NBER WORKING PAPER SERIES

SEARCH ENGINES AND DATA RETENTION:
IMPLICATIONS FOR PRIVACY AND ANTITRUST

Lesley Chiou
Catherine Tucker

Working Paper 23815
http://www.nber.org/papers/w23815

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2017

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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Search Engines and Data Retention: Implications for Privacy and Antitrust
Lesley Chiou and Catherine Tucker
NBER Working Paper No. 23815
September 2017
JEL No. K21,K24,K40

ABSTRACT

This paper investigates whether larger quantities of historical data affect a firm's ability to maintain market share in Internet search. We study whether 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 policymakers. 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 an advantage in market share 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."

Lesley Chiou
Occidental College
1600 Campus Road
Los Angeles, CA 90041
lchiou@oxy.edu

Catherine Tucker
MIT Sloan School of Management
100 Main Street, E62-533
Cambridge, MA 02142
and NBER
cetucker@mit.edu

I. Introduction

Currently, Internet search attracts legal scrutiny on both sides of the Atlantic (Goldfarb and Tucker, 2011a). In this heavily concentrated market, one firm, Google, accounts for 70% of the search market in the U.S. and over 90% of the search market in the European Union.[1] Public and legal controversy surround why and how such dominance in the market may arise.

One argument presented in the policy debate is that the ability of search engines to store historical data on its users' searches may confer long-term advantages. These advantages subsequently allow a dominant search engine to maintain its market share in the long-term. This practice of "data retention" has been quite controversial. Proponents indicate that the storage of data is necessary to provide high quality searches to users in the future. Critics allege that any benefits from such "network effects" in search are minimal and are outweighed by a loss in privacy and data security and accompanied by an increase in antitrust concerns.

This antitrust debate reflects how data retention is deeply intertwined with legal developments in privacy and data security. 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.

Despite the policy debate and interest surrounding search engines and data retention, no empirical work exists to date on the effects of data retention on the accuracy or quality of search results. When establishing the legal framework for data retention, policymakers must

[1] Pouros (2010) reports Google's market share for the five most populous countries in the European Union: United Kingdom (93%), France (96%), Germany (97%), Spain (97%), and Italy (97%). Population measures were obtained from nationsonline.org, and the list of countries within the European Union is obtained from the official European Union website.

weigh the benefits and costs of data retention to firms, private citizens, and society, so it is important to establish first whether and how much benefit exists from the practice of data retention.

We report on the results of our empirical study to measure the benefits that companies may receive from having large quantities of data. Specifically, we use variation in 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.[2] 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 long-term advantage in market share bestowed by longer periods of data retention. Some potential explanations exist for the lack of an 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, a substantial fraction of searches are unique: 20% of searches that Google receives each day are searches that Google has not received in the last 90 days (AdWords, 2008). Of course, we also recognize

[2] The term "exogenous" shifter refers to how differences in the length of data retention policies are independent of the outcome of the policy.

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 have implications for 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)[3]. As pointed out by Korenhof et al. (2014) the timing of data retention plays a part in this debate as longer periods of data retention make it difficult for digitally recorded actions to be forgotten. As US policymakers, companies, and consumers keep an eye towards developments in the EU, concerns exist over whether legal actions abroad could "take over the American Internet, too" (Dewey, 2015).

Part II provides the background for this debate, including context on the existing regulatory landscape, controversies over search data, and the changes in data retention policies that we study. Part III describes our study design and methodology and presents our empirical results. Part V discusses our results and their implications. Finally, Part VI concludes with recommendations for future study.

II. Background and Institutional Setting

A. Existing Regulatory Landscape

Firms' policies on data retention are deeply intertwined with broad legal and policy concerns over privacy, security, and antitrust. Privacy laws encompass any policy or legislation that governs the use and storage of personal information about individuals whether by the government, public, or private entities. As Hetcher (2001) points out, the Internet can often lead to a "threat to personal privacy" due to the "ever-expanding flow of personal data online." This notion of privacy and security of personal data has become one of the more significant public policy concerns generated by the Internet, leading to "legal and regulatory

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

challenges" (Salbu, 1998).

