Oxford University Press, Inc., publishes works that further Oxford University's objective of excellence in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi Kuala Lumpur Madrid Melbourne Mexico City Nairobi New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece Guatemala Hungary Italy Japan Poland Portugal Singapore South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

Oxford is a registered trademark of Oxford University Press
Oxford University Press is a registered trademark of Oxford University Press, Inc.

Library of Congress Cataloging-in-Publication Data
Lessons from the identity trail : anonymity, privacy and identity in a networked society / Editors : Ian Kerr, Valerie Steeves, Carole Lucock.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-19-537247-2 ((hardback) : alk. paper)
1. Data protection—Law and legislation. 2. Identity. 3. Privacy, Right of. 4. Computer security—Law and legislation. 5. Freedom of information. I. Kerr, Ian (Ian R.) II. Lucock, Carole. III. Steeves, Valerie M., 1959-
K3264.C65L47 2009
342.08'58—dc22

2008043016

1 2 3 4 5 6 7 8 9

Printed in the United States of America on acid-free paper

Note to Readers
This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is based upon sources believed to be accurate and reliable and is intended to be current as of the time it was written. It is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If legal advice or other expert assistance is required, the services of a competent professional person should be sought. Also, to confirm that the information has not been affected or changed by recent developments, traditional legal research techniques should be used, including checking primary sources where appropriate.

(Based on the Declaration of Principles jointly adopted by a Committee of the American Bar Association and a Committee of Publishers and Associations.)

LESSONS FROM THE IDENTITY TRAIL

ANONYMITY, PRIVACY AND IDENTITY IN A NETWORKED SOCIETY

EDITED BY IAN KERR, VALERIE STEEVES, AND CAROLE LUCOCK

OXFORD UNIVERSITY PRESS

exit node can prove, without the node owner's knowledge or intervention, that it did not originate the communication and does not harbor information that could be linked to the sender.[32] Exit node repudiation provides a method of retaining the anonymity of the sender while presenting a response to the pertinent question of legal liability for the exit node as well as the practical hassles of equipment seizure. We hope this innovation is helpful in preserving the legality of anonymity networks and decreasing the aversion to volunteer as operators of servers in these networks.

32. This approach is also consistent with the need to establish "probable cause" in preparing a warrant to search and seize a computer under U.S. law. "Probable cause" has been defined by the U.S. Supreme Court as the establishment of "a fair probability that contraband or evidence of a crime will be found in a particular place." (*Illinois v Gates* [1983] 462 U.S. 213 at 238).

23. TRACKMENOT: RESISTING SURVEILLANCE IN WEB SEARCH

DANIEL C. HOWE AND HELEN NISSENBAUM*

TrackMeNot (TMN) is a Firefox browser extension designed to achieve privacy in Web search by obfuscating users' queries within a stream of programmatically generated decoys. Since August 2006, when the initial version of TMN was made publicly available free of charge, there have been over 350,000 downloads. TMN protects Web users against data profiling by simulating HTTP search requests to search engines with queries extracted from the Web. In an attempt to mimic users' search behavior, this basic functionality is augmented with several technical mechanisms: dynamic query lists (with RSS-based initialization),

* Many individuals and institutions have contributed in essential ways to this paper. For critical feedback on earlier versions of this paper, we thank audiences at the Haifa Center of Law and Technology Conference on Law of Search Engines (2006); the Conference on Computer Ethics: Philosophical Equiry (2007); the Annual Meeting of the American Association, Eastern Division (2007); the A2K2 Conference, Information Society Project, Yale University (2007),; the Poynter Center, Indiana University, Bloomington (2007); and the Santa Fe Institute (2007). Additional thanks to Jinyang Li, Robb Bifano, the Mozilla foundation, MissingPixel™, and the NYU Media Research Lab. Thanks also to the reviewer for this volume, who guided us toward several key improvements. We are indebted for help with TrackMeNot itself to innumerable users around the world who cheered us, critiqued us, and generously offered marvelous tips. We extend a special thanks to Michael Zimmer for all these contributions, and more. Support for this project came from the NSF award CCR-0331542: *Sensitive Information in a Wired World* (or PORTIA: Privacy, Obligations, Rights in Technologies of Information Assessment). The idea for TrackMeNot was hatched in a series of stimulating conversations with PORTIA colleagues at a project retreat.

real-time search awareness, live header maps, burst-mode queries, and selective click-through. We describe each of these mechanisms, evaluate its strengths and weaknesses, and demonstrate how the consideration of values directly informed design and implementation. In the discussion section we conceptualize TMN within a broader class of software systems serving ethical, political, and expressive ends. Finally, we address why Web search privacy is particularly important and why TMN's approach, for the present, is both legitimate and necessary.

