Complementarities in Learning from Data: Insights from General Search

Maximilian Schaefer
Yale University*

Geza Sapi
European Commission†

September 16, 2022

Abstract

The ability to make accurate predictions relating to consumer preferences is a key factor of a digital firm's success. Examples include targeted advertisements and, more broadly, business models relying on capturing consumers' attention. The prediction technologies used to learn consumer preferences rely on consumer generated data. Despite the importance of data-driven technologies, there is a lack of knowledge about the precise role that data-scale plays for prediction accuracy. From a policy perspective, a better understanding about the role of data is needed to assess the risks that "big data" might pose for competition. This article highlights potential complementarities in algorithmic learning, which suggest data-scale advantages might be substantial. We analyze our hypothesis using search engine data from Yahoo! and provide evidence consistent with locally increasing returns to scale.

JEL Codes: L12, L41, L81, L86
Keywords: Antitrust, Competition, Learning from Data, Search Engines

*Corresponding author: Tobin Center for Economic Policy, Yale University, 37 Hillhouse Ave., New Haven, CT 06511, USA. The author is also a research associate at the Department of Economics, University of Bologna, Piazza Scaravilli 2, 40125 Bologna, Italy. Please direct inquiries to: maximilian.schaefer@yale.edu.

†Chief Economist Team of DG COMP, European Commission, Place Madou 1, 1210 Saint-Josse-ten-Noode, Belgium. The author is also a research fellow at the Duesseldorf Institute for Competition Economics, Universitaetsstr. 1, 40225 Duesseldorf, Germany. Please direct inquiries to: sapi@dice.uni-duesseldorf.de.

We thank Dirk Bergemann, Emilio Calvano, Giacomo Calzolari, Tomaso Duso and Fiona Scott-Morton for insightful comments and advice. We are particularly grateful to Hannes Ullrich for his extensive and continued support. All errors are our own. An earlier version of this paper circulated under the title "Big Data and Recommendation Quality: The Example of Internet Search" (Schaefer *et al.*, 2018).

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.

1 Introduction

The theoretical economic literature has studied the externalities inherent in the collection of user data and their potentially adverse effects on market outcomes (Acemoglu *et al.*, 2019; Bergemann *et al.*, 2021). These externalities rely on combining two data dimensions: Data collected "within users" and data collected "across users" (Hagiu and Wright, 2020; Lee and Wright, 2021). This article provides one of the first empirical analyses that studies the combined effect of both data dimensions for the competitiveness of firms using data from an industrial-scale application of algorithmic learning technology.

Our analysis provides new insights relevant for the question whether exclusive control of consumer data can grant significant competitive advantages (Sokol and Comerford, 2015; Tucker, 2019).[1] The answer depends upon the extent to which returns to data are diminishing. If returns diminish fast, then small competitors and new entrants can easily reach the efficient scale. If, on the other hand, returns diminish slowly, or do not diminish at all, then data control can represent a substantial barrier to entry, and regulators may have more grounds for concern. Studying the role of data complementarities helps contributing to understand how data unfold value at scale.

We analyze a unique one-month sample of 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 the logs of users' clicks on the search result page. The collected data allow the search engine to learn from experiments with different search results across searches *and* to build richer user profiles based on the observed interaction between the user and the results.

From our data, we can construct two variables that capture the across-user and within-user dimension of data accumulation: The first variable, which captures the across-user dimension, counts the number of times a keyword is searched (typically, the majority of

[1]Recent antitrust proceedings have focused on the potential anti-competitive effects of data, see, for example, European Commission (2018), recitals 111, 114, 458, 514, 739, 860(3), 1318, 1348.

searches stems from different users). The second variable counts the average number of times users have been observed prior to entering a keyword. We call this second variable the *average user history*. A longer user history indicates that the search engine observed a user more often and, hence, could collect more user-specific data. If a keyword is searched by users with longer histories, this leads to a longer *average user history* for the keyword, which indicates that the search engine could *on average* rely on more user-specific data.

Our empirical analysis documents a positive relationship between search quality (defined as the likelihood that a user selects the top displayed search result returned from entering a keyword) and the number of searches a keyword experiences. Additionally, we find that the increase in search quality from additional searches is more pronounced as the *average user history* increases. This suggests that more user-specific information makes learning in the across-user dimension more efficient, i.e. the search engine learns faster from additional searches on a keyword when more user-specific data is available about the users searching the keyword.

Our findings are consistent with the notion that adding additional observable characteristics to a prediction model should generally improve its prediction performance. In the well-known ordinary least squares (OLS) setting, our results could be thought of as capturing the property that using a larger set of explanatory variables will generally lead to a larger R-squared. Within this analogy, the search engine algorithm corresponds to the OLS model, the number of searches corresponds to the number of observations (the N-dimension), and a longer average user history corresponds to the number of explanatory variables (the K-dimension) used to estimate the OLS model. In this setting, our results translate to the statement that an OLS model with more variables (larger K-dimension) will have a better fit once the parameters have been consistently estimated (exploiting the N-dimension).

While the OLS-analogy might be helpful to illustrate our findings, it should be noted that search engine algorithms face more complex data environments. For instance, the number of observable characteristics is likely to vary across different users. Additionally, the amount

of data collected about each user is constantly increasing as users interact with the search engine. Both facts constitute peculiar challenges which render our findings non-trivial.

The outlined mechanism relies on the assumption that user-specific data is used by the search algorithm: If user-specific data does not constitute an input into the algorithm, technological complementarities between user-specific data and across-user learning are not possible. We rely on this insight to design a test to assess whether the observed pattern of faster across-user learning from more user-specific information is likely to capture a genuine complementarity effect (instead of spurious correlation).

To do so, we classify keywords in two groups: "Personalizable" keywords, i.e. keywords for which we find indication that the search engine relies on user-specific information, and "non-personalizable" keywords for which we find no such indication. Our classification method, which we describe in more detail later, relies on the insight that search results should change more frequently for personalizable keywords and that, additionally, these changes should depend on the information contained in the user-specific data.

We find no evidence for faster across-user learning for non-personalizable keywords. Instead, this pattern is only observed for the group of personalizable keywords. The fact that faster across-user learning is *only* observed for the group of personalizable keywords is compelling evidence that our findings are caused by genuine technological complementarities.

Finally, we also explore the intensive margin relationship between the average user history and the speed of across-user learning using generalized random forests (Athey *et al.*, 2019). Our results are indicative of a S-shaped relationship between the average user history and across-user learning efficiency gains. This is consistent with *locally* increasing returns to scale when the average user history is short. To the best of our knowledge, we are the first study providing empirical evidence for increasing returns to scale in a specific data dimension.

One caveat of our empirical analysis is that the nature of our data hinders an exact identification of effect sizes. It is inherently difficult to relate the search quantities we observe in our sample to the real search traffic. We note that precise effect sizes are likely to vary

across firms and to depend on the quality of the employed algorithm. Our main contribution is to highlight a mechanism, which is based on plausible interaction effects between different aggregation levels of consumer data.

The interaction between the user-specific dimension (K) and the across-user dimension (N) of data has implications for competition policy as well as newly shaping digital regulation.[2] First, with respect to data sharing, our results suggest that personalized data might be particularly valuable because they increase the efficiency from learning across different users. Our results also call for awareness from antitrust policy regarding firms seeking to deepen knowledge about their existing customer base. For instance, merging databases across different services with large overlap in the user base might grant firms significant data advantages.

Second, our findings indicate that sharing search, query and keyword data may not be sufficient to restore a level playing field among competing search engines, if those data cannot be connected to individual users. Our results show that overly cautious anonymization standards may have serious implications on market outcomes: De-personalization of data that does not allow to link searches to user profiles might render data sharing significantly less effective for fostering competition.[3]

We emphasize that our results are likely to generalize to applications other than search engines. The idea of using data generated through the interaction between the service and the consumer is at the core of modern recommendation systems, which power the content of social media feeds, streaming services, and the product recommendations of online retailers. Each time a user interacts with one of the recommendations generated by these systems, she

[2]Regulators around the world are considering the mandated sharing of click and query data to level the playing field between competing search engines, see, for example, ACCC (2021), p. 22; European Commission (2020), Article 6(1)j; and CMA (2020), p. 365 for Australia, Europe, and the UK, respectively. In the US, following an investigation, lawmakers have proposed several bills to curb the market power of Amazon, Facebook, Google, and Apple: The ACCESS (last accessed: February 15, 2022) bill aims at enabling consumers to take control over their personal data.

[3]Striking the right balance between privacy and data-sharing may impose significant challenges, and the anonymization of data is far from trivial as testified by the recent publication of the European Data Protection Supervisor highlighting common misunderstandings related to data anonymization, see EDPS (2021a).

contributes to across-user learning *and* reveals information about herself. Thus, our findings are likely to apply to a broader class of use-cases of algorithmic learning technology.

The remainder of the article proceeds as follows: Section 2 briefly locates our article within the related economic literature. Section 3 provides background information on web search and on how user data can be used for learning. Section 4 introduces the data and explains our empirical strategy. Section 5 presents the main results. Section 6 concludes.

2 Related Literature

On the theory side, Argenton and Prüfer (2012), Prüfer and Schottmüller (2017), Farboodi *et al.* (2019), and De Corniere and Taylor (2020) study competition in data-driven markets with across-user learning. Hagiu and Wright (2020) are the first to consider within-user and across-user learning simultaneously. Close to our study is Lee and Wright (2021), who theoretically model the value created by both data dimensions in recommender systems.

Our results can be thought of as providing evidence for data externalities (Acemoglu *et al.*, 2019; Bergemann *et al.*, 2021) in a real world application of algorithmic learning technology: Since data collected on a specific user also lead to learning about other users, user-specific data have the potential to make learning across users more efficient.

Bajari *et al.* (2019) analyze the impact of data on the predictive performance of Amazon's retail forecast system. Claussen *et al.* (2019) and Yoganarasimhan (2020) document the important role of personalized data for predictive performance. Our study adds to this empirical literature by studying complementarities between different data dimensions.

Bajari *et al.* (2019), Claussen *et al.* (2019), Azevedo *et al.* (2020), and Yoganarasimhan (2020) find evidence for decreasing returns from data. We find evidence for decreasing returns in the across-user dimension and for locally increasing returns from more user-specific data. Our article also speaks to the hypothesis outlined in Posner and Weyl (2019) who stress the importance of considering the *overall* system performance when studying returns from data.

Our findings indicate that user-specific data, through their effect on across-user learning efficiency, are an important driver of system performance.[4]

3 Web Search and User Data

The data we use stem from Yahoo! (2010) and contain fully anonymized search logs spanning a period of 32 days from July 1, 2010 until August 1, 2010, inclusive.[5] An observation in our database contains a keyword identifier, a cookie identifier (which identifies the device on which the search was conducted), the precise time the keyword was entered in the search bar, the ordered list of the top ten organic result URLs and the sequence of clicks performed by the user. In total, we observe approximately 80 million searches performed by 29 million distinct users (identified by the cookies) searching 67 thousand different keywords.

Figure 1 illustrates the structure of the typical search result page at Yahoo! around 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. We do not observe the original sequence of characters but only an identifier allowing us to identify the same keyword over time. Organic search result URLs are highlighted in yellow in the search result list. Paid advertisements are displayed on the north and east edges of the result list.[6]

[4]Several contributions discuss the role of data for competition from a policy perspective. Lerner (2014), Lambrecht and Tucker (2015), and Tucker (2019) argue that the era of digitization poses no special challenge for antitrust authorities and that anti-competitive effects from data should be expected to be weak. By contrast, Newman (2014) argues that data play an important role for firms in securing competitive advantages over rivals. Grunes and Stucke (2015) call for a reorientation of antitrust policy to better account for the role of data. Schepp and Wambach (2015) highlight the role of data in understanding dynamics in digital marketplaces. Sokol and Comerford (2015) emphasize the lack of evidence regarding the role of data for the success of firms

[5]The data that support the findings of this study are available for research purposes on request from https://webscope.sandbox.yahoo.com/. The authors are not allowed to distribute the data directly.

[6]The general layout of the search result page has remained largely intact up to today, with search engines typically devoting the top and east edges of the result page to ads.

3.1 Learning from User Click Behavior

We now discuss how click data can be used to learn relevant search results. Using data resulting from the interaction between the users and the web service to improve the prediction accuracy of recommendations is the core principle of modern recommender systems.[7] The click data the search engine collects can be used to train the algorithm to learn relevant search results.