One challenge faced by the US legal system is that currently most privacy laws at the federal level predate the technologies, such as the Internet, that "raise privacy issues" (Salbu, 2014). In recent years, innovations such as behavioral advertising, location-based services, social media, mobile apps, and mobile payments lead to heated debates over an individual's privacy and security. The issue is pressing among lawmakers, as the GAO prepared a report in conjunction with the inquiry by Senator Rockefeller over data collection for marketing purposes.[4] According to Salbu (2014), the report suggests that the "US privacy debate will increasingly look to international standards and privacy concepts." For instance, the report cites the Fair Information Practice Principles as the de facto international standard.

Consequently, the need for understanding the effects of data retention on search quality is a crucial component for the debate. Given that most innovations and regulations occur in the EU, we study here the effects of changes in those policies abroad and their implications for the US Internet.

Our study is related to a privacy concern that began abroad and quickly spread to US policy debate: the right to be forgotten. The right to be forgotten "soared into public view" internationally recently when the European Court of Justice "ordered Google to grant a Spanish man's request to delete search results that linked to 1998 news stories about the man's unpaid debts" (Roberts, 2015).[5] While at present no formal right to make requests to delete data from the Internet exist in the US, proponents of privacy laws argue that such a right to be forgotten exists in the US through privacy torts and credit reporting rules.

As a result, companies are often left to determine their own policies for the storage and use of data. Differences in policies across companies may reflect external pressure such as court rulings and public sentiment. In our empirical study below, we will use variation in

[4] United States Government Accountability Office, "Information Resellers: Consumer Privacy Framework Needs to Reflect Changes in Technology and the Marketplace," December 18, 2013.

[5] See *Google Spain SL, Google Inc. v. Agencia Espanola de Proteccion de Datos.*

data-retention policies from public pressure by the European Commission.

B. 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 Trust 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." [6]

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.[7]

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

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

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

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

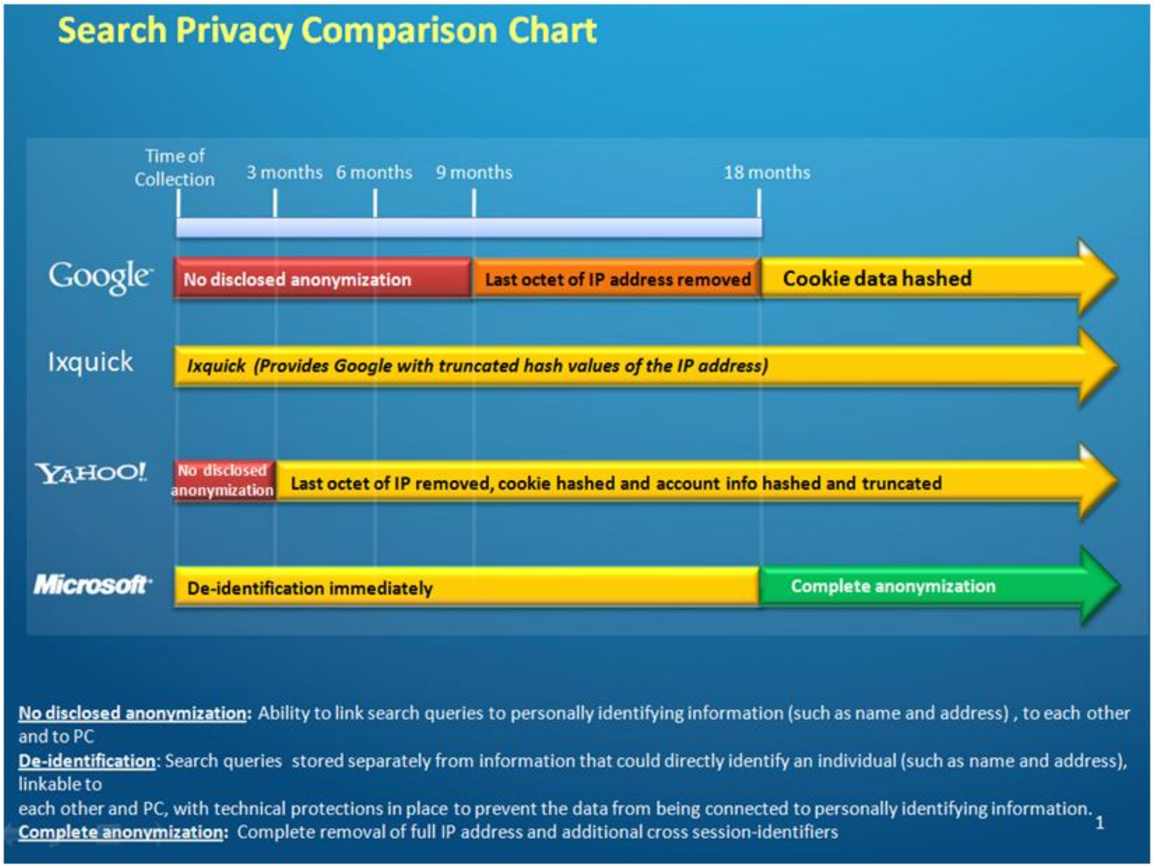
highly personalized services that include shopping recommendation as well as customized news pages and search tools that "can anticipate what users are looking for." According to Anne Toth, Chief Trust Officer at Yahoo!, "To pick out patterns for such personalization, Yahoo needs to analyze a larger 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.[8]

