The Problem of Anonymous Vanity Searches

Christopher Soghoian*

School of Informatics
Indiana University Bloomington
csoghoia@indiana.edu

Abstract

This paper explores privacy problems related to search behavior conducted using public search engines. Specifically it exposes problems related to unintentional information leakage through a vanity search - which is a search for information about one's self. We begin by discussing recent events which have made this problem extremely topical. We introduce a number of existing technologies, such as Tor and TrackMeNot, that aim to protect users' privacy online and explain how each of these programs fails to protect users against the specific risks related to self-search. We highlight the inherent information asymmetry in the relationship between search engines and their users which makes it almost impossible to create cover traffic good enough to blend into. We conclude by exploring other avenues for protecting user privacy online.

1 Introduction

On August 4, 2006, America Online (AOL) released the search records of 650,000 of its users to the Internet. The stated goal of this release was to aid the research community at large. In an effort to protect user privacy, the records were "pseudonymized" by replacing each individual user account I.D. with a unique random number. While this severed the connection between the search records and AOL account information such as a user's name and address, the ability to link individual's searches over multiple sessions remained. Almost a quarter of all Internet users engage in the common practice of searching for their own name or online nickname (which can include an email address, instant messenger ID, or MySpace address) on the Internet [25]. Due to this behavior, often called a vanity search or self-googling [20], it was possible for journalists from the New York Times to reveal the identity of user 4417749 to be Thelma Arnold, a 62-year-old widow from Lilburn, Georgia after linking together all of her vanity searches contained in AOL's pseudonymized records [12].

In court papers filed on January 18 2006, the US Justice Department revealed that Google refused to co-operate with a previous year's subpoena for user search records. Data requested included one million random indexed web page addresses and records for all searches performed during a one week period. In its refusal, Google cited the privacy rights of its customers and the risk of revealing company trade secrets [35]. Lawyers from the Justice Department argued that they needed the information to prepare their defense of the 1998 Child Online Protection Act, a law which the courts had previously struck down, stating that it was too broad and had the potential to prevent adults from accessing legal pornography sites. Both Yahoo and MSN silently complied with the request, and had Google not publicly refused to do so, it is likely that the subpoenaed information would have been handed over without the public being notified.

Testifying in front of a Senate panel on September 19, 2006, US Attorney General Albert Gonzales urged the Senate to pass legislation requiring Internet service providers (ISPs) to keep two years worth of detailed log data on their customers' online browsing habits. He stated that the growing threat of child pornography made it necessary to keep such information for subsequent law enforcement investigations [43].

Search engine logs are increasingly being sought in

*The author of this work was partially funded by a scholarship from Google and the Hispanic Scholarship Fund. This paper is not part of any research conducted as part of the author's internship with Google. It is not sponsored or authorized by Google, and it does not reflect or take into account Google's views in this area. This paper was written using publicly available information and is based upon the personal opinions of the author.

other cases. In one case, a murder suspect's search records were produced in court to prove that he had searched for the words "neck," "snap," and "break" before killing his wife [3]. Others have speculated that is only a matter of time before search records are subpoenaed in civil cases, including divorces [36].

2 Protecting privacy online

Shortly after the AOL search records were released, the Electronic Frontier Foundation released a set of recommended best practices for safe search behavior on the Internet [26]. These include:

- Don't put personally identifying information in your search terms
- Don't use your ISP's search engine
- Don't login to your search engine or related tools
- Block "cookies" from your search engine
- Vary your IP address
- Use web proxies and anonymizing software like Tor [22]

While their first piece of advice is to never search for information on one's self, the EFF seems to accept that even though conducting vanity searches may be risky that people will do them anyway. They suggest that such users should heed the rest of their advice, or at the least, do those sensitive searches from another computer than the one used for normal search activity.

In this paper, we argue that certain terms placed together in a search log, even without identifying information such as an IP address, cookie, or user-account still reveal far too much information about the user. There is a considerable difference between surfing the Internet privately, and not revealing your identity through your search behavior. Had the AOL customers who were identified turned off their cookies, and used an anonymizing proxy such as Tor, it is extremely unlikely that the New York Times would have been able to link together their individual searches from the released log data. Yes, analysis of the search search logs would reveal the fact that *someone* was searching for "Britney Spears", but the identity of that someone would remain a secret.

