A Survey of Query Log Privacy-Enhancing Techniques from a Policy Perspective

ALISSA COOPER
Center for Democracy & Technology

As popular search engines face the sometimes conflicting interests of protecting privacy while retaining query logs for a variety of uses, numerous technical measures have been suggested to both enhance privacy and preserve at least a portion of the utility of query logs. This article seeks to assess seven of these techniques against three sets of criteria: (1) how well the technique protects privacy, (2) how well the technique preserves the utility of the query logs, and (3) how well the technique might be implemented as a user control. A user control is defined as a mechanism that allows individual Internet users to choose to have the technique applied to their own query logs.

Categories and Subject Descriptors: K.4.1 [**Computers and Society**]: Public Policy Issues—*privacy, regulation, Ethics*; K.5.2 [**Legal Aspects of Computing**]: Governmental Issues—*Regulation*; A.1 [**Introductory and Survey**]

General Terms: Design, Legal Aspects

Additional Key Words and Phrases: Privacy, query, log, search, policy

ACM Reference Format:

Cooper, A. 2008. A survey of query log privacy-enhancing techniques from a policy perspective. ACM Trans. Web, 2, 4, Article 19 (October 2008), 27 pages. DOI = 10.1145/1409220.1409222 http://doi.acm.org/10.1145/1409220.1409222

1. INTRODUCTION

Web search has unquestionably become the most essential tool for finding information online. With billions of users generating tens of thousands of queries per second, search engines find themselves with an immense body of data unparalleled in its potential to describe the interests, thoughts, and behaviors of individuals everywhere. While query logs—which generally consist of the user's IP address, a time stamp, the query content, the user's browser and operating system information, the user's cookie ID, and the result clicked—may be extremely useful as a research or marketing tool, they also present

Author's Address: Center for Democracy & Technology, District of Columbia 2006; email: acooper@cdt.org.

serious privacy risks to those doing the searching. By searching for anything and everything associated with their daily lives, individuals provide search engines with detailed information about themselves which may then be subject to theft or public disclosure, available to civil litigants or government authorities, and shared with marketers and data brokers. Thus search engines face an amalgam of competing goals and interests: improving search functionality, combating fraud, increasing marketing capability, supporting academic research, protecting privacy, abiding by numerous laws and legal frameworks, and so on. Determining how to collect, store, and share query logs requires a delicate balancing act among all of these interests.

For the search engine companies navigating this landscape, it is likely that a combination of technical and policy measures will ultimately be required to develop a strategy that both protects privacy and maintains the utility of query logs for many different purposes. A variety of technical measures have been proposed that alter query logs in some way for the purpose of protecting privacy. This article will explore seven of these techniques, all of which are either currently employed by a commercial search engine or have been developed as research prototypes: deleting entire query logs, hashing query log content, deleting user identifiers like IP addresses and cookie IDs, hashing user identifiers, scrubbing personal information from query log content, grouping queries based on short sessions rather than maintaining a persistent identifier for each individual, and deleting queries that occur infrequently in the log data set. While there are a variety of privacy-enhancing tools that may be deployed by the user, without the involvement of the search engine (proxy servers or browser extensions like TrackMeNot [Howe and Nissenbaum 2008], for example), the focus here is on techniques that search engines themselves can apply.

In order to gain a comprehensive understanding of these techniques, they are assessed in Section 4 against three sets of criteria: (1) how well the technique protects privacy, (2) how well the technique preserves the utility of the query logs, and (3) how well the technique might be implemented as a *user control*. A user control is defined as a mechanism that gives individual Internet users the choice of having the technique applied to their own query logs. Although query log privacy is often viewed as an all-or-nothing proposition—a search engine applies a particular technique either to all of its query logs or none of them—viewing privacy-enhancing techniques as user controls, even in a theoretical context, may help to distinguish particular techniques that could be suitably applied on an individual basis. Empowering users with tools to protect their own privacy is a key element of a successful privacy strategy in the Internet age, and the query log context provides an environment rife with possibilities to help users help themselves.

This article takes a holistic approach, attempting to highlight most or all of the interests and goals that a search engine must balance, rather than focusing on a particular implementation scenario. Although much work in this area has focused on protecting privacy in releasing query logs to researchers, that scenario forms only one facet of the complex privacy tradeoffs that search engines are facing. In order to serve as a useful guide for the broad space of privacy challenges that search engines confront, all privacy threats, and all motivations for retaining query logs, must be taken into account.

Section 2 outlines the privacy threats that arise from query log retention. Section 3 describes the types of rationales that search engine companies may have for retaining query logs. In Section 4, privacy-enhancing technical approaches are assessed based on their privacy protectiveness, their ability to maintain the utility of the query logs, and their capacity as user controls. Section 5 provides an overview of some of the policy measures that can complement technical approaches, and Section 6 concludes.

2. QUERY LOG PRIVACY THREATS

2.1 Data Types that Pose Potential Threats

Understanding the privacy threats that arise from query log retention requires an understanding of what query logs may reveal about individuals. Traditional privacy concerns have focused on *identifying information*—name, address, phone number, email address, Social Security number and other government-issued identification numbers, and financial information such as credit card numbers—that may be linked back to a specific person. There is no question that this type of information appears throughout query log data, as users query their own or other people's information. But the privacy concerns go far beyond identifying information as narrowly defined.

Nonidentifying personal information that can be gleaned or inferred from query logs, such as birth date, zip code, and gender, may be used to link queries to an individual if combined with other publicly available data, such as census or voter registration databases [Sweeney 2000]. This correlation may be easier or harder depending on an individual's circumstances and the information available—for individuals living in sparsely populated areas, for example, queries that contain or suggest merely their zip codes and one other factor may be sufficient to identify them.

Query logs are rife with information about potentially sensitive subjects, including health conditions, sexual orientation, political affiliation, religious affiliation, adult material, and criminal activity. In societies where free speech is restricted, and discussing certain topics may be punishable by law, query logs may suggest access to or interest in material that governments deem to be objectionable, and individuals may wish to disassociate themselves from their queries to avoid persecution.

Query logs may also provide insight into an individual's physical location, which could allow the individual to be physically tracked, either by a malicious actor or by government authorities. The specificity of location information varies, since it generally must be extrapolated from IP address information or from the query content itself.

In cases where IP addresses, cookie IDs, or other identifiers are stored as part of query logs, they may be used to link together information about a particular user, and possibly to create user profiles, which may be pseudonymous or identifiable. If the identifiers are also employed outside of the search context (for behaviorally targeted ads, purchase histories, etc.), they may be useful for linking search data to other kinds of data. New query log data sets can also be combined with sets of query logs previously made public (such as those released by AOL in 2006 [Nakashima 2006]) to link queries in the new set back to individuals who may have been identifiable from the old set.

For all of these data types, privacy threats may arise either from individual queries or from groups of queries. Single queries may pose privacy problems if they contain identifiable information or identifiers that can be combined with other information to identify a particular individual. Individual queries that may have otherwise been innocuous on their own may expose private information when joined in a group with other queries related to the same user.

2.2 Privacy Risks of Query Log Retention

AOL's disclosure of twenty million search queries in 2006 spurred lengthy discussion about search query privacy by demonstrating, in a rather public manner, how query logs can be linked to individuals, and the sensitivity of the information that they contain [Barbaro and Zeller 2006]. The risks of query log retention are by no means limited to public disclosure, however. The risks fall into four general categories, each of which should be considered when crafting an overall strategy for protecting query log privacy:

Accidental or malicious disclosure. Disclosure of information that users intended to keep private, or that may harm them when released, is an obvious risk of query log retention. Even for a search engine that does not intentionally disclose query logs (to researchers or otherwise), the risk of accidental disclosure remains, as we have seen with the series of high-profile data breaches over the last several years resulting from security flaws, stolen laptops, and the like. Accidental disclosure may also arise from mistakes relating to a purposeful disclosure, as in the AOL case [Nakashima 2006]. Individuals also face threats from malicious disclosure, where an attacker or a rogue employee or researcher purposely discloses data that was meant to be kept private or that may cause harm to others in some way.

Compelled disclosure to third parties. Query logs may be subject to subpoena as part of civil litigation between individuals or organizations. In a divorce lawsuit, for example, a search engine company may be compelled to disclose queries related to an individual involved in the case as part of the evidence provided to the court. This kind of disclosure could be compelled in almost any kind of civil dispute. Search engine companies face the anger of their users if they do comply, and potentially lengthy and costly litigation by the requesting party if they resist.