In sum, our study focuses on changes in the data retention policies. We observe changes in the length of data retention for Yahoo! and Bing. Since Google did not change its data retention policy, we do not observe changes in Google's policy.

It is also important to highlight that not all de-identification and anonymization procedures were the same. Figure 1 is a 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.[9]

[8] 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/

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



Search Privacy Comparison Chart

No disclosed anonymization: Ability to link search queries to personally identifying information (such as name and address) , to each other and to PC

De-identification: Search queries stored separately from information that could directly identify an individual (such as name and address), linkable to each other and PC, with technical protections in place to prevent the data from being connected to personally identifying information.

Complete anonymization: Complete removal of full IP address and additional cross session-identifiers

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*

III. Empirical Analysis

A. Study Design and Methodology

Our study design relies on natural experiments to exploit changes in the data-retention policies of major search engines. A natural experiment is a situation in which entities are randomly exposed to a treatment or control policy.[10] In our study, we compare companies with different policies of the length of data retention due to external pressure from policy-makers. The treatment group here is the search engine with the change in the length of data retention, and the control group consists of search engines with no change in data-retention policies. The idea is that the control group will allow us to control for other seasonal patterns in users' search behavior that are unrelated to the change in data retention. In this way, we will not attribute spurious factors to the change in data retention. In other words, the control group describes the counterfactual of how we would expect the treatment group to behave in the absence of the policy change.

Given recent changes in data-retention policies at major search engines, this methodology provides us with several experiments that we study. We describe the search data that we use below, and then we report in the raw data as well as regression analysis of the policy changes. Our regression analysis models the outcome variable (a measure of the quality of search) as a function of other explanatory variables.

B. 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.[11] Since we study policy changes that affect search engines in Europe, we

[10] See The New Palgrave Dictionary of Economics, "natural experiments and quasi-natural experiments."

[11] For further details, Chiou and Tucker (2012) also use this data.

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.

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.

