The Information Society

Who Searches the Searchers? Community Privacy in the Age of Monolithic Search Engines

Jonathan A. Poritz[a]

[a] Department of Mathematics and Physics, Colorado State University, Pueblo, Colorado, USA

PLEASE SCROLL DOWN FOR ARTICLE

Routledge
Taylor & Francis Group

PERSPECTIVE

Who Searches the Searchers? Community Privacy in the Age of Monolithic Search Engines

Jonathan A. Poritz

Department of Mathematics and Physics, Colorado State University, Pueblo, Colorado, USA

Privacy has largely been equated with every individual's right to privacy. Accordingly, current efforts to protect privacy on the Internet have sought anonymity by breaking, where possible, links with personally identifiable information (PII)—all uses of aggregated data stripped of PII are considered legitimate. This article argues that we need to use a broader concept, general or group identifying information (GII), because even aggregated data stripped of PII violate privacy at the community level. The search engine companies, or anyone else with access to their log files, can use these data to generate a moment-by-moment view of what is on the collective mind. Such a view can be used in a variety of ways, some with deep economic and even political impact. In order to frame this discussion, it is necessary to examine some of the realities of the search engine-mediated associative interface to the World Wide Web. While this interface has enormous benefits for the networked world, it also fundamentally changes a number of issues underlying various current debates about Internet governance.

We usually find the particular resource we are seeking on the Internet by querying one or at most a handful of search engines. This method of access is clearly very convenient, but results in a concentration of power in the hands of the search engine operators. Google, Inc., has drawn a good part of the public's attention in this regard: partly because of its market dominance, but perhaps also because of the

public unease at the disconnect between the corporation's "do no evil" slogan and its complicity in network censorship in some parts of the world and, more personally, the targeted advertisement inserted in every query response.

A different but related nervousness is induced by the U.S. Department of Justice subpoenaing search engine log files. Since the USA PATRIOT Act authorized government subpoenas of library records to investigate suspicious activities, the U.S. government has shown a willingness to exercise the unprecedented powers it now has to monitor the intellectual interests of U.S. residents. Access to easily mined electronic data sources such as the searches Internet users submit to the major search engines is logical extension of earlier surveillance activities, and recent subpoenas have shown this is what the government is now after.

Public discussion of these issues has mostly centered on two themes, one of preserving privacy and another of preserving the broad decentralization of control that has been a feature of the Internet since its inception.

Privacy has largely been equated with every individual's right to privacy, and the concern is personally identifiable information, or PII. I argue in this article, however, that PII or like notions are inappropriately restricted concepts when it comes to what we might call *community privacy*. The failure of PII to be a satisfactory framework for ensuring privacy is clear in the context of search engine log files, as I explain later in this article. I also show that invasions of community privacy could have profound consequences and so are a topic to which we should pay attention.

The defenders of the decentralized structure of the Internet have focused on issues related to network governance, such as "network neutrality" in the United States, or over the future role of ICANN (Internet Corporation for Assigned Names and Numbers) in the global arena. In the opinion of this author, it behooves the champions of an open Internet to consider the influence and implicit

control which lies with the search engine oligarchy; for example, I argue later in this article that ICANN in reality has far less relevance to most current issues of network governance than what is generally believed.

ICANN'T: SEARCHING AS A GLOBAL ASSOCIATIVE MEMORY

Users want data *not* (usually) by location (address), they want them because they are *kind of like* some other data, which can only be referenced in terms of keywords or search terms or phrases related to the target data. In other words, users want the Web to be an *associative memory*, that long-sought-after target of computer science. In the 1990s it was difficult to imagine a system that constantly crawled the Web, digested the harvested data, and then offered the results via clean, well-crafted user interfaces and cleverly designed prioritization schemes. Such a system would require immense network bandwidth, processing power, and storage. But today the search engines pretty much do what was only a decade ago difficult to imagine.

What was also not imagined was the requested data returning with various other links that the user could follow or ignore, perhaps even with targeted advertisements. It was certainly not understood that these advertisements could become a stupendous source of revenue, where no similar revenue stream attached to the Web had existed previously. Interestingly, there is a striking similarity with the situation at the beginning of the television broadcast era, when it was not clear that TV would be an incredible source of advertising revenue and, in fact, that advertising would drive the entire business model of television distribution.

