MIT Sloan School of Management

MIT Sloan School Working Paper 5094-14

Search Engines and Data Retention: Implications for Privacy and Antitrust

Lesley Chiou, Catherine Tucker

Search Engines and Data Retention: Implications for Privacy and Antitrust

Lesley Chiou* and Catherine Tucker†

May 27, 2014

Abstract

This paper investigates whether larger quantities of historical data confer a competitive advantage to firms that offer Internet search. We study how the length of time that search engines retained their server logs affected the apparent accuracy of subsequent searches. Our analysis exploits changes in these policies prompted by the actions of the European Commission. We find little empirical evidence that reducing the length of storage of past search engine searches affected the accuracy of search. Our results suggest that the possession of historical data confers less of a competitive advantage than is sometimes supposed. Our results also suggest that limits on data retention may impose fewer costs in instances where overly long data retention leads to privacy concerns such as an individual's "right to be forgotten."

JEL classification: L86
Keywords: digitization, information, consumer search, network effects, privacy, right to be forgotten

*Economics Department, Occidental College, CA
†MIT Sloan School of Management, MIT, Cambridge, MA and National Bureau of Economic Research.
‡We thank Christopher Hafer, Anton Grutzmacher, and James Murray of Experian Hitwise. We also thank Katherine Eriksson for excellent research assistance. While this research has not received financial assistance, in the past Lesley Chiou has received financial support for other research from the Net Institute and the National Bureau of Economic Research. Catherine Tucker has received financial support for other research from Google, the National Bureau of Economic Research, the National Science Foundation, the Net Institute, and WPP.

1 Introduction

Currently, Internet search is attracting antitrust scrutiny on both sides of the Atlantic (Goldfarb and Tucker, 2011a). In this heavily concentrated market, one firm, Google, dominates search in both Europe and the US. However, it is not clear the extent to which this concentration is entrenched. One potential mechanism for entrenchment is "network effects" in search where historic data on past searches conveys benefits in the provision of accuracy of current searches.

In this paper, we exploit a natural experiment to obtain empirical evidence regarding the benefits that companies may receive from having large quantities of data. Specifically, we use variation in European guidelines surrounding the length of time that search engines can store an individual's data as an exogenous shifter of the amount of data available to a search engine. We then study how the accuracy of search results changes before and after the policy change. We measure the accuracy of search results by whether the customer navigates to a new website or whether the customer had to repeat the search either on that search engine or another search engine.

We find no empirical evidence of a negative effect from the reduction of data retention on the accuracy of search results. Our findings are apparent in the raw data as well as in a regression analysis of panel data with fixed effects to control for changes over time and across search engines. Our regression analysis suggests not only insignificance but also that the likely economic effects of the imprecisely measured coefficients are small.

We believe that absence of a decline in the accuracy of searches suggests little competitive advantage bestowed by longer periods of data retention. Some potential explanations exist for the lack of competitive advantage. First, historic data may be less useful for accurately predicting current news than is sometimes supposed. Given that recent developments in search have highlighted consumers' desire for more current and recent news, large of amounts

of historic data may not be useful for relevancy. Second, the precise algorithms that underly search engines algorithms are shrouded in secrecy. Third, 80% of searches are unique. Of course, we also recognize the possibility that our measure of search accuracy may be too direct to pick up nuances in the precise quality of search results.

Our results are important from a public policy perspective. Increased anti-trust scrutiny on Internet search has focused on whether or not access to large swathes of historical data could lead to market power and incumbent entrenchment. Indeed, such concerns has led theoretical research such as Argenton and Prfer (2011) to propose mechanisms for sharing these logs among search engine providers.

This paper also has implications for privacy and data security regulation. At the moment, much privacy regulation focuses on obtaining informed consent, and less emphasis exists over how long data may be stored after a person's consent has been acquired. However, the length of time of data storage is key for both privacy protection and the security of an individual's data. Successful attempts at de-anonymizing clickstream or search engine log data have relied on providing a history or time series of people's searches or web browsing behavior that did not reveal an identifiable pattern. Our finding of little effect contrasts with other work that has found significant costs from different types of privacy regulation on commercial outcomes (Miller and Tucker, 2009; Goldfarb and Tucker, 2011b, 2012). We recognize that the difference may reflect the importance of data recency and current results to the search engine business model.