Some search terms are so sensitive that their mere presence in logs are likely to cause alarm to the user. Examples of this type of query may include the combination of a user's name or online nickname with words such as "HIV", "rape", "sexual harassment", "sex offender" and "homosexual". After-the-fact analysis of search logs will not confirm that the subject of the sensitive search query was also the person issuing the search. Neverless the mere fact that someone was looking for such combinations is interesting and extremely sensitive in itself.

3 A conflict of interest in the advertising business

The major search engines depend upon advertising for their revenue. Google provides free email, free wireless Internet access [9] and mobile-phone based GPS mapping services [38] all so that it can offer targeted advertisements to customers most likely to respond to them. If Google can build a higher-quality data set of customer information, they can charge more per advertisement, whilst also gaining a significant market advantage over the other search engines.

Some have claimed that loss of privacy on the Internet has allowed merchants to perform more effective price discrimination on their customers: where the lack of privacy allows the merchants to learn exactly how much each customer is willing to pay for a product [39]. While Google's search engine and email products are clear market leaders in terms of quality and functionality, users are not given the choice between a subscription charging (yet privacy preserving) model and the more common advertising supported (and privacy denying) model. Instead, users are given the binary choice of either using the products, with the advertisements and potential intrusions into their privacy, or not using them at all. If users wish to attempt to preserve their privacy and avoid advertising, they must take matters into their own hands. They cannot ask the search engines to take care of this for them, nor can they financially compensate the search engines for the potential lost revenue due to the loss of accurate user data.

Users have struck a Faustian bargain of sorts with the major search engines. They seem to be willing to put up with advertising and a wholesale loss of privacy, assuming that they are even aware that it is happening, for free access to the services that search engines offer. Just as a generation of American college students have shared their personal financial data with credit card companies in order to get a free t-shirt or

pizza [40], Internet users seem to be engaged in a similar mass exchange of their personal information for access to accurate search results.

Most Internet users place blind trust in search engines to present only the best or most accurate unbiased results to them [33]. In the vast majority of cases, users are not seeking out advertisements when they go to a search engine. They are instead trying to locate the results that most accurately match their search query. Three out of five searchers have no idea that search engines are paid for some of the results in their listings [33] and only one in six Internet users are able to tell the difference between paid advertisements and unbiased search results. Thus, the line between paid advertisements and genuine search content can be extremely blurry, a feature that some search engines take advantage of to increase the click-through rate of their advertisements [42].

The search engines must be very careful to ensure that advertisements do not become too intrusive or disruptive. In particular, flash-based or javascript advertisements that take control a user's screen can be irritating enough that some users then seek ways to disable them [47]. Popup advertisements have proven to be annoying enough that over 80% of users with high-speed Internet connections now employ popup ad-blocking technology — a 100% increase during the past two years [18].

Even Google's relatively un-intrusive text-based advertising has inspired a number of avoidance technologies [1, 2]. Savvy people have been using technologies that allow them to skip advertisements for a number of years. Use of such programs by a tiny percentage of users does not have a significant, or even measurable impact upon the search engines' revenue. However, consider the cases of television advertisement skipping with Replay TV/Tivo and the P2P file-sharing technologies made mainstream by Napster. When this kind of easy to use technology can enable an average user to bypass the copyright/advertising systems in place, it threatens to destabilize the entire business model that companies' revenue streams are built upon.

Google would no doubt prefer that each user sign into one of their services before searching - thus providing them with the data (should they wish to use it) to link together searches, advertising clicks and browsing habits to an individual person across multiple sessions and from different computers. Likewise, simply letting Google store a persistent cookie on one's computer allows them achieve this, albeit with a significantly lesser quality of user data. It is not surprising that each of the EFF's private searching recommendations threaten Google's bottom line, should these measures be employed by the masses, by denying Google the ability to mine the search logs for detailed information on users.

As of September 2006, the highest valued search terms in Google's AdWords advertising system are related to medical class action lawsuits and other legal problems [45, 5]. Yet, these are the same kinds of searches that users are most likely to want to protect from prying eyes, or worse, a subpoena after the fact. One of our main goals in this paper is to highlight this relationship - that Google's ability to serve fine-grained advertising (and thus achieve higher revenues) directly competes with the methods by which a user can achieve anonymity and preserve what little is left of their privacy online.