Serving targeted ads with search results is in fact a far better use of advertising monies than TV advertisements for many markets, as the readers of such ads are preselected to be fairly well off (they are surfing) and definitely interested in topics related to the submitted search terms. This economic payoff promises to provide a continuous and growing stream of resources for technological innovations leading to the development of a fully functioning associative interface to the Web. In effect, the search engine oligarchy is a self-sustaining phenomenon—absent catastrophes in the networked community or major, unforeseen changes such as a near-monopoly operating system vendor somehow squeezing out all competing search engine providers (although this would merely concentrate searching in *their* hands, not remove any of the features I am discussing in this article)—where the advertising revenues allow excellent search responses which in turn allow greater advertising revenues, and so on.

This perspective sets current debates within a new frame. For example, whether ICANN should require the DNS system to accept non-ASCII character sets, whether the influence of the United States on the global Internet through ICANN is too strong, and whether some gov-ernment could control content available within their geographic territory (by partitioning pornography into an "*.xxx*" top-level domain [TLD], or by blocking access to certain Internet protocol [IP] addresses with politically objectionable content). These issues need to be reexamined and the power of search engine operators taken into consideration.

Search engine providers now have much more invasive power than ICANN ever did: Privacy-conscious users of earlier Web access technologies could always go directly to the IP number, and then (aside from their packets being copied in transit) they were fairly safe from someone monitoring their activities. Even going directly to the IP number was not particularly indicative of users' interests, since what the system provided was the domain name-to-IP address translation, and domain name server (DNS) caching rendered this via an intermittent side channel with great latency. Now, however, most users have very little interest in the URL; they simply type the target pages' associations (keywords) into a search engine, and they do this *constantly*. So the search engine log files track, moment by moment, exactly what Internet users are interested in. This leads to a whole host of privacy questions, many of which have received a lot of public attention, and one of which has somehow escaped the spotlight but is potentially far more devastating: What about traffic analysis on the collective consciousness of the networked community? It is this question that we examine at length in the rest of this article.

ONLINE PRIVACY: FROM PII TO GII

What are the expectations *individuals* have for privacy in their online lives? Every network transaction has a source, destination, and content. Along the chain of machines relaying the data, one or more may be compromised or simply unscrupulous hosts that may copy down source, destination, and content information. In addition, the destination site may keep records of the transaction with the intent of using some transaction-related data for future profit, or perhaps to comply with the policies of governments of the United States, China, Saudi Arabia, and others with similar laws about Internet content monitoring.

Privacy against in-transit eavesdropping can be achieved by encryption, as is well known. Protection against most types of misuses by the destination web site obviously requires a certain amount of trust on the part of the source that the destination will act in accordance with announced policies of data usage, or the user must have faith that the countervailing legal threats of violation of contract or false advertisement litigation will compel respect for privacy. A prerequisite for this trust, then, is such an announced privacy policy regarding the use of sensitive information.

Sometimes simply the source and destination information are sensitive, even if the content is encrypted;

this is *traffic analysis*. Perhaps the best solution available at this time to users who are concerned about this issue is low-latency mix networks such as Tor (Dingledine, Mathewson, & Syverson, 2004), I2P (I2P Team, 2006), Morphmix (Rennhard & Plattner, 2002), Tarzan (Freedman & Morris, 2002), and JAP (Federrath, 2006) (and there are others). These systems seek to maximize anonymity, sometimes even the exact network location of the sender and the recipient, by various cryptographic techniques; for example, each server receives only the minimum necessary information and does not have any logs so that whatever information is received is quickly forgotten.

To summarize, then, source, destination, and content information cannot simply be withheld because the transmitted information has to transverse numerous network nodes as it moves toward the intended destination and, once there, it has to carry meaning. Only the records of the information transaction can be minimized, in the interests of privacy. So, current efforts are directed at anonymizing these records by breaking their linkage with PII. While this is handled somewhat differently in traffic analysis-resistant technologies, the basic point nevertheless always remains that *aggregated information—information stripped of its PII—is always acceptable*. In fact, privacy policies (see Word Wide Web Consortium, 2005) accept "data aggregation," keeping statistics on transactions but not the specifics of those transactions and their sources, as the simplest and purest process to remove PII.