Table 1 provides an illustrative example: The ones stand for positive search experiences users had with a specific result for a specific keyword (i.e. the user clicked on the result), the zeros stand for negative experiences (i.e. the user ignored the result). Table 1 illustrates the stylized scenario in which the algorithm observes the implicit user feedback for all user-keyword combinations, except one. The objective is to accurately predict the click behavior for the missing user-keyword combination (user $i = N$ and keyword $j = K$).

Note that in the stylized example of Table 1, the dimension K captures the amount of user-specific data available about each user, while N captures the amount of searches observed for each keyword. Table 1 serves to illustrate why both dimensions of data are likely to be relevant for statistical learning. To build intuition, we can think of the algorithm as training a prediction model for each keyword. For the sake of exposition, consider the case in which the model used to train the algorithm for keyword K is a simple OLS regression:

$$Y_i = \beta_0 + \beta_k X_i + \epsilon_i \tag{1}$$

Where Y_i denotes the preference realization of user i for keyword K in Table 1, and X_i denotes the preference realization over the other keywords (1 to $K-1$). As N increases, more observations become available to estimate the regression parameters. A larger set of observable characteristics, K, will lead to a larger R-squared of the OLS-regression. Therefore, for the same N, the OLS-model with larger K will have a higher prediction accuracy.

[7] This method is known as collaborative filtering, we refer the reader to Adomavicius and Tuzhilin (2005) and Lu *et al.* (2015) for a survey on recommender systems.

As a result, the average prediction accuracy for user N will be weakly higher if the algorithm can rely on more observable characteristics about users when training the model.[8]

Algorithms used by real-world recommendation engines are more sophisticated than simple OLS-models. Additionally, the typical prediction tasks are usually more complex and the data environment less well-behaved (in reality, the algorithm will typically be confronted with a different length of K across users). Nevertheless, the example serves to illustrate why it is reasonable to investigate if more user-specific data might lead to more pronounced across-user learning.

3.2 Narrative Evidence for Learning from User Click Behavior

If the search algorithm does not rely on user-specific information, like click data, the outlined mechanism would not apply. As far back as 2005, Yahoo! engaged in research related to recommendation systems (Decoste *et al.*, 2005).[9] A conference article published in 2011, and coauthored by a senior Yahoo! researcher, states that "Today, we can also use the collective wisdom of users, reflected in weblogs as clicked pages or in query logs as queries and clicked results." and "Contextualization and personalization are an essential ingredient of modern search [...]" (Baeza-Yates *et al.*, 2011, see p. 28 and p. 29).

The best evidence we could find for the use of user-specific information in the search industry around the period our data were collected stems from a research article using Bing search traffic data from September 2010. In the data description, the article states

[8]We omit a discussion of the difference between in-sample accuracy, as measured by the R-squared, and out-of-sample accuracy, as measured by the prediction error. Technically, a large in-sample R-squared does not automatically imply a better out-of-sample prediction accuracy. It might be that increasing the number of observable characteristics reduces the prediction accuracy. The fact that a larger in-sample fit can correspond to a worse out-of-sample prediction accuracy is known as the bias-variance trade-off. Methods, such as cross-validation techniques, have been developed to account for the phenomenon. As a result, an algorithm that can rely on more observable characteristics should not perform worse than an algorithm that relies on less. Intuitively, if increasing the number of observable characteristics reduces the prediction accuracy, the algorithm will simply learn to ignore the corresponding variables by applying cross-validation techniques. As a result, the main intuition of our simplified exposition should remain intact.

[9]Some early personalization features were also publicly announced and introduced around 2005, see https://www.cnet.com/news/yahoo-debuts-personalized-search/ (last accessed: February 15, 2022). While these early personalization attempts were not a default setting of the search engine, they nevertheless demonstrate an active agenda towards using user-specific information to recommend search results.

that "To isolate the impact of long-term personalization, we did not use any other form of personalization from the Bing search engine over the time period for which the data were collected." (Sontag *et al.*, 2012, see p. 438 at the beginning of Section 5). While this evidence does not pertain to Yahoo!, it corroborates the use of user-specific information in the industry.

4 Empirical Strategy

In this Section we define our main variables. Additionally, we discuss important characteristics of our data which help understand the choices we make in the empirical analysis.

4.1 Variable Description

Search Quality Measure

We observe a log that records the sequence and position of the clicks performed for each search. The log ends if a user clicks on, i.e. visits, an URL and does not return to the search result page within a specified amount of time. Figure 2 shows the distribution of the last click position across searches. Position 0 identifies URLs above the first organic URL such as ads, spelling suggestions and "also try" recommendations. Positions 1 to 10 identify the organic URLs. Position 11 identifies clicks below the last organic URL (next result page). The character "*o*" encodes other clicks (such as closing the browser), "*s*" encodes cases in which the user entered a new keyword in the search bar, and "*nc*" identifies instances where no click was performed.

We rely on the notion that a user not returning to the search result page after clicking a URL means she was satisfied with the provided content. Our click based quality measure assumes that the search experience is better if the URL visited *last* is displayed further up on the result page. Based on this, we classify the quality of each search experience as good (encoded as 1) or bad (encoded as 0). For each keyword, we aggregate the individual search

experiences to an average quality score, which is known as the click-through-rate (ctr).[10] Click-based quality measures, such as the click-through-rate, are widely used in the search engine literature (Joachims, 2002; Jain and Varma, 2011). We use the following quality measure:

$$ctr_j^{\{1\}} = \frac{\sum_{s \in S_j} \mathbb{1}[lcp_s = 1]}{\sum_{s \in S_j} \mathbb{1}[lcp_s \neq 0]}$$

S_j denotes the number of searches over which the click-through-rate for keyword j is calculated. lcp_s denotes the last click position for search s. $\mathbb{1}$ denotes the indicator function. The quality measure only counts a search as successful if the last click was performed on the top displayed organic URL (position 1).

$ctr_j^{\{1\}}$ ignores searches ending with a click on position 0 URLs because their meaning in terms of quality is ambiguous. Besides ads, position 0 URLs also capture spelling suggestions or "also try" recommendations. If a user clicks on an "also try" recommendation, we do not systematically observe her subsequent click behavior, which could either indicate good or bad quality. Discarding clicks on position 0 only leads to a drop in the number of observations "within" a keyword. Most keywords also experience clicks on other URLs. As a consequence, we only loose a marginal fraction of keywords from discarding position 0 clicks. In other words, the sample of keywords considered remains essentially the same (also see the notes in Table 2).

Explanatory Variables

For the empirical analysis, we define two variables that capture the two data dimensions which are at the center of our analysis. The first data dimension captures the amount of data accumulating across users. The second dimension captures the amount of user-specific

[10] Note that if a user first clicks on the top displayed URL and returns to the result page afterwards to choose another URL further down the result list, this is not counted as a good search experience. We only use the last performed click in a session.

data. The data dimensions are defined relative to the keywords that we observe in our sample, more precisely:

- S_j denotes **the number of searches** for keyword j.

- $\overline{H_j}$ denotes **the average user history** for keyword j. It captures the average number of times users were observed before they entered the keyword j in the search bar of the search engine. Denote by H_{jt} the length of the search history of a user searching keyword j at time t during the period of our sample, $T = [\underline{t}, \overline{t}]$, then $\overline{H_j} = \frac{\sum_{t \in T} H_{jt}}{S_j}$.

Figure 3 illustrates the computation of both variables. S_j, the number of searches, captures the number of observations (the N-dimension) the algorithm could rely on to train the statistical model for keyword j. $\overline{H_j}$ is the average number of times users have been observed before entering keyword j in the search bar of the search engine. We use $\overline{H_j}$ as a proxy for the average number of observable characteristics that the search engine can use to train its statistical model for keyword j (i.e. the K-dimension). Users who have been observed more often are likely to reveal more observable characteristics about themselves.[11]

We rely on the keyword as the fundamental unit of analysis because keywords are generally observed more often than users. Thus, keywords lend themselves better to analyze quality changes as a function of data: We do not observe the typical individual often enough to study how the search experience of individuals changes as the number of searches increases.[12]

Table 2 shows the summary statistics for the main variables used in the analysis. It becomes apparent that keywords are generally observed much more often than users. This implies that the different searches observed for a keyword are generally performed by different

[11] How the search engine extracts characteristics from the observed search behavior of users is not known to us. However, it is plausible to assume that a longer history generally corresponds to more observable characteristics. We are aware that users might delete their cookie cache. In this sense, the term cookie history might be more appropriate. The chosen terminology underscores the notion that a cookie is generally associated with a single user.

[12] Keyword might also be considered a more "homogeneous" unit of analysis. When following individuals, the experienced quality will heavily depend on the "difficulty" of the different keywords entered by the user. By contrast, different individuals entering the same keyword are interested in a similar topic.

users. Thus, the number of searches indeed captures the dimension relevant for across-user learning. Note that we observe substantial variation in the average user history despite observing many users only once.

4.2 Additional Considerations

Relationship to Real Search Traffic

A back of the envelope calculation using available data on the search engine market suggests that the size of our sample corresponds to approximately two percent of the worldwide search traffic on the Yahoo! search engine during July 2010.[13] This suggests that, on average, the search quantity variables constructed from our sample are 50 times smaller than in reality. In light of this, the apparently small search quantities reported in Table 2 are likely to capture substantially larger values in reality.

Search Traffic Prior to the Sample Period

Most keywords likely experienced search traffic already prior to our sample period. Similarly, most users likely already used Yahoo! before we observe them in our sample. This source of unobserved heterogeneity is an important factor that we need to account for. In the first part of our analysis, we will therefore use the observed variables as proxy measures for past quantities in an attempt to gauge the long-run impact of data. The insights gained from this analysis will motivate the subsequent analyses. In Appendix A.1, we explain in more detail why the observed search quantities are likely to be good proxy-measures for the past search traffic.

[13] It is estimated that Google's total search volume in 2010 amounted to roughly 1 trillion searches (see https://www.internetlivestats.com, last accessed: 15 February 2022). Google's share in worldwide searches in 2010 was roughly 90%. Yahoo's share in searches during the same period amounted to four percent (see https://www.statista.com, last accessed: 15 February 2022). From this, we obtain that Yahoo's total monthly search traffic in 2010 was approximately equal to 3.7 billion searches. Our sample corresponds to roughly 2 percent of 3.7 billion searches. We have no indication on the geographic scope of the sampling. The factor 50 is therefore an upper bound. If we assume that the sampling region is the US, we can rely on the figure of 2.7 billion monthly searches in October 2010 (see https://www.comscore.com, last accessed: 15 February 2022). In this case, our sample would cover roughly 3 percent of all Yahoo! searches in the US.

July 20th Anomaly

We observe an abnormal drop in the click-through-rate in the period from July 19 to July 21. To the best of our knowledge, this anomaly is related to the testing of a new algorithm in the wake of the Yahoo!-Bing merger. While the anomaly appears to be transitory, we cannot exclude the possibility that it might mark the permanent transition to a new algorithm. For our main analysis we therefore rely on quality measures obtained using observations prior to the anomaly. The same is true for the number of searches. For the average user history, we use all the information available in the sample.

We have no indication that the anomaly affects the explanatory variables. To understand why we prefer keeping all observations when computing the average user history, note that keywords would barely be differentiated with respect to this variable if we would rely on a very short time frame for computation. This is the case because most users are only observed very rarely (see Table 2). As a result, in the extreme scenario with only one day of data, most keywords would have a average user history close to one. The issue is less pronounced for the quality measure and the number of searches for which one day of data will generally result in informative variation across keywords.

We provide an in-depth account of the anomaly in Appendix A.2; there we also report the results obtained when including observations after the anomaly. Since most keywords accumulate searches evenly over time (see Appendix A.1), dropping observations during and after the anomaly does not impact the sample of keywords considered, instead it only affects the number of observations available per keyword.

5 Results

The results section is structured in three parts. The first part treats the variables observed during the sample period as proxies for the unobserved search histories in the past. We document a pattern consistent with faster across-user learning from more user-specific data

and offer a first interpretation of these results. Building on this, the second part gathers evidence on the short-run impact of data. By focusing on within-keyword variation in quality, i.e. quality *changes*, the second part addresses identification concerns related to the proxy-variable approach, which focuses on quality *levels*.