It is important to put our results in the context of the new debate in the legal literature on the right to be forgotten (Rosen, 2012). In the European Union in particular, this "right to be forgotten," has been gaining increasing traction as a potential foundation of privacy regulation Bennett (2012)[1]. As pointed out by Korenhof et al. (2014) the timing of data retention plays a part in this debate. Our study focuses on blanket policies by firms towards

[1]See also "Europe's 'Right to be Forgotten' Clashes with U.S. Right to Know," Forbes, May 16, 2014.

data retention policies and finds little observable effects on search accuracy as measured by the need to repeat searches. However, we do want to highlight that the kind of policies studied in this paper are very different from the recent cases concerning the right to be forgotten in the European Union such as the ECJ case where a Spanish man requested to have details of his foreclosure deleted from Google which have focused on the individual rather than blanket data retention policies.[2]

2 Background and Institutional Setting

2.1 Changes in Data Retention Policies

Table 1 summarizes the variation in data-retention policies that we use in our study. The first two changes in search data retention that we study were prompted by pressure from the European Commission's data protection advisory group, the Article 29 Working Party. In April 2008, the group recommended that search engines reduce the time they retained their data logs.

The first search engine to respond to this challenge was Yahoo!. Yahoo's chief privacy officer Ann Toth declared that its decision to anonymize its user personal information after 90 days "set a new industry standard for protecting consumer privacy. This policy represents Yahoo!'s assessment of the minimum amount of time we need to retain data in order to respond to the needs of our business while deepening our trusted relationship with users." [3]

In January 2010, the chief privacy strategist at Microsoft announced that Microsoft would delete the Internet protocol address associated with search queries at six months rather than 18 months.[4]

In the last example, we study a change in Yahoo! policy where they increased the amount

[2]Google Spain SL, Google Inc. v Agencia Espanola de Proteccion de Datos. Accessed at http://curia.europa.eu/jcms/upload/docs/application/pdf/2014-05/cp140070en.pdf.

[3]http://www.ft.com/cms/s/0/f6776768-cc6b-11dd-9c43-000077b07658.html#axzz1JyhQBZ2u

[4]http://blogs.technet.com/b/microsoft_on_the_issues/archive/2010/01/19/microsoft-advances-search-privacy-with-bing.aspx

Table 1: Timeline of policy changes

Date	Search Engine	Change in Storage Policy
December 2008	Yahoo!	13 to 3 months
January 2010	Bing	18 to 6 months
April 2011	Yahoo!	3 to 18 months

of data they kept. Yahoo claimed that "going back" to 18 months was required in order to "keep up" in the competitive environment against other search engines. Yahoo! offers highly personalized services that include shopping recommendation as well as customized news pages and search tools that "can anticipate what users are looking for." According to Anne Toth, chief Trust officer at Yahoo!, "To pick out patterns for such personalization, Yahoo needs to analyze a larger set of data on user behavior." Since this change was prompted by internal competitive motivations rather than exogenous changes in the strictness of EU enforcement of the data directive, we use this policy as a robustness check to our main analyses.[5]

It is also important to highlight that not all de-identification and anonymization procedures were the same. Figure 1 is a (likely slanted) representation of Search Engine policies as of February 2009 by Microsoft. The figure makes a distinction between de-identification (where the ability to match search queries with other identifying information is removed) and anonymization which involves the removal of IP addresses. In general the policies we studied were targeted towards anonymization. The policies come in the wake of the release of the AOL search engine log query data for 658,000 users within the US that demonstrated how a series of search engines queries over time could reveal an individual's identity. For example, reporters were able to identify Thelma Arnold, a 62-year-old widow who lives in Lilburn, Georgia as AOL searcher "No. 4417749" from the content of her searches.[6]

[5]For more details see http://www.ypolicyblog.com/policyblog/2011/04/15/updating-our-log-file-data-retention-policy-to-put-data-to-work-for-consumers/

