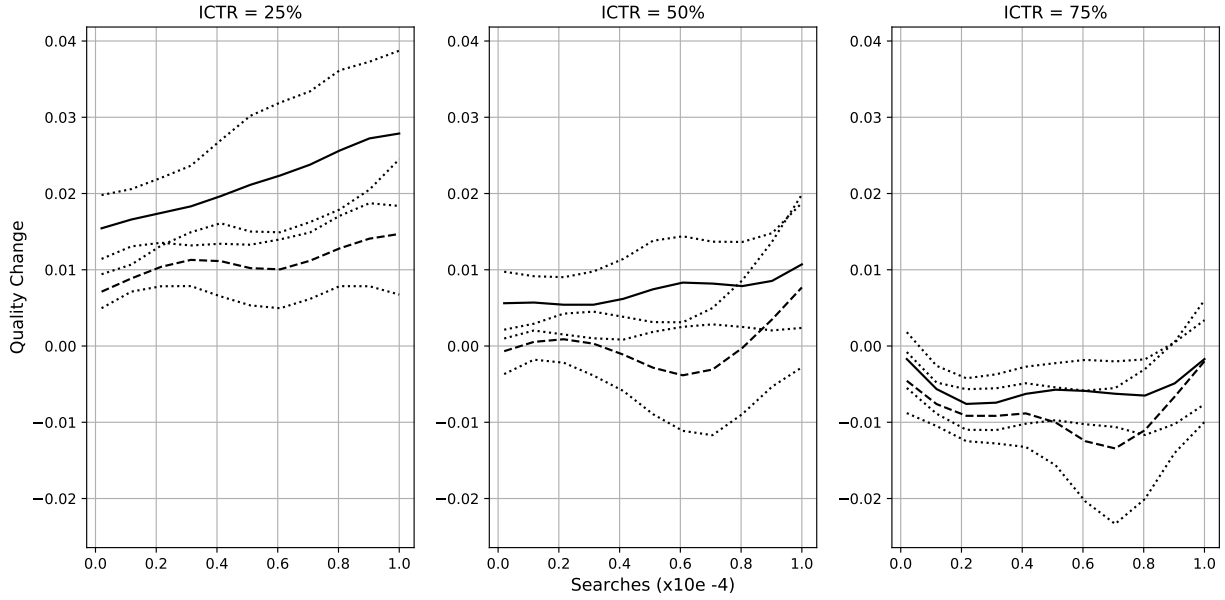


Figure 15: Results for estimation of $\hat{f}$ on the grid $ictr \times s \times h$. Each column stands for a different value of $ictr$. The solid lines map $\hat{f}$ as a function of s for $h = Q_3$. The dashed lines map $\hat{f}$ as a function of s for $h = Q_1$. Dotted lines: respective 95% confidence intervals. Dependent variable: $rank^3$.

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 16 shows the results of estimating Equation 3 on the grid $s \times h$ with ctr^3 as the dependent variable. The results in Figure 16a are based on all the keywords in the sample. Figure 16b shows the results for the population of keywords that deviate no more than ten percentage points from the even accumulation criterion. Figure 17 repeats the same exercise for the ctr^{all} quality measure. Figures 18 and 19 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.

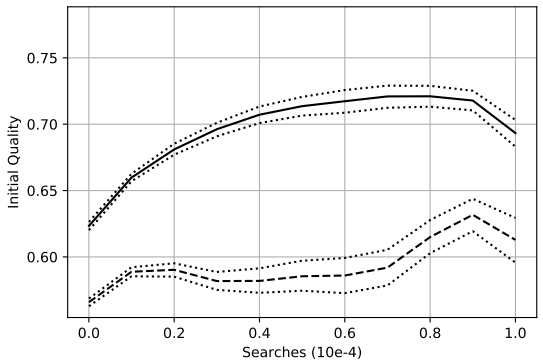
Figures 20 and 21 show the results of estimating Equation 4 on the grid $ictr \times h \times s$ for ctr^3 and ctr^{all}, respectively. The results are qualitatively identical to the ones obtained based on the ctr^1 quality measure.

Its 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 deemphasizes the value of personalized information.

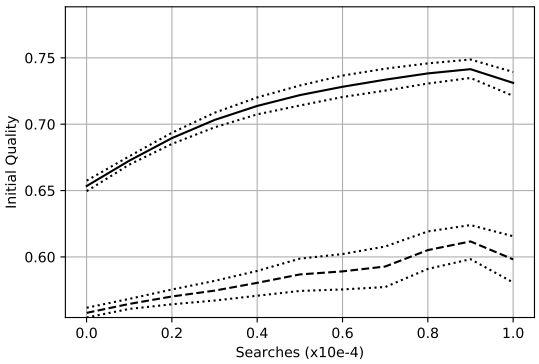
Another noteworthy difference to the results in the main section is that the quality

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.