4 TrackMeNot

Shortly after the AOL incident became public, New York University researchers Daniel C. Howe and Helen Nissenbaum released a tool named TrackMeNot (TMN) [31] that aims to to protect user search privacy. Their tool is an extension to the Firefox web browser, and initiates randomized search queries in the background to a number of commercial search engines. These searches, issued over random periods of time, aim to lose the user's real searches in a cloud of "ghost queries" and as the authors describe, *"significantly [increase] the difficulty of aggregating such data into accurate or identifying user profiles."*

4.1 The difficulty of faking traffic

> *The current version of TMN begins only with a small seed list (taken from lists of most-frequent search terms published by the search companies). Using these terms as 'seeds', each TMN client dynamically 'evolves' its query list by parsing likely search terms from the results of each query and swapping these back into its 'Current-Query' list.* – The TrackMeNot FAQ.

There is very little risk to the user if someone else learns that they are searching for the hot topic, or immensely popular search term, of the day. Any hot topic is sought by millions of other users, thus one is unlikely to be embarrassed or suffer otherwise if

searches for those terms become public. As a concrete example, one wants privacy when searching for information on breast cancer, and not when searching for Britney Spears.

TMN creates cover traffic, or "ghost queries" as the authors describe them, by submitting queries to search engines containing terms from its Current-Query list at random intervals. The actual level of plausible deniability provided to users could prove to be rather problematic, due to the fact that users do not submit their own legitimate queries at random intervals. They tend to be very bursty, sending multiple searches over a short period of time, followed by longer periods of web-browsing. It should be fairly easy for the search engine companies to focus on users' real search activity, by filtering out all searches that match the behavior exhibited by TMN.

The major search engines have extremely accurate data on the search frequency of various terms. Portions of this information, albeit generalized to remove specific traffic figures, are even made public through services like Google Trends [4], and Google Zeitgeist [7]. TMN will find itself sticking out if the search queries it issues significantly diverge from the norm and deviate from the standard frequency of search terms in the population at large - or even those other users in one's geographic location.

Users often do not find the information they want from the first clicked upon link that they reach via the search engine. A rather limited eye-tracking study by the firm Enquiro found that users returned to the search page and then clicked on additional items listed by the search engine over 49% of the time [30]. This behavior, which they dubbed "pogo sticking", suggests that the first link returned is often not enough to meet the user's needs. Researchers analyzing the AOL search data have also noted a strong tendency for people to refine their searches. If the first page of results does not deliver what they're after, they will refine the search terms in an effort to produce better results [24].

While only Google, Yahoo and other search engines will have true data for this kind of behavior, it is logical to assume that any kind of automated search behavior through tools such as TMN will diverge from the norm, even if they attempt to replicate this behavior (which they currently do not), simply due to the fact that they lack accurate data on the "norm".

4.2 Unintended consequences

Search engines such as Google monitor the click-through-rates (CTR) of the advertisements they display - in web pages containing search results and elsewhere. Advertisements that consistently suffer from low click through rates will be punished, and in time, no longer be displayed to the user, no matter how much the advertiser pays per click [6]. In attempting to mask the user's real search behavior, TMN will in fact be inadvertently performing a specific kind of denial of service against Google's advertisers: impression spam [37].

In addition to performing this denial of service, TMN will no doubt stand out due to the fact that its ghost queries will have a 0% click through rate. Were TMN to attempt to fix this behavior by clicking on advertisements at random (yet with the same approximate rate that real users do), they would soon find themselves engaged in a different, yet equally unfriendly behavior towards Google: they would then be engaging in mass click-fraud and would be defrauding advertisers for each phantom search and click.

Google's terms of service [8] clearly state that the kind of behavior that TMN engages in is forbidden:

> *You may not send automated queries of any sort to Google's system without express permission in advance from Google.*

At the very least, Google would be perfectly within its rights to terminate the user accounts of customers who install TMN. Given that one can use Google's search engine without a user account, this is probably not a problem for the vast majority of Internet users. However, for those 8 million plus [29] Gmail users who have entrusted Google with their email data - being kicked off the service could prove to be extremely painful.

The legality of TMN's techniques is not clear, and use of it and similar tools may expose users to legal risks. The tort claim of trespass to chattel has been successfully used against resource-hogging Internet tools in the past [44, 15]. While the laws surrounding click-fraud are not yet mature, Google has brought a number of cases to trial against people for attempting to defraud their AdSense advertising system [23, 49]. The legal issues that surround TMN's behavior are beyond the scope of this paper, although they merit thorough analysis. Researching them is left as an exercise to the legally inclined reader.