This kind of disclosure gained a lot of attention in 2006 when the U.S. Department of Justice issued subpoenas to AOL, Google, Microsoft, and Yahoo as part of its litigation of an Internet child safety law (although the Department of Justice was one of the parties in the case, it is still considered a civil dispute, as opposed to a criminal or intelligence investigation conducted by the government). The Department of Justice was seeking, among other things, several months' worth of query logs to use as evidence that Internet filters were not adequately protecting children from adult content—a request that was largely viewed as massively overbroad and irrelevant to the case [Rasch 2006]. Although Google refused to comply with the subpoena and ultimately handed over a much narrower set of information than what was originally requested, the other search engines agreed to disclose what the government was originally seeking.

Disclosure to the government. Query logs may be subject to government demands in the context of law enforcement or intelligence investigations. As with many other kinds of information, the advent of the Internet and digital technologies has precipitated a glut of new data that may now be of interest to the government in these situations. Government authorities often have valid, compelling, and even urgent needs to examine query logs in pursuit of criminals and terrorists, but the standards for turning over this sort of information should be calibrated to avoid surveillance that is overbroad, unjustified, or erroneous.

Unfortunately, laws in countries such as the United States have not kept pace with technological advances, and thus the criteria for government access to information such as query logs are ambiguous, allowing the government to seek access under standards that are quite weak [Center for Democracy & Technology 2006]. Indeed, in the United States and other countries lacking strong general privacy laws, governments or search engine companies may contend that query logs are totally unprotected and may be turned over to the government with no legal process at all.

Even if there is a clear, strong standard in law, the U.S. executive branch has claimed in recent years that it is not bound by legal standards when the President is acting in the name of national security [Government Accountability Office 2007]. So long as this claim persists, it is possible that the government could seek the cooperation of search engine companies to disclose data in violation of whatever standard is legislatively established. This disclosure would be similar to the arrangements between telephone companies and the U.S. National Security Agency (NSA) that were revealed in 2005 [Roberts 2007].

Misuse of user profiles. The retention of query logs may allow the creation of detailed profiles of individuals' interests, preferences, and behaviors. As search becomes more pervasive, the depth of information encapsulated by search logs grows. These profiles may be particularly appealing for marketing purposes, both internal to the search engine (e.g., sponsored links), and as data sets provided to third-party marketers. They may also be used as a tool to calibrate price discrimination or to make decisions about a particular user's applications for insurance, credit, or other services. The privacy concerns with respect to user profiles will depend on whether users are informed about the profiling and what choices they have about it, whether they provide consent, whether the inferences drawn from query logs are valid, and users' rights to access their own profiles, delete profile information, or dispute decisions made based on the profile data.

All of these risk types are compounded by the potential for logs to be erroneously linked to the wrong individuals, whether through database errors or the assumption that multiple users of the same computer are all the same person. As with identity fraud, attributing one person's query logs to someone else can have adverse consequences, though in the case of query logs, either individual may be affected. For example, if one user conducts several searches for sensitive medical conditions, and those logs are attributed to someone else, it is the latter individual who may suffer. This is true whether the logs are accidentally or maliciously disclosed, provided to civil litigants or to the government, or used to build individual profiles.

3. QUERY LOG RETENTION RATIONALES

Previous investigations of how privacy-preserving techniques may affect query log analysis have tended to focus on specific applications, usually in the context of releasing log data to researchers for particular purposes. But search engines have many other reasons for retaining query logs beyond technical academic research, and all of these competing rationales must be evaluated in order to develop a comprehensive policy for protecting privacy.

This article will explore seven categories of reasons a search engine may want to retain query logs: improving ranking algorithms, language-based applications, query refinement, personalization, combating fraud and abuse, sharing data for academic research, and sharing data for commercial purposes. One of the benchmarks used to assess the privacy-enhancing techniques discussed in the next section will be how the techniques may reduce the utility of the query logs for these seven purposes.

Improving ranking algorithms. Query logs are an invaluable resource when it comes to improving the quality of search results. Knowing which search results get selected most frequently in the aggregate helps to improve result rankings and fine-tune the ranking algorithm [Agichtein et al. 2006; Joachims 2002; Spink et al. 2002]. Understanding common sequences of queries issued by individual users can help improve rankings for later queries in the sequence [Radlinski and Joachims 2005].

This also holds true for algorithms used to generate search advertisements. Analyzing ad performance can help search engines improve the way they select which ads to show for particular queries. Technologists and consumers may not regard this rationale very highly as compared with improving actual search ranking algorithms, since improvement of search-related ads is primarily a self-interested pursuit for a search engine trying to increase its bottom line. Because serving the right ads has the potential to improve search engines' profits, however, it is likely high on the companies' lists of reasons to retain query logs.

Language-based applications. Query logs provide a great deal of information about how Internet users employ language, and can thus be very useful in improving language-based features offered by search engines. These include features such as query spelling correction (e.g., having the search engine suggest or search for "Prague" when a user searches for "Prage") [Fleischer 2007], or helping the search engine to recognize when a user's query is posing a question [Spink and Ozmutlu 2002].

Query refinement. There are many different ways that a query can be refined to generate better search results by making suggestions to the user, or by adjusting the query behind the scenes. Several commercial search engines suggest related queries to users as they type their initial queries, or as part of the search results page. Studying past query logs can help inform these suggestions [Cucerzan and White 2007]. Analyzing query logs can also help to reformulate queries or to add specificity to the user's initial query in order to generate better search or advertising results [Cui et al. 2002; Jones et al. 2006].

Personalization. The kinds of analyses previously described can be compiled and applied at the macro level for all users, but they may also be used to improve search results for specific users—what many search engines call *personalization*. Using query log analysis to determine that a particular user's search for "apple" refers to the company, rather than the fruit, can help to tailor search results and rankings in the future [Google 2007]. And just as query logs may be used to personalize the search experience, so too may they be used to personalize the ads that users are served, and the way the ads are ranked or displayed.

Combating fraud and abuse. Query log analysis may help search engines detect and respond to many kinds of fraud and abuse, targeted both at their own systems and the Web at large [Fleischer 2007]. Web site hosts looking to pump up their own sites in search rankings may use techniques such as link bombing or Web spam. Query log analysis can be useful in identifying the suspicious query patterns that may result from these behaviors.

Query logs also help search engines respond to click fraud and other abuses of their advertising systems often undertaken by advertisers competing fiercely for clicks. Advertisers may contractually require search engines to retain some log data, at least until advertising billing cycles are complete, to allow the advertisers to assess the veracity of the ad clicks and displays that they pay for. Search engines likely view this particular application of query log analysis as crucial to their business, since they are likely to lose advertising clients who feel they are being cheated by click fraud.

Finally, retaining and analyzing query logs may help search engines detect Web threats such as phishing, scripting attacks, and Web bots that troll search results for malicious code hosts. While these threats may not impact a search engine directly, search engines may view it as in their best interest to limit their products' utility for malicious purposes as much as possible.

Sharing data for academic research. Any log analysis applications that appeal to the search engine companies themselves—and likely many more—provide compelling research questions for academics. Query logs retained by popular commercial search engines are an unmatched data source for researchers across disciplines. Not only are they invaluable to the technical fields of information retrieval and natural language processing, but they serve as a unique window into individuals' intentions, desires, and behaviors for researchers in social science disciplines. As an archive of user interests and activities on the Web over time, the query logs held by the large commercial search engines are unparalleled.

Sharing data for marketing and other commercial purposes. Query logs are useful to third party marketers for some of the same reasons they are useful to search engines—they provide insight into what a particular user is seeking.

Retaining query logs allows search engines to share (and possibly profit from) the data for marketing or any other commercial purpose.

An additional rationale that is important to note, but will not be discussed further here, is complying with applicable laws. Although no countries currently have laws that mandate the retention of query logs for the purposes of aiding government investigations, proposals for these kinds of laws are a constant subject of debate, and such laws do exist for other types of information (e.g., the European Union Parliament [2006] directive mandating the retention of communications-related information). Since this rationale is only theoretical, it is not factored into the analysis of the techniques in the next section; but the way that search engines approach query log privacy in future will undoubtedly see changes if data retention laws covering query logs become a reality.