The third part of the analysis directly asks if the observed pattern of faster across-user learning is *caused* by user-specific data. The causal mechanism we have in mind relies on the assumption that the algorithm relies on user-specific data as an input. If this is not the case, longer user histories cannot *cause* faster across-user learning through technological complementarities, because this presupposes that user-specific data constitute an input into the technology. We design a test that leverages this idea and show that the pattern of faster across-user learning is only observed for the subset of keywords for which we have indication that the algorithm relies on user-specific data. In the third part, we also gather evidence on the intensive margin relationship between user-specific data and efficiency gains in across-user learning and show that this relationship appears to be S-shaped.

5.1 Proxy-Variable Approach: A Long-Run Perspective

We treat S_j and $\overline{H_j}$ as proxy variables for $\mathcal{S}_j$ and $\overline{\mathcal{H}}_j$, the unobserved number of searches and average user history realized before our sample. Under the assumption that data matters, more searches and a longer average user history in the past should lead to a higher quality level during the period of our sample.

Denote by b_s the bins defined by the deciles of the distribution of S and denote by b_h the bins defined by the median of the distribution of $\overline{H}$. To study the impact of past data, we estimate the following regression:

$$ctr_j^{\{1\}} = \sum_{s=1}^{10} \sum_{h=1}^{2} \lambda_{sh} \mathbb{1}\{S_j \in b_s\} \mathbb{1}\{\overline{H_j} \in b_h\} + \epsilon_j \tag{2}$$

$ctr_j^{\{1\}}$ denotes the click-through-rate of keywork j using the observations before the July

20th anomaly. λ_{sh} captures the average click-through-rate conditional on keywords j belonging to the bin s and the bin h.

Figure 4 maps the estimated values of λ_{sh} against the base ten logarithm of the number of searches defining the left edges of the bins b_s.[14] The black curve stands for keywords with an average user history below the median, the gray curve for keywords with an average user history above the median.

Summarizing the results shown in Figure 4, we observe a positive concave relationship between the number of searches and the average click-through-rate.[15] Additionally, this positive relationship is more pronounced for keywords with a longer average user history. Thus, keywords with a similar number of searches in the past attained a higher quality level if we observe a longer average user history.[16] The results are therefore consistent with more user-specific data (longer average user history) leading to efficiency gains in across-user learning (number of searches).

For a low number of searches, we observe no difference between keywords as a function of the average user history. This is consistent with complementarities between both dimensions of data. Lacking data in one dimension reduces the benefits of additional data in the other dimension. In the OLS-analogy used throughout this article, a lack of searches corresponds to a lack of observations to estimate the OLS parameters consistently (small N). As a consequence, the potential benefits form a larger K-dimension cannot be leveraged. We also note that the positive concave relationship between the number of searches and the increase in the click-through-rate is consistent with statistical learning theory, which predicts decreasing returns in the N-dimension (Lerner, 2014; Bajari *et al.*, 2019).

[14] For the proxy-variable analysis, we use the number of searches observed over the entire sample period. This is motivated by the fact that the observed total number of searches does not appear to be affected by the anomaly. Using all the available information about the popularity enhances the proxy-variable property of our popularity measure. However, our results do not change if we only use the number of searches observed until the anomaly.

[15] Note that the x-axis has a logarithmic scale, the range with apparently increasing returns in Figure 4 has, in fact, strongly decreasing returns.

[16] We refer the reader to Appendix A.1 for a discussion on the properties of S_j and $\overline{H_j}$ as proxies for $\mathcal{S}_j$ and $\overline{\mathcal{H}}_j$.

If we interpret the learning curve from additional searches as a quality-production function, more user-specific data appear to act as a "technology shifter" leading to more efficient across-user learning. This offers a novel perspective on the value of data for firms: The firm with deeper user profiles has an inherent *ceteris paribus* advantage when confronted with a novel prediction task (such as a new search keyword) because it can learn faster across-users. This potential complementarity effect has not yet been documented in empirical applications using real-world data.[17]

5.2 Exploiting Short-Run Variation in the Sample

In this subsection, we focus on within-keyword changes in the click-through-rate and the number of searches during the period of our sample. The question we ask is whether the observed quality change for a given number of searches is larger if we observe a longer average user history. This allows us to directly investigate if there is evidence that keywords with a longer average user history learn *more*. Exploiting first differences in the quality and the number of searches helps addressing endogeneity concerns related to the previous analysis which focused on quality levels instead of quality changes.

From the proxy-variable analysis, we know that the quality level of a keyword is correlated with its observable characteristics. Keywords with a longer average user history and a larger number of searches experience, on average, a higher quality level in our sample. Since the maximum click-through-rate is bounded from above (the click-through-rate lies in the interval between 0 and 1), ignoring the initial quality is likely to distort the true relationship between our explanatory variables and the observed quality change.

Since keywords with a longer average user history are more likely to experience a higher quality level, they have mechanically less scope for quality improvement compared to key-

[17]The article of Bajari *et al.* (2019) investigates a similar effect but not in the context of user-generated data. The mechanism of action they propose resembles economies of scope, where a larger variety shifts the quality production function from data scale. Bajari *et al.* (2019) find no effect consistent with this type of data externality. The pattern we document is related to economies of scope because a larger variety of keywords helps generating more user-specific data.

words starting from a lower quality level. Thus, ignoring the quality level is likely to lead to a downward bias when estimating the relationship between the observed quality change and the average user history.[18]

In our analysis of quality changes, we therefore condition on the initial quality level by computing the initial click-through-rate over the first 100 searches of a keyword. We use a rather large number of initial searches to mitigate the potential impact of *regression to the mean*.[19] To obtain two non-overlapping windows of 100 searches, we therefore have to restrict the analysis to keywords with at least 200 searches.This reduces our sample to 23637 keywords.

To analyze the relationship between the observed quality increase and the average user history of keywords, we calculate the following statistic for different sub-samples of keywords:

$$\overline{\Delta ctr_j^1(s)} = \frac{1}{N_J} \sum_{j \in N_J} \left(ctr_j^1(s) - ictr_j^1 \right) \tag{3}$$

Where $s \in \{1,101,...,4901\}$ denotes the number of searches at the left edge of each 100 searches window, $ictr^1$ denotes the click-through-rate over the first 100 searches, and $\Delta ctr_j^1(s)$ denotes the increase in click-through-rate observed for keyword j after s searches. In words, Equation 3 describes the average change in quality between the first window of 100 searches, $ictr^1$, and all subsequent windows, $ctr^1(s)$. With a slight abuse of notation, N_J denotes the sub-sample of keywords used in the computation of Equation 3.

In Figures 5b to 5d, each curve shows the results obtained for keywords above (gray) and below (black) the median average user history. Figure 5b shows the results obtained for all keywords, irrespective of the initial click-through-rate. Figure 5c shows the results obtained for the sub-sample of keywords starting from a quality level below the median initial click-

[18]The necessity of controlling for the initial quality, which is a variable that does not vary within keywords, is the main reason we do not perform a conventional regression analysis with keyword fixed-effects.

[19]Regression to the mean occurs whenever units are classified based on an initial outcome. Intuitively, the problem arises because subjects are "erroneously" allocated to a category based on a single (or few) observation(s), which is not representative of the truth. This erroneous allocation leads to a reversion to the mean in subsequent observations. Barnett *et al.* (2004) provide an accessible discussion of the phenomenon.

through-rate. Figure 5d shows the results for keywords starting from a click-through-rate above the median. The distribution of the initial click-through-rate is shown in Figure 5a.[20]

Figure 5 highlights the importance of controlling for the initial click-through-rate. On average, keywords starting form a high initial click-through-rate do not experience noticeable quality changes. Since these keywords already reached a high quality level at the beginning of the sample, they have less scope for learning. By contrast, keywords with an initial click-through-rate below the median experience sizeable quality changes. Furthermore, for a low initial quality, the quality increase is more pronounced for keywords with a longer average user history (approximately five percentage points), which provides further evidence that user-specific data increase across-user learning efficiency.

When ignoring the initial quality (Figure 5b), there is no apparent difference between keywords with a longer and shorter average user history. This is the result of the positive correlation between the average user history and the initial click-through-rate, which leads to a downward bias when estimating the relationship between the average user history and the observed quality changes without accounting for the initial click-through-rate.

We note that accounting for the initial quality and focusing on quality changes address concerns related to reverse causality. For example, one might be worried that more experienced users search specific keywords more often because they learned that entering these keywords provides a better search experience. However, the keywords in Figure 5c all start from a similar *low* level of quality. Thus, it appears not plausible that there is an initial quality difference in keywords that leads more experienced users to learn to use these keywords.

[20] The plots show the quality evolution until 5000 searches because dropping observations after the anomaly results in a low number of observations with more than 5000 searches. As is explained in more detail in Appendix A.1, keywords accumulate searches evenly in our sample. As a result, a keyword with 10000 searches from the first to last day of our sample, will typically have accumulated approximately 5625 searches until July 18th (our sample spans 32 days, therefore $\frac{18}{32} = 0.5625$).

5.3 Assessing Causality

In this Subsection we ask whether the observed pattern of faster across-user learning from more user-specific data is causal. In other words, we ask the question whether the pattern is a result of technological complementarities between different aggregation dimensions of user data. A trivial prerequisite for complementarities to become effective is that both dimensions of data enter the algorithm (i.e. the quality production technology) as an input. If we could be certain that the algorithm does not exploit user-specific information, we could dismiss the results presented so far as spurious because technological complementarities could be ruled out.

The test that we present in this section exploits this insight by classifying keywords in two groups: The first group consists of keywords for which we have indication that the search algorithm exploits user-specific information. The second group consists of keywords for which we have no such indication. If the hypothesis of technological complementarities is correct, we would expect to observe a pattern of faster across-user learning from more personal keywords *only* for the first group of keywords. By contrast, we would not expect to see faster across-user learning from additional user-specific data for the second group of keywords.

Our classification method combines two approaches: The first approach relies on the notion that, if user-specific information plays a role, we should observe changes in the distribution of search results as a function of the length of the user-history of searchers. The second approach relies on the notion that search results should change more frequently if personalization is used for a keyword.

For the first approach, we test whether the frequency distribution of search results varies between users with long histories and users with short histories. Based on the overlapping set of search results shown to both group of searchers, we use a Chi-Square Test to test the null hypothesis of an equal frequency distribution of search results across both groups. The Chi-Square Test provides a method to assess whether user-specific information is likely to be

leveraged by analyzing differences in the result distribution. We combine the results of the Chi-Square Test with a second approach that is based on a variance criterion: If a keyword personalizes search results we would expect the search results to change frequently. We build a measure that captures the typical length of search sequence without changes in content. If the typical sequence is short this indicates that results "rotate" frequently.

We call a keyword "personalizable" if the Chi-Square Test rejects the null hypothesis *and* if we find a high variance of search results. All other keywords are classified as "non-personalizable". Out of 23637 keywords with at least 200 searches, 4496 are classified as "personalizable" (19%). We provide a comprehensive description of both approaches used for classification in Appendix A.3.

Figure 6 shows the average ctr-increase (using Equation 3) for keywords with a long average user history and keywords with a short average user history for the group of personalizable keywords (Figure 6a) and the group of non-personalizable keywords (Figure 6b). The analysis relies on keywords starting from an initial click-through-rate below the median.

We view the results of Figure 6 as providing strong support for the causal mechanism of data complementarities: Faster learning as a function of user-specific data is only observed for the group of personalizable keywords. Non-presonalizable keywords do not exhibit a similar pattern. The observed quality increase for non-personalizable keywords is likely to capture generic learning, such as popularity based ranking methods, which aim to learn the most popular results across users. The absence of a pattern consistent with differential learning in Figure 6b suggests that confounding factors unrelated to data complementarities are not likely to play a major role in explaining Figure 6a.[21]

Since non-personalizable keywords appear unaffected by the average user history, we from now on treat them as a stable benchmark (control group) which allows us to assess the impact of the average user history on personalizable keywords. To study how the learning difference

[21]For instance, if differential across-user learning could be explained by keywords with longer average user histories being "easier" (a confounding factor), we would expect to see faster across-user learning with a longer average user history in *both* Figures.

between personalizable and non-personalizable keywords is, *ceteris paribus*, related to the average user history, we use the method of generalized random forests (Athey *et al.*, 2019; Nie and Wager, 2021), which allows us to estimate $\beta(x)$ in the following model:

$$\begin{aligned} \Delta ctr_j &= b_j W_j + \epsilon_j \\ \beta(x) &= \mathbf{E}[b_j | x] \end{aligned} \tag{4}$$

Where Δctr_j denotes the change in click-through-rate between the first and last 100 searches of a keyword and W_j is an indicator variable that takes the value of one if keyword j is personalizable. b_j is the keyword specific marginal effect of personalization, and $\beta(x)$ is the average marginal effect, conditional on x, which the generalized random forest estimates. If, conditional on x, the indicator variable W_j is independent of the error term ϵ_j, then $\beta(x)$ is identified.