4.3 The lack of a feedback loop

One of the major problems for TrackMeNot's developers may be that they will never be told if they have succeeded. The major search engines have very little incentive to share information with them. If the search engines are indeed able to detect the presence of TMN, or worse, filter out the automated searches from those legitimate queries, TMN's authors and users will never know. Google and others will be able to successfully use this data, and potentially gain even more value from it - with the knowledge that it is data that users value enough to go out of their way to protect. Needless to say, if the National Security Agency or some other government agency were to get their hands on Google's log data, they could also perform the same analysis — silently, of course.

This absence of a feedback loop will make it extremely difficult for the TMN creators to evolve their technology, as they will be denied knowledge of which particular behavior is giving them away.

If TMN is widely adopted, and actually becomes a significant burden on the search engines' network resources, the search engines may have to adopt a more active, and verbose approach by banning those users of the product. In the past, Google has blocked technologies such as Tor when known Tor IP addresses were submitting too many queries [48]. In this case, they offered two options to users wishing to search from a Tor IP address: Solve a CAPTCHA [50], or be blocked. One can easily imagine Google deploying a similar system for all the queries sent by a TMN user (those automated and those legitimate), if they are able to easily detect the presence of the extension through search log analysis. This could instantly break TMN, unless of course, users are willing to solve a CAPTCHA for every fake search submitted by the extension - a rather unlikely scenario.

4.4 Information asymmetry

The relationship between search engines and their customers can be seen as a classic case of information asymmetry [11]. The search engines know how often each and every word is searched for, how many searches any one user issues on average per session, how much time there is between sessions and individual searches, how many advertisements are clicked on per session and how often advertisements are expected to be clicked by a given user. Furthermore, they keep the vast majority of this information to themselves, since their competitors, those actors trying to actively defraud them (e.g. those committing click fraud), and those users trying to hide their search behavior, would all love to have this data.

It is because of this huge gap in information, that technologies such as TrackMeNot are doomed to failure. They lack accurate Internet behavior data that is essential in helping a user mask her searches in a cloud of effective cover traffic. Their attempts to maintain privacy without the necessary information on what cover traffic "should" look like may cause their users stand out even more than if they had not attempted to evade the search engines' watchful eyes in the first place.

5 Tor

> *"Tor is a network of virtual tunnels that allows people and groups to improve their privacy and security on the Internet... Individuals use Tor to keep websites from tracking them and their family members, or to connect to news sites, instant messaging services, or the like when these are blocked by their local Internet providers."* [10]

Tor [22] allows users to mask the link between their own network activity (search behavior, web browsing or instant messaging) and any logs kept by webmasters, or worse, oppressive governments. Servers receiving web requests see only a Tor exit node, and are unable to learn the actual IP addresses of the users initiating queries.

Use of the Tor anonymity proxy to sever the link between a user and an identifying IP address (and in tandem, disabling cookies in the browser) restricts the search engines' abilities to link an individual's searches together. Users who are solely concerned with protecting their search behavior against log analysis by the search engines do not necessarily have to use Tor - any proxy server will work. Ironically, simply being an AOL customer will provide them with this protection, as long as they do not also use AOL's search service; AOL funnels all of their users' Internet traffic through a few proxy IP addresses.

The list of Tor exit nodes is public, and as Google's past behavior of selectively blocking Tor servers at times has shown, they subscribe to this list. Presumably, if a user issues a non-common search via a Tor server, even with their cookies turned off, and then a few minutes later issues a refined but similar search query via a different Tor exit node, Google can link

those two searches together. While this link between the two queries is not certain, as is the case when cookies are present, it still has the potential to reveal information that the user expected to remain private. This technique only works for uncommon search queries. Yet, as we have previously explained, these are often the searches that users wish to protect the most.

5.1 Why Tor alone cannot protect vanity searches

The combination of an anonymizing proxy such as Tor and a cookie-less browser session does much to protect user search activity on the Internet, and in particular, the linking of multiple queries during one or more sessions. As we have discussed in earlier sections, a search for one's own name combined with culturally or politically delicate terms can be extremely sensitive search information. While Tor will deny the search engine the knowledge of who issued the search, the mere fact that such a search was issued is extremely valuable information, and as such, the use of Tor is not enough to protect these kind of queries.

While Tor does much to hide users' network information from the websites hosting content, it introduces a few other problems. Nefarious Tor exit node operators have the ability to view, or worse, modify the data that they relay. In at least one published case [17], a server operator placed flash based "web bugs" into web pages served in order to reveal the true source of the web requests. At the very least, an exit node operator has the ability to view users' search requests. Tor users must worry about the exit node operators keeping and later disclosing potentially sensitive searches in addition to search engine logging.