I. INTRODUCTION

In August 2005, public awareness of the ubiquitous practices of logging and analyzing users' Web search activities was heightened when front-page articles in the mainstream press revealed that the United States Department of Justice (DOJ) had issued a subpoena to Google for one week's worth of search query records (absent identifying information) and a random list of one million URLs from its Web index. These records were requested in order to bolster the government's defense of the constitutionality of the Child Online Protection Act (COPA), then under challenge. When Google refused the initial request, the DOJ filed a motion in a federal district court to force compliance. In March 2006, swayed by Google's arguments that the request imposed an unreasonable burden and would compromise trade secrets, undermine customers' trust in Google, and have a chilling effect on search activities, the court granted a reduced version of the first motion, ordering Google to provide a random listing of 50,000 URLs, and denied the second motion seeking the query records. One year later, however, the illusion that our Web searches are a private affair was further eroded when a news investigation revealed that in anonymized search query logs provided to the research community, the identities of certain searchers had been extracted from personal information embedded in search terms.[1] Other media reports followed detailing how the major search companies (Yahoo!, AOL, MSN, and Google) log, store, and analyze individual search query logs.

Setting aside the details of these two highly publicized cases, a few disquieting facts are evident: one, that search queries are systematically monitored, scrutinized, and indefinitely stored by search service providers; two, that for all we know, they are shared with third parties; and three, that policies governing these practices are unilaterally set by search companies with little indication,

1. Saul Hansell, "Marketers Trace Paths Users Leave on Internet," *New York Times*, September 15, 2006; Michael Barbaro and Tom Zeller, Jr., "A Face Is Exposed for AOL Searcher No. 4417749," *New York Times*, August 9, 2006.

or control, provided to individuals about what is done with their search records.[2] Since then, interest in the issue of search privacy has greatly expanded, drawing attention from citizens and advocacy organizations, scholars, and government agencies in the United States and beyond.[3] Responding to concerns surrounding the handling of search-query logs, search companies have offered several compromises, few of which, with the possible exception of those offered by Ask.com, have proved adequate or fully transparent. We believe these policies and practices challenge foundational moral and political principles of our society.

In Western liberal democracies, freedom of expression and of association are among a set of core values protected directly through laws (for example, the U.S. Constitution) and indirectly in the design of public institutions. Protection of liberties is also extended to activities considered supportive of these values, such as education, research, reading, and communication. As many of these activities have moved online, so the recognition has grown that robust civil rights protections are required online as well. It is no great leap to compare the role of public libraries and town squares in promoting core freedoms with that of the Web, functioning as it does not only as a repository of information, but also as a public and personal medium for communication and association. Just as we expect freedom and autonomy in the former, "brick and mortar" venues, so we should in the latter, digital electronic version. Information search and retrieval behaviors are part and parcel of these activities, profoundly reflecting who we are, what we care about, with whom we associate, and how we live our lives. For dealing with behaviors that open a window to the personal and political commitments of individuals, existing practices and policies of search engine companies seem clearly inadequate. Less clear, however, is how to pursue reforms to achieve necessary levels of protection, and who should or will lead the way.

Among potential agents of reform, the evident structure of incentives indicates that two with the greatest power to effect change—government, by pursuing new laws and regulations, and search companies, by revising internal policies—would be the least likely to support such change. Intransigence and inaction in the face of early challenges has borne this expectation out. For the first potential source of reform, government, search logs are an obvious and potentially important repository of information about individuals' interests and transactions, a valuable component of the vast stockpile of personal information assembled under the more lenient terms governing the collection and uses of

2. For example, court documents indicate that AOL, Yahoo!, and Microsoft had not been issued subpoenas because they had complied with the government's request.

3. See, for example, Michael Zimmer, *"The Quest for the Perfect Search Engine: Values, Technical Design, and the Flow of Personal Information in Spheres of Mobility" (unpublished dissertation, New York University, 2007)*; "Privacy Issues and Behavioral Advertising," (Federal Trade Commission Town Hall Meeting, Washington, D.C., November 1–2, 2007).

information by the private sector.[4] Actions that might constrain access to such information or limit its availability are not likely to be attractive.

As for the second potential source of reform—search engine companies—we predicted that they would be unlikely to welcome external restraints on how their logs are treated and used. For a start, there is the general suspicion corporate actors hold for any imposition of third-party regulation. With their interests best served by as little oversight as possible, search companies attempt to mollify worried users and regulators by insisting that unconstrained access to and use of query data is an essential necessity for running their businesses, as, for example, explained by Eric Schmidt, CEO of Google: "the data helps us to improve services and prevent fraud."[5] Although there is no reason to doubt this explanation, it masks a story that is never front and center in search companies' public rhetoric, but lies behind concerns of critics and privacy advocates, namely, the ways unconstrained assembly and use of detailed search query logs factor into the massive profit engine of personalized advertising.