While Figure 6b suggests that confounding factors are unlikely to play a major role in the complementarities we observe in 6a , correlation between W_j and ϵ_j cannot be ruled out. It is worthwhile emphasizing, that even under confoundedness, generalized random forests remain an interesting method to descriptively explore the effect of the average user history in a non-parametric manner. As is explained in Athey *et al.* (2019), generalized random forests can be considered efficient kernel estimators, which, compared to conventional non-parametric methods, offer substantial computational advantages.[22]

By estimating $\beta(x)$ for various x, we can assess how the differences in quality changes between both groups vary with the observed characteristics. In the subsequent analysis, we include the average user history, the total search quantity and the initial click-through-rate as variables in the x-vector.[23] For instance, by estimating $\beta(\overline{H}, S = \mathbf{c}, ictr = \mathbf{c})$ for different

[22]The standard implementation of the generalized random forests in *R* (*grf* library), which we use, automatically prevents overfitting. Compared to more conventional non-parametric estimation methods, such as local polynomial regressions, the generalized random forest offers substantial savings in computation costs because it avoids the costly tuning of bandwidth parameters.

[23]We leave all parameters of the *grf* library at their default values, except for the number of trees, which

values of $\overline{H}$, we can assess how the difference in ctr-increases between personalizable and non-personalizable keywords varies as a function of the average user history, when holding fixed the total search quantity and the initial click-through-rate ($ictr$).

Before estimating $\beta(x)$, we perform a matching procedure to balance the observable characteristics between personalizable and non-personalizable keywords. This is necessary to apply the generalized random forest method successfully as it requires propensity scores with good overlap properties.[24] We use standard nearest neighbor mahalanobis matching. In Appendix A.4, we provide an analysis of the covariate balance and describe the matching procedure in more detail.

Figures 7a to 7d show the results obtained when applying the random forest methodology to the matched sample of keywords. Each curve shows the estimates for $\beta(x)$ as a function of the average user history. Each panel conditions on a different total search quantity. All panels condition on the same initial click-through-rate of 25 percent. According to the results shown in Figure 7d, personalizable keywords with an average user history of seven and a total search quantity of 5476 searches experience a six percentage point larger increase in the click-through-rate than comparable non-personalizable keywords. No statistically significant difference is observed when the average user history is short.

For a low number of total searches, we observe no indication for differential learning as a function of the average user history. With an increasing number of searches, the positive relationship between the length of the average user history and the differential quality change increases. This is further evidence for complementarities between the number of searches and the average user history: A large number of searches is required to leverage the potential gains from user-specific data.

Interestingly, the results from Figures 7c and 7d suggest that the effect of longer average

we set equal to the number of observations in the sample, as is recommended by the authors of the *grf* library.

[24] Generalized random forests fail to recover even obvious patterns in the data if the groups are not well balanced, see the "troubleshooting" section on the website of the authors of the *grf* package: https://grf-labs.github.io (last accessed on 15 February 2022).

user histories on efficiency gains from across-user learning is S-shaped. This would imply locally increasing returns to data. To the best of our knowledge, this is the first empirical evidence consistent with increasing returns to scale in a particular data dimension.

6 Conclusion

Our results call for awareness from antitrust policy regarding firms seeking to deepen knowledge about their existing customer base. For instance, merging databases across different services with large overlap in the user base might lead to substantial gains in across-user learning efficiency.

The documented S-shaped relationship between the amount of user-specific data and efficiency gains from across-user learning indicates that lack of data might constitute a serious barrier to entry. Contrary to the scenario with rapidly diminishing returns, a potential entrant might need to first reach a certain data threshold before starting to benefit from learning effects.

Our findings suggest that sharing search, query and keyword data may only help restoring a level playing field among competing search engines if those data can be connected to individual users. Overly cautious anonymization standards might render data sharing significantly less effective for fostering competition. This is of relevance because current regulatory suggestions mandating the sharing of data tend to focus on the size of the user base and less on personal data.[25] With relation to the sharing of personal data, commentators tend to emphasize primarily the need to protect privacy.[26]

Careful consideration should be given to striking the right balance between preserving privacy and improving market outcomes. For example, this may mean that search engines create common user identifiers, based on which search histories can be shared and connected,

[25]For example, the EU Digital Markets Act explains that "*the value of online search engines to their respective business users and end users increases as the total number of such users increases*"(See European Commission (2020), paragraph 56).

[26]See, for example, EDPS (2021b), par 32.

but any personal information subject to privacy regulation would be kept in separate data-silos, not directly connected to these common user identifiers.

References

ACCC (2021) Digital Platform Services Inquiry – September 2021, Report on market dynamics and consumer choice screens in search services and web browsers: issues paper, https://www.accc.gov.au/focus-areas/inquiries-ongoing/digital-platform-services-inquiry-2020-2025/september-2021-interim-report (July 2020).

Acemoglu, D., Makhdoumi, A., Malekian, A. and Ozdaglar, A. (2019) Too much data: Prices and inefficiencies in data markets, *National Bureau of Economic Research, No. w26296*.

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.

Athey, S., Tibshirani, J. and Wager, S. (2019) Generalized random forests, *The Annals of Statistics*, **47**, 1148–1178.

Azevedo, E. M., Deng, A., Montiel Olea, J. L., Rao, J. and Weyl, E. G. (2020) A/b testing with fat tails, *Journal of Political Economy*, **128**, 4614–000.

Baeza-Yates, R., Boldi, P., Bozzon, A., Brambilla, M., Ceri, S. and Pasi, G. (2011) Trends in search interaction, in *Search computing*, Springer, pp. 26–32.

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.

Bergemann, D., Bonatti, A. and Gan, T. (2021) The economics of social data, *RAND Journal of Economics, Forthcoming*.

Claussen, J., Peukert, C. and Sen, A. (2019) The editor vs. the algorithm: Returns to data and externalities in online news, *SSRN*, https://ssrn.com/abstract=3479854 (January 11, 2021).

CMA (2020) Online platforms and digital advertising market study, https://www.gov.uk/cma-cases/online-platforms-and-digital-advertising-market-study (July 2020).

De Corniere, A. and Taylor, G. (2020) Data and competition: a general framework with applications to mergers, market structure, and privacy policy, *SSRN*, https://ssrn.com/abstract=3547379 (February 2020).

Decoste, D., Gleich, D., Kasturi, T., Keerthi, S., Madani, O., Park, S.-T., Pennock, D. M., Porter, C., Sanghai, S., Shahnaz, F. *et al.* (2005) Recommender systems research at yahoo!

research labs, *Beyond Personalization 2005*, p. 91.

EDPS (2021a) AEPD-EDPS joint paper on 10 misunderstandings related to anonymisation, https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf (June 30, 2020).

EDPS (2021b) Opinion 2/2021 on the Proposal for a Digital Markets Act, https://edps.europa.eu/system/files/2021-02/21-02-10-opinion_on_digital_markets_act_en.pdf (February 2021).

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).

European Commission (2020) Proposal for a regulation of the European Parlament and the Council on contestable and fair markets in the digital sector (Digital Markets Act), https://ec.europa.eu/info/sites/default/files/proposal-regulation-single-market-digital-services-digital-services-act_en.pdf (December 20, 2020).

Farboodi, M., Mihet, R., Philippon, T. and Veldkamp, L. (2019) Big data and firm dynamics, in *AEA papers and proceedings*, vol. 109, pp. 38–42.

Grunes, A. P. and Stucke, M. E. (2015) No Mistake About It: The Important Role of Antitrust in the Era of Big Data, *SSRN*, https://ssrn.com/abstract=2600051 (April 28, 2015).

Hagiu, A. and Wright, J. (2020) Data-enabled learning, network effects and competitive advantage, *Working Paper*.

Jain, V. and Varma, M. (2011) Learning to re-rank: query-dependent image re-ranking using click data, in *Proceedings of the 20th international conference on World wide web*, pp. 277–286.

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).

Lee, G. and Wright, J. (2021) Recommender systems and the value of user data, *National University of Singapore Working Paper*, https://scholarsite-production.s3.eu-west-2.amazonaws.com/ (September 2021).

Lerner, A. V. (2014) The role of'big data in online platform competition, *SSRN*, https://ssrn.com/abstract=2482780 (August 2014).

Lu, J., Wu, D., Mao, M., Wang, W. and Zhang, G. (2015) Recommender system application

developments: a survey, *Decision Support Systems*, **74**, 12–32.

Newman, N. (2014) Search, Antitrust, and the Economics of the Control of User Data, *Yale J. on Reg.*, **31**, 401.

Nie, X. and Wager, S. (2021) Quasi-oracle estimation of heterogeneous treatment effects, *Biometrika*, **108**, 299–319.

Posner, E. A. and Weyl, E. G. (2019) *Radical Markets*, Princeton University Press.

Prüfer, J. and Schottmüller, C. (2017) Competing with Big Data, *SSRN*, https://ssrn.com/abstract=2918726 (February 16, 2017).

Schaefer, M., Sapi, G. and Lorincz, S. (2018) The effect of big data on recommendation quality: The example of internet search, *DIW Berlin Discussion Paper No. 1730*.

Schepp, N.-P. and Wambach, A. (2015) On Big Data and its Relevance for Market Power Assessment, *Journal of European Competition Law & Practice*, **7**, 120–124.

Sokol, D. D. and Comerford, R. (2015) Antitrust and Regulating Big Data, *Geo. Mason L. Rev.*, **23**, 1129.

Sontag, D., Collins-Thompson, K., Bennett, P. N., White, R. W., Dumais, S. and Billerbeck, B. (2012) Probabilistic models for personalizing web search, in *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 433–442.

Tucker, C. (2019) Digital Data, Platforms and the Usual [Antitrust] Suspects: Network Effects, Switching Costs, Essential Facility, *Review of Industrial Organization*, **54**, 683–694.

Yahoo! (2010) L18 anonymized yahoo! search logs with relevance judgments, *Yahoo! Research*, https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=50 (last accessed: November 27, 2021).

Yoganarasimhan, H. (2020) Search personalization using machine learning, *Management Science*, **66**, 1045–1070.

7 List of Tables

Table 1: Learning from User Click Behavior – Illustrative Example

	User 1	...	User i	...	User N
Keyword K	0	0	1	1	-
Keyword K - 1	0	0	0	1	0
...	1	0	1	0	1
Keyword j	0	1	0	1	0
...	1	0	1	1	1
Keyword 1	0	0	1	1	1

Stylized example in which the search engine has to predict the taste of user $i = N$ for search results presented in response to entering keyword $j = K$. The search engine observes whether previous users clicked the results for keyword K. Additionally, we assume the search engine observes the clicks each user left before entering keyword K.

Table 2: Summary Statistics

	count	mean	min	p25	p50	p75	max
Searches by Keyword (S_j)	67,652	1,194.04	3.00	19.00	115.00	851.00	10,000.00
Searches by User	29,664,490	2.72	1.00	1.00	1.00	3.00	516.00
Average User History ($\overline{H_j}$)	67,652	4.06	1.00	2.79	3.56	4.78	67.33
Quality Measure ($ctr_j^{\{1\}}$)	67,473	0.47	0.00	0.21	0.45	0.75	1.00

Note: The quality measure is computed using all the available searches observed for a keyword. Using the ctr^1 quality measure results in a minimal loss of observations because for some keywords with very few searches, we only observe clicks on position 0. We observe 67,652 keywords and 29 million users. Our analysis is performed at the keyword level. The number of searches by user is reported because it forms the basis of the computation of $\overline{H_j}$.