In the next section, however, I discuss how aggregated data, particularly in the case of search engine log files, represent a deeply "personal" picture of the community. This information remains accessible even after data is anonymized, as mentioned earlier, by the removal of all PII. Hence we must introduce a broader concept to understand the implications for privacy at the community level. Accordingly, I then suggest that the usual definition of PII be expanded to:

> *General or group identifying information* [*GII*], information that can be used to identify uniquely, to contact, or to locate a single person *or group of persons*.

Please note that the italicized portion is the only addition to a standard definition of PII.

The new thing about GII is that it allows for the formulation of privacy polices that detail how agents who have acquired information that has implications for a community's privacy can use this information. Also, best practices documents can be produced for the proper use of GII. It may be argued that this would result in too stringent controls on this information, but then requirements could be made merely for full disclosure of the uses to which GII may be put, which would allow consumers an opportunity to exercise free choice.

AGGREGATED SEARCH ENGINE LOGS: ANONYMOUS TRAFFIC ANALYSIS FOR FUN AND PROFIT

Having made the point that we need to think beyond PII in order to protect community privacy, let us now consider a few specific examples in which data mining of even (conventionally) anonymized search engine log files yields interesting applications. Here I speak as if there is only one global search engine used by all Web surfers, which I shall denote **G** for brevity. This is of course not entirely true, but it is fairly close, and since the conclusions are of a statistical nature, they will remain mostly true for any individual search engine that has sufficient traffic to analyze.

1. A corporation conducting a marketing campaign could go to **G** for information about correlates of searches related to its offering. Features that interest users could be found, as could be geographic or other real-world variables which influence buying habits. These results from the queries **G** has received could also be used in a dynamic way, giving timely feedback to various components of a marketing campaign.
2. Before an election in a first-world nation or community, **G** could correlate geographic information about IP addresses with topics of potential concern to voters.
3. **G** itself could monitor qualitative words which correlate with the names of corporations or major products, with an eye to predicting changes in the values of publicly traded stocks.
4. **G** could proactively approach corporations and offer them suggestions on marketing or even on products and/or services which clients expect the corporation to offer (e.g., if **G** often saw the search phrase "anti-virus for IBM Thinkpad", it could suggest to IBM that it would have an easy client base for AV software for its Thinkpad computers).
5. Even more disconcerting, **G** could proactively offer to political candidates an election-season "hot issues for your constituents, and what they feel about these issues" summary.

Whether these or related "attacks" are being or will be tried, it is at least clear that the search log files have many interesting uses, often because they continue to be laden with GII after the PII has been aggregated away. I discuss various reactions to the possibilities 1–5 mentioned, and to other such traffic analyses of the collective consciousness, in the next section.

Let us emphasize what is astonishing here: **G** has a moment-by-moment view of what is on the collective mind, by seeing the frequency distribution of search

terms arriving at its portals across the globe. Even more impressive is the correlative information: Frequency distributions on searches with several keywords simultaneously allow **G** to build a weighted network whose nodes are currently interesting search terms and whose links are between terms that appear together, with the weight being the frequency of these simultaneous searches. Evidently this dynamic mental map of the Internet community has numerous interesting applications, of which the community should at least be aware.

It is interesting to consider what happened in *Attorney General of the United States v. Google Inc.*, heard before the U.S. District Court for the Northern District of California in late 2005 (see Bylund, 2006, and Price, 2006, for overviews). Here, for a period of time, the attorney general was after fairly extensive search log information, supposedly to enforce the *Child Online Protection Act* (*COPA*) of 1998. Google, Inc., fought this request on the grounds that the selected data were not relevant to COPA, were an invasion of users' privacy, were too close to Google's most important trade secrets, were redundant given that the government already had data from other search engines, and were an undue burden to collect and submit. What is quite fascinating is that Judge Ware publicly stated prior to his formal ruling that he would require some selected and anonymized data to be turned over to the Department of Justice (DoJ). Government lawyers said they would use this data to show how easy it is to circumvent blocking software intended to protect children from inappropriate material on the Web. Hence the court is ordering Google to comply with the government's request for data to back up some social science research it intends to do, despite the fact that "Now Google could face hundreds of university professors [saying], 'I've got a study I'd like you to conduct,'" Judge Ware admitted (McCullagh, 2006). This is exactly my point in this article.