A third possible source of reform is new government regulation or legislation steered by direct citizen action or advocacy organizations such as the Electronic Privacy Information Center, Privacy International, the Center for Democracy and Technology, and the Electronic Frontier Foundation.[6] Although this approach has already borne fruit—see, for example, the widely publicized report "A Race to the Bottom"[7]—it will require an orchestrated effort of diverse parties, including many (government actors, search companies, advertisers, etc.) with a stake in maintaining unrestricted access to search logs. Ultimately, however, this is our soundest hope for lasting change, with measurable success most likely a long-term prospect.

TrackMeNot (TMN), a lightweight Firefox browser extension designed to ensure privacy in Web search by obfuscating a user's actual searches amidst a stream of programmatically generated decoy searches, represents a fourth alternative. Since August 2006, when the first version of TMN was made publicly accessible free of charge, there have been over 350,000 downloads.[8] Overcoming some

4 Michael D. Birnhack and Niva Elkin-Koren, "The Invisible Handshake: The Reemergence of the State in the Digital Environment," *Virginia Journal of Law & Technology* 6 (2003): 1–57.

5. "Eric Schmidt on Global Privacy Standards," *Peter Fleischer: Privacy . . . ?* September 19, 2007, http://www.peterfleischer.blogspot.com/ (accessed January 2, 2008).

6. Electronic Privacy Information Center, http://www.epic.org; Privacy International, http://www.privacyinternational.org; The Center for Democracy and Technology, http://www.cdt.org; Electronic Frontier Foundation, http://www.eff.org.

7. Privacy International, "A Race to the Bottom: Privacy Ranking of Internet Service Companies," http://www.privacyinternational.org/article.shtml?cmd[347]=x-347-553961 (accessed December 30, 2007).

8. TrackMeNot (http://trackmenot.org) may be downloaded from the Web site http://mrl.nyu.edu/~dhowe/TrackMeNot/ or from the Mozilla add-on Web site https://addons.mozilla.org/en-US/firefox/addon/3173.

of the obstacles inherent in similar software, TMN offers control directly to those most motivated to seek reform, providing a relatively near-term if imperfect solution. The hope, too, is that alternatives like TrackMeNot will bring reluctant parties into meaningful dialogue about search privacy.

II. DESIGN CONSTRAINTS

> The constraints of technique, resources, and economics *underdetermine* design outcomes. To account fully for a technical design one must examine the technical culture, social values, aesthetic ethos, and political agendas of the designers.[9]

Our approach to the development of TrackMeNot builds on prior work that has explicitly taken social values into consideration in software design.[10] Throughout the planning, development, and testing phases, we have integrated values-oriented concerns as first-order "constraints" in conjunction with more typical engineering concerns such as efficiency, speed, and robustness. Specific instances of values-oriented constraints include transparency in interface, function, code, and strategy; personal autonomy, where users need not rely on third parties; social protection of privacy with distributed/community-oriented action; minimal resource consumption (cognitive, bandwidth, client and server processing, etc.); and usability (size, configurability, ease-of-use, etc.). Enumerating values-oriented constraints early in the design process enabled us to iteratively revisit and refine them in light of the specific technical decisions under consideration.[11] Where relevant in the following section, we discuss ways in which TMN's technical mechanisms benefited from this values-oriented approach.

III. TECHNICAL MECHANISMS

TrackMeNot, written in Javascript, C++, and XUL, is a Firefox browser extension designed to hide users' Web searches in a stream of decoy queries. Query-like phrases are harvested by TMN from the Web and sent, via HTTP requests, to search engines specified by the user. To augment this basic functionality and

9. Bryan Pfaffenberger, "Technological Dramas," *Science, Technology & Human Values* 17, no. 3 (1992): 282–312.

10. Batya Friedman, Daniel C. Howe, and Edward Felten, "Informed Consent in the Mozilla Browser: Implementing Value Sensitive Design," *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* 8 (2002): 247; Mary Flanagan, Daniel C. Howe, and Helen Nissenbaum, "Values at Play: Design Tradeoffs in Socially-Oriented Game Design," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2005): 751–760.

11. Flanagan, Howe, and Nissenbaum, "Values at Play" (n. 9).

frustrate attempts by search engines to distinguish between actual and generated queries, a range of mechanisms were implemented to simulate users' actual search behaviors more effectively. These mechanisms and the design constraints informing their implementations are described in the following sections.

A. Dynamic Query Lists

To keep control in the hands of users TMN operates solely on the "client," with no dependence on centralized servers or third-party sites during its operation. To support this design constraint while maintaining unique query lists for each instance of TMN's operation, we employed a mechanism we called *dynamic query lists*, which functions as follows. When downloaded, each instance of TMN is equipped with two mechanisms for creating an initial seed list of query terms: (1) a set of RSS feeds from popular Web sites (e.g., the New York Times, CNN, Slashdot) and (2) a list of popular queries gathered from publicly available lists of recent search terms.