8 List of Figures

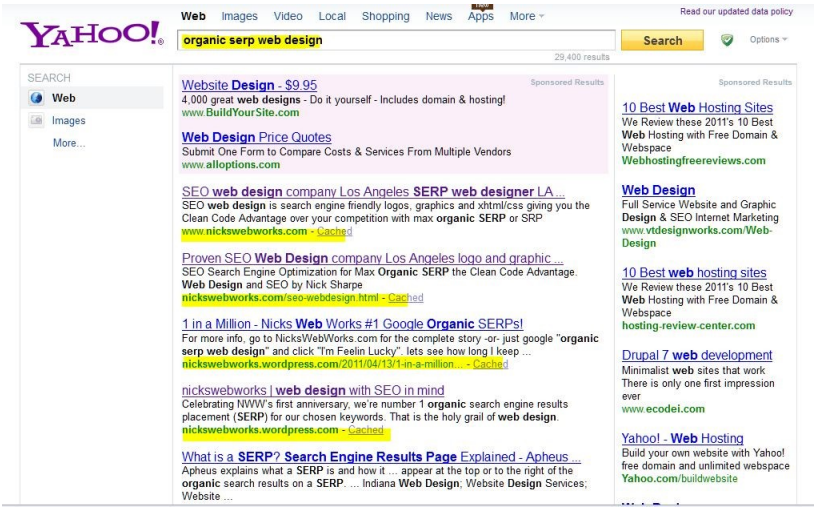


Figure 1: Search Result Layout at Yahoo!, 2011

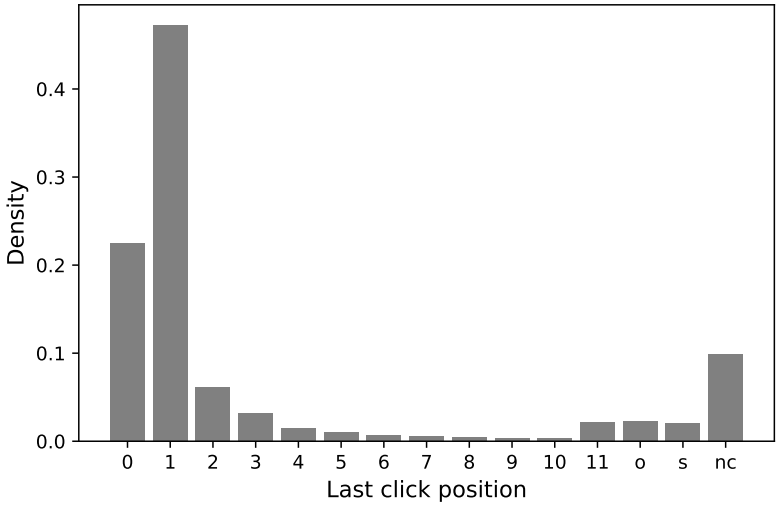


Figure 2: Distribution of Final Clicks

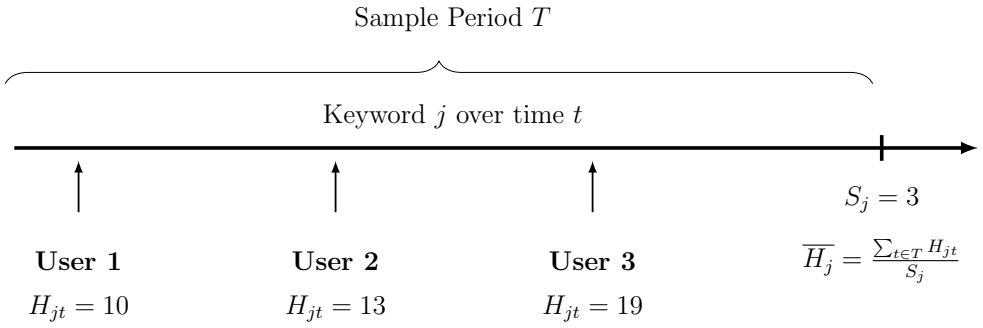


Figure 3: Computation of Explanatory Variables

Note: We label the individual user histories by t to indicate that we use the length of the user history at the time of search. We also do not discriminate between users that search the keyword once or several times, which is the reason why we do not use a user-specific subscript (i). In general, the contribution of one single user in the total number of searches a keyword experiences is small (see the discussion of the summary statistics of Table 2).

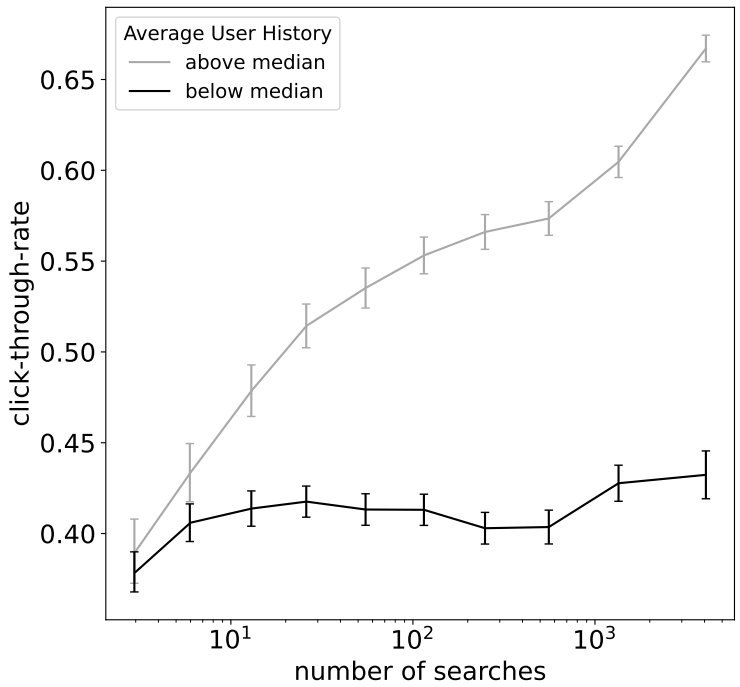
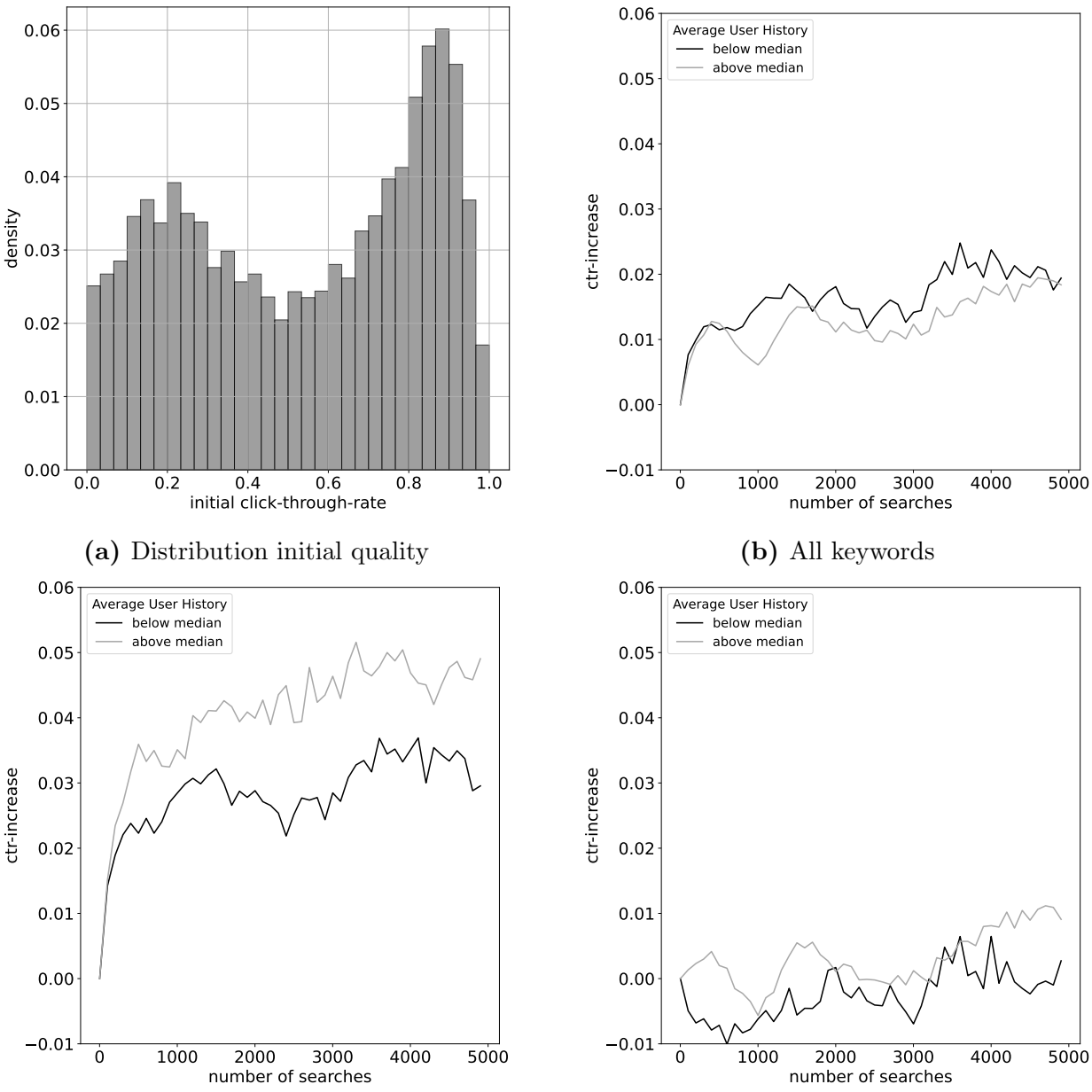


Figure 4: Average Click-Through-Rate

Note: The error bars denote the 95% confidence interval using robust standard errors.



(a) Distribution initial quality

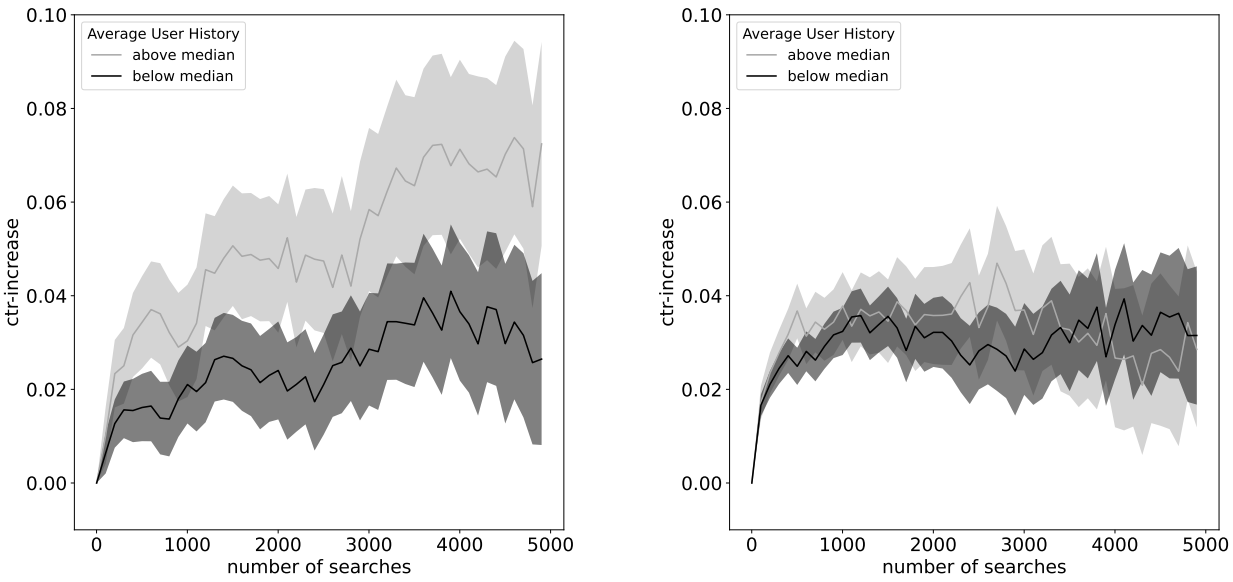
(b) All keywords

(c) Keywords with init. quality below median

(d) Keywords with init. quality above median

Figure 5: Analysis of Quality Changes

Note: Only keywords with at least 200 searches are considered. The upper right panel shows the results unconditional on the initial quality. The lower left and lower right panels show the results for keywords below and above the median initial click-through-rate, respectively.