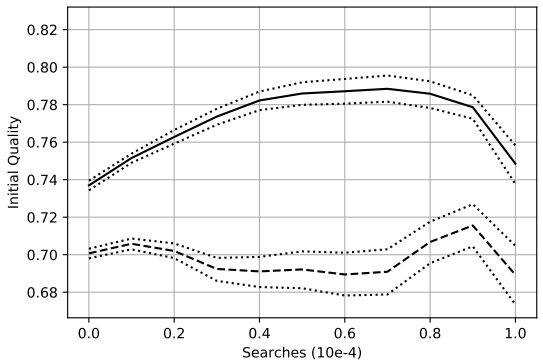


(a) Results of local linear regression for all keywords in the sample

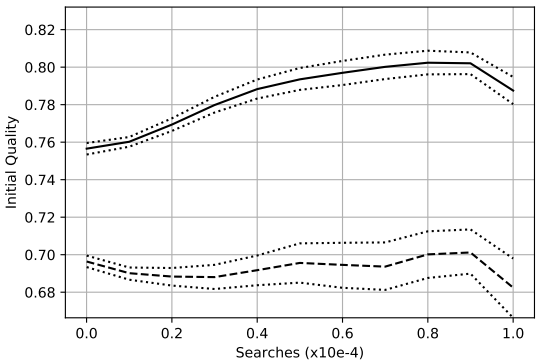


(b) Results of local linear regression for keywords that deviate no more than ten percentage points from the even accumulation criterion

Figure 16: Estimated average ctr^3 on grid $s \times h$. Solid line: $h = Q_3$, Dashed line: $h = Q_1$. Dotted lines: respective 95% confidence intervals.



(a) Results of local linear regression for all keywords in the sample



(b) Results of local linear regression for keywords that deviate no more than ten percentage points from the even accumulation criterion

Figure 17: Estimated average ctr^{all} on grid $s \times h$. Solid line: $h = Q_3$, Dashed line: $h = Q_1$. Dotted lines: respective 95% confidence intervals.

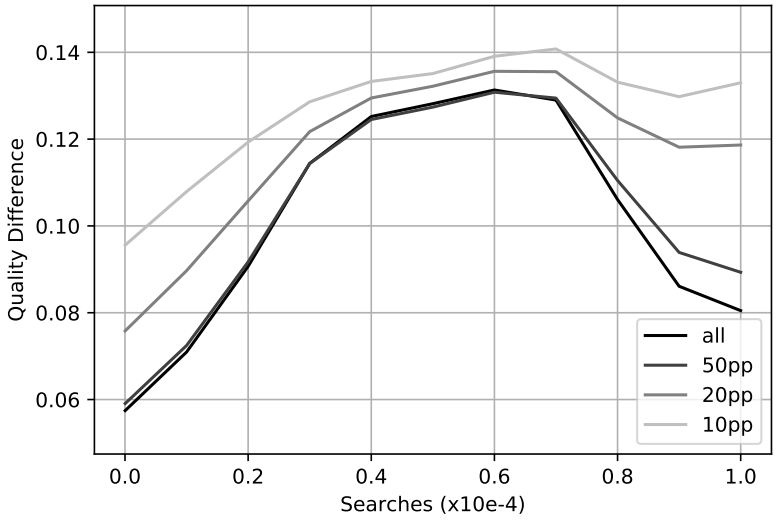


Figure 18: Measured network effect when reducing the tolerance levels for the even accumulation criterion. The legend denotes the tolerated percentage point (pp) deviation from the even accumulation criterion. Dependent variable: ctr^3.

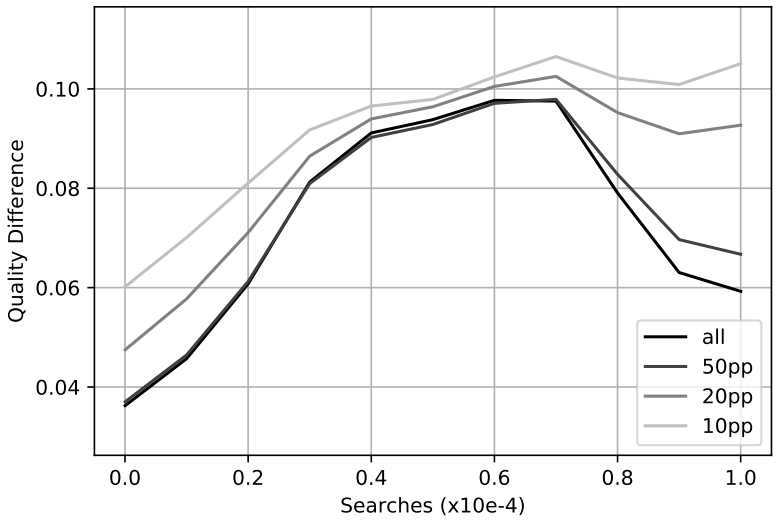


Figure 19: Measured network effect when reducing the tolerance levels for the even accumulation criterion. The legend denotes the tolerated percentage point (pp) deviation from the even accumulation criterion. Dependent variable: ctr^{all}.