When TMN is first enabled, an initial query list is constructed from both the results of requests to the RSS feeds and the list of popular terms.[12] From this list of seed terms (100 to 200 per client, as illustrated in Figure 23.1), TMN issues its initial queries. As operation continues, individual queries from this set are randomly marked for substitution. When a marked query is sent, TMN intercepts the search engine's HTTP response and attempts (nondeterministically) to parse a suitable "query-like" term from the HTML returned. If, according to a series of regular-expressions tests, the substitution is successful, this new term replaces the original query in the query list and the substitution mark is removed. This new term is now a member of the current query list (visible to users via the options panel described later) and included as a potential future substitution candidate. Additionally, each time the browser is started, a randomly selected RSS feed is queried and some subset of its terms are substituted into the seed list in the same manner. Over time, each client "evolves" a unique set of query terms,

```
fashion, tv guide, barbie, neopets, bit torrent, xbox, angelina jolie, nin-
tendo, jennifer lopez, jennifer aniston, local weather, anime, jokes, reci-
pes, music lyrics, games, iraq, global warming, north korea, hillary clin-
ton, barack obama, dick cheney, zodiac, music and lyrics, bone cancer, lena
katina, iran, canada, veronica mars, lost, the constitution, valerie
plame, karl rove, halliburton, Iceberg, global warming, world map, earth
day, southern cross, spiderman 3, 300 movie, borat, shrek, bill of rights,
ghost rider, Hawaii, dubai, mexico, freedom of speech, Chelsea, London,
kurt vonnegut, shaha riza, yuri Gagarin, knut, Virginia tech, wellness,
copyright law, health, yoga, fishing, golf, Israel, Syria, Iraq, Pakistan
```

FIGURE 23.1 SAMPLE FROM A TMN SEED LIST

12. The default list of popular search terms is included primarily for the rare case where some or all of the RSS feeds may be unavailable.


```
Turning carbon dioxide into fuel, Online Student Services, free
essential software, business globalization solutions, National
Pasta Association, Share your life with friends, Demand Financial
Suite, este calitatea produselor, Chicago Symphony Orchestra, This
film contains violence, Expects below Average, Emergency Contact,
bodies have been established, residential real estate, American
Heritage Month, Manhattan Athletic Club, healthcare support
occupations, people cannot realize their dreams, green chemistry
breakthroughs, Free online versions, Also find tools, Hope Press
```

FIGURE 23.2 SAMPLE FROM AN "EVOLVING" QUERY LIST

based in part on the random selection of queries for substitution, in part on the nondeterministic query extraction from HTML responses, in part on new terms gathered from continually updating RSS feeds, and in part on the continually changing nature of Web search results (generally yielding different results for the same search on different days). Figure 23.2 shows examples from the query list of Figure 23.1 after several weeks of TMN operation. With dynamic query lists, TMN is able to avoid the use of any central or shared (and necessarily trusted) repository of query terms while still frustrating the filtering schemes to which a static list is vulnerable.

B. Selective Click-Through

"Click-through" refers to the behavior of following one or more additional links on a results page after an initial search query. Although versions of TMN with this functionality were tested from early on, we chose not to release any until we were confident we could minimize potential impacts on existing business practices, specifically on those advertisers who paid search engines on a per-click basis. Current versions of TMN (since 0.6), however, employ what we call *selective click-through*, in which a series of regular-expression tests are used to identify and avoid potentially revenue-generating ads. Clicks are then simulated on one or more of the remaining links on the results page—either a "more results" button, a returned link to an external Web site, or a link internal to the search engine (e.g., "news" or "images"). Assuming that the search engines continue to format ad-related links in a relatively consistent manner, this appears to be an adequate solution for the time being.

C. Real-Time Search Awareness

Real-time search awareness (RTSA) is a second mechanism developed to improve TMN's capacity to mimic searchers' actual behavior. As TMN evolved, it became clear that it would need to "know," in real-time, when a user had initiated a search at one of the engines the user had selected. To facilitate this, the RTSA module examines each outgoing request from the browser and, via a series of regular expressions unique to each search engine, alerts TMN when the user is

initiating a search. This feature has proved increasingly important, by enabling the development of several other mechanisms (described later) that require knowledge of the user's current behavior, whether it be initiating a search, performing a series of searches, or engaging in other, nonsearch activities.

D. Live Header Maps

Initially, development efforts focused on simulating the behavior of searchers in general. In later versions, however, several features were introduced that enabled TMN to adapt to the behavior of specific users. In addition to the *TMN Control Panel* (described later), which allows users to manually configure TMN to more closely mimic their own search behavior, *live header maps* operate automatically to adapt TMN-generated queries to specific data sent by the client browser. This data generally varies according to browser version and operating system, as well as the search habits of specific users. To facilitate this adaptive behavior, TMN maintains a set of variables (per search engine) representing the header fields and URLs for the search most recently issued by the browser (see Figure 23.3). These dynamically updating variables allow TMN to reproduce, in its own requests, the exact set of headers the browser last used.