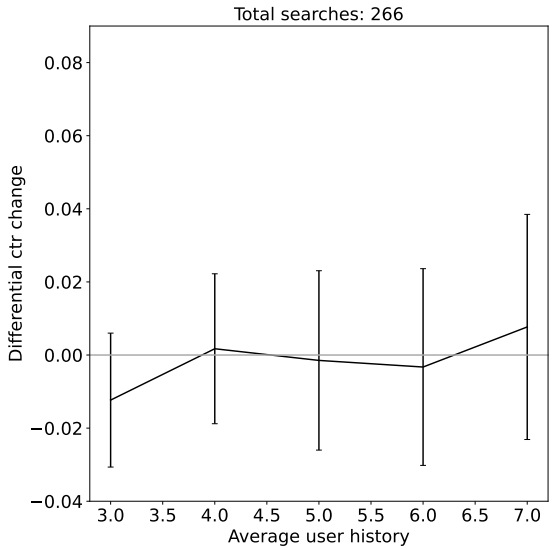


(a) Personalizable keywords

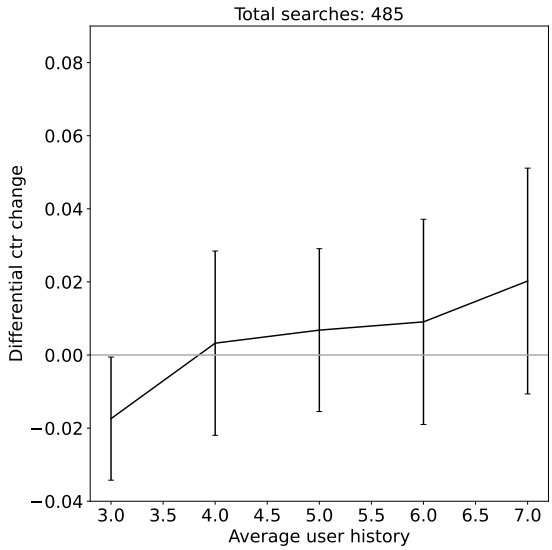
(b) Non-personalizable keywords

Figure 6: Analysis of Quality Changes - Personalizable vs Non-personalizable Keywords

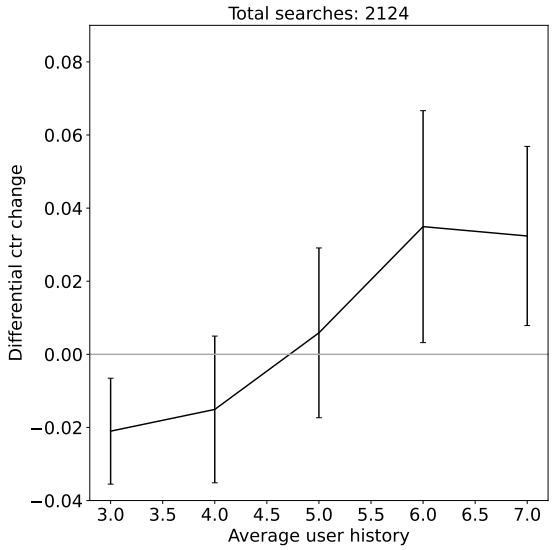
Note: Only keywords with at least 200 searches and an initial click-through-rate below the median are considered. The shaded areas denote the 95% confidence intervals.



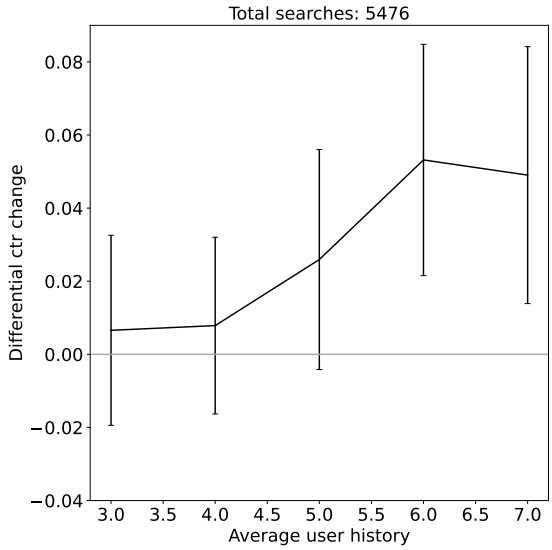
(a) First decile of search quantity



(b) Third decile of search quantity



(c) Seventh decile of search quantity



(d) Ninth decile search quantity

Figure 7: Generalized Random Forest Results

Note: The range of the x-axes are determined by the first and ninth decile of the distribution of the average user history. Each panel stands for a selected decile of the total search quantity. The initial click-through-rate is set to 25% throughout. Error bars denote 95% confidence intervals.

A Appendix

A.1 Proxy-Variable Property of Explanatory Variables

In this Appendix, we provide evidence that most keywords in our sample accumulate searches linearly over time. We also provide a more formal exposition for why this is likely to improve the proxy-variable property of our explanatory variable in the proxy-variable analysis of Subsection 5.1.

Consider the cumulative number of searches of a keyword, $\mathcal{S}$. Note that the cumulative number of searches can always be rewritten as $\mathcal{S} = \overline{S} \times T$, where T denotes the number of periods (for example months) the keyword existed. $\overline{S}$ denotes the average per period popularity (the monthly popularity) of the keyword. In our analysis, we observe one realization of the monthly popularity of a keyword, S.

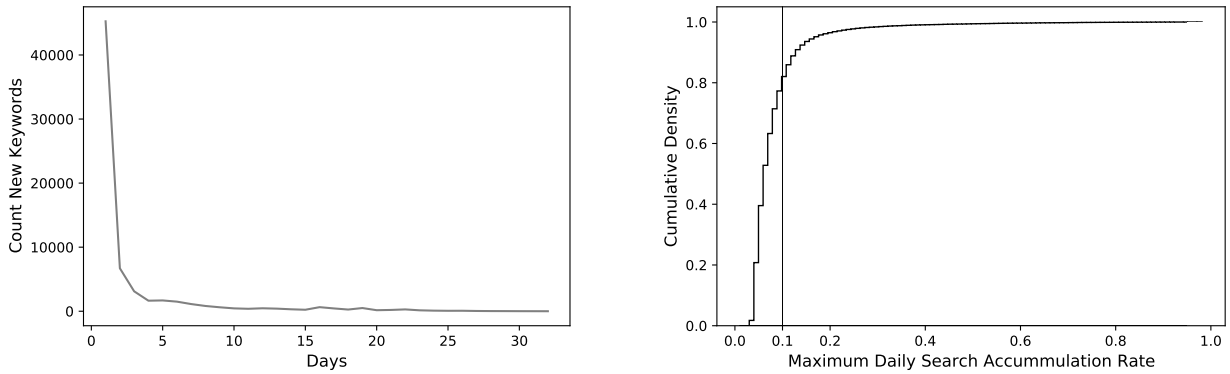
We are interested in the quality of S as a proxy for $\mathcal{S}$. We can assess the quality by analyzing $\text{Var}(\overline{S} \times T|S)$. The smaller the variance of $\mathcal{S} = \overline{S} \times T$ given S, the better the proxy-variable property of S for $\mathcal{S}$. Note that for a population of keywords with steady popularity we have that $\text{Var}(\overline{S}|S)$ is small because observing S is informative about $\overline{S}$ for all keywords in this population. In the extreme case in which the monthly popularity is perfectly constant, we have $\text{Var}(\overline{S}|S) = 0$. It holds that $\text{Var}(\overline{S} \times T|S) = Cov(\overline{S}^2,T^2|S) + \big(\text{Var}(\overline{S}|S) + E(\overline{S}|S)^2\big)\big(\text{Var}(T|S) + E(T|S)^2\big) - \big(Cov(\overline{S},T|S) + \text{E}(T|S)E(\overline{S}|S)\big)^2$. Clearly, reducing $\text{Var}(\overline{S}|S)$ reduces the overall variance.

We can use the above formalization to reason under which conditions the proxy variable preserves the ordinal ranking of the variable of interest, on average. S preserves the ordinal ranking of $\mathcal{S}$, on average, if $\partial \text{E}(\overline{S} \times T|S)/\partial S > 0$. For exposition, consider the case in which the monthly popularity is perfectly steady, i.e. $S = \overline{S}$. Then $\partial \text{E}(\overline{S} \times T|S)/\partial S = \big(\text{E}(T|S)/\partial S \times S\big) + \text{E}(T|S)$. Clearly, the second term is always positive. Thus, under constant monthly popularity, a sufficient condition for preserving the ordinal ranking is that the monthly popularity of keywords is not negatively correlated with their "age".

The average user history depends on the type of users searching a keyword. High intensity types lead to a longer average user history. Thus, the quality of the proxy variable cannot be argued for in the same way. The evolution of the average user history over time is not only determined by the popularity. However, the assumption that a keyword with a steady popularity has also reached a steady state in terms of the type of users searching the keyword appears reasonable. If this is the case, the type of users we observe is informative of the type of users searching the keyword before our sample.

We now proceed by providing evidence that most keywords existed already prior to our sample and that the majority reached a steady popularity level, i.e. accumulate searches linearly with time. To do so, we first analyze how many new keywords appear each day in our sample. A new keyword is simply a keyword that has not been observed previously. Figure A.1a shows the number of new keywords appearing each day. Clearly, most keywords appear within the first days of our sample. This is in line with the notion that the majority of keywords already originated prior to the sample. Otherwise, one would expect a larger share of keywords appearing during the period of our sample.

Figure A.1b shows the cumulative distribution of the maximum daily accumulation rate of keywords. For each keyword with at least 32 searches, we compute the percentage of total



(a) Number of new keywords by day

(b) CDF of max. daily accumulation

Figure A.1: Proxy-Variable Property of Observed Variables

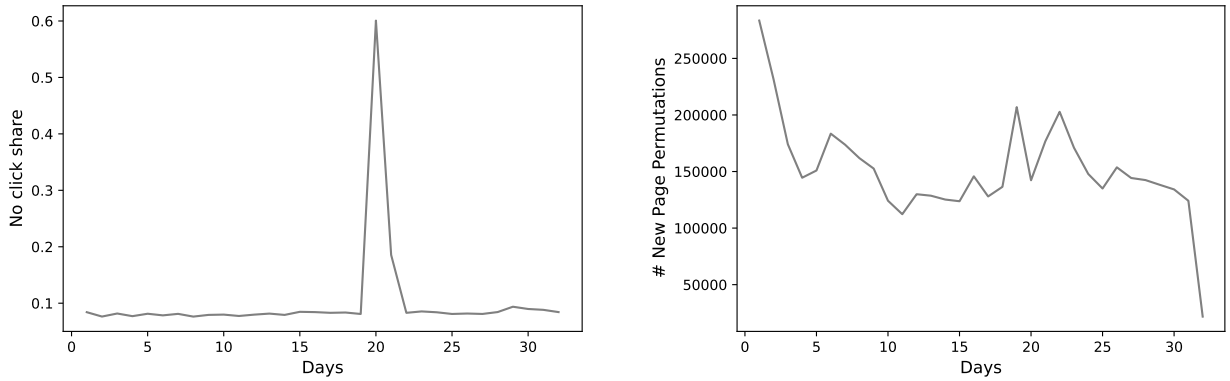
searches a keyword accumulated each day and take the maximum value. For example, a value of 50 percent indicates that a keyword accumulated more than half of the total searches it experienced in the sample in one day. Figure A.1b shows that for more than 80 percent of keywords the maximum accumulation rate is smaller than 10 percent. This is clearly not in line with a strongly oscillating or trending popularity for most keywords and rather suggest that most keywords accumulate searches steadily.

We focus on keywords with at least 32 searches because daily intervals offer an intuitive measure to assess how evenly searches accumulate over time. Over our 32 day sample period, a keyword with perfectly steady daily popularity would accumulate 1/32 searches per day. For keywords with less searches this intuitive criterion fails. Note that, for a keyword with exactly 32 searches, four searches in one day would already constitute more than 10 percent of all total searches. Since keywords with few searches dominate numerically in the sample, it is remarkable that so few keywords exceed the ten percent threshold in Figure A.1b.

A.2 July 20th Anomaly

There is an anomaly in our data which is characterized by a significant drop of the observed click-through-rate on July 20, 2010. To the best of our knowledge, this drop is likely to capture a technical issue related to the testing of a new algorithm in the context of the Yahoo!-Bing merger. Figure A.2a reveals that the day of July the 20th is characterized by an exceptionally high share (60 percent) of searches ending without any recorded user action (i.e. the log records no click).

It is unlikely that the large share of searches without any recorded click can be entirely explained by a change in the content presented to the searchers: Firstly, it is unlikely that searchers would not even try to click on some links, even if these links appear unreasonable. Secondly, and less speculative, our analysis of the content presented to users in Figure A.2b reveals that the number of new results pages was not particularly high on the day of July 20th. This suggests that the *magnitude* of the anomaly might be best explained by a bug in the logging technology, i.e. the technology used to record the actions of users.



(a) Share of searches without click. **(b)** Number of new search result pages per day.

Figure A.2: July 20th Anomaly.