5.2 Information leakage

It is commonly accepted amongst computer security experts that encrypting one's most sensitive communications is not enough, and in fact, can be extremely dangerous. The mere act of encrypting only sensitive messages leaks valuable information to outsiders watching the wire. They can see there are some messages that are important enough to try to protect. Techniques such as traffic analysis [19], when employed against a user who only encrypts important messages can prove to be extremely effective.

Applying this idea to the problem of sensitive searches, it is quite clear that to achieve privacy one must apply the protection methods to all searches, and not just those that one deems to be sensitive.

Many privacy enhancing technologies impose rather steep costs on the user, such as lack of convenience due to the absence of cookie tracking across sessions or in the case of Tor, a significant increase in traffic latency as encrypted packets bounce across the globe before they reach their final destination. While users may be willing to put up with this in order to gain privacy protections for sensitive searches, they may be less willing to put up with this for the bulk of their less sensitive traffic. This selective use of Tor and other technologies will unfortunately leave users vulnerable to traffic analysis by those with wiretap or network level access to users' communication data.

5.3 Traffic analysis and pornography

It has been noted by some observers that pornography drives technology [32]. Porn consumers are often the early adopters of technology, and are often willing to put up with beta quality products that other users refuse to use. One example of this is the initial attempts to stream video on the web. Users of adult content were the main audience willing to put up with excruciatingly slow downloads of jittery, low quality videos. These early adopters wanted better video quality, and their demand arguably drove the market to develop better technology that eventually reached the masses.

There is a notable absence of good information on the traffic that the Tor network carries, primarily because collecting such data in the US could put researchers into legal jeopardy [27]. One recent study performed outside the US suggests that one of the primary uses of Tor is to transfer pornographic content [17]. If the anecdotal evidence presented in this report accurately describes the Tor network, it is thus likely that Tor traffic has a higher porn-per-packet ratio than "normal" Internet data. Given the assumption that a Tor user is probably interested in adult content, Google could allow advertisers to bid on keywords displayed to users coming from Tor exit nodes. "Tor targeting" would surely seem valuable to pornographic advertisers and would be a way for them to guess user intent without knowing anything else about a market segment that fiercely guards its privacy. This is just one example of the kind of user data that be inferred from the use of a privacy-preserving system, even when encryption is used.

6 Other Options

TrackMeNot and Tor are not enough to protect vanity searches. We now explore a few other options.

6.1 Searching on encrypted data and PIR

There has been a significant amount of research into the areas of searching on encrypted data [46] [14] [28] and private information retrieval (PIR) [13] [16]. Search engines are focused on collecting *more* information on customers, and not in protecting their privacy. Their business models are built on the practice of mining and exploiting user online behavior data. Thus, while research in these areas is interesting from an academic perspective, there is very little incentive for a service provider to expend the significant resources required to support PIR or encrypted data searching, especially given that these technologies would hinder their primary goal of collecting customer data. The burden of attempting to protect their privacy thus falls on the user.

6.2 Doing the search yourself

Local searching, a surprisingly simple technique, may prove to be extremely useful at maintaining user privacy online. For fairly uncommon names, a user can simply request every single web page containing their name or online nickname from a search engine, preferably, using an anonymizing proxy such as Tor. They can then download a copy of each of these web pages to their own computer, and perform a local search on those web pages for the sensitive terms they are looking for. This technique also has the added property of being "information theoretic" secure [34].

This method has several shortcomings. A person with a unique name, but a fairly major web presence, may find that there are far too many web pages citing their name to download. Likewise, someone with a common enough name may encounter too many false positives when attempting to save a local copy of every page referencing them. In both of these cases, a complete local search may prove to be impossible. Finally, while most major operating systems include the ability to search through a large number of directories and files for one or more phrases, in many cases, it is not yet easy to use. The technology required to download every instance of a user's name from the Internet requires automation software, something not readily available to the masses. Thus, effective local search is sadly not yet an option for the vast majority of users.

Local search lacks the bells, whistles, and ease of use that Google and the other search engines provide. Yet, it remains far safer in terms of user privacy than sending a sensitive vanity search out onto the Internet.