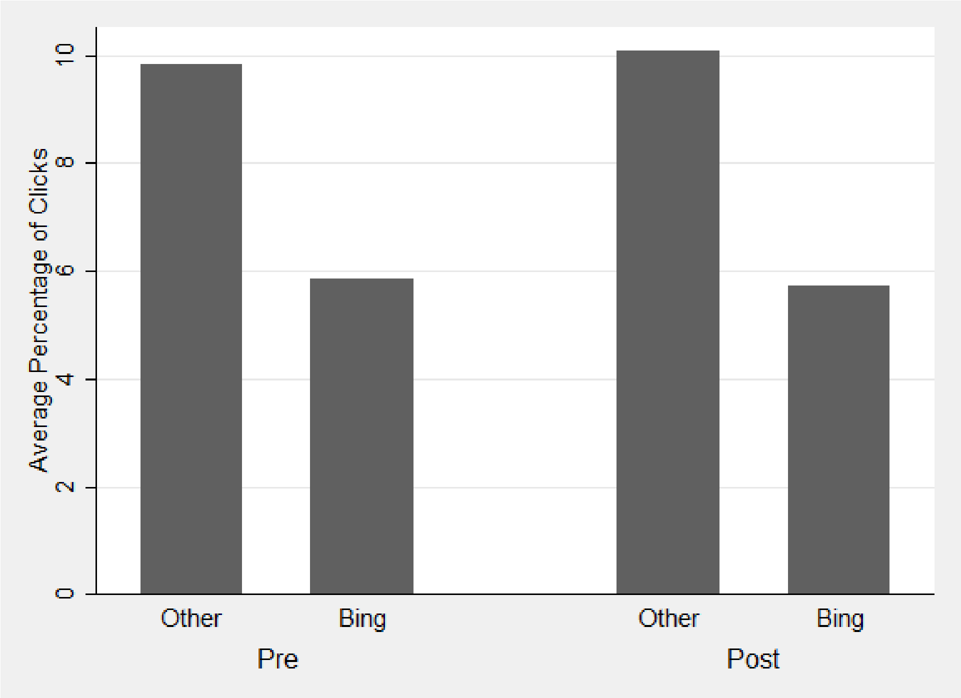
Table 2 reports the summary statistics for the downstream search sites in our sample. Each observation in our final sample represents a search engine-website-week combination. For instance, we can observe the percent of outgoing traffic from Yahoo! Search that navigated to a particular search site during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

C. 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.

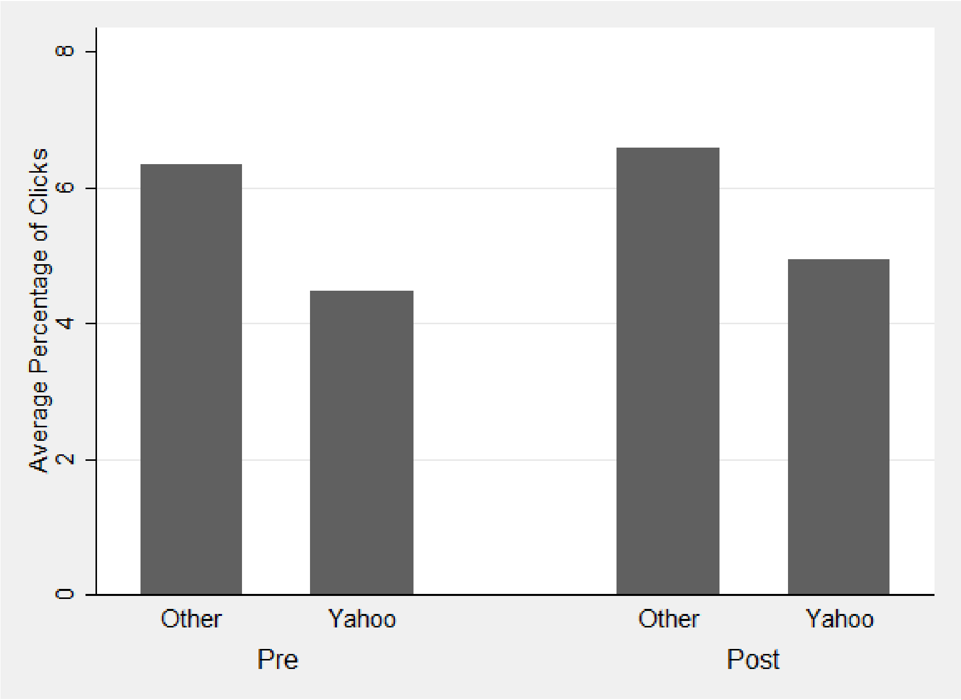
In Figure 3, we summarize the fraction of traffic to all downstream search sites before and after Yahoo's policy change from 13 to 3 months of data retention. Total traffic to search sites from Yahoo! remained relatively unchanged over this period compared to traffic from other search engines.