4. ANALYSIS OF PRIVACY-ENHANCING TECHNIQUES

A wide variety of privacy-enhancing techniques have been suggested in the context of query log retention. The seven techniques analyzed in this article were chosen because at least one implementation of each technique currently exists either as a research prototype, or in the commercial search engine sector. Other techniques that have gained attention in the privacy research community, such as k-anonymity [Sweeney 2002], present intriguing concepts but have yet to be formally applied in the context of query logs, and thus are not addressed.

The techniques analyzed in this article also share the characteristic of being applicable both within the search engine itself and after the search engine decides to disclose query logs for research or commercial purposes. Some previously suggested privacy-enhancing approaches, such as grouping a user's queries based on their content [Adar 2007], may help to protect privacy in the disclosure of logs to researchers, but would either not be feasible for search engines to implement on their internal data storage, or would have little practical privacy benefit in such cases.

It is important to note that search engines may not be the only ones collecting search query logs. Internet Service Providers (ISPs) have access to their subscribers' full clickstreams, and thus could be recording (and even disclosing or selling) search logs as well [Reimer 2007]. The privacy benefits users may derive when search engines employ the techniques described in this article do not automatically transfer to other entities collecting and storing search query data.

Each technique is analyzed based on three criteria. The first criterion is how well the technique protects privacy. Each technique's effectiveness in this respect is assessed against the four types of privacy risks outlined in Section 2. Any potential attacks on the technique that would reduce its privacy effectiveness are also considered.

The second benchmark is how well the technique preserves the utility of the query logs for the seven categories of retention rationales outlined in Section 3. This evaluation assumes that each technique is applied broadly to most or all of the query logs stored both previously and on a going-forward basis by commercial search engines. It is important to note that applying some of the techniques would prevent certain kinds of research from occurring. The analysis that follows describes the types of research that may no longer be possible with a particular technique, and balances this drawback against the benefits of the technique.

The final criterion is how well the technique might be implemented as a user control that provides individual Internet users with the choice of having the technique applied to their own query logs. This analysis takes into account both the practicality of an individualized implementation, and the user education challenges associated with offering the techniques as choices to average Internet users who may not be tech-savvy.

Implementing any of these techniques as user controls will require the search engine to maintain the user's preference across searches and sessions. If a user is given the choice of having the search engine apply a particular privacy protection, the user will likely want to have that choice maintained for all queries on that search engine until he or she decides otherwise. Thus, some mechanism—perhaps a cookie, login, or browser extension or plug-in—will be necessary to retain the user's preference.

This mechanism should not itself require the user to disclose more information than necessary to retain the preference (e.g., users should not have to disclose identifying information just to ensure that the search engine will stop retaining their identifying information in query logs). But even if the mechanism does not require the disclosure of additional information, some implementations may create potential privacy risks for other users who do not activate the user controls. This may be the case if creating the kind of structure to be able to recognize particular individuals' choices entails grouping each individual's queries together on the back end—whether those individuals are making use of individualized controls or not. Although the necessity of this will depend on the particular privacy-enhancing technique, it could present an additional complexity for search engines to grapple with in developing user controls.

Typically user controls can only be applied on a forward-going basis, since it may be difficult to identify past queries belonging to the particular user activating a control. This is very different from applying the techniques across the board for all users, where it may be much more straightforward to implement the techniques on past query logs.

Search engines would likely also need their own mechanisms for distinguishing particular queries that have had privacy-enhancing techniques applied to them, from other queries. This distinction may be important if they intend to share queries for research or commercial purposes, for example, where the fact that certain techniques have been applied may be material to the recipient of the query logs.

Implementing any of these techniques as user controls raises the concern that the users who choose to employ them will become free riders who benefit from search improvements that are only made possible because search engines can conduct analysis on the logs of other users who do not employ the privacy-enhancing techniques. The extent to which this is true depends on the techniques employed and how many users decide to use them—as discussed subsequently, many search improvements are still possible even after these techniques have been applied. But search engines will need to be conscious of free riders if uptake of these techniques as user controls swells.

Although the following evaluations deal with one technique at a time, multiple techniques could be combined to provide a package of privacy protections to users. Indeed, search engines will likely find that a combination approach will be most effective in both protecting privacy and maintaining the utility of the query logs.

A note on terminology: In the discussion that bellow, the term *internal identifier* refers to identifiers that are generated and retained strictly within a search engine's storage, whereas *external identifier* refers to IP addresses, cookie IDs, and any other individual identifiers that may be transmitted or stored in locations external to the search engine.

4.1 Log Deletion

Log deletion involves the erasure of users' complete query logs—the query content, user identifiers, and all other log components—from a search engine's storage. This deletion may occur as early as when the search engine returns search results to the user.

Log deletion is the most privacy-enhancing technique available, since it allows for all query log data to be completely erased from a search engine's storage. All four privacy threats are essentially eliminated if an individual's logs are deleted promptly (or never retained in the first place). This is contingent upon the search engine either not disclosing the logs to third parties prior to deleting them, or requiring that any third parties that do obtain the logs maintain the same deletion policy as the search engine itself.

The flip side of log deletion is that the utility of the logs drops to zero after they are erased. For this reason, search engines may be wary of adopting a policy of deleting all query logs in short order (after storing them for a matter of days or hours, for example). However, search engines may seek to gain some of the benefits of log analysis and storage over a longer period, and then delete the logs at the end of that time. Consider the case of logs that are retained to help combat fraud and abuse. If logs are being kept for the purpose of investigating later incidents of advertising click fraud, for example, then they could be deleted at the end of the billing cycle for the search engine's advertisers (although this may not work for other kinds of abuse like Web spam and scripting attacks). Similarly, if the logs are being used for personalization, old logs that no longer provide insight about a user's current interests could be deleted—it may become clear, for instance, that a user was searching for engagement rings, but several weeks or months after the ring is purchased, those queries may no longer be useful in personalizing search results.

Search engines may be able to retain what they learn about users for improving ranking algorithms, language-based applications, and query refinement, without retaining the queries themselves. For example, maintaining aggregate statistics about the number of "Prage" searches that were meant to be for "Prague" is possible without retaining actual queries. However, these applications may still suffer from log deletion, since it makes identifying new query

IP Address	Cookie ID	Query Content	Timestamp	Browser/OS	Result Clicked
76.26.159.134	359b81298e37g	HIV test	2008-06-18 11:54:41	Firefox 2.0; Windows XP	http://www. cdc.gov/hiv/

Fig. 1. An example query log for the query "HIV test."

IP Address	Cookie ID	Query Content	Timestamp	Browser/OS	Result Clicked
76.26.159.134	359b81298e37g	wd8fy0972j9kv	2008-06-18 11:54:41	Firefox 2.0; Windows XP	http://www. cdc.gov/hiv/

Fig. 2. The query log for "HIV test" after applying a one-way hash.

trends very difficult, and it negates the possibility of ever analyzing historical data. Log deletion also essentially prevents the sharing of log data for research or commercial purposes, since the logs no longer exist to be shared.

Log deletion is, however, a very good candidate technique to be implemented as a user control, particularly because it presents a fairly straightforward choice for users—they can either have their logs retained or deleted. In fact, one commercial search engine is already allowing its users to choose to not have their logs retained beyond a few days [Ask.com 2007]. This kind of control allows privacy-conscious users to obtain all of the privacy-enhancing benefits of log deletion, while leaving other users' logs intact for use in various forms of log analysis. Implementing log deletion as a user control also leaves open the possibility of personalization for users who want it, while allowing other users to forego personalized results.

Offering log deletion as a user control will likely require some extra adjustments on the part of the search engine. For example, search engines may not want to serve search ads or sponsored links to users who opt to have their logs deleted, since the search engine will not be able to tabulate and retain clicks and impressions for those users.

4.2 Hashing Queries

Secure one-way hashes, which take an input string and produce a hash value that is difficult or impossible to reverse-engineer to produce the original input, are used in a variety of contexts for privacy protection. There are several different ways query logs could be hashed to help reduce privacy threats. Entire queries could be hashed, so that the resulting log contains a hash value rather than the query content. Another approach would be to tokenize the query, and hash each token, resulting in a set of hash values in lieu of the original query content (known as *token-based hashing* [Kumar 2007]). Because these approaches have a great deal in common, they are considered together in the following analysis.

Applying a one-way hash helps to protect privacy because the original query is removed from the logs and becomes difficult to deduce from its replacement. Consider the example query in Figure 1, which many might consider to be highly sensitive based on its content:

After applying a one-way hash, the query content is replaced with the hash value for "HIV test" (Figure 2).