6.3 Pre-announcing your strategy

Technologies such as TrackMeNot pose a specific threat to the advertising dependant search engines. Instead of merely leeching Google's network and computing resources as those who search without viewing ads, TMN's method of achieving anonymity has the potential to cause significant collateral damage (via click-fraud) to Google's advertising system by requesting web pages with advertisements that will never be clicked. While TMN's goals are noble, their methods can cause unintentional harm to Google and others.

One simple technique that could solve this problem of collateral damage would be for TMN users to disclose their intentions ahead of time. By marking all search requests - both genuine and automated - with an additional argument in the query sent to Google's servers, they could significantly reduce Google's incentive to locate and neutralize TMN traffic. All TMN-user originating queries could then be easily excluded from the advertising system. While TMN's network activity will probably not be too difficult to differentiate from real user traffic, this simple technique at least reduces Google's incentive to do so.

Just as webmasters can currently include a "robots.txt" file on their websites to notify web-crawlers of their desire to not be crawled, adding an additional flag to the search query would be a polite and reasonable way for TMN and other privacy preserving systems to communicate intent to Google.

The downside to this flagging technique is that by adding the flag, a user instantly announces himself as a TMN user. This then reduces his anonymity set [41] to that of all TMN users, a small minority of all clients. Conversely, without the flag, he could potentially be any of Google's millions of search users. TMN's current behavior is anything but covert, and thus, he has probably already reduced his anonymity set by using TMN, even if he has not explicitly announced it.

6.4 Be your own proxy

As others have noted, anonymity loves company [21]. Users gain privacy and plausible deniability when they

can blend into a large crowd of other users. TrackMeNot allows users to claim, "Google knows who I am, but if I send enough fake queries, they won't know which are real, and which are not." Tor and other anonymizing proxies instead adopt the philosophy of "If I can keep my network location secret from Google, then while they'll know exactly which searches are being issued, they won't know who is initiating them." Additionally, Tor users not only reveal their search information to Google, but also reveal it to the operator of a Tor exit node, who might not be trustworthy.

Falsifying search queries well enough to make them seem like genuine user queries can be much harder than it initially seems. Instead of trying to create a fake stream of queries, why not use legitimate queries from real users: Become a Tor exit node.

When using the Tor network, users risk nefarious exit-node operators seeing their search queries. By assuming the role of an exit-node operator, and using one's own exit node for queries to Google and the other search engines, a user can be sure that no other proxy administrator will learn of her search data. Using one's own Tor exit node is probably a very bad idea in terms of overall network anonymity and is not recommended for general web surfing. However, for vanity searches, perhaps it is not such a bad idea.

7 Conclusion

In this paper, we explored the problem of sensitive information leakage due to vanity searches on the Internet. We highlighted the inherent conflict of interest in the advertising/search engine business in which Google's ability to serve fine-grained advertising (and thus achieve higher revenues) directly competes with the methods by which users can achieve anonymity, preserving what little is left of their privacy online. We also highlighted the state of information asymmetry between the search engines and users which makes it almost impossible to create artificial search queries that are indistinguishable from those submitted by real users. We show that technologies such as TrackMeNot may expose their users more through their attempts to create cover traffic than if they had not been used in the first place. We further identify how anonymizing proxies such as Tor are themselves not enough to protect vanity searches, and discuss several other potential solutions, none of which are ideal or 100% foolproof.

Effective privacy protection for vanity searches is a difficult problem. Current privacy-preserving systems, although appearing to solve the problem, will only exacerbate it. Existing technologies cannot be trusted alone to provide private searching functionality to users. While still an unresolved issue, we are enthusiastic that future research in the area will be promising and fill this dire need for privacy-preserving searches.

Acknowledgements

Many thanks to Kelly Caine, L. Jean Camp, Allan Friedman, Markus Jakobsson, Sid Stamm and Katherine Townsend for their helpful comments.

References

[1] Customize Google - Firefox Browser Extension. http://www.customizegoogle.com/.

[2] Userscripts.org: Hide Google Adsense Ads. http://userscripts.org/scripts/show/675.

[3] WRAL News: Petrick Googled 'Neck,' 'Snap,' Among Other Words, Prosecutor Says, November 9 2005. http://www.wral.com/news/5287261/detail.html.

[4] About Google Trends, 2006. http://www.google.com/intl/en/trends/about.html.

[5] CyberWire: Updated: Highest Paying AdSense Keywords, March 23 2006. http://www.cwire.org/2006/03/23/updated-highest-paying-adsense-keywords/.

[6] Google AdWords: Learning Center, 2006. http://www.google.com/adwords/learningcenter/text/18754.html.