[6]http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all&_r=0

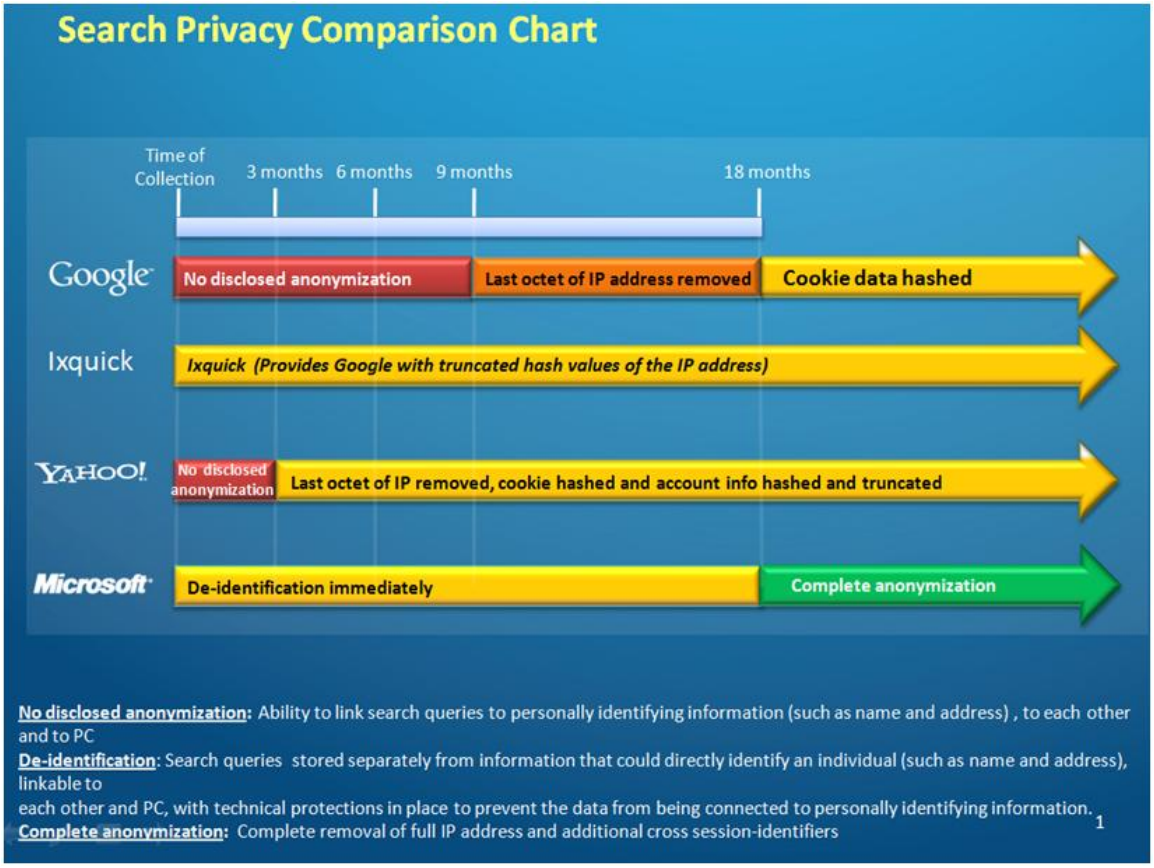


Figure 1: Microsoft comparison of Search Data Retention Policies of Major Search Engines in February 2009

Source: `http://blogs.technet.com/b/microsoft_on_the_issues/archive/2009/02/10/comparing-search-data-retention-policies-of-major-search-engines-before-the-eu.aspx`

3 Empirical Analysis

3.1 Search Data

Our analysis relies on data from Experian Hitwise. Hitwise assembles aggregate data using the website logs from Internet Service Providers. The information is combined with data from opt-in panels to create a geographically diverse sample with usage data from 25 million people worldwide. For further details, Chiou and Tucker (2012) also use this data. Since we study policy changes that affect search engines in Europe, we use data from Hitwise on the search behavior of UK residents.

We are interested in whether a change in policies of data retention affected the accuracy of search. As a measure of accuracy, we examine whether a consumer repeats a search or navigates to a new site. Hitwise reports the top 20 sites that users navigate to after visiting a particular site. We observe the fraction of outgoing traffic to each of these "downstream" sites from each of the major search engines during a given week.