This eliminates the sensitive information from the query log content.

In the specific case of query logs, however, it may be possible to reverse-engineer particular queries by using other publicly available data sets, such as previously released query logs, or any aggregate statistics the search engine might retain about queries in unhashed form. These kinds of data can be leveraged to form a statistical analysis of query frequencies that can be used to determine the original contents of hashed queries. If the query "HIV test" is the 50th most frequent query in a large set of query logs, for example, that statistic may be used to determine what the hash of "HIV test" is in the set of hashed query logs. With token-based hashing, a great deal of further information may be gleaned based on which words occur together in a query. Knowing how often the query "Michael Jordan" appears in a previously released log can help decipher the separate hashes for "Michael" and "Jordan," which may allow the individual doing the analysis to glean identifying information from the hashed logs [Kumar 2007]. These kinds of attacks on the one-way hashing system require a determined individual applying sophisticated techniques, but they are nonetheless important in considering how privacy-protective hashing queries may be.

Despite this potential weakness, hashing queries greatly reduces the threat of accidental or malicious disclosure. Even if some hashes can be reverse engineered, the likelihood that all hashed queries could be revealed is slim.

The impact that hashing queries has on civil litigant and government requests depends on what information the requesting party is after. Hashing queries greatly diminishes the value of asking a search engine to disclose all queries associated with a particular user, IP address, or cookie ID, because the query content is hashed and essentially unrecoverable. Government authorities conducting unauthorized surveillance may still find the queries useful if they can apply the reverse-engineering attacks described previously, but at the very least, the surveillance would be much more cumbersome than if the queries were not hashed. However, civil litigants and government authorities may be able to inquire about particular query terms. They may be able to request that a search engine disclose the IP addresses or cookie IDs associated with all queries for "bomb-making," for example, and by hashing that query and looking it up in the logs, the search engine would be able to comply. It may even be possible for the government to compel a search engine to look up and disclose large volumes of query terms in this way, or to create a reverse-lookup table that maps hashes to their query content on a going-forward basis. Thus, although hashing provides some protections from disclosure through legal process, the protections are largely based on what kinds of requests the courts deem to be permissible, and what kinds of requests search engines are willing to comply with.

Hashed queries would not be useful for third-party marketers and others for the purposes of profiling applications concerned with a user's interests or behaviors, since this information would no longer be available to profilers unless they are able to reverse-engineer the hashes. Profiling is still possible only if the search engine shares queries with third parties before hashing them, or if the search engine itself maintains user profiles by recording notes about users' queries before they get hashed (e.g., a user who searches for "airline fares" and "Tuscany" might be noted as a user interested in travel).

Many of a search engine's rationales for retaining logs are still viable, to some extent, with hashed queries. This is particularly true for token-based hashing, or if the search engine retains aggregate statistics about queries (or some subset of queries) in an unhashed form. Some improvements to ranking algorithms, if they are based on aggregate data, may still be possible if statistics about the frequency of queries and the success of particular search results are available (but ranking improvements based on analyzing individual sequences of queries would no longer be possible). The same is true for some language-based applications and query refinement analyses—if aggregate statistics are sufficient to conduct the analysis, hashed queries may still be useful. Certain kinds of fraud and abuse investigations may also be possible if particular individuals or queries are known targets of the investigation, and if query logs can be searched for particular combinations of identifiers and query content.

Hashing queries may be more detrimental to other log analysis applications. Hashed queries do not provide much basis for personalization or sharing for commercial marketing purposes. Some forms of technical academic research are likely still possible, although social science research that relies on the query content is not.

Offering to hash queries as a user control may have a particular benefit not enjoyed by many of the other privacy-enhancing techniques: the technique's effectiveness in protecting privacy may actually increase for those who choose to adopt it if not all individuals make use of it. Because reverse-engineering attacks on the hashing algorithm are dependent on a statistical analysis of query frequency, hashing a smaller number of queries reduces the accuracy of the frequency statistics by reducing the number of hashes that contribute to the frequency calculations. This effect is dependent on how many users actually sign up to have their queries hashed—if half of a major search engine's users decide to participate, for example, this benefit likely disappears.

The concept of hashing may be somewhat familiar to users who understand how encryption can help protect their privacy in other contexts. But given that only the most technically savvy and privacy-conscious users fall into this category, offering hashing as a user control would likely require a substantial investment in consumer education.

4.3 Identifier Deletion

Identifiers, such as IP addresses and cookie IDs, are standard components of query logs. Because they may be used to tie queries both together and to particular users, it may benefit user privacy if the identifiers are deleted, in whole or in part. All of the major search engines already have policies in place to delete partial or complete IDs after storing them for between 13 and 18 months [Center for Democracy & Technology 2007].

The exact benefits of identifier deletion depend on what the identifiers are and what portion remains after deletion. The first few octets of an IP address may reveal some information about where its associated computer is physically located and which ISP assigned the IP address, so even when the last octet or two are removed, the query log may still reveal some information about the

IP Address	Cookie ID	Query Content	Timestamp	Browser/OS	Result Clicked
76.26.159.134	359b81298e37g	HIV test	2008-06-18 11:54:41	Firefox 2.0; Windows XP	http://www.cdc.gov/hiv/
76.26.91.2	711k03296e86g	whitman-walker clinic	2008-06-18 11:59:03	Firefox 2.0; Windows XP	http://www.wwc.org/

Fig. 3. Two example query logs issued by different users.

IP Address	Cookie ID	Query Content	Timestamp	Browser/OS	Result Clicked
76.26		HIV test	2008-06-18 11:54:41	Firefox 2.0; Windows XP	http://www.cdc.gov/hiv/
76.26		whitman-walker clinic	2008-06-18 11:59:03	Firefox 2.0; Windows XP	http://www.wwc.org/

Fig. 4. Two example query logs issued by different users, with cookie IDs and the last two IP octets erased.

user conducting the search. However, identifier deletion makes it much more difficult to distinguish between separate users who all share the same partial IP address. Consider the two example queries in Figure 3 conducted by different users:

Even though these two query logs have the same browser/OS information, related query content, and timestamps in close proximity to each other, a search engine viewing these two logs in full could reasonably conclude that they were issued by different users, since they have different cookie IDs and IP addresses. If the search engine had a policy of deleting cookie IDs and the last two octets of IP addresses, however, it would become impossible to know whether the two queries were issued by the same user (Figure 4).

In the case where only the last IP address octet is deleted, the level of privacy protection depends on how many devices are assigned to a particular combination of the first three octets. Because the range of IP addresses available to be assigned within certain jurisdictions may far outnumber the quantity of Internet-connected devices in those jurisdictions, it is possible for a particular combination of the first three octets to be assigned only once (for example, imagine that 76.26.159.134 is the only assigned address in the 76.26.159 block). In such a situation, removing the last octet may still allow some queries to be tied back to a single individual.

The conditions for cookie IDs may be similar depending on how the search engine assigns them—if ID prefixes have a specific meaning to the search engine, then retaining the prefix allows the search engine to retain whatever information it encodes. The difference is that, to outside observers, the ID prefix is likely meaningless, whereas a partial IP address can provide information to any observer, since the correlations between IP address and location/ISP are publicly available.

Deleting all identifiers in their entirety leaves less information that may tie a query back to an individual—essentially just the query itself and information about the user's browser and operating system configurations. Removing all identifiers also makes it much more difficult to associate multiple queries with the same user—had the queries shown in Figure 3 been issued by the same user, for example, it would be impossible to determine that fact if the full IP addresses and cookie IDs were removed. Deleting internal identifiers but keeping IP addresses may allow third parties to correlate queries to specific people, but is dependent upon whether the same individual uses the same IP address often enough to make that correlation. Deleting IP addresses but maintaining internal identifiers may allow one user's queries to be tied together if the same individual is consistently assigned the same identifier.

Deleting identifiers can be a powerful tool in protecting user privacy. Although user data remains in the queries themselves – leaving open some of the threats from accidental or malicious disclosure, especially if users query their own personal information—this technique makes it difficult or impossible for civil litigants, government authorities, and other third parties to request all query logs corresponding to a specific user, IP address, or cookie ID. It also means that any user profiles based on remaining partial identifiers cannot likely be correlated to specific individuals, computers, or browsers, thus reducing the potential for misuse of user profiles.