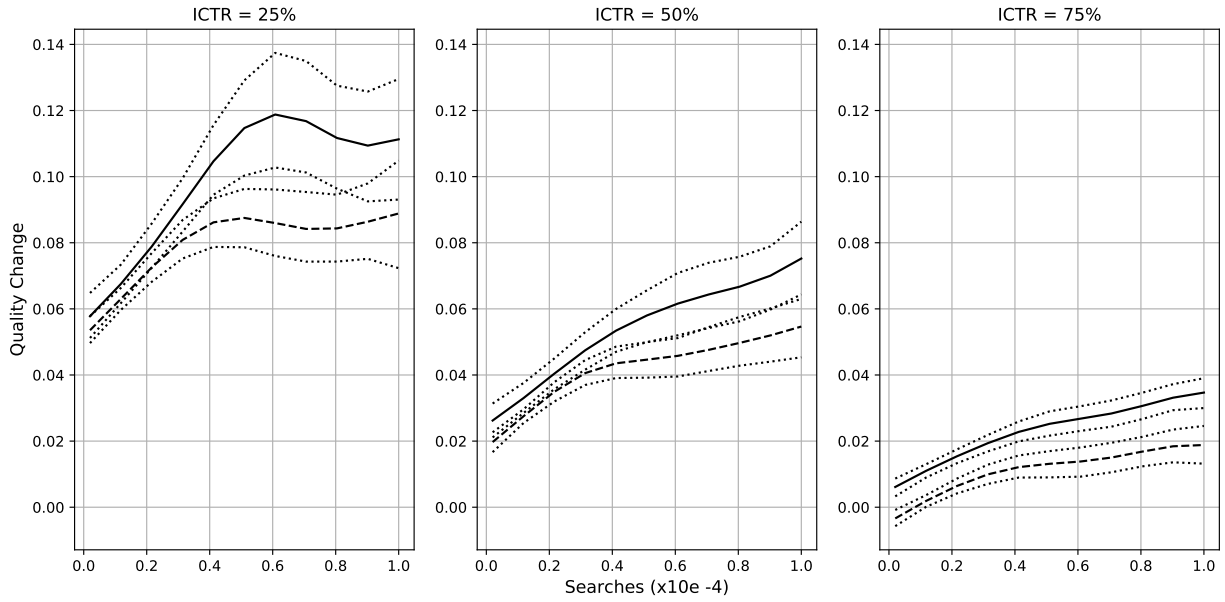


Figure 20: Results for estimation of $\hat{f}$ on the grid $ictr \times s \times h$. Each column stands for a different value of $ictr$. The solid lines map $\hat{f}$ as a function of s for $h = Q_3$. The dashed lines map $\hat{f}$ as a function of s for $h = Q_1$. Dotted lines: respective 95% confidence intervals. Dependent variable: ctr^3.