Figure 2: Downstream search sites visited after Bing and other search engines before and after Bing reduced its length of data retention



Note: This figure shows the average percentage of visits to "downstream" search websites after users visited Bing and other search engines (Yahoo! and Google) before and after Bing reduced its length of data retention from 18 to 6 months in Juanary 19, 2010.

Figure 3: Downstream search sites visited after Yahoo! and other search engines before and after Yahoo! reduced its length of data retention



Note: This figure shows the average percentage of visits to "downstream" search websites after users visited Yahoo! and other search engines (Bing and Google) before and after Yahoo! reduced the length of its data retention from 13 to 3 months in December 17, 2008.

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:

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

where δ 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, which allow each website to have a specific intercept in the regression line. The controls ρ_t are weekly fixed effects to allow each week to have a specific intercept, since variation in the volume and interest of searches may occur across weeks. 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).

To summarize, the regression model describes the relationship between the dependent or outcome variable and a group of explanatory variables. Our coefficient of interest β_1 measures the extent to which storage policy may increase or decrease subsequent visits to search sites. If the coefficient is positive, this suggests that reducing the length of data retention increased the number of repeat searches on the search engine, i.e., the quality of search results decreased. If the coefficient is negative, this suggests that reducing the length of data retention decreased the number of repeat searches on the search engine, i.e., the quality of search results increased.

Table 3: Downstream traffic to search websites before and after Bing's reduction of the length of its data retention from 18 to 6 months in January 2010

	(1) 2 months	(2) 4 months	(3) 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 to search websites. The

We test for whether a positive or negative coefficient is statistically significant. As described in Schwartz and Seaman (2013), "statistical significance is the probability that an observed relationship is not due to chance."[12] If a coefficient is not statistically significant, this means that we cannot reject the hypothesis that the coefficient is equal to zero, i.e., the policy had no effect on the outcome variable.

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.

[12] See Getting Started with Statistics Concepts. "A p-value of less than 0.05 is usually considered statistically significant....("When a result has less than a 5 percent change of having been observed but is observed anyways, it is said to be statistically significant.") A 5% probability is equal to a p-value of 0.05 or less. Results with a p-value of less than 0.01 are considered highly statistically significant...(a 1% chance "represents a 'higher' level of significance because it indicates a less probable outcome and hence a more rigorous statistical test."

Table 4: Downstream traffic to search websites before and after Yahoo!'s reduction of the length of its data retention from 13 to 3 months in December 2008

	(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 to search websites.

Table 5: Downstream traffic to search websites before and after Yahoo!'s increase in the length of its data retention from 3 to 18 months in April 2011

	(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 to search websites.

D. 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.

IV. Discussion and Policy Implications

Our findings suggest that long periods of data storage do not confer advantages in search quality, which is an often-cited benefit of data retention by companies. Of course, a few caveats exist. First, our study focuses on blanket policies by firms towards 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 which have focused on the individual rather than blanket data retention policies.[13]

[13] For instance, in an ECJ case, a Spanish man requested to have details of his foreclosure deleted from Google. Google Spain SL, Google Inc. v Agencia Espanola de Proteccion de Datos. Accessed at

Furthermore, our finding of little effect of longer periods of data retention 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.

Our findings also suggest important policy implications. Unlike the EU, the US does not have a "single overarching privacy law."[14] If long periods of data retention do not generate higher quality of searches, this suggests that the costs of privacy laws for users and companies may be lower than otherwise presumed. The debate thus far has centered on whether more privacy is worth the cost. Our results suggest that the costs of privacy may be lower than currently perceived.