Whether or not particular log retention rationales are still viable after removing identifiers depends on what identifier information is kept. Deleting all identifiers impedes many applications that rely on tying query logs to particular users, including some forms of ranking algorithm improvement, anti-fraud measures, personalization, log data sharing for any commercial application that relies on user profiles, and some social science research. When partial identifiers are retained, language-based applications and query refinement may be more workable, since those analyzing the data can use other query information (e.g., timestamps and query content) to guess when a group of queries belongs to a particular user. Partial identifiers may also still be useful in identifying fraudulent activity attributable to groups of similar IP addresses. Sharing deidentified query logs for technical academic research is likely still useful in answering some research questions.

Offering identifier deletion as a user control would be a fairly straightforward process once a search engine has architected its system to store logs without IDs, or with only partial IDs. The biggest challenge would likely be educating users about the technique and its benefits. Many users may not even be aware that search engines collect and store IP addresses and cookie IDs, or they may not realize how this allows search engines and third parties to link search queries back to them. Gaining an understanding of these concepts will be necessary for users to evaluate whether they would like to have their identifiers deleted from their logs.

4.4 Hashing Identifiers

Hashing identifiers involves applying a secure, one-way hash function to all external identifiers associated with a query log. Applying a one-way hash makes it essentially impossible to correlate a log's internal identifier (i.e., the hash value) to the external identifiers originally associated with the log.

Hashing identifiers mitigates some of the risks that stem from accidental or malicious disclosure, by removing identifiers that could otherwise be used to correlate search logs to other information. Consider the example of a user who purchases a bicycle and helmet online, and later conducts searches on an

PURCHASE HISTORY				
IP Address	**Item**	**Quantity**	**Merchant**	**Timestamp**
76.26.159.134	carbon fiber road bike	1	Bikes R Us	2008-06-18 14:09:30
76.26.159.134	bike helmet	1	Bikes R Us	2008-06-18 14:09:30

QUERY LOGS					
IP Address	**Cookie ID**	**Query Content**	**Timestamp**	**Browser/OS**	**Result Clicked**
76.26.159.134	359b81298e37g	HIV test	2008-06-19 20:47:21	Firefox 2.0; Windows XP	http://www.cdc.gov/hiv/
76.26.159.134	359b81298e37g	whitman-walker clinic	2008-06-18 21:00:50	Firefox 2.0; Windows XP	http://www.wwc.org/

Fig. 5. Example purchase history and query logs belonging to the same individual.

PURCHASE HISTORY				
IP Address	**Item**	**Quantity**	**Merchant**	**Timestamp**
76.26.159.134	carbon fiber road bike	1	Bikes R Us	2008-06-18 14:09:30
76.26.159.134	bike helmet	1	Bikes R Us	2008-06-18 14:09:30

QUERY LOGS					
IP Address	**Cookie ID**	**Query Content**	**Timestamp**	**Browser/OS**	**Result Clicked**
m98b3hd444	359b81298e37g	HIV test	2008-06-19 20:47:21	Firefox 2.0; Windows XP	http://www.cdc.gov/hiv/
m98b3hd444	359b81298e37g	whitman-walker clinic	2008-06-18 21:00:50	Firefox 2.0; Windows XP	http://www.wwc.org/

Fig. 6. Example purchase history and query logs belonging to the same individual, with hashed IP addresses in the query logs.

unrelated topic. If the bicycle merchant was storing online purchase histories indexed by IP address, and gained access to the search logs, he or she may be able to reasonably conclude that the purchases and searches were conducted by the same person by linking the two data sets (shown in Figure 5).

If the query logs contained hashed IP addresses, however, gaining access to the search logs would not allow the bicycle merchant to automatically correlate the query logs and the purchase histories (Figure 6).

However, as the AOL leak demonstrated, whether or not identifiers are hashed has no bearing on the utility of the identifier in linking a single user's queries together [Barbaro and Zeller 2006]. Thus the creation (and misuse) of user profiles is still possible. The fact that a user's queries may still be linked together, combined with the possibility that identifying information may still reside in the contents of the query, means that hashing identifiers does not eliminate the risk of breaching individual privacy.

Query logs with hashed identifiers may also still be subject to a subpoena or court order. If civil litigants or government authorities possess the input into the hash for a particular user (e.g., the user's IP address and/or cookie ID), then they may request that the search engine determine the hash value for that information and turn over all query logs associated with that value, which corresponds to the input.

Hashing identifiers has little impact on many of the log retention rationales, since it preserves the correlation between individual users and their queries. Improving ranking algorithms, language-based applications, query-refinement, personalization, and sharing for academic purposes are all largely unaffected.

Sharing for commercial purposes may be more difficult, since users are no longer associated with well-established external identifiers, but it may be possible to provide the third-party recipient of the query logs with a mechanism for hashing external identifiers in the same way that the search engine executes the hashing.

The log retention rationale most affected by hashing identifiers is combating fraud. Pinpointing the exact users responsible for past fraud and abuse incidents becomes much more difficult since there is little chance of reverse engineering a fraudulent user's external identifiers just by looking at the query logs. Search engines would need to discover a pattern of abuse and wait for the abuser to conduct a new search in order to correlate particular behaviors to specific external identifiers. Hashing also eliminates the possibility for search engines to identify groups of similar IP addresses that seem to be generating fraudulent activity.

Implementing the hashing of identifiers as a user control faces similar challenges to identifier deletion: once the search engine's storage is set up to accommodate hashed identifiers, the main task will be to educate users about the technique. The education challenge with hashing is perhaps even greater, since the benefits of the technique are not as clear.

4.5 Scrubbing Query Content

Scrubbing query content involves removing identifying information such as phone numbers, Social Security numbers, credit card numbers, addresses, and names from query logs. Although it is impossible to guarantee that all identifying information has been removed after scrubbing, some techniques have been suggested to better distinguish identifying information from the remainder of the query content [Xiong and Agichtein 2007]. These techniques must also contend with the problem of over-inclusiveness—removing query information (such as names of celebrities entered as search keywords) that is not in fact identifying. Yahoo currently employs this technique [Center for Democracy & Technology 2007].

Scrubbing query content provides some protection from accidental or malicious disclosure, because it reduces the likelihood that queries can be traced back to individuals. The strength of this protection depends on how much identifying information can be scrubbed. Even after scrubbing, it may be possible to link queries back to individuals by using other publicly available information. Consider the set of queries in Figure 7, all conducted by an imaginary user, Alice Birdsboro, before being scrubbed.

Suppose that the search engine used for these queries had a policy of scrubbing first and last names. After scrubbing, the content of the second query would be empty, which in theory would mean that the other three queries could no longer be tied to Alice. But suppose that the page that was viewed after searching for "sun valley triathlon" (http://www.sunvalleytri.com) contains the results of the Sun Valley Triathlon, in which Alice took part. Imagine that the results are displayed publicly, and that Alice's result appears as follows (Figure 8).

IP Address	Cookie ID	Query Content	Timestamp	Browser/OS	Result Clicked
76.26.159.134	359b81298e37g	sun valley triathlon	2008-06-18 20:47:21	Firefox 2.0; Windows XP	http://www.sunvalleytri.com/
76.26.159.134	359b81298e37g	Alice Birdsboro	2008-06-19 21:00:50	Firefox 2.0; Windows XP	<none>
76.26.159.134	359b81298e37g	sports stores 20009	2008-06-20 07:22:33	Firefox 2.0; Windows XP	http://www.fleetfeetdc.com/
76.26.159.134	359b81298e37g	Stanford class of 2003 reunion	2008-06-21 12:15:12	Firefox 2.0; Windows XP	http://www.stanfordalumni.org/

Fig. 7. A set of queries prior to scrubbing.

Number	Name	Home City	Home State	Age	Time	Rank
86	Alice Birdsboro	Washington	DC	26	1:34:08	47

Fig. 8. Publicly available race result for Alice Birdsboro.

Anyone with access to both this public information and the scrubbed query set—which still contains a Washington, DC zip code and the "Stanford class of 2003 reunion" query, indicating that the searcher is likely in the 25-27 year old age range—could reasonably conclude that the queries and the race result both belong to Alice, assuming that none of Alice's competitors had similar ages and hometowns.