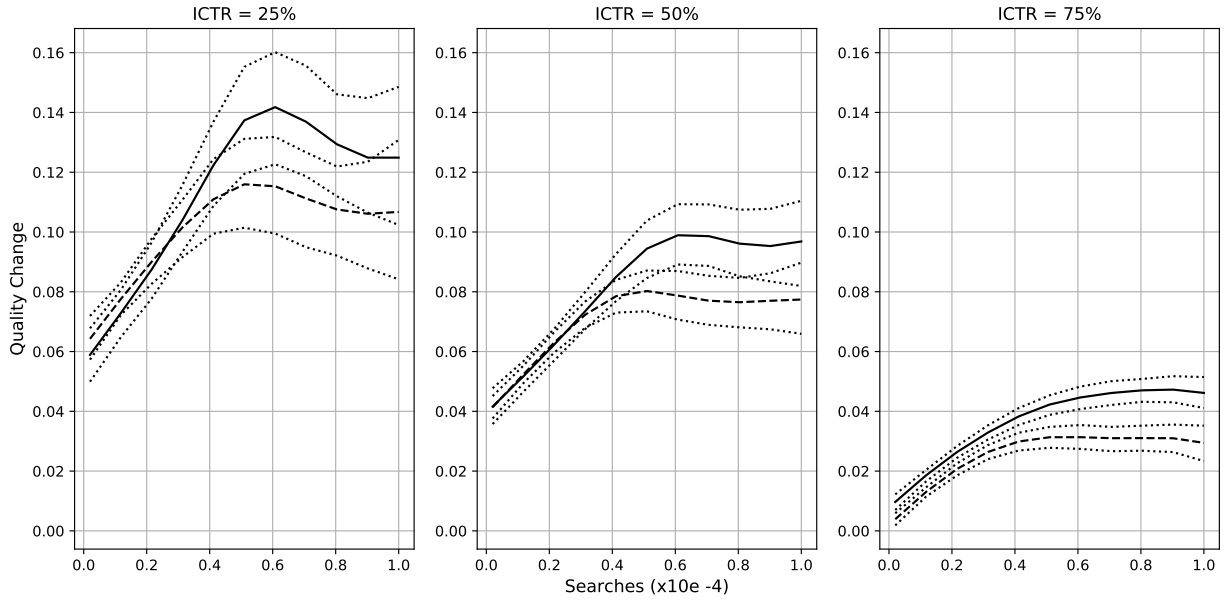


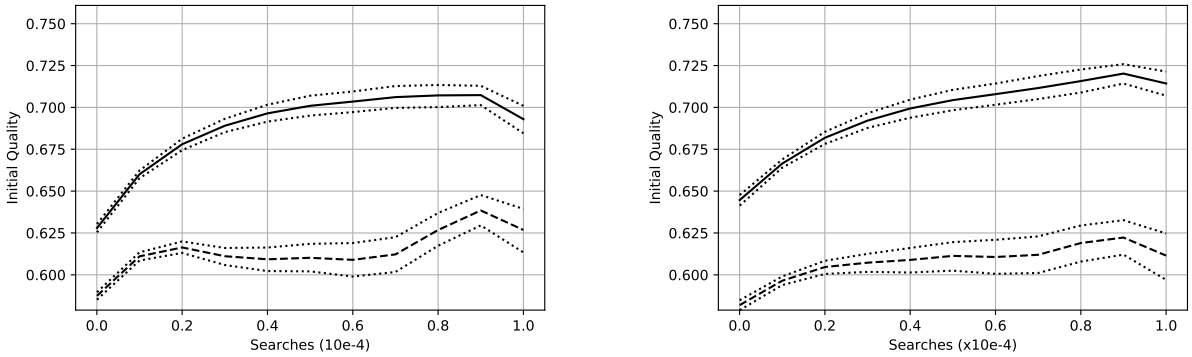
Figure 21: Results for estimation of $\hat{f}$ on the grid $ictr \times s \times h$. Each column stands for a different value of $ictr$. The solid lines map $\hat{f}$ as a function of s for $h = Q_3$. The dashed lines map $\hat{f}$ as a function of s for $h = Q_1$. Dotted lines: respective 95% confidence intervals. Dependent variable: ctr^{all}.

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:

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

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 22, 23 and 24 show that our conclusions remain unaffected by whether or not, we count these ignore these clicks or not.



(a) Results of local linear regression for all keywords in the sample

(b) Results of local linear regression for keywords that deviate no more than ten percentage points from the even accumulation criterion

Figure 22: Estimated average ctr^{1ad} on grid $s \times h$. Solid line: $h = Q_3$, Dashed line: $h = Q_1$. Dotted lines: respective 95% confidence intervals.

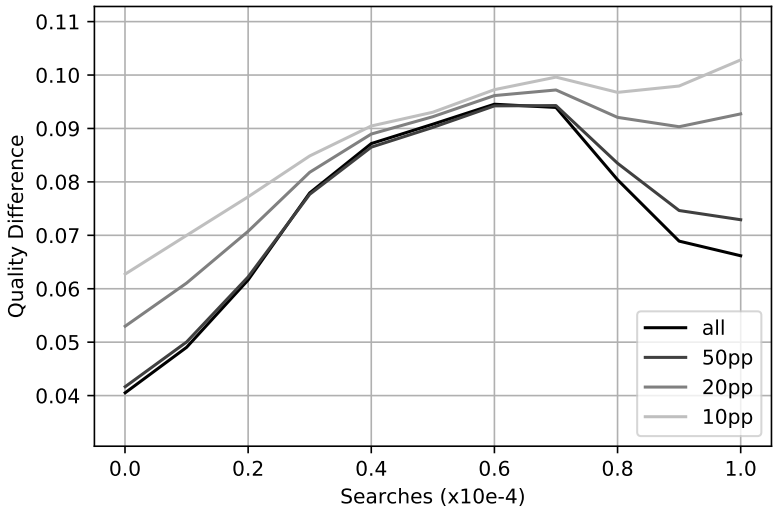


Figure 23: Measured network effect when reducing the tolerance levels for the even accumulation criterion. The legend denotes the tolerated percentage point (pp) deviation from the even accumulation criterion. Dependent variable: ctr^{1ad}.