Privacy concepts differ between the US and other legal regimes in the EU (Laux, 2007). In the EU, the user owns a "set of legal rights entitling him to control data that are describing him, regardless of who had access to the data." In the US, whoever has rightfull access to the data "owns" the data. While our results do not necessarily suggest that privacy concepts in the US need to change, our results suggest that other policy innovations such as consent use may be useful. In the EU, a consent requirement allows a user to prevent any use of the data that he or she does not agree to. This notion is similar to intellectual property rights, e.g., Copyright, Patent, and Trademark rights. Alternatively, policymakers may choose to adopt blanket policies that directly govern the length of data retention.

In addition, our empirical results contribute to the antitrust debate over search engine dominance. We do not find evidence of an advantage in search quality for search engines that adopt longer periods of data retention. This suggests that a dominant search engine with a large fraction of market share does not necessarily maintain its dominance in the long-

http://curia.europa.eu/jcms/upload/docs/application/pdf/2014-05/cp140070en.pdf.

[14] "Differences between the privacy laws in the EU and the US", Management, Compliance, & Auditing, January 10, 2013.

term due to its access of historical data on users. Of course, we do not rule out that other antitrust concerns may exist as to why market concentration remains so high in Internet search markets.

V. Conclusion and Recommendations for Future Study

This paper investigates whether retention of large sets of data by firms that offer Internet search provide measurable changes to their performance from the perspective of consumers. Specifically, 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 an advantage to firms who own the data than is sometimes supposed.

Our results also suggest that restrictions 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). More generally, the length of data retention has become an issue in this debate over privacy (Korenhof et al., 2014). The question is whether the benefits of privacy (less data retention) for consumers outweigh any potential costs to consumers (lower quality search results). Our study suggests that retaining data for shorter periods of time does not lead to lower quality searches; in other words, we do not find a cost to privacy in our setting.

Several avenues beyond the scope of this study exist for future research. 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. If in the future, Google changes its data retention policy, this would a useful exercise for study. The third avenue for future research is that we do not know whether longer term effects exist due to the change in retention policies. Our data is truncated partly because Yahoo! reversed its previous data retention policy.

Given several interesting directions for future study, we believe that our study is a useful first step in measuring the effect of data retention policies on consumer behavior.

References

AdWords, G. I. (2008). Reach more customers with broad match.

Bennett, S. C. (2012). The right to be forgotten: reconciling EU and US perspectives. *Berkeley Journal of International Law 30*, 161–195.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). 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.

Dewey, C. (2015). How the 'right to be forgotten' could take over the American internet, too. *The Washington Post*.

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, pp. 65–89. National Bureau of Economic Research, Inc.

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

Hetcher, S. (2001). Changing the Social Meaning of Privacy in Cyberspace. *Harvard Journal of Law and Technology 15*(1), 150–206.

Korenhof, P., J. Ausloos, I. Szekely, M. Ambrose, G. Sartor, and R. Leenes (2014). *Reforming European Data Protection Law: Law, Governance and Technology Series*, Volume 20, Chapter Timing the Right to Be Forgotten: A Study into "Time" as a Factor in Deciding About Retention or Erasure of Data, pp. 171 – 201.

Laux, C. (2007). Privacy Concepts: US v EU. *The Center for Internet and Society at Stanford Law School*.

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.

Pouros, A. (2010). Search Engine Market Shares. *Econsultancy*.

Roberts, J. (2015). The right to be fogotten from Google? *Fortunte Magazine*.

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

Salbu, S. (1998). Who Should Govern the Internet?: Monitoring and Supporting a New Frontier. *Harvard Journal of Law and Technology 11*(2), 429–480.

Salbu, S. (2014). The US Data Privacy Debate, in a Nutshell. *Mondaq*.

Schwartz, D. and C. Seaman (2013). Standards of Proof in Civil Litigation: An Experiment from Patent Law. *Harvard Journal of Law & Technology 26*(2), 430–480.