Although this is a small-scale example of how an individual could be reidentified despite query scrubbing, reidentification research has shown that even having a limited set of data points about individuals (such as zip codes and birth dates) can be sufficient to reidentify large portions of the population using other publicly available data [Sweeney 2000]. Although query scrubbing can reduce this risk if many identifiers are scrubbed, it cannot eliminate it altogether.

Compelled disclosure to civil litigants or government authorities as part of a judicial process is mostly unaffected by scrubbing. Since the remainder of query content stays intact, and queries can still be tied to an individual via an identifier, the parties conducting surveillance or requesting the subpoena or court order can still obtain the majority of the information they were looking for, minus any identifying search terms. Sensitive information related to health, political affiliation, religious affiliation, adult material, criminal activity, and the user's general interests and behaviors remains intact.

Scrubbing has minor effects on user profiling. While it may remove information that would allow a marketer or other party to more accurately pinpoint the exact identity of the searcher, all the other information of value—the searcher's interests, tastes, and behaviors—remains available to be used (and abused) in user profiles.

Unless logs are being retained with the specific purpose of analyzing identifying information (perhaps for personalization or social science research purposes), scrubbing likely has only minor effects on a search engine's log retention rationales. Although the technique removes some information that may have been useful for tracking down those committing fraud or abuse, it is unlikely that search engines rely solely on such information for that purpose.

Search engines could offer scrubbing as a fairly straightforward user control, though doing so raises an important question: if a user is privacy-conscious enough to request that their query logs be scrubbed, are they conducting identifying queries in the first place? To be fair, privacy-conscious users might want the ability to conduct some identifying searches if they could be confident that scrubbing would remove those search terms. But given the search engine's reluctance to guarantee exactly which future queries will be scrubbed, privacy-conscious users are unlikely to have confidence in the mechanism. Offering scrubbing, either as a user control or across-the-board, presents an educational challenge; search engines would need to explain to users that typing identifying queries may impact their privacy, and how scrubbing identifying information may benefit them.

Conceptually, there are two different ways search engines could offer scrubbing as a user control. One way would be to scrub everything the search engine deemed to be identifying from a particular user's logs. In this case, if user Jane Smith searches for "123 Main Street," "Jane Smith," "John Smith," and John Smith's Social Security number, all of those searches would be scrubbed, even though some of them identify people other than Jane. The main shortcoming of this approach is that one person's identifying information can exist in other users' queries. Consider the case where John Smith elects to have his queries scrubbed, while Jane Smith does not. In this scenario, John's Social Security number remains in the logs, even though he had his own queries scrubbed. Thus, offering scrubbing in this manner may provide users with a false sense of security.

Alternatively, search engines could offer users the option of having their identifying information scrubbed from all queries (not just their own). However, this would require users to first provide their identifying information to the search engine so that it could be located in others' queries, creating a privacy problem in and of itself. In the end, the fact that scrubbing one user's queries does not rid the logs of that user's information creates a conundrum that may provide justification for search engines to apply the scrubbing technique across-the-board, rather than as a user control.

4.6 Deleting Infrequent Queries

Adar [2007] has suggested that deleting queries that appear infrequently in the logs may be an effective way of removing queries that contain identifying information—it is unlikely for the same legitimate credit card number to appear in a search engine's logs thousands and thousands of times, for example. Some fine-tuning is required to determine exactly how to set the infrequency threshold to maximize the number of identifying queries that are eliminated, while deleting as few other queries as possible. This may be a challenging endeavor, since studies of large query log data-sets have indicated that the vast majority of queries occur a small number of times [Beitzel et al. 2004; Spink et al. 2001], meaning that even an extremely low threshold may end up eliminating substantial amounts of nonidentifying queries.

This technique also requires some mechanism for identifying new query trends. Just because a particular query appears for the first time does not mean that it will never become a frequent query (one can imagine this happening with the names of new professional athletes or celebrities, new slang, new product names, etc.). Thus a search engine employing this technique will likely need to retain all queries, however infrequent, for some period of time before determining that certain queries are truly infrequent and should be deleted.

Since the goal of this technique is to eliminate identifying queries, its effects are very similar to those of the scrubbing technique. Deleting infrequent queries mitigates the risks that come from accidental or malicious disclosure, since less identifying information is available to be disclosed. These risks are unaffected, however, during the period when infrequent queries are retained. Because there is no way to guarantee that identifying queries have been removed, and because of the existence of other publicly available data, it may still be possible to track queries back to individuals even after infrequent queries have been deleted. The scenario illustrated in Figures 7 and 8 could still occur, for example, if only the "Alice Birdsboro" query is considered infrequent and the other three queries remain in the search engine's logs.

Query logs are still likely useful in litigation and for government requests or surveillance, even after infrequent queries have been removed. The case where this technique may provide some protection is when the litigant or government authority is explicitly looking to find unusual queries, which is certainly a plausible scenario. In that case, the effect of deleting infrequent queries depends on the timing of the subpoena or court order—if the request is made when the queries of interest are still being buffered to determine whether their infrequent status is valid, then the search engine may be compelled to turn them over to the requesting party.

Deleting infrequent queries may reduce the threats from user profiling in two ways. It may eliminate identifying data, such as a user's address, that otherwise could have been used to link profiles to specific individuals, and it may remove information about users with unusual interests. However, the remaining query logs may still be used to profile user behaviors and preferences, since the logs can still be linked together via IP address, cookie ID, or internal identifiers.

The log retention rationales that may see the most negative impact from deleting infrequent queries are language-based applications and query refinement. Some aspects of these applications depend on being able to recognize rare queries and learn how to offer the searcher suggestions or make adjustments behind the scenes. Because deleting infrequent queries will inevitably cause the removal of some nonidentifying queries, any analytical value that could have been gleaned from those queries is lost. Personalization and social science research applications that make use of identifying queries may also suffer from the use of this technique.

Many of the other rationales—improving ranking algorithms, combating fraud and abuse, and sharing for commercial purposes—are based in part on the value of analyzing popular or high-volume queries, so deleting infrequent queries may have less of an impact. Whether or not sharing for technical academic research will still be viable depends on how necessary rare queries are to the specific research question at hand.

It is difficult to imagine a simple way to implement the deletion of infrequent queries as a user control. Although a single user may generate many infrequent

IP Address	Cookie ID	Query Content	Timestamp	Browser/OS	Result Clicked
	359b81298e37g	sun valley triathlon	2008-06-18 20:47:21	Firefox 2.0; Windows XP	http://www.sunvalleytri.com/
	038k81767m35g	Alice Birdsboro	2008-06-19 21:00:50	Firefox 2.0; Windows XP	<none>
	444w81540p30g	sports stores 20009	2008-06-20 07:22:33	Firefox 2.0; Windows XP	http://www.fleetfeetdc.com/
	247h81798q33g	Stanford class of 2003 reunion	2008-06-21 12:15:12	Firefox 2.0; Windows XP	http://www.stanfordalumni.org/

Fig. 9. Search queries from Figure 7 with a new cookie ID assigned each day, and with IP addresses deleted.

queries of his or her own, deleting all of these misses the central idea of the technique in the first place—to eliminate identifying queries—and may cause logs to be erased that otherwise would have been useful for analysis. If a search engine offered users the option of making their queries eligible for deletion if they were deemed to have infrequent status, this would add a huge layer of complexity to the determination of which logs should be deleted. Not only would the search engine need to keep track of the frequency of queries and whether rare queries seem to be gaining popularity, but it would need to constrict these calculations based on which users' queries could contribute to frequency counts, and these counts would have to be adjusted every time users change their minds about whether or not they want their queries to be eligible for deletion. This technique also suffers from the scrubbing conundrum—just because one user makes his or her queries eligible for deletion does not mean that the user's identifying information will not appear in other users' queries. If it were possible to offer this technique as a user control, it may thus give users a false sense of security.

4.7 Shortening Sessions

The identifiers associated with many users' queries remain persistent over time, allowing a large number of query logs to be associated with a single user (or browser or device). By shortening the length of time that any identifier is associated with an individual, some privacy threats can be mitigated [Xiong 2007]. In practical terms, shortening sessions would likely require deleting IP addresses (since static addresses can persist indefinitely) and setting shorter cookie lifetimes, or replacing persistent internal identifiers with identifiers that get replaced after a short period. These shorter sessions may be based on a certain amount of time passing; a user may be assigned a new identifier every month, day, or hour. Alternatively, the session length could be based on browsing behavior; sessions could end when users close their browsers or when they navigate away from the search engine's site.