Similarly, the specific URL last used to access a search engine is maintained, so that, for example, if one user searches via the Google toolbar and another via the Google home page, TMN requests will mimic the header values for each. The RTSA module facilitates this functionality by allowing updates to these variables only when the user is initiating a new search at one of the selected engines.

E. Burst-Mode Queries

Another functionality enabled by RTSA is termed *burst-mode querying*. In initial versions of the software, semi-random intervals were used to temporally space TMN requests, with the average of these intervals set by the user. To more closely mimic actual user behavior, burst mode triggers a batch of queries within close proximity to an actual user search (as detected by RTSA). By using this mode in

```
Url -> http://www.google.com/search?hl=en&client=firefox-
a&rls=org.mozilla%3Aen-US%3Aofficial&hs=nxM&q=hello&btnG=Search
User-Agent -> Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US;
rv:1.8.1.11) Gecko/20071127 Firefox/2.0.0.11
Accept -> text/xml, application/xml, application/xhtml+xml,text/html;
q=0.9, text/plain; q=0.8, image/png, */*; q=0.5
Accept-Language -> en-us,en; q=0.5
Accept-Encoding -> gzip, deflate
Accept-Charset -> ISO-8859-1, utf-8; q=0.7, *; q=0.7
Keep-Alive -> 300
Connection -> keep-alive
Referer -> http://collection.eliterature.org/1/
```

FIGURE 23.3 AN EXAMPLE OF A HEADER MAP FOR A GOOGLE WEB SEARCH

conjunction with randomized intervals, users can "blend" the two behaviors, employing more or less of each as desired. Further, by limiting bandwidth and processing use (for both client and search engine) while dynamically adjusting to the use patterns of the individual user, burst-mode operation allows TMN to meet another design constraint, namely, lowered client (and network) resource-use. To further mimic user behavior, a subset of burst-mode queries are selected as variants on a "search theme." Specifically, a longer query is selected and permuted into a set of smaller, related queries constituting a "burst." For example, a search from a recent RSS feed, "dancing with the stars," was permuted into the following set of queries: "dancing," "stars," "dancing with," "with the stars," and "dancing with the stars"—all of which were sent sequentially over the course of a 60-second burst.

F. TMN Control Panel

The Control Panel provides a range of user-configurable parameters allowing users to customize TMN's behavior further (see Figure 23.4). These include options to enable/disable TMN itself, the status bar display, query bursting, and RSS feed management. Additionally, users may select which search engines

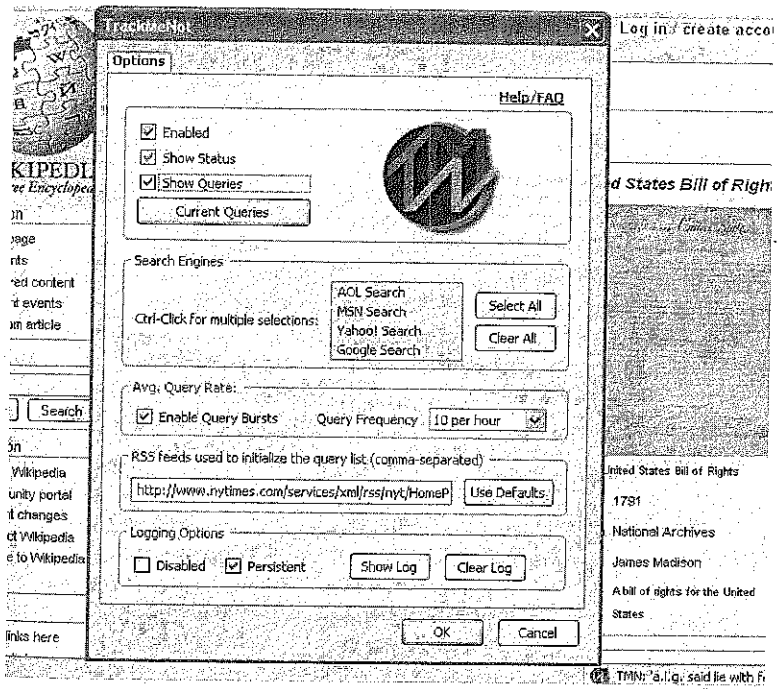

FIGURE 23.4 THE TRACKMENOT CONTROL PANEL

they wish to target, select an average query frequency, and manage TMN's logging options. Finally, the Control Panel features buttons enabling users to view the current query list and action logs within the browser itself and to access the TMN Web site for additional information.