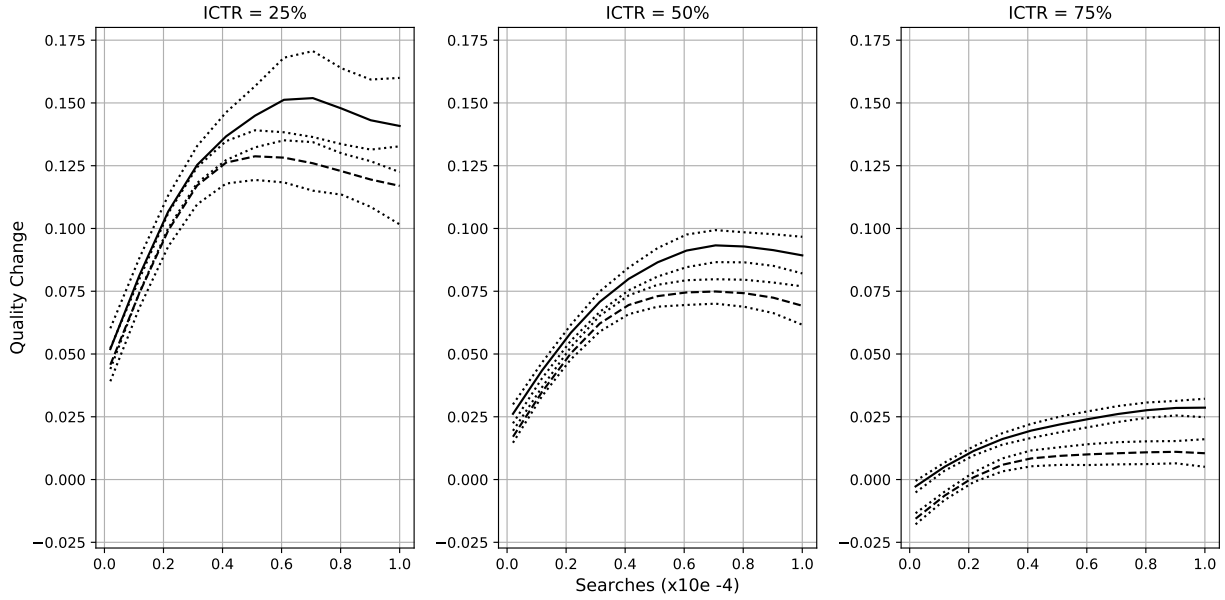


Figure 24: Results for estimation of $\hat{f}$ on the grid $ictr \times s \times h$. Each column stands for a different value of $ictr$. The solid lines map $\hat{f}$ as a function of s for $h = Q_3$. The dashed lines map $\hat{f}$ as a function of s for $h = Q_1$. Dotted lines: respective 95% confidence intervals. Dependent variable: ctr^{1ad}.

B Discussion Data Generating Process

In this Appendix, we provide a more formal discussion of **DGP 1** and **DGP 2**. We start by discussing **DGP 1**.

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 $S_{i\underline{t}} = \mathrm{T}_i \times s_i$. 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 $\mathrm{Var}[S_{i\underline{t}}|s,h] = \mathrm{Var}[s_i \times \mathrm{T}_i|s,h]$. If we make the assumption of conditional independence between s_i and T_i, the conditional variance of the product $s_i \times \mathrm{T}_i$ is given by $\sigma_s^2\sigma_{\mathrm{T}}^2 + \sigma_s^2\mu_{\mathrm{T}}^2 + \mu_s^2\sigma_{\mathrm{T}}^2$. Where σ denote conditional variances and μ conditional expectations. For a population of keywords that follows **DGP 1**, we have $\sigma_s^2 = 0$ and, consequently, the variance reduces to $\mu_s^2\sigma_{\mathrm{T}}^2$, which is obviously smaller.

The assumption that we must make above, in addition to conditional independence, is that μ_s and σ_{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 μ_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 $\overline{H_{i\underline{t}}} = \sum_1^{\mathrm{T}_i} w_\tau \times h_\tau$, where τ denote months previous to the sample period and h_τ is the average user history in month τ. w_τ is the monthly weight given by $s_\tau/S_{i\underline{t}}$, where s_τ is the number of searches in month τ. 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 $\overline{H_{i\underline{t}}} = h$ and by **DGP 2**, $\mathrm{Var}[\overline{H_{i\underline{t}}}|s,h] = 0$. While we find it realistic to assume that the monthly popularity of keywords s_τ 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: $\{h_1, ..., h_\tau, h_{\tau+1}, ..., h_{\mathrm{T}_i}\}$. 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 $\mathrm{T}_i \neq \mathrm{T}_j$, we have that $h_{\tau i} = h_{\tau j}$ for all $\tau \leq min\{\mathrm{T}_i, \mathrm{T}_{ij}\}$. 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 $\tau \leq min\{\mathrm{T}_i, \mathrm{T}_{ij}\}$. It is logical that this additional heterogeneity can only increase $\mathrm{Var}[\overline{H_{i\underline{t}}}|s, h] = 0$.

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.

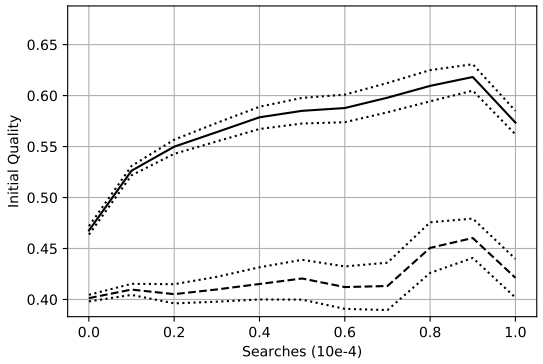
Figures 25 and 26 show the results for the analysis performed in subsection 4.1 when employing a bandwidth of 0.05, which is close the optimal bandwidth determined by the cross validation procedure. The irregular patterns observed in Figures 25a and 25b indicate that the selected optimal bandwidth tends to overfit the data. Figures 25 and 26 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

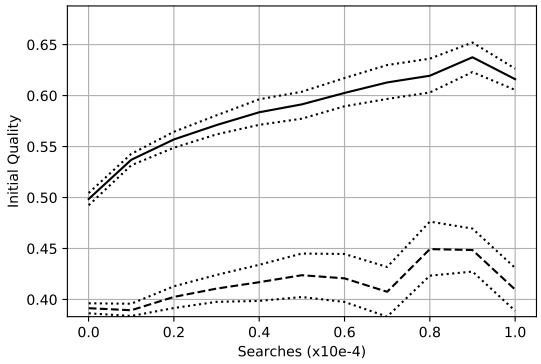
	all	dev:50pp	dev:20pp	dev:10pp
ctr^1	0.03	0.03	0.04	0.04
ctr^3	0.03	0.03	0.04	0.05
ctr^{all}	0.03	0.03	0.04	0.05
$rank^1$	0.04	0.04	0.06	0.1
$rank^2$	0.04	0.04	0.04	0.05
$rank^3$	0.04	0.06	0.1	0.08

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) Results of local linear regression for all keywords in the sample



(b) Results of local linear regression for keywords that deviate no more than ten percentage points from the even accumulation criterion

Figure 25: Estimated average ctr^1 on grid $s \times h$. Solid line: $h = Q_3$, Dashed line: $h = Q_1$. Dotted lines: respective 95% confidence intervals.