WHAT IS TO BE DONE?

We know that an enormous amount of information has fallen into the hands of the search engine operators through their associative interface to the World Wide Web. The power they possess has generated much concern and some responses have been suggested. What I believe has not been sufficiently discussed is the issue of collective privacy invasion that I pointed out in the last section. There are, again, a whole range of possible reactions, some of which we consider now.

Caveat Quaestor

The simplest response would be simply to do nothing. Users of the search engine interface to the Web are aware of giving information to **G** that is valuable both with PII and without, in an aggregated form that yet retains significant GII—or, at least, I hope they are aware of this now. A free-market optimist might then believe that surfers will preferentially use search engines whose privacy policies are better in their guarantees regarding GII.

Note that the definition of GII is so broad that it is essentially impossible to anonymize GII in any way, but this has attractive consequences: Current privacy policies often describe specific allowed uses of PII and then say that aggregated/anonymized data will be used in any way the service provider desires. In future policies that use the concept of GII, this open-ended clause applying to anonymized data would have to be abandoned and providers would have to commit to a complete listing of uses to which they will put collected information. This transparency would presumably allow market forces to be more effective.

We should remark here that privacy policies such as we have described here are impossible to express in languages such as P3P, the Platform for Privacy Preferences (see World Wide Web Consortium, 2005). P3P needs an extension to language features that address collective privacy, probably by reworking the basics to build on the concept of GII rather than merely PII.

Finally, one problem with relying on competition is that any individual who considers this issue has a very small individual economic motivation to do anything about it: The increment of information granted to the search engine providers by the searches of any one person is minuscule, and the loss to the individual is possibly zero. Hence we seem to be in a Prisoner's Dilemma-like situation where the free-market model would predict no individual should act, even though everyone's common action would be in the community's interest.

Proactive Government

The Internet community could press for new laws. Public policy about ownership and appropriate use of information is complex (to say the least) and does, in certain circumstances (think of insider trading or blackmail), impose restrictions upon the use of information for profit. The data mining of search engine logs would pose a very interesting test case for such laws: It is information legally acquired, used in a way that does not harm any particular individual, yet it seems to be an affront to the collective privacy, and it seems also to give the engine operators some sort of unfair advantage merely because of their position in the network.

The discomfort we (perhaps) feel over the collective privacy issue, and the feeling that users did not intend their search terms to drive anyone's investment strategy (or political or marketing campaign, or any of the other attacks identified earlier), might motivate a law similar to the prohibition of insider trading. The precise legislative basis

for such a law is somewhat unclear to this author: There is a train of legal thought on related topics (see Boyle, 1992) that finds a crucial difference between information that is typed as "public" and that which is viewed as "private," but the privacy of aggregated search log files is very unclear (at least from a viewpoint dominated by PII, not GII).

Some justifications of insider trading laws are based upon a notion of a "level playing field" (although Boyle, 1992, and others criticize this), which might also be a rallying cry of concerned network users looking for regulatory intervention. Publication of statistics on keywords and associations appearing in Internet searches could be required of the engine providers. One flaw here, however, is that the relevant files in **G**'s hands are so enormous, and useful data mining on them would be so specialized (one could go after particular "positive" terms that are statistically associated with terms related to a particular company or its products, for example), and therefore so resource-intensive, that this would be a very large imposition on the engine operators, and perhaps simply an impossibly large task.

A specific law that could be considered, beyond simply forbidding the search engineers from using the aggregated, anonymized data in the disquieting ways already described, would be to require full disclosure. For example, current securities laws sometimes require corporate officers to make public their stock trades, so as to discourage insider trading and to allow other investors to follow with the officers even when they are not illegally acting upon inside information. A similar requirement could be imposed upon the personnel of search engine providers, with regard for example to public disclosure of financial transactions such as stock trades, or public disclosure of actions such as the second and the fifth points listed in the section on aggregated search engine logs.