Imagine a search engine that assigned new identifiers each day, and did not retain IP addresses. Using the queries from Figure 7 as an example, the search engine would no longer be able to link all four searches to Alice Birdsboro, or to each other (Figure 9).

The length of the session will have an impact on both how much privacy protection the technique offers, and how much utility the logs retain for certain applications. For example, some users hardly ever close their browsers, so basing query log session intervals on browser sessions may not be providing them with a great privacy benefit. Setting the session length to one hour may eliminate the search engine's ability to do a lot of personalization, whereas setting it to one month may leave some personalization options open.

Shortening sessions has the potential to be highly privacy-protective. Because shorter sessions can remove the link between a user and his or her entire query history, it makes government and civil litigant requests much less likely to obtain identifying data, since search engines can only provide one session's worth of logs (if anything at all). In the example in Figure 9, if the requesting party wanted all queries associated with a particular cookie ID, the search engine could only provide, at maximum, one query. Shortening sessions also drastically reduces the viability of user profiling, since profiles could only be based on data from a narrow window. However, because the query content may still contain identifying information, and because queries within the same session may still be linked together, the risks from accidental and malicious disclosure are not entirely resolved by this technique.

Shorter sessions may undermine some log analysis applications. The technique limits personalization to using query information gleaned within the session. Identifying which users committed fraud in the past (prior to the session interval) is also difficult or impossible with shorter sessions. Some of the reasons why search engines might have shared data for commercial purposes may no longer be viable if they are based on sharing historical profiles of users.

The consequences for other applications may be less drastic, however. Some ranking improvements, language-based applications, and query refinement techniques may be largely based on queries that occur in close proximity to each other, so shorter sessions have less of an impact. Fraud and abuse analysis that is based on queries in close proximity is also still possible. Sharing for academic research is likely still useful for the types of analyses that do not rely on complete historical user profiles.

Implementing shorter sessions as a user control has an advantage over many of the previously described techniques; it has an analog in another context. Some users may already be familiar with the notion of clearing their browser cookies, either manually or by using a browser setting that clears them at a particular interval. Having the option of clearing identifiers stored on the search engine's side may thus appear to be a logical step for users accustomed to managing their cookies. However, the overhead for managing which users have requested shorter sessions and when their sessions expire may be substantial.

5. POLICY-BASED PROTECTIONS

In addition to all of the privacy-enhancing technical measures available to search engines, a plethora of policy-based protections are also at their disposal. Indeed, the major search engines are already employing a combination of technical and policy strategies to help protect user privacy, and such combinations will likely continue to be necessary as advances are made in both the technical and policy realms. In many cases where technical measures are insufficient, policy strategies may help to fill the void. This section explores five policy-based protections, some of which are already in wide use, and others that have yet to be employed or are applied differently in different jurisdictions: privacy laws, privacy policies, confidentiality and licensing agreements, user consent, and institutional review boards.

5.1 Privacy Laws

Laws form the ultimate backstop to privacy abuses, but in many jurisdictions the threshold question is whether privacy laws apply to query log data. The European legal framework provides a prime example. Over a decade ago, the European Parliament [1995] passed a harmonized Data Protection Directive to be applied across all EU member states. The directive declares that personal data should not be processed, except when a specific legal basis explicitly allows it or when individuals consent to the data processing. This provides a very protective framework for "personal data," defined in the EU as "any information relating to an identified or identifiable natural person." Do query logs (or some portions thereof) constitute personal data? The EU Article 29 Data Protection Working Party [2008] recently published a 29-page opinion that explains how different components of a query log (content, IP address, etc.) may or may not constitute "personal data" depending on the situation. Add in the fact that different EU member states each have their own interpretations of the directive and the definition of personal data, and the task of determining how the laws apply to query logs becomes extremely complex.

Privacy laws in the United States have followed a much more sectoral approach, with distinct regulations for many different kinds of consumer data, but no overarching framework to secure privacy across-the-board (aside from the Fourth Amendment, which protects against unreasonable government search and seizure but has been determined not to apply to much data disclosed to businesses providing services). In the United States today there are separate privacy laws for medical information, financial information, library records, video rental records, and numerous other classes of data. Courts have determined that decades-old communications privacy laws apply to Internet communications, but the status of search query logs in particular under these laws is unclear, since little judicial precedent exists for determining how search queries fit into a legal framework developed long before the advent of the commercial Internet [Foley 2007]. Although no privacy law clearly addresses query log data, privacy violations outside of the regulated sectors may be handled under the authority of the Federal Trade Commission and the state Attorneys General to pursue unfair and deceptive practices affecting commerce.

5.2 Privacy Policies

All of the major search engines, and most of the Web's most popular sites, display their privacy policies publicly. As far as search is concerned, privacy policies are useful for explaining what information a search engine collects when a user conducts a search, how logs are used, with whom logs are shared, how log data is secured, and how users may view their logs. Arming users with this information (if they are able to find it and understand it, which is the subject of some debate [Cranor 2007]) helps them to make better choices about which search engines to use. Establishing a privacy policy also provides a guide for a search engine's employees to understand what they are and are not allowed to do with the data they collect.

In the United States, the Federal Trade Commission has held that privacy policies posted by online service providers are binding, and therefore failing to comply with stated policies constitutes a violation of the Federal Trade Commission Act. This kind of enforceability helps to build trust among consumers that search engine companies are following their privacy policies. Publishing the results of internal or third-party privacy audits or signing up for a privacy certification program (through organizations like TRUSTe) can also help search engines demonstrate that their privacy policies meet high standards and that the companies are adhering to those standards.

5.3 Confidentiality and Licensing Agreements

Search engines that share query log data with third parties can contractually require those third parties to abide by the same policies and procedures used by the search engines to safeguard query log data. These contracts may contain provisions that hold the third parties responsible for privacy violations as an additional incentive to comply. This model is well-established in other contexts, and has already been used in the context of releasing query logs for research purposes [Microsoft 2007].

5.4 Consent

The spectrum of consent options for the collection of data is broad, ranging from no consent on one end, to explicit opt-in consent on the other. Many of the ways that search engine companies already use query logs fall on the low end of the spectrum in their reliance on *implied consent*—the fact that users input queries into a particular search engine, given that the search engine discloses its uses of query logs in its privacy policy, implies that the users have consented to those uses of the logs. In most cases, users have no way to opt out of this data collection while continuing to use the search engine; if they do not want their data collected, their only option is to stop conducting searches. While obtaining implied consent is practical for a large volume of users, certain uses of query logs that create greater privacy risks likely require opt-in consent instead, where users are presented with an explicit choice about whether their query logs can be used, and they must affirmatively agree to such use before it can occur. For example, users might be offered the opportunity to affirmatively consent to making their logs available to researchers for specific projects, much as they do in some forms of medical research. Laws in certain jurisdictions may require that opt-in consent be obtained for certain data uses.

One of the big challenges in gaining informed opt-in consent is presenting choices to users in an understandable and timely fashion, given the fact that many users' only relationship with a search engine consists of conducting Web searches. For these users, search engines will need to provide some mechanism for displaying the relevant option, gaining consent, and preserving users' choices. Given the huge volume of users that major search engines currently have, obtaining informed consent may prove to be particularly challenging.

5.5 Institutional Review Boards

Bar-Ilan [2007] has suggested that the Institutional Review Board (IRB) structure—frequently employed for medical research [United States Department of Health and Human Services 2005]—may be useful in developing ways for search engines to supply query logs to the research community. IRBs consist of panels of experts who review specific research study plans and evaluate them based on the risks to the individuals participating in the studies, the anticipated study benefits, and how consent is obtained from participants, among other criteria. Although IRB review already occurs at U.S. universities for some research using query logs from certain sources, the major commercial search engines have yet to adopt it as a mechanism for supplying search data to researchers.

6. CONCLUSION

As search becomes an increasingly essential part of so many Internet users' daily lives, the breadth and depth of information contained in query logs grows to unparalleled levels. As a body of data that can reveal the interests, preferences, search strategies, and linguistic behaviors of entire populations, query logs are a true bounty for research of all kinds, conducted both internally, at the search engine companies, and externally, by academics and others. But the great promise of query logs as a research tool is bound by the privacy risks that arise for some of the very same reasons that the logs are so useful in the first place—the richness of detail that they offer about individuals' lives. Achieving the right balance between protecting privacy and promoting the utility of the logs is thus difficult but necessary to ensure that Internet users can continue to rely on Web search without fear of adverse privacy consequences.