The capacity for direct, real-time access to system operation (logs and query lists) was directly informed by the design constraint of transparency. Further, TMN was released as free and open-source under a Creative Commons license[13] with all source files included (in plain-text format) in every download, allowing technically sophisticated users to examine the inner workings of the code and verify that it has functioned as described. Similarly, our intentions and the specific technical decisions made to realize them are included in straightforward, nontechnical language on the TMN Web site (accessible directly from within the software itself).

IV. EVALUATION: STRENGTHS AND WEAKNESSES

Evaluation of TMN was performed iteratively throughout development and relied on solicited and unsolicited feedback from a range of groups, including users, developers, software reviewers at Mozilla (where TMN is cohosted), and a range of privacy and security advocates. Although the question of whether TMN in fact "worked" seemed at first, a simple one, we soon noticed how users' behaviors, goals, expectations, and perceived risks shifted the meaning of the question. Analysis of feedback was thus often a two-part process: first determining the users' orientations and then examining their feedback in light of their respective goals, concerns, and other priorities. Through such analysis we were able to identify at least four distinct groups of users, although individuals often identified with more than one group.

One group was interested in TMN's ability to cloak a searcher's identity and thus prevent all search activity from being traced back to the user. We recognized that there were at least four mechanisms through which a search could be identified: (1) identifying information included in search queries (name, zip code, phone number, Social Security number, etc.), (2) IP addresses linking searches across sessions, (3) explicit login to search engines (often for mail or other services), and (4) persistent cookies linking any of the preceding items to users' search activities.

Although various prototype versions of TMN had included code to generate arbitrary personal information to mask actual identifying information, this strategy was not energetically pursued. This was largely because TMN was not designed to mask IP addresses and thus could not prevent identification via the

13. Creative Commons, http://creativecommons.org/.

IP addresses logged by search engines with every query, or those maintained by users' ISPs. We pointed users interested in such capabilities to various proxy-based solutions[14] linked from the TMN FAQ. Contrary to the assertions of some critics,[15] TMN was presented not as a lightweight replacement for proxy-style solutions, but rather as a very different approach (with a distinctive set of strengths and weaknesses). For one thing, proxies generally require users to grant some degree of trust to a third party, whether a centralized server or some "exit node" representing the last hop in a "distributed" solution. In the past, such exit nodes have been abused for a variety of purposes, or simply blocked by those not wishing to receive their traffic (Google and Wikipedia being prime examples).[16] Although specific TMN users could also easily be blocked once they were identified by a search engine (at least for the duration of the activity connected with their IP address), it would take a very different kind of effort to block all such users. With Tor, for example, the identification of a single proxy node could result in the blocking of many thousands of user requests.

While a full discussion of the relative strengths and weaknesses of proxy-based solutions (including problems with internationalization and potential vulnerabilities to traffic analysis attacks[17]) is beyond the scope of this chapter, it is worth noting the relative "user-friendliness" of such solutions in comparison with TMN. At least at the time of this writing, proxy-based solutions have been notoriously difficult for nonexperts to set up, configure, and use. They generally involve multiple components (e.g., a local executable and a browser plug-in) that must be installed and configured to communicate correctly, at which point it is

14. Tor: Anonymity Online, http://www.torproject.org/.

15. Schneier on Security, http://www.schneier.com/blog/archives/2006/08/trackmenot_1.html.

16. In September 2007 Dan Egerstad, a Swedish security consultant, revealed that he had intercepted usernames and passwords for a large number of e-mail accounts by operating and monitoring Tor exit nodes. On November 15, 2007, he was arrested on charges stemming from discovering and publishing this information. As Tor does not, and by design cannot, encrypt the traffic between an exit node and the target server, any exit node is in a position to capture any traffic that is not encrypted at the application layer (e.g., by SSL). Although this does not inherently violate the anonymity of the source, it affords more opportunities for data interception by self-selected third parties, greatly increasing the risk of exposure of sensitive data by users who are careless or who mistake Tor's anonymity for security. From http://en.wikipedia.org/wiki/Tor_(anonymity_network). See also http://www.boingboing.net/2006/09/07/google-blocking-priv.html, by Cory Doctorow, discussing Google's blocking of Tor nodes, and http://simple.wikipedia.org/wiki/Wikipedia:Bans_and_blocks, regarding Wikipedia's policy of blocking all requests from anonymizing proxies (including Tor).

17. See Steven J. Murdoch and George Danezis, "Low-Cost Traffic Analysis of Tor," *Proceedings of the 2005 IEEE Symposium on Security and Privacy* (2005): 183–195, http://dx.doi.org/10.1109/SP.2005.12.

often still unclear exactly what the proxy is doing. This situation has been considered serious enough that privacy advocates (and third-party companies) have begun providing versions of popular software already containing, say, a Tor configuration to eliminate these difficulties for users.[18] This differs noticeably from TMN's one-click-and-restart installation and subsequent transparency of operation.