We restrict our sample to outgoing traffic from the three major search engines: Yahoo!, Google, and Bing. We identify which downstream sites are search sites by examining sites that contain the domain of any major search engine. Our category of search sites excludes mail, book, or wiki sites, which serve a different purpose than general search. We collect data for the two months before and after each policy change in our sample.

Table 2 reports the summary statistics for the downstream search sites in our sample. Each observation in our final sample represents a search engine-website-week combination. For instance, we can observe the percent of outgoing traffic from Yahoo! Search that navigated to a particular search site during the first week of February 2009. The average search site received 0.85 percent of all outgoing clicks from a search engine.

Table 2: Summary statistics

	Mean	Std Dev	Min	Max	Observations
% clicks	0.85	1.29	0	9.08	2882
Google	0.31	0.46	0	1	2882
Yahoo!	0.51	0.50	0	1	2882
Bing	0.18	0.39	0	1	2882
Observations	2882				

Notes: We observe the fraction of traffic to each "downstream" search website from a major search engine. Each observation in our final sample represents a search engine-website-week combination.

3.2 Graphical and Regression Analysis

As a preliminary analysis, we explore the change in traffic to search sites before and after each major policy change. Figure 2 summarizes the fraction of traffic to search engines among the top 20 downstream sites from Bing and other search engines. The pre- and post-periods refer to the time before and after the Bing's policy change from 18 to 6 months of data retention. As seen in the figure, traffic to search sites remained relatively constant over this period of time.

In Figure 3, we summarize the fraction of traffic to all downstream search sites before and after Yahoo's policy change from 13 to 3 months of data retention. Total traffic to search sites from Yahoo! remained relatively unchanged over this period compared to traffic from other search engines.

Figure 2: Bing January 19, 2010: 18 to 6 months

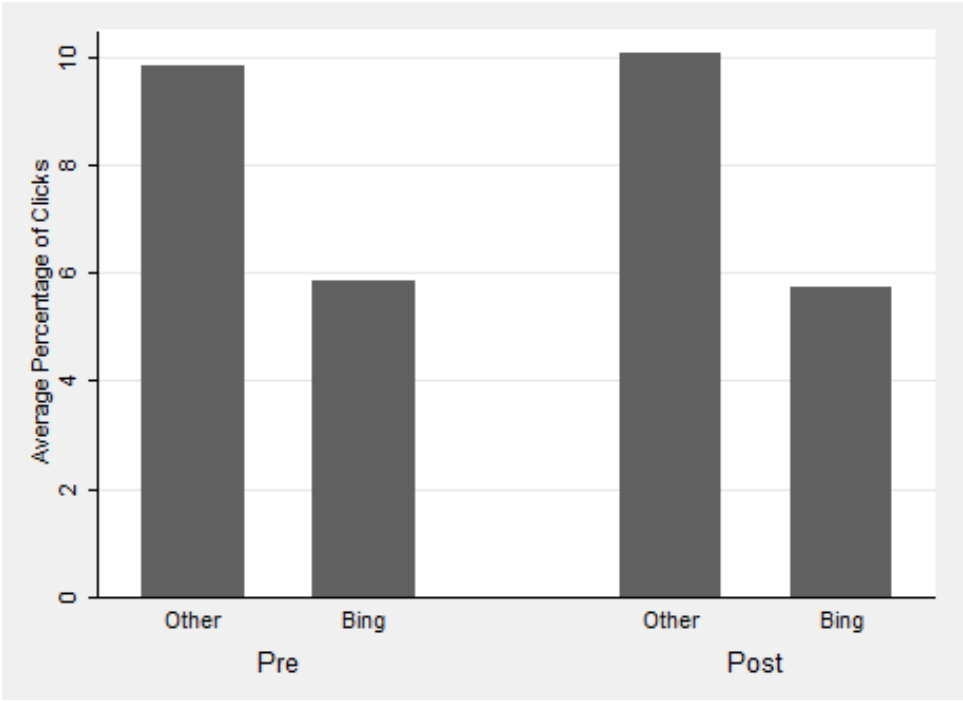
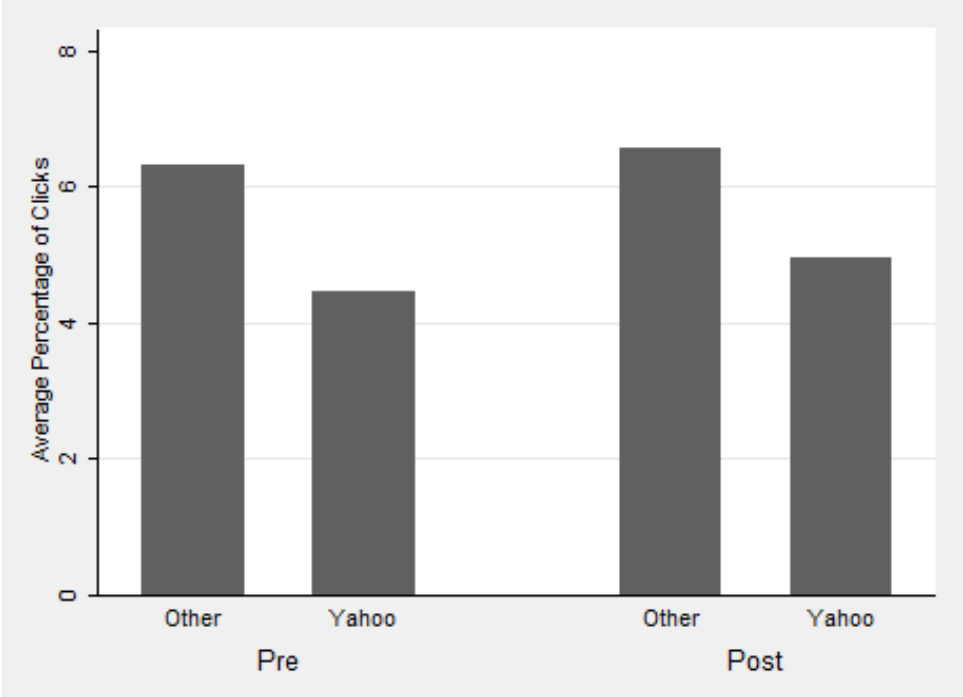