While the *full magnitude* of the anomaly is unlikely to be explained by a change in intrinsic quality, our analysis also reveals that there are two spikes in the amount of new content appearing around the day of the anomaly (July 19th and July 21st). Additionally, we have narrative evidence that Yahoo! was planning to test a new algorithm around the same time. In fact, shortly before July 20, 2010, Yahoo! publicly announced its intention to begin testing Microsoft's search engine online:

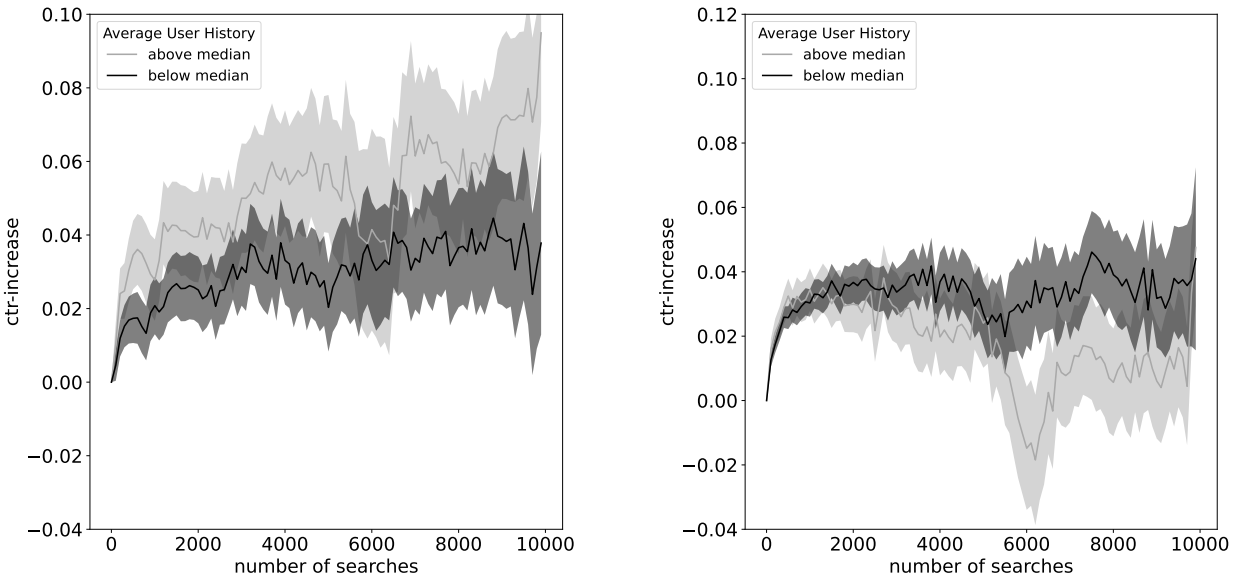
> *"Though much of our testing is already happening offline, this month we'll also test the delivery of organic and paid search results provided by Microsoft on live Yahoo! traffic"* - Yahoo Website, July 15, 2010[27]

Given this publicly available information, it appears cautious to assume that the anomaly could be indicative of the beginning of a longer testing period. To safeguard against the

[27] See https://web.archive.org/web, last accessed on 15 February 2022.

possibility that our results might be affected by a more permanent change, we restrict the analysis presented in the main text to searches observed until July 18th.

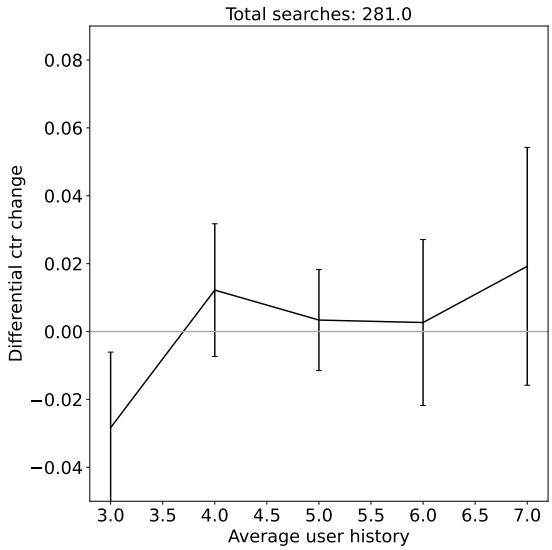
Figures A.3 and A.4 replicate the results from Subsection 5.3 using the full sample period to compute quality changes. We now only exclude observations for the three-day period from July 19th to July 21st instead of dropping all observations after July 18th. For example, when computing the average quality change in the click-through-rate between the first window of 100 searches and subsequent windows of 100 searches shown in Figure A.3, we remove all windows which contain searches overlapping with one day of the three day anomaly. The results in Figure A.3b indicate that, despite excluding these observations, searches after the anomaly might be affected by non-transitory changes, as we observe peculiarities in the data starting at around 6000 searches, which roughly marks the transition from the period before to the period after the anomaly for very popular keywords: Since keywords accumulate searches linearly over time (see Appendix A.1), the day of July 20th typically corresponds to approximately $\frac{19}{32} = 0.59$, i.e. 59 percent of total searches, i.e. almost 6000 searches for the group of keywords with 10000 searches in total. This group numerically dominates the sample of keywords reaching large search quantities.



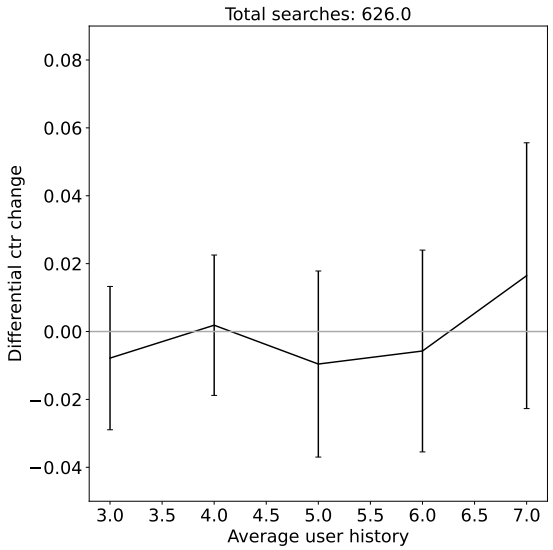
(a) Personalizable keywords **(b)** Non-personalizable keywords

Figure A.3: Analysis of Quality Changes - Personalizable vs Non-personalizable Keywords

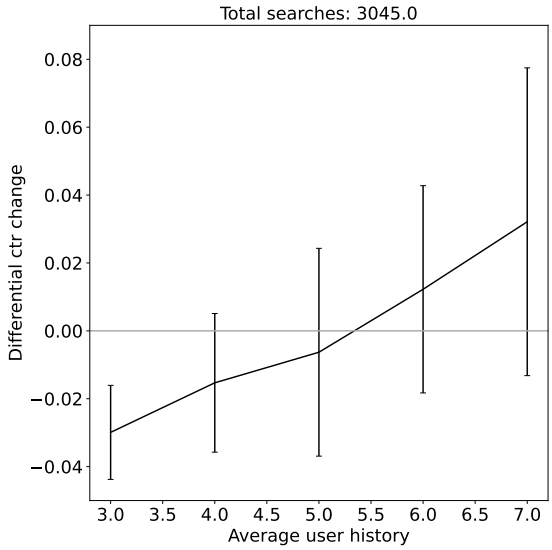
Note: Only keywords with at least 200 searches and an initial click-through-rate below the median are considered. The shaded areas denote the 95% confidence intervals.



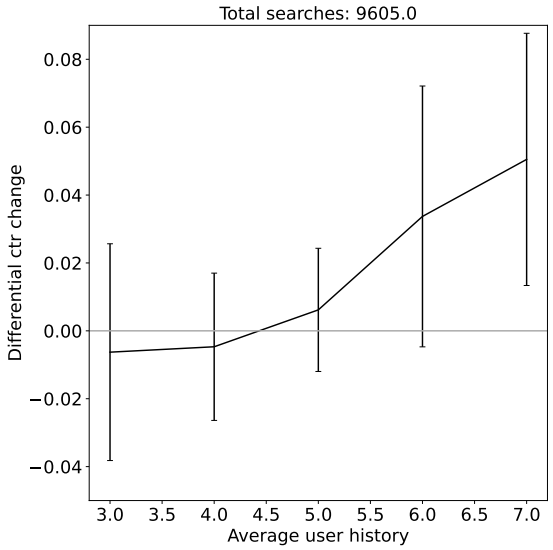
(a) First decile of search quantity



(b) Third decile of search quantity



(c) Seventh decile of search quantity



(d) Ninth decile of search quantity

Figure A.4: Generalized Random Forest Forest Results

Note: The range of the x-axes are determined by the first and ninth decile of the distribution of the average user history. Each panel stands for a selected decile of the total search quantity. The initial click-through-rate is set to 25% throughout. Error bars denote 95% confidence intervals.

A.3 Description of Keyword Classification Method

In this Appendix, we describe the procedure we employ to classify keywords into personalizable and non-personalizable keywords. Our classification method is based on a mixture of two criteria. The first criterion relies on the notion that additional information should lead to different search results. The second criterion relies on the notion that personalization should be reflected in a higher variation of search results displayed.

Criterion I: More information should lead to different search results

The first criterion relies on the notion that, if user-specific information plays a role, changes in search results should be related to the amount of user-specific data we observe for a user. We now describe how we develop a classification method based on this insight.

First, for each keyword, we determine the distribution of user history lengths for all users searching a keyword. We use the user history observed over the entire sample. This allows us to group users in two groups for each keyword: Users with a user history above the upper quartile and users with user histories below the lower quartile. The quartiles are determined based on the user history distribution observed for each keyword. Second, for each keyword and top-ranked search result combination, we compute the number of times this search result was shown to searchers in both groups. Third, for the overlapping set of top-ranked search results shown to both groups of users, we test the null hypothesis that the frequency distribution of search results is the same across both groups using the Chi-Square Test.

We use the distribution obtained for a short average user history as distribution under the null hypothesis. If the null hypothesis is rejected at a five percent level of significance, we classify the respective keyword as potentially "personalizable".[28]

Criterion II: Personalizable keywords should have a higher variance

The criterion described above relies on the assumption that observing systematically different distributions of search results as a function of the length of the user history is a strong indicator for the reliance on user-specific information.

The Chi-Square Test is based on the overlapping set of search results shown for low- and high-intensity users. Note that the Chi-Square Test will never lead to classify a keyword as perosnalizable if this keyword exhibits no variance in search results. This leads us to the second criterion we use for classification: The purely variance-based criterion, which we now describe in more detail.

The variance-based criterion relies on the number of times the top -ranked search result "rotates".[29] More precisely, for each keyword, we compute the median "periodicity" of the

[28]For the Chi-Square Test, it is recommended that the absolute frequency of every single search result should exceed four under the null hypothesis. We therefore drop search results that are shown weakly less than four times under the null hypothesis. For some keywords, this procedure results in only one search result remaining after we drop search results shown less than four times. If this is the case, we test whether the relative frequency of the single search result is the same. To do so, we employ a simple T-Test instead of a Chi-Square Test.

[29]It is important to note that a personalization does not necessarily require a great number of different

top-ranked URL. We define the periodicity as the number of consecutive searches the same search result is shown. The median periodicity therefore captures the typical length of sequences of searches without result variation. For example, a median periodicity of two says that the typical consecutive number of times the same search result was shown is two.

We normalize the periodicity by the number of total searches to make the measure comparable across keywords with different popularity levels: If a keyword has only 100 searches, a median periodicity of 100 corresponds to no change in search results, while for a keyword with 10000 searches, a median periodicity of 100 indicates relatively frequent changes in the top-ranked URL. We call the normalized measure the relative periodicity.

The problem with the relative periodicity measure, when used as the sole criterion, is that there might be keywords with frequent changes in the top-ranked URL that are not personalizable, such as news-related keywords leading to constantly updated newspaper URLs with the most recent news relating to a VIP ("US president" or "California Governor") or events of elongated public interest ("war in Afghanistan" or "US Midterms"). Such keywords might exhibit frequent changes, even without personalization.

Combination of Both Criteria

Figure A.5 shows the histograms of the relative periodicities for keywords for which the null hypothesis of the Chi-Square Test was rejected and the complementary set of keywords. As can be seen, the Chi-Square Test naturally select keywords with a shorter periodicity. However, we also observe a small fraction of keywords with a long periodicity for which the null was rejected.