Of course, there is no reason these different approaches cannot be used together to additive effect. In fact, we believe this to be a rich area for future research. On the other hand, there are difficulties faced by Web users that are equally troublesome for all proposed solutions. An example is the increasingly common occurrence where a user wishes to search the Web while being explicitly logged on to a search engine, say for its free e-mail services. Here none of the proposed solutions, TMN, proxy server, or any other, offers much help.

Another group of users were worried about being targeted in connection with "hot-button issue" searches, that is, stigmatized or taboo subject descriptors such as "anarchy," "HIV," or "drug-use." To protect such users adequately, TMN would need to generate a range of similarly "hot" query terms, a capability with which we experimented throughout development, and that showed particular promise with the addition of dynamic queries in version 0.4. Having found that dynamically evolved queries tend to stay within general topic areas, we reasoned that with hot terms in the initial seed list, TMN would generate some number of extreme, or at the very least unwelcome, surprises—potentially offensive or NSFW (not safe for work) queries, for example, which would be displayed publicly in the browser's status bar. An open question, since the release of version 0.6, is whether, and to what extent, users in this group will be able to customize their RSS feeds (via the options panel) to maintain a query list with a significant number of "hot" queries. It would seem that such users could find, or even create, one or more RSS feeds matching the types of "hot" queries in which they were interested, though how such a list would "evolve" is unclear. Although a similar tactic has worked quite well for moving between languages (e.g., the queries in a French user's list will tend to stay in French as they evolve, assuming that their initial RSS feeds are in French), it will require significantly more testing before we can claim the same for "hot-button" queries.

The same issue was of concern to users in a third group, who instead wanted TMN to avoid hot-button issues entirely, citing worries relating to social stigma, job loss, and even potential arrest. This group, preferring generic and innocuous noise, was primarily interested in TMN as a way to mask the true nature of their online searches in order to avoid wholesale aggregation and profiling,

18. See OperaTor, a preconfigured bundle including the Opera Browser, Tor, and Privoxy, http://archetwist.com/opera/operator, and the XeroBank Browser, a Firefox derivative with an integrated Tor configuration, http://xerobank.com/xB_browser.html.

advertising and marketing purposes being the most salient. This clash between groups 2 and 3 is evident in the following two excerpts:

> Some of them would have to have HIV, some of them would have to be contemplating suicide, some of them would have to be anarchists, etc. Maybe you wouldn't want to have pedophiles and terrorists in the mix . . . or people growing hydroponic marijuana? (anonymous user from group 2)

> I downloaded and installed the plug-in you developed. I just turned it off when I noticed that the search term it had generated was "free russian porn boys." I'm a little confused. I understand the rationale of TrackMeNot, or I thought I did, but how does associating my IP with searches for gay porn fit in? If my employer logged this search, it could put my job at risk. (anonymous user from group 3)

The concerns of a fourth group of users stemmed from potential civil rights violations due to Web search monitoring. Although, like the third group, they also worried about the logging and aggregation of search queries for the purpose of profiling, they were interested in TMN mainly as a disruptive tool for protecting citizens against agents of government who might be engaged in various search surveillance and aggregation practices.

As we saw these groupings emerge and considered how they might guide the iterative process of feedback and development, it was clear (as per the popular saying) that "you can't please all the people all the time." Accordingly, it made sense to focus on what we believed to be TMN's greatest strength, that is, providing protection against aggregation and profiling of individual search queries. This meant anticipating various ways that TMN-generated queries might be detected and filtered, a process in which we were aided greatly by the helpful feedback of critics and enthusiasts alike who volunteered their insights and pointed out potential weak links. We conjecture, in large part due to the many iterations of the software emerging from such discussions, that considerable effort is now necessary to "defeat" TMN and successfully filter user queries from TMN queries. We further surmise that such filtering efforts would require significant resources and would still be likely to generate a number of false positives, that is, user queries mistakenly judged to be TMN generated.

The critiques we considered drew attention to various ways in which TMN queries could be different enough from user queries to inform a filtering algorithm able to distinguish between the two. We have worked most recently to address variations of this critique based on three aspects of TMN's operation: query timing, click-through behavior, and query term analysis. Timing-based critiques have argued that the timing patterns of TMN's queries, even when randomized, were different enough from those of human-generated queries to be detectable. Our solution to this critique was to add burst-mode querying so that TMN queries can occur only when users are actually searching at a targeted engine.