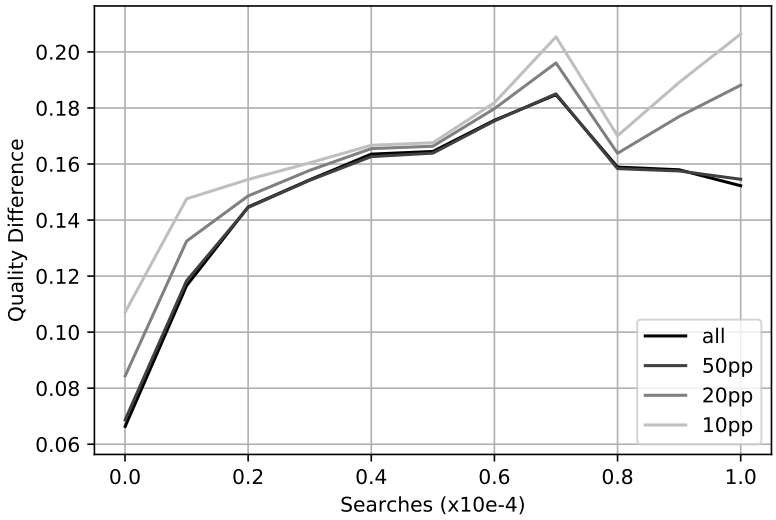


Figure 26: Measured network effect when reducing the tolerance levels for the even accumulation criterion. The legend denotes the tolerated percentage point (pp) deviation from the even accumulation criterion. Dependent variable: ctr^1.

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 27 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 27, 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 28 displays the approximated RTM effect for the results presented in Figure 27. As shown, the RTM effect seems to explain much of the observed quality changes observed for small search quantities in Figure 27. The right column of Figure 28 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 28 has roughly the same magnitude as the RTM effect in the right column of Figure 28. 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.

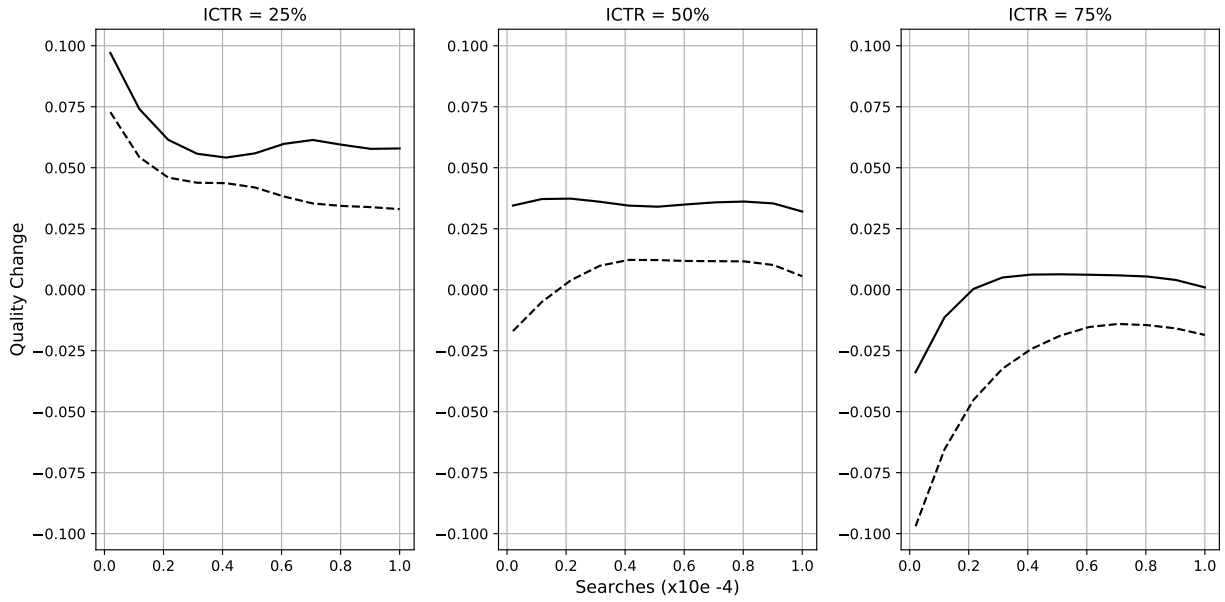


Figure 27: Analogue results to Figure 6 of subsection 4.2 if analysis is performed based on quality differences between first and last day and the initial quality is assessed based on the click through rate on the first day.