"Do No Evil"

Google has an announced policy of "doing no evil." One way to actualize this ideal would be for this operator or others simply to offer their log files to the public in a completely aggregated, anonymized form. This would not only be a generous step toward making the playing field more level, but it would also allow research into the fascinating information contained in the detailed, timely, weighted network of memes inhabiting the minds of the online part of the first-world citizenry (recall my definition of this weighted network). There must be great treasures of many kinds there. Would it not be great if the public were offered access to this data?

More concretely, a very public-spirited search engine provider could offer a sophisticated application programming interface (API) that could perform standard statistical functions on the data in suitably anonymized log files; the actual computation to return these results would probably have to be done on the provider's machines, as the relevant files will most likely be distributed over many data centers, and only the providers have access to, and the technical resources to handle, these files. So that any outside (or inside) party does not make their own tilted playing fields with sophisticated applications of this API, a usage agreement might require all output to be displayed on publicly available web pages.

A fairly modest market follower (see "Alexa Web Services" from Amazon.com, at http://developer.amazonwebservices.com) has provided an extremely limited API of the type we are considering, as a paid service.

Finally, the mechanisms for setting up and refining advertising campaigns which the search engine providers offer to their client base are in some sense a primitive API of the sort we are discussing here. However, the information provided in this way is more limited and perhaps may not be used as I am suggesting because of the providers' terms of service.

Who?ogle

Scientists love technological solutions to social problems, and it is not particularly hard to imagine one for the problem exposed in this paper (also see Poritz, 2007). A mixnet devoted to searching would be necessary, in order to prevent exposure of PII. This "searching mixer" could also break compound search queries into their separate constituents and submit them via different paths as different simple queries to the search engine servers, in order to break the associations whose importance we have pointed out above. Data returning from single-term queries must be reassembled into the composite query results, by intersecting the returning URL lists for the separate terms. Clearly, some prioritization scheme must be chosen to return these intersected hit lists, and experience shows that the order here is extremely important in the eventual success of the search service.

What turns out to be harder is to hide the basic single search term frequency counts. Typically in situations like this, one would like to generate random additional (single keyword) searches at a fairly high rate—in fact, this is necessary to prevent timing attacks on the anonymity of the infrastructure—in such a way that spikes in the frequency counts of terms that are actually on the mind of the collective get hidden in the noise. However, there are so many possible search terms that a flat distribution on all such terms would give each term an extremely small probability, so we would never see the clustering of repeated queries on one term that probably occurs as the public's interest is caught by some issue or other. There are various ad hoc approaches that depend upon finding sources of noise

that are difficult for **G** to distinguish from the actual time series of single-term frequency counts; this is particularly tricky to accomplish in the general world outside of the search engine provider firms, since the statistics of search queries are not made public.

Who?ogle Spelled with a "G"

As one last possibility, I should mention the idea that **G** might cooperate with some consortium of privacy advocates to assist the development of a system like the one I just described in the previous section. This seems quite unlikely in the world today but it would certainly have tremendous technological advantages: Data could be passed in intermediate form among the mixnet servers; the algorithms they use to recombine the composite query results could be those very ones **G** uses now; and so on.

It is not even clear that providing this level of security and anonymity would impact **G**'s current business model: The individual keyword queries would return results to the anonymizing network that had targeted advertising attached, and Who?ogle would presumably forward these, and intersect them when recombining composite query results at the client terminus, so the same click-through revenue stream would result. In fact, if this process were in the hands of the search engine providers—if **G** operated Who?ogle—then the elaborate fee structures for advertisers, based on the order of their advertising appearing next to results for certain searches, could be more perfectly justified and maintained.

There is a problem with giving Who?ogle into the hands of **G**: Why should we network users trust this corporate Who?ogle any more than just plain **G** as it is now? The likely answer is that no additional trust would accrue were the entirety of Who?ogle to be on **G**'s machines. If, instead, **G** were to design an API for single-keyword searches that returned the answers in easy-to-use form—easy to intersect, and easy to keep and sort the accompanying targeted advertisements—and perhaps were to give technical assistance to the coding of the anonymization infrastructure, open-sourcing the code and helping anyone who wanted to contribute their machine's spare cycles to this great anonymization network, then much additional trust might accrue.