"Click-through" critiques pointed out that TMN queries were never followed by clicks to outgoing links on a search results page. The current version of TMN, however, implements "selective click-through" for nonadvertising links, as described previously. Yet even with this mechanism in place, a search company might be able to identify real users whenever an ad link *was* clicked. However, while one might "know" that queries leading to click-throughs on ads were user generated, inferring the converse—that queries yielding either unclicked results or those with click-throughs to non-ad pages were TMN generated—would surely result in a significant number of false positives (user-generated searches discarded along with those generated by TMN). Missing user queries of this type might be particularly costly to search companies aiming to improve their performance through personalized query log analysis.

A final version of the filtering argument focused on the nature of the query terms themselves, claiming that these were not "real" enough to fool a sophisticated learning algorithm with access to the vast amounts of data that search engines have already collected. Although dynamic query generation from actual Web pages has clear virtues, it is difficult to state how effective this strategy would be if search engines were willing to allocate significant resources to overcoming it. It is possible that a machine-learning algorithm focusing on query content, perhaps in conjunction with other factors, could be trained to identify a high percentage of TMN users, possibly even a high percentage of specific user queries themselves. The primary obstacles to defeating TMN are the costs of human and material resources (engineers, hardware, software), the cost of false positives (discarded user queries), potential costs to one's reputation (as a result of user and/or media outcry), and the potentially increasing maintenance costs necessary to handle past and future versions of TMN equipped with different behaviors. Although the extent of such costs are difficult to assess, especially in light of the vast resources available to search companies, we hope they are high enough to make other collaborative, trust-oriented compromises more attractive.

Unfortunately, we have little insight into the countermeasures that may be taken by targeted search engines. Unlike tactics such as URL format changes or IP address blocking, which will be readily apparent, approaches such as the filtering strategies discussed earlier might go unnoticed. We have thus far benefited from the "many eyes" of the developer and open-source communities, which have prodded us to consider such countermeasures, as well as from the evaluations of users and critics.

V. POLITICS THROUGH TECHNOLOGY

Conceiving of technologies as forms of political action builds on an intellectual tradition that includes figures such as Langdon Winner and Bruno Latour, who have argued that technical devices and systems may embody political and moral qualities. Lawrence Lessig and others[19] have explored these ideas in the context of information technologies and digital networks. Allied with this academic tradition, though not necessarily in direct dialogue with it, activist designers, software developers, and digital artists have leveraged the malleability of IT and the openness of network protocols to develop utilities that are expressive of particular political commitments or that mediate transactions in politically charged ways.[20]

The placement of control in the hands of users, which we adopted as one of TMN's design constraints, is not the only thing that makes it political. Its political character stems also from the way it enters into and attempts to reshape a particular aspect of individuals' relationships with social actors far more powerful than themselves on nearly every measurable dimension—including wealth, mastery over technology, and access to power. TMN, by allowing individuals to set limits on the flow of personal information, belongs in a class of technical tools that serve as amplifiers of social resistance or political voice. Relying on neither the largesse nor the permission of others, most notably those with potentially clashing interests, TMN provides for some users a means of expression, akin to a political placard or a petition. For others it provides a practical means of resistance similar to that described by Gary T. Marx, where individuals take advantage of the blind spots inherent in large-scale systems of surveillance.[21]

VI. IS TRACKMENOT MORALLY DEFENSIBLE?

In previous sections we addressed some of TMN's technical limitations, many drawn to our attention by critics. Here we will discuss challenges to TMN's moral standing. Those we will *not* discuss, however, are accusations that TMN makes life easier for the likes of pedophiles and terrorists by enabling them to hide from public view. Although these are important concerns, we believe they call attention to the more general challenge of living in a free society where protecting the rights of speech, association, and action inevitably creates space for exercising these liberties in ugly and hurtful ways. In order to remain free,

19. Friedman, Howe, and Felten, "Informed Consent in the Mozilla Browser," 247 (n. 9); Lucas Introna and Helen Nissenbaum, "Shaping the Web: Why the Politics of Search Engines Matters," *Information Society* 16, no. 3 (2000): 186–189.

20. Examples include GNU, http://www.gnu.org/; Creative Commons tools, http://creativecommons.org/; P3P, http://www.w3.org/P3P/; Adrian Ward's AutoIllustrator, http://www.medienkunstnetz.de/works/autoillustrator/; Wikipedia, http://www.wikipedia.org/; and the Radical Software Group's Carnivore, http://r-s-g.org/carnivore/.

21. Gary T. Marx, "A Tack in the Shoe: Neutralizing and Resisting the New Surveillance," *Journal of Social Issues* 59, no. 2 (2003): 369–390.

a society strives to minimize or prevent the hurt and ugliness without diminishing the relevant liberties. Thus, this is not a problem related to TMN only, nor is it one that we can make progress on here.

Instead we focus on criticisms addressing specific features of TMN. One such criticism accuses TMN of being no different from "spamware" or "denial of service" (DoS) attacks, generally wasting network bandwidth and clogging the servers of search engines. Naturally, we resist these critiques. By