Figure 3: Yahoo December 17, 2008: 13 to 3 months



The figures suggest that changes in data retention policies did not shift downstream traffic from search engines. To formalize the analysis, we run difference-in-differences regressions at the website level for downstream traffic to the top 20 firms for each of the policy changes in our sample. For instance, to analyze Bing's policy change, we estimate the percentage of visits to website i after visiting search engine j in week t:

$$\%visits_{ijt} = \beta_0 + \beta_1 Post_t \times Bing_j + \delta_j + \alpha_i + week_t + \epsilon_{ijt}$$

where *delta* is fixed effect for the originating search engine j, and $Post$ is an indicator variable equal to 1 for the weeks of Bing's change in storage policy. The controls α are downstream-website fixed effects. The vector $week_t$ contains weekly fixed effects to capture variation in the volume and interest of searches in that week. The coefficient β_1 on the interaction term $Post \times Bing$ measures the effect of change in Bing's storage policy on subsequent visits to search sites with the corresponding change in search sites from traffic originating on Yahoo! or Google as a control. We estimate this specification using ordinary least squares and cluster our standard errors at the website level to avoid the downward bias reported by Bertrand et al. (2004).

We report our results in Table 3 for the specification as described by equation (1). We run a similar regression analyzing the effect of Yahoo!'s policy change, and we report those results in Table 4. Both tables indicate that the change in storage policy did not have an effect on downstream visits to search sites. The estimated effect is small and statistically insignificant. To rule out possible delays in implementation, we run our regressions using varying windows of 2, 4, and 6 months.

Table 3: Bing January 19, 2010: 18 to 6 months

	(1)	(2)	(3)
	2 months	4 months	6 months
Post × Bing	-0.0516	-0.0373	-0.0463
	(0.0405)	(0.0978)	(0.140)
Website Fixed Effects	Yes	Yes	Yes
Search Engine Fixed Effects	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes
Observations	464	928	1392
R-Squared	0.952	0.833	0.790

Notes: Robust standard errors clustered at website level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is the percentage of visits.