The technical and policy measures discussed in this article represent this extraordinarily complex landscape faced by popular search engines today. Although it is sometimes useful to focus on a single query log analysis technique or privacy-enhancing measure, search engines must ultimately base their decisions about privacy on a holistic analysis of the myriad competing interests involved in the retention of query logs. In the end, the frameworks that achieve the best balance of these interests will consist of a number of different technical features coupled with an array of policy measures that can fill in the gaps where technology is insufficient.

Ultimately, the search industry, academia, consumers, and regulators all play important roles in determining how to strike this balance. Search engine companies can use the techniques and policies described in this article, and they can also leverage their position in the marketplace to promote best practices among their partners and competitors. By offering user controls, search engines can put the power in the hands of consumers and empower individuals to develop their own tailored privacy preferences. Academics can continue to develop innovative technologies and solutions that make the job of protecting privacy easier and more effective. Regulators can work to enforce against the truly bad actors in the marketplace, and they can simplify the process of privacy compliance by providing comprehensive, baseline standards for companies to follow. With online search becoming such an essential tool, contributions in all of these areas will be paramount to its continued success.

ACKNOWLEDGMENTS

Many thanks to my colleagues Jim Dempsey, John Morris, Ari Schwartz, David Sohn, Peter Swire, and Paul Otto for their helpful feedback in developing this analysis.

REFERENCES

ADAR, E. 2007. User 4XXXXX9: anonymizing query logs. In *Proceedings of the 16th International World Wide Web Conference (WWW'07)*. IW3C2, Banff, Alberta, Canada.

AGICHTEIN, E., BRILL, E., AND DUMAIS, S. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA.

ARTICLE 29 DATA PROTECTION WORKING PARTY. 2008. Opinion on data protection issues related to search engines. http://www.cbpweb.nl/downloads_int/Opinie%20WP29%20zoekmachines.pdf.

ASK.COM. 2007. Ask.com puts you in control of your search privacy with the launch of 'AskEraser'. http://www.irconnect.com/ask/pages/news_releases.html?d=132847.

BARBARO, M. AND ZELLER, T. 2006. A face is exposed for AOL searcher no. 4417749. In *The New York Times*. http://www.nytimes.com/2006/08/09/technology/09aol.html?ex=1312776000.

BAR-ILAN, J. 2007. Position paper: Access to query logs—an academic researcher's point of view. In *Proceedings of the 16th International World Wide Web Conference (WWW'07)*. IW3C2, Banff, Alberta, Canada.

BEITZEL S., JENSEN, E., CHOWDHURY, A., GROSSMAN D., AND FRIEDER, O. 2004. Hourly analysis of a very large topically categorized web query log. In *Proceedings of the 27th Annual International ACM SIGIR Conference*, Sheffield, South Yorkshire, UK.

CENTER FOR DEMOCRACY & TECHNOLOGY. 2006. Digital search & seizure: Updating privacy protections to keep pace with technology. http://www.cdt.org/publications/digital-search-and-seizure.pdf.

CENTER FOR DEMOCRACY & TECHNOLOGY. 2007. Search privacy practices: A work in progress. http://www.cdt.org/privacy/20070808searchprivacy.pdf.

CRANOR, L. 2007. Making privacy disclosures to consumers more usable. http://www.ftc.gov/bcp/workshops/ehavioral/presentations/6lcranor.pdf.

CUCERZAN, S. AND WHITE, R. 2007. Query suggestion based on user landing pages. In *Proceedings of the 30th Annual International ACM SIGIR Conference*, Amsterdam, Netherlands.

CUI, H., WEN, J., NIE, J., AND MA, W. 2002. Probabilistic query expansion using query logs. In *Proceedings of the Eleventh International World Wide Web Conference*, Honolulu, HII.

EUROPEAN PARLIAMENT. 1995. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:NOT.

EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION. 2006. Directive on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks and amending Directive 2002/58/EC. http://www.ispai.ie/DR%20as%20published%20OJ%2013-04-06.pdf.

FLEISCHER, P. 2007. Google response to Data Protection Working Party. http://64.233.179.110/blog_resources/Google_response_Working_Party_06_2007.pdf.

FOLEY, J. 2007. Are Google searches private? An originalist interpretation of the fourth amendment in online communication cases. *Berkeley Techn. Law J. 22*, 1, 447–472.

GOOGLE. 2007. Google search privacy: Personalized search. http://youtube.com/watch?v=UsUBnPRtTbI.

GOVERNMENT ACCOUNTABILITY OFFICE. 2007. B-308603, presidential signing statements accompanying the fiscal year 2006 appropriations acts. http://www.gao.gov/decisions/appro/308603.htm.

HOWE, D. AND NISSENBAUM, H. 2008. TrackMeNot: Resisting surveillance in web search. In *On the Identity Trail: Privacy, Anonymity and Identity in a Networked Society*. Oxford University Press, Oxford, UK, To appear.

JOACHIMS, T. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada.

JONES, R., KUMAR, R., PANG, B., AND TOMKINS, A. 2007. "I know what you did last summer"—query logs and user privacy. In *Proceedings of the ACM 16th Conference on Information and Knowledge Management (CIKM)*, Lisbon, Portugal.

JONES, R., REY, B., MADANI, O., AND GREINER, W. 2006. Generating query substitutions. In *Proceedings of the 15th International World Wide Web Conference*, Edinburgh, Scotland.

KUMAR, R., NOVAK, J., PANG, B., AND TOMKINS, A. 2007. On anonymizing query logs via token-based hashing. In *Proceedings of the 16th International World Wide Web Conference (WWW'07)*. IW3C2, Banff, Alberta, Canada.

MICROSOFT. 2007. Microsoft privacy principles for live search and online ad targeting. http://download.microsoft.com/download/3/7/f/37f14671-ddee-499b-a794-077b3673f186/Microsoft%E2%80%99s%20Privacy%20Principles%20for%20Live%20Search%20and%20Online%20Ad%20Targeting.pdf.

MICROSOFT. 2006. Microsoft live labs: Accelerating search in academic research. http://research.microsoft.com/ur/us/fundingopps/RFPs/Search_2006_RFP.aspx.

NAKASHIMA, E. 2006. AOL takes down site with users' search data. *The Washington Post*. http://www.washingtonpost.com/wp-dyn/content/article/2006/08/07/AR2006080701150.html.

RADLINSKI, F. AND JOACHIMS, T. 2005. Query chains: Learning to rank from implicit feedback. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illinois, IL.

RASCH, M. 2006. Google's data minefield. In *The Register*. http://www.theregister.co.uk/2006/01/31/google_subpoena_us_government/.

REIMER, J. 2007. Your ISP may be selling your web clicks. *Ars Technica*. http://arstechnica.com/news.ars/post/20070315-your-isp-may-be-selling-your-web-clicks.html.

ROBERTS, C. 2007. Transcript: Debate on the foreign intelligence surveillance act. In *El Paso Times*. http://www.elpasotimes.com/news/ci_6685679.

SPINK, A., JANSEN, B., WOLFRAM, D., AND SARACEVIC, T. 2002. From e-sex to e-commerce: Web search changes. *Computer 35*, 3, 107–109.

SPINK, A. AND OZMUTLU, H. C. 2002. Characteristics of question format web queries: An exploratory study. *Inform. Proc. Manag. v38*, n4, 453–71.

SPINK, A., WOLFRAM, D., JANSEN, B., AND SARACEVIC, T. 2001. Searching the web: The public and their queries. *J. Amer. Soc. Inform. Sci. Techn. 52*, 3, 226–234.

SWEENEY, L. 2000. Uniqueness of simple demographics in the U.S. population, LIDAP-WP4. Laboratory for International Data Privacy, Carnegie Mellon University.

SWEENEY, L. 2002. k-anonymity: A model for protecting privacy. *Int. J. Uncertainty, Fuzziness Knowl.-based Syst. 10*, 5, 557–570.

UNITED STATES DEPARTMENT OF HEALTH AND HUMAN SERVICES. 2005. Protection of human subjects. http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm.

XIONG, L. AND AGICHTEIN, E. 2007. Towards privacy-preserving query log publishing. In *Proceedings of the 16th International World Wide Web Conference (WWW'07)*. IW3C2, Banff, Alberta, Canada.

Received December 2007; revised June 2008; accepted August 2008