[7] Google Press Center: Zeitgeist, 2006. http://www.google.com/press/zeitgeist.html.

[8] Google Terms of Service for Your Personal Use, 2006. http://www.google.com/terms_of_service.html.

[9] Google WiFi Mountain View, 2006. https://wifi.google.com/support.

[10] Tor: Overview, August 22 2006. http://tor.eff.org/overview.html.en.

[11] George A Akerlof. The market for 'lemons': Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, August 1970. http://ideas.repec.org/a/tpr/qjecon/v84y1970i3p488-500.html.

[12] Michael Barbaro and Tom Zeller Jr. A Face Is Exposed for AOL Searcher No4417749. *The New York Times*, 9 August 2006. http://www.nytimes.com/2006/08/09/technology/09aol.html.

[13] Amos Beimel, Yuval Ishai, Eyal Kushilevitz, and Jean-Francois Raymond. Breaking the $O(n^{1/(2k-1)})$ Barrier for Information-Theoretic Private Information Retrieval. In *IEEE Symposium on Foundations of Computer Science*, pages 261–270, 2002.

[14] Dan Boneh, Giovanni Di Crescenzo, Rafail Ostrovsky, and Giuseppe Persiano. Public Key Encryption with Keyword Search. In C. Cachin, editor, *Proceedings of Eurocrypt 2004*, volume LNCS 3027, pages 506–522, 2004. http://crypto.stanford.edu/~dabo/papers/encsearch.pdf.

[15] Dan L. Burk. The Trouble with Trespass. *Journal of Small and Emerging Business Law*, 4(1), 2000. http://www.isc.umn.edu/research/papers/trespass-ed2.pdf.

[16] Benny Chor, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. Private Information Retrieval. In *IEEE Symposium on Foundations of Computer Science*, pages 41–50, 1995.

[17] Andrew Christensen. Practical Onion Hacking: Finding the real address of Tor clients, October 2006. http://www.packetstormsecurity.org/0610-advisories/Practical_Onion_Hacking.pdf.

[18] Thomas Claburn. Consumer Use Of Ad-Blocking Technology Doubles. *Information Week*, 5 December 2006. http://www.informationweek.com/internet/showArticle.jhtml?articleID=196601694.

[19] George Danezis. Introducing Traffic Analysis: Attacks, Defences and Public Policy Issues, December 2005. http://homes.esat.kuleuven.be/~gdanezis/TAIntro.pdf.

[20] Daniel Dasey. A quick self-Google once a day to guard your reputation. *The Sydney Morning Herald*, May 23 2004. http://www.smh.com.au/articles/2004/05/22/1085176043551.html.

[21] Roger Dingledine and Nick Mathewson. Anonymity loves company: Usability and the network effect. In *Proceedings of the Fifth Workshop on the Economics of Information Security (WEIS 2006)*, Cambridge, UK, June 2006. http://freehaven.net/doc/wupss04/usability.pdf.

[22] Roger Dingledine, Nick Mathewson, and Paul Syverson. Tor: The second-generation onion router. In *Proceedings of the 13th USENIX Security Symposium*, August 2004. http://tor.eff.org/tor-design.pdf.

[23] Dan Elgin. BusinessWeek. *The Vanishing Click-Fraud Case*, 4 December 2006. http://www.businessweek.com/technology/content/dec2006/tc20061204_923336.htm.

[24] Geoffrey Faivre-Malloy. Search Engine Optimizer Blog: AOL Search Data, August 22 2006. http://www.seomoz.org/blogdetail.php?ID=1320.

[25] Deborah Fallows. *Internet Search Users*, page 8. Pew Internet and American Life Project, January 23 2005. http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf.

[26] Electronic Frontier Foundation. Six tips to protect your online search privacy, 2006. http://www.eff.org/Privacy/search/searchtips.php.

[27] The Electronic Frontier Foundation. Legal FAQ for Tor Server Operators, April 25 2005. http://tor.eff.org/eff/tor-legal-faq.html.en.

[28] Philippe Golle, Jessica Staddon, and Brent Waters. Secure conjunctive keyword search over encrypted data. In M. Jakobsson, M. Yung, and J. Zhou, editors, *Proceedings of the Second International Conference on Applied Cryptography and Network Security (ACNS-2004)*, pages 31–45. LNCS 3089, 2004. http://crypto.stanford.edu/~pgolle/papers/conj.pdf.