Figure A.5 also illustrates that there is a substantial fraction of keywords with a short relative periodicity for which the null hypothesis of the Chi-Square Test is not rejected. As explained above, news-related keywords are likely candidates.

We therefore combine both approaches and consider keywords as personalizable only if the null hypothesis of the Chi-Square test has been rejected *and* the relative periodicity is below 0.1 (indicated by the black vertical line in Figure A.5). While the periodicity threshold is arbitrary, it appears desirable to remove keywords with a very large periodicity as it appears implausible that those keywords indeed personalize search results. Based on this procedure, out of 23637 keywords with more than 200 searches, 4496 keywords are classified as personalizable.

We performed robustness checks using either one of the criteria as sole classification method and found qualitatively consistent, albeit less clear results.

search results. To better understand why this is the case, consider the example of the keyword "mouse": In a hypothetical world with only two potential search results, the Wikipedia article about the rodents and the Wikipedia article about the computer hardware, there might still be substantial gains from personalizing search results to animal or computer enthusiasts.

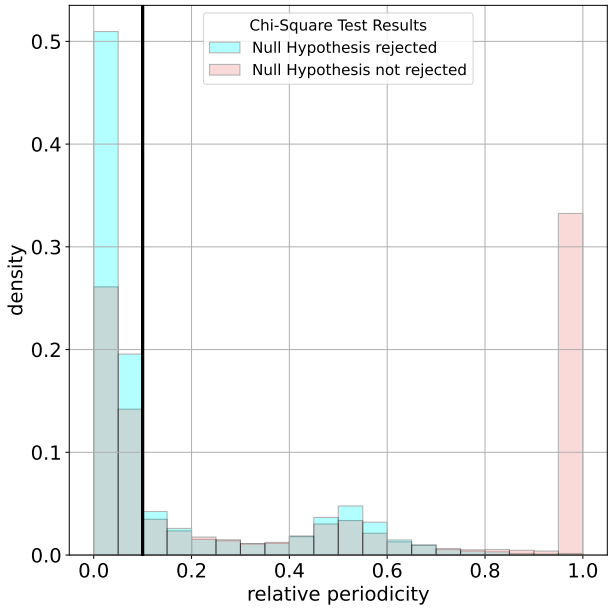


Figure A.5: Relative Periodicities

A.4 Matching Procedure and Covariate Balance

This Appendix describes the matching procedure used to create the matched sample for the generalized random forest estimation and provides details on the covariate balance between the group of personalizable and non-personalizable keywords for both the original and matched sample. For details on the classification procedure, see Appendix A.3.

Matching is performed using greedy one-to-one nearest neighbor matching. For each personalizable keyword, we find the nearest non-personalizable neighbor. The distance metric used is the mahalanobis distance. Matching is performed without replacement. The matched sample consists of 8992 keywords.

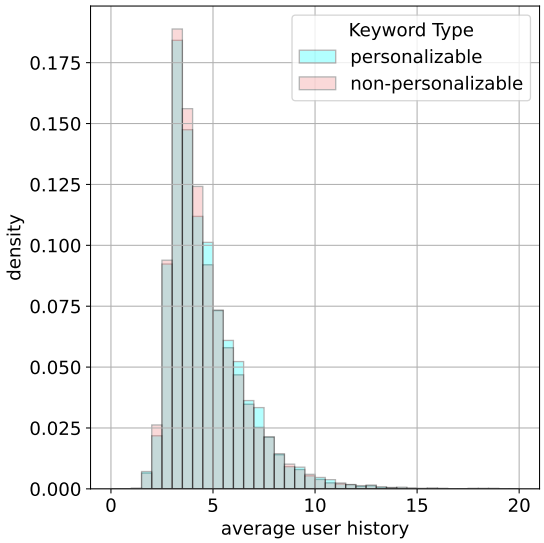
Table A.1 shows the means and the differences in means between both groups of keywords for the key explanatory variables. Figure A.6 shows the corresponding distributions. The matching is successful in that it achieves balance in the covariates (both in terms of means and distributions).

The last row of Figure A.6 shows the estimated propensity scores for the original (Subfigure A.6g) and matched sample (Subfigure A.6h). The propensity scores are automatically obtained when estimating the generalized random forest in *R*. The propensity scores of the original sample are not bounded away from zero. This results in poor overlap and is known to adversely affect the performance of the generlaized random forest. Subfigure A.6h reveals that the matched sample has substantially better overlap.

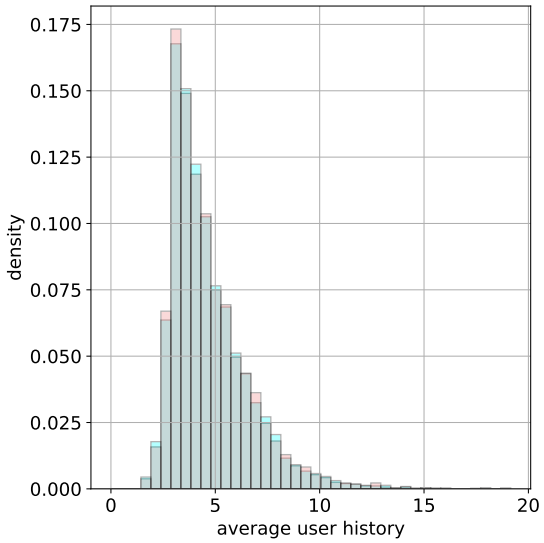
Table A.1: Covariate Balance - Original and Matched Sample

	Original sample			Matched sample		
	pers. $= 0$	pers. $= 1$	Diff.	pers. $= 0$	pers. $= 1$	Diff.
Average user history	4.59	4.69	0.1	4.68	4.69	0.01
			(0.03)			(0.04)
Initial click-through-rate	0.56	0.42	−0.14	0.43	0.42	−0.01
			(0.00)			(0.01)
Number of searches	1730.54	2488.32	757.78	2470.95	2488.32	17.37
			(35.26)			(45.93)

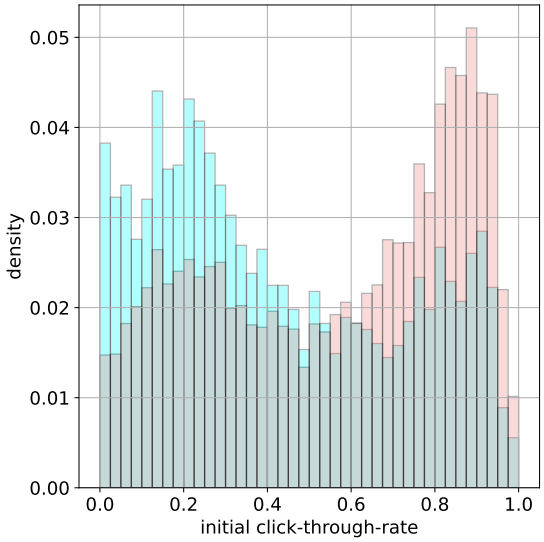
Note: The table shows means and mean differences between non-personalizable (pers = 0) and personalizable (pers = 1) keywords. The values in parantheses denote standard deviations.



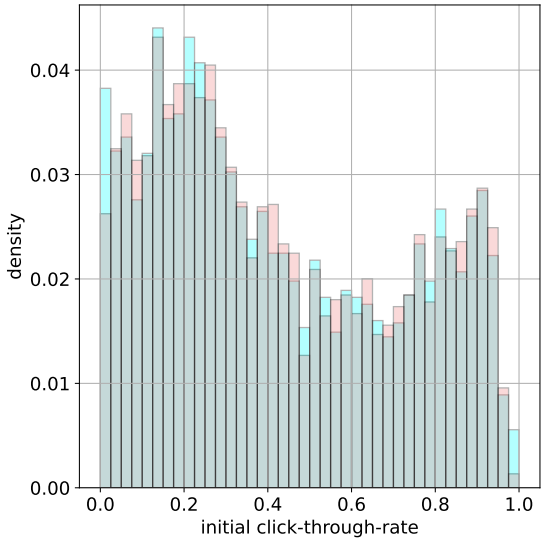
(a) Average user history – unmatched



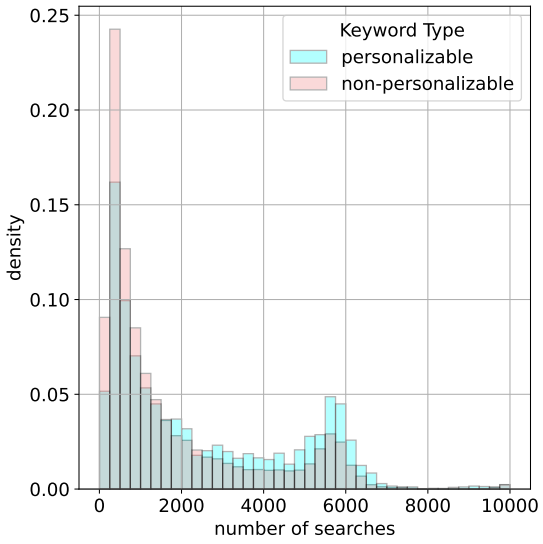
(b) Average user history – matched



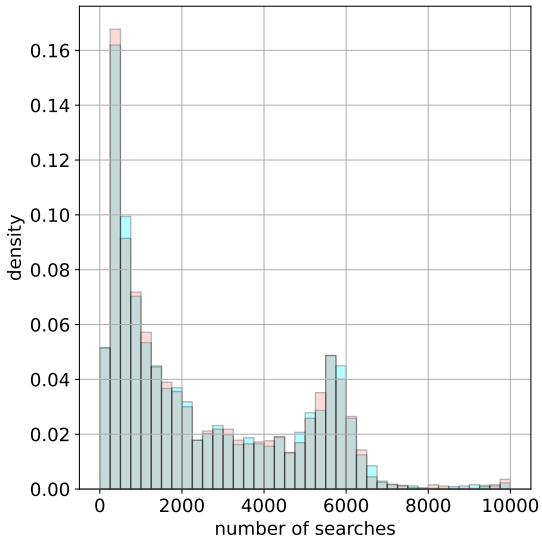
(c) Initial click-through-rate – unmatched



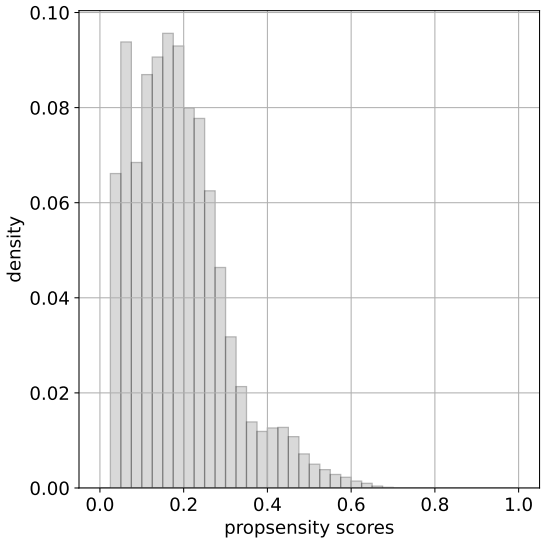
(d) Initial click-through-rate – matched



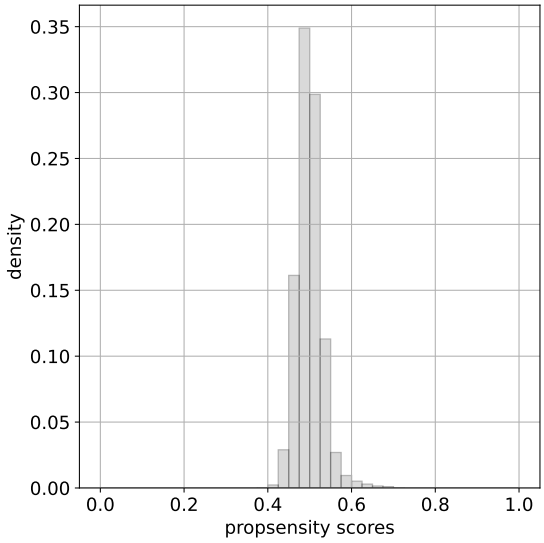
(e) Number of searches – unmatched



(f) Number of searches – matched



(g) Propensity scores – unmatched



(h) Propensity scores – matched

Figure A.6: Covariate Balance in Original and Matched Sample

Note: Each panel shows the distribution of variables for personalizable and non-personalizable keywords. Panels in left column refer to the original sample. Panels in the right column refer to the matched sample. Propensity scores were estimated using random forests.