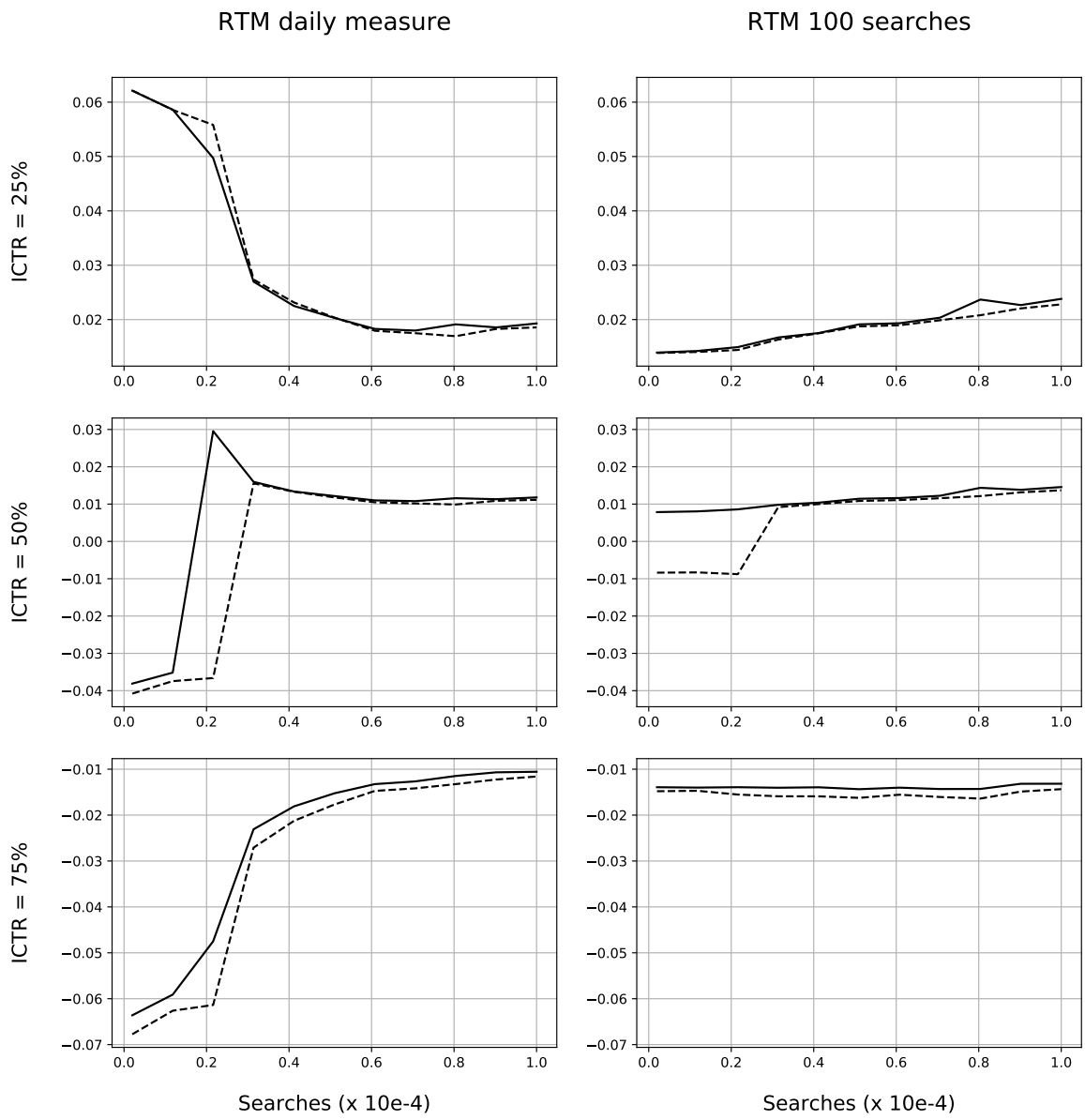


Figure 28: Left column: Estimated Regression to the mean effect when analysis is based on difference in the click through rate between first and last day. Right column: Estimated Regression to the mean effect for results in 6 of subsection4.2

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 modeled by $f(\mu_i, \mathrm{T}_i|s,h) = f(\mu_i|s,h) \times f(\mathrm{T}_i)$. 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 $\mathrm{T}_i = \mathrm{T}$ is given by $E[\mu|s,\Delta h]/(s \times \mathrm{T})^{\delta(h)}$, because $\mathrm{T}_i \perp h, s$. Assume that $\mathrm{E}[\mu|\tilde{s},\Delta h] < 0$ 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 $E[\mu|\bar{s},\Delta h]/(s \times \mathrm{T})^{\delta(h)} < 0 \quad \forall \mathrm{T}$, the integral of the latter expression over T, which yields $\mathrm{E}[ctr_i^1|s,\Delta h]$ must be negative. Hence, absent network effects, it is not possible that the model generates a pattern consistent with our data if $\mathrm{E}[\mu|s,\Delta h] < 0$ for some s. Now assume that $\mathrm{E}[\mu|s,\Delta h] >= 0 \quad \forall s$, and consider a convergent pattern, i.e. $\frac{\partial \mathrm{E}}{\partial s}[\mu|s,\Delta h] <= 0 \quad \forall s$. Any $\overline{s}$ and $\underline{s}$, such that $\overline{s} > \underline{s}$, implies that $\mathrm{E}[\mu|\underline{s},\Delta h]/(\underline{s} \times \mathrm{T})^{\delta(h)} > \mathrm{E}[\mu|\overline{s},\Delta h]/(\overline{s} \times \mathrm{T})^{\delta(h)} \quad \forall \mathrm{T}$. Consequently, we must have that $\mathrm{E}[ctr_i^1|\Delta h,\underline{s}] > \mathrm{E}[ctr_i^1|\Delta h,\overline{s}]$ for any $\overline{s}$ and $\underline{s}$ such that $\overline{s} > \underline{s}$. ■