[29] Saul Hansell. In the Race With Google, It's Consistency vs. 'Wow'. *The New York Times*, 24 July 2006. http://www.nytimes.com/2006/07/24/technology/24yahoo.html.

[30] Gord Hotchkiss. Tales of pogo sticks, bouncy serps and sticky pages, September 11 2006. http://www.searchengineguide.com/hotchkiss/2006/0911_gh1.html.

[31] Daniel C. Howe and Helen Nissenbaum. TrackMeNot, 2006. http://mrl.nyu.edu/~dhowe/TrackMeNot/faq.html.

[32] Peter Johnson. Pornography drives technology: Why not to censor the internet. In *Federal Communications Law Journal*, volume 49:1, page 217, November 1996. http://www.law.indiana.edu/fclj/pubs/v49/no1/johnson.html.

[33] Leslie Marable. Consumer Reaction to Learning the Truth About How Search Engines Work. *Consumer Reports WebWatch*, 30 June 2003. http://www.consumerwebwatch.org/dynamic/search-report-false-oracles-abstract.cfm.

[34] Ueli Maurer. Information-theoretic cryptography. In Michael Wiener, editor, *Advances in Cryptology — CRYPTO '99*, volume 1666 of *Lecture Notes in Computer Science*, pages 47–64. Springer-Verlag, August 1999.

[35] Declan McCullagh. CNET News.com: Google, feds face off over search records, March 14 2006. http://news.com.com/Google,+feds+face+off+over+search+records/2100-1028_3-6049262.html.

[36] Declan McCullagh. CNET News.com: When Google is not your friend, February 3 2006. http://news.com.com/FAQ+When+Google+is+not+your+friend/2100-1025_3-6034666.html.

[37] Rob McGann. Impression spam worries google advertisers – ClickZ News, 24 February 2005. http://www.clickz.com/showPage.html?page=3485386.

[38] Alex Medina. Official Google Blog: Get Lost!, November 9 2006. http://googleblog.blogspot.com/2006/11/get-lost.html.

[39] Andrew Odlyzko. Privacy, economics, and price discrimination on the internet. *Proceedings of the 5th international conference on Electronic commerce*, pages 355–366, 2003. http://www.dtc.umn.edu/~odlyzko/doc/privacy.economics.pdf.

[40] Lynn O'Shaughnessy. Credit cards offer college students early danger lesson. *The San Diego Union-Tribune*, 24 September 2006. http://www.signonsandiego.com/uniontrib/20060924/news_lz1b24oshaugh.html.

[41] Andreas Pfitzmann and Marit Hansen. Anonymity, unobservability, and pseudonymity: A consolidated proposal for terminology, July 2000. http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.

[42] Associated Press. Users Confuse Search Results, Ads, 24 January 2005. http://www.wired.com/news/culture/0,1284,66374,00.html.

[43] Associated Press. Gonzales Wants New Web Rules, Attorney General: ISPs Should Preserve Customer Info To Help Fight Kid Porn, September 19 2006. http://www.cbsnews.com/stories/2006/09/19/politics/main2023209.shtml.

[44] Pamela Samuelson. Unsolicited communications as trespass? *Commun. ACM*, 46(10):15–20, 2003. http://www.ischool.berkeley.edu/~pam/papers/acm_vol46_p15.pdf.

[45] John Battelle's Searchblog. Highest Paying AdWords, March 26 2006. http://battellemedia.com/archives/002444.php.

[46] Dawn Xiaodong Song, David Wagner, and Adrian Perrig. Practical Techniques for Searches on Encrypted Data. In *IEEE Symposium on Security and Privacy*, 2000. http://www.ece.cmu.edu/~dawnsong/papers/se.pdf.

[47] Tom Spring. Net Watchdog: The Most Annoying Online Ads. *PC World*, September 26 2006. http://pcworld.com/article/id,127207-page,1-c,topics/article.html.

[48] Danny Sullivan. Search Engine Watch: More On Google & Blocking Privacy Proxies, September 8 2006. http://blog.searchenginewatch.com/blog/060908-080437.

[49] David A. Vise. The Washington Post. *Clicking To Steal*, 17 April 2005. http://www.washingtonpost.com/wp-dyn/articles/A58268-2005Apr16.html.

[50] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. CAPTCHA: Using Hard AI Problems For Security. In *Proceedings of Eurocrypt*, pages 294–311, 2003. http://www.cs.cmu.edu/~biglou/captcha_crypt.pdf.