Table 4: Yahoo December 17, 2008: 13 to 3 months

	(1)	(2)	(3)
	2 months	4 months	6 months
Post × Yahoo	-0.0148	-0.123	-0.173
	(0.122)	(0.195)	(0.229)
Website Fixed Effects	Yes	Yes	Yes
Search Engine Fixed Effects	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes
Observations	210	322	434
R-Squared	0.948	0.904	0.885

Notes: Robust standard errors clustered at website level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is the percentage of visits.

Table 5: Yahoo April 20, 2011: 3 to 18 months

	(1) 2 months	(2) 4 months	(3) 6 months
Post × Yahoo	0.0133	0.0648	0.0687
	(0.121)	(0.110)	(0.104)
Website Fixed Effects	Yes	Yes	Yes
Search Engine Fixed Effects	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes
Observations	352	704	1056
R-Squared	0.910	0.928	0.933

Notes: Robust standard errors clustered at website level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is the percentage of visits.

3.3 Robustness Check

As a robustness check, we examine a third policy change by Yahoo!, which lengthened the data retention period from 3 to 18 months. The policy change contrasts with the two policy changes in the prior section, which decreased the length of data retention. Reassuringly, we find that our results are also statistically insignificant.

4 Conclusion

This paper investigates whether larger quantities of historical data confers a competitive advantage on firms that offer Internet search. We study how the length of time that search engines retained their server logs affected the apparent accuracy of subsequent searches. Our analysis exploits changes in these policies prompted by the actions of the European Commission. We find little empirical evidence that reducing the length of storage of past search engine searches affected the accuracy of search. Our results suggest that the possession of historical data confers less of a competitive advantage than is sometimes supposed. Our results also suggest that limits to data retention provoked by privacy concerns may impose fewer costs if directed at limits on the recency of data (e.g, "right to be forgotten" policies).

Some limitations of this research exist. The first is that it is not clear that search

engine responsiveness to a search query is the only area where consumer might benefit from a search engine retaining data. Other benefits may include testing new algorithms or fraud prevention. The second is that the policy changes we study occurred in Bing and Yahoo!. Unsurprisingly, these two search engines lacked the market share of Google and were experimenting with differentiating themselves via user privacy in order to try and regain market share. Consequently, we study the effects of a reduction in data retention for firms that were not the market leader. The third limitation is that we do not know whether longer term effects exist of the change in retention policies. Our data is truncated partly because Yahoo! reversed its previous data retention policy.

Notwithstanding these limitations, we believe that our study is a useful first step in measuring the effect of data retention policies on consumer behavior.

References

Argenton, C. and J. Prfer (2011). Search Engine Competition with Network Externalities. Discussion Paper 2011-024, Tilburg University, Tilburg Law and Economic Center.

Bennett, S. C. (2012). Right to be forgotten: Reconciling eu and us perspectives, the. *Berkeley J. Int'l L. 30*, 161.

Bertrand, M., E. Duflo, and S. Mullainathan (2004, February). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Chiou, L. and C. Tucker (2012). How does the use of trademarks by third-party sellers affect online search? *Marketing Science 31*(5), 819–837.

Goldfarb, A. and C. Tucker (2011a). Substitution between offline and online advertising markets. *Journal of Competition Law & Economics 7*(1), 37–44.

Goldfarb, A. and C. Tucker (2012). Privacy and innovation. In *Innovation Policy and the Economy, Volume 12*, NBER Chapters. National Bureau of Economic Research, Inc.

Goldfarb, A. and C. E. Tucker (2011b). Privacy regulation and online advertising. *Management Science 57*(1), 57–71.

Korenhof, P., J. Ausloos, I. Szekely, M. L. Ambrose, G. Sartor, and R. E. Leenes (2014). Timing the right to be forgotten: A study into'time'as a factor in deciding about retention or erasure of data. *Available at SSRN 2436436*.

Miller, A. R. and C. Tucker (2009, July). Privacy protection and technology adoption: The case of electronic medical records. *Management Science 55*(7), 1077–1093.

Rosen, J. (2012). The right to be forgotten. *Stanford law review online 64*, 88.