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 $\mathrm{E}[\mu|\underline{s},\Delta h]/(\underline{s} \times \mathrm{T})^{\delta(h)} < \mathrm{E}[\mu|\overline{s},\Delta h]/(\overline{s} \times \mathrm{T})^{\delta(h)}$ for $\mathrm{T} < \mathrm{T}^\star$. I.e. if keywords are not too old, for example if all T in the age distribution are below $\mathrm{T}^\star$, 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 $\mathrm{E}[\mu|s,\Delta h]/(s \times \mathrm{T})^{\delta(c)}$ illustrates that differences in average type vanish for larger T, hence, if the age distribution puts more weight on older queries, $\mathrm{E}[ctr_i^1|\Delta h,s]$ 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 $\mathrm{T}_i = \mathrm{T}_j = \mathrm{T}$ caused by differences in types is given by $|\mu_j - \mu_i|/(s \times \mathrm{T})^{\delta(h)}$. 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 $\big((1-\mu) \times |\mathrm{T}_j^{\delta(h)} - \mathrm{T}_i^{\delta(h)}|\big)/\big(s \times \mathrm{T}_i \times \mathrm{T}_j\big)^{\delta(h)}$. 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 μ, the difference in the expected initial quality between h^l and h^s can be written as $\int \big(1 - \frac{1-\mu}{\bar{s}\times\mathrm{T}}\big) f(\mathrm{T}|\bar{s}, h^l) d\mathrm{T} - \int \big(1 - \frac{1-\mu}{\bar{s}\times\mathrm{T}}\big) f(\mathrm{T}|\bar{s}, h^s) d\mathrm{T}$. The latter expression is integrated over $f(\mu)$ to obtain $\mathrm{E}[ctr_i^1|\Delta h, \bar{s}]$. If $F(\mathrm{T}|\bar{s}, h^s)$ FODs $F(\mathrm{T}|\bar{s}, h^l)$, which implies $E[\mathrm{T}|\bar{s}, \Delta h] \leq 0$, because $1 - \frac{1-\mu}{\bar{s}\times\mathrm{T}}$ is a strictly increasing function in T, it follows that $\int \big(1 - \frac{1-\mu}{\bar{s}\times\mathrm{T}}\big) f(t|\bar{s}, h^l) d\mathrm{T} - \int \big(1 - \frac{1-\mu}{\bar{s}\times\mathrm{T}}\big) f(t|\bar{s}, h^s) d\mathrm{T} \leq 0 \quad \forall \mu$ and therefore $\mathrm{E}[ctr_i^1|\Delta h, \bar{s}] \leq 0$. Hence, absent network Effects, the model cannot generate a pattern consistent with our data if $E[\mathrm{T}|\bar{s}, \Delta h] \leq 0$ for some $s = \bar{s}$. Now assume that $\mathrm{E}[t_{i0}|s, \Delta h] >= 0 \quad \forall s$, and consider a convergent pattern, i.e. $\frac{\partial \mathrm{E}}{\partial s}[t_{i0}|s, \Delta h] \leq 0 \quad \forall s$. From $\int \big(1 - \frac{1-\mu}{s\times\mathrm{T}}\big) f(\mathrm{T}|s, h^l) d\mathrm{T} - \int \big(1 - \frac{1-\mu}{s\times\mathrm{T}}\big) f(\mathrm{T}|s, h^s) d\mathrm{T}$, it is easy to see that for larger s, the difference between the integrals reduces if $f(\mathrm{T}|s, h^l)$ and $f(\mathrm{T}|s, h^s)$ do not depend on s. Furthermore, reduction in the FOD of $F(\mathrm{T}|s, h^l)$ over $F(\mathrm{T}|s, h^s)$ also reduces the difference between the integrals for each s. Finally, if we replace $f(\mathrm{T}|s, h^l)$ and $f(\mathrm{T}|s, h^s)$ by $f'(\mathrm{T}|s, h^l)$ and $f'(\mathrm{T}|s, h^s)$, such that each F' FOD F and $F'(\mathrm{T}|s, h^l) - F'(\mathrm{T}|s, h^s) \leq F(\mathrm{T}|s, h^l) - F(\mathrm{T}|s, h^s) \quad \forall s$, the difference between the integrals also reduces. Thus, if an increase in s is accompanied by a decrease in the FOD of $F(\mathrm{T}|s, h^l)$ over $F(\mathrm{T}|s, h^s)$, which implies $\frac{\partial \mathrm{E}}{\partial s}[t_{i0}|s, \Delta h] \leq 0 \quad \forall s$, a divergent pattern cannot be generated, unless $F(\mathrm{T}|s, h^l)$ and/or $F(\mathrm{T}|s, h^s)$ 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 $S^\star$, which is increasing for $S < S^\star$ and decreasing for $S > S^\star$. Keywords with a larger monthly popularity s reach the region $s \times \mathrm{T} = S > S^\star$

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 $S > S^{\star}$. This is true $\forall \mu$. Therefore $\mathrm{E}[ctr_i^1 | s, \Delta h]$ first decreases for larger s.