CONCLUSION

It seems unlikely to this author that any of the solutions presented in this article for the great search engine invasion of collective privacy will come to pass in a recognizable form. My goal in this article is to bring certain facts to the attention of the networked community and to offer examples of the range of reactions which could be considered.

The facts are as follows:

The Internet search engine oligarchy provides a great benefit to the user community, that of turning the Web into an associative memory of giant scope. One consequence is that certain institutions (like ICANN and its entire DNS system, or, for that matter, the whole infrastructure of human-readable URLs) become much less important to the Internet as it is actually used. As long as there is enough of a hierarchical network structure for search robots to crawl, most people will continue to use the associative web interface and ignore what is going on underneath.

Another consequence is that this small group of service providers has good expectation of a solid future revenue stream, but they have also access to an entirely unexpected source of potential revenue (and *power*): the data-mined contents of their log files. Simple mining yields for them a timely, subtle, and fine-grained mind map of the collective consciousness of the entire networked world. Interestingly, this data depends not at all on PII, so it is privacy-invasive only against the more inclusive concept of GII, which turns certain standard assumptions of privacy advocates on their heads.

The moral of this story, which I hope we will learn, is as follows:

GII is important, and its misuse can be as disturbing as that of PII. Privacy frameworks should therefore be built out of primitives and syntax using GII. Privacy policies should be based on white-lists, in that they explicitly list all that is to be allowed. Proactive steps on the part of one or more of the search engine providers in offering their data to the public or in sponsoring sophisticated new privacy technologies or simply in offering a public API to some of their internals and so permitting those concerned about GII to build their own solutions would be of enormous benefit. The top few search engines have such a strong position that they could be entirely neutral with regard to all proposals for changing DNS or other named structures on the Internet, knowing that their associative interface provides a complete end run around such structures. It would be nice if they, in providing this usable overlay on network layout, built in at the lowest levels a respect for privacy and confidentiality. And the government could have a role to play in this area, either feeding off this fantastic new data source or as a defender of the public's freedoms and collective privacy and benevolent director of future features and structure of the associative interface to the World Wide Web.

REFERENCES

Boyle, J. 1992. A theory of law and information: Copyright, spleens, blackmail, and insider trading. *California Law Review* 80:1413–1540.

Bylund, A. 2006. Attorney General v. Google going to court. *Ars Technica,* February 27, 2006. http://arstechnica.com/news.ars/post/20060227-6275.html (accessed April 7, 2006).

Dingledine, R., Mathewson, N. and Syverson, P. 2004. *Tor: The second-generation onion router.* Proceedings of the 13th USENIX Security Symposium, San Diego, CA, August.

Federrath, H. 2006. JAP project web pages. http://anon.inf.tu-dresden.de (accessed April 7, 2006).

Freedman, M., and Morris, R. 2002. Tarzan: A peer-to-peer anonymizing network layer. *Proceedings of the 9th ACM Conference on Computer and Communications Security* (CCS '02).

I2P Team. 2006. I2P project web pages. http://www.i2p.net (accessed April 7, 2006).

McCullagh, D. 2006. Judge to help feds against Google. *CNET news.com,* March 14, 2006. http://news.com.com/2100-1028_3-6049493.html (accessed April 7, 2006).

Poritz, J. 2007. *Who?ogle: An associative interface to the web which preserves individual and community privacy.* Manuscript under preparation.

Price, G. 2006. Court documents & summary of United States versus Google over search data. *Search Engine Watch.* January 19, 2006. http://blog.searchenginewatch.com/blog/060119-161802 (accessed April 7, 2006).

Rennhard, M., and Plattner, B. 2002. Introducing MorphMix: Peer-to-peer based anonymous internet usage with collusion detection. *Proceedings of the Workshop on Privacy in the Electronic Society*, Washington, DC, November.

World Wide Web Consortium. 2005. P3P 1.1 1 July 2005 Public Working Draft. *Platform for Privacy Preferences (P3P) Project web site*, July 1, 2005. http://www.w3.org/TR/2005/WD-P3P11-20050701/ (accessed April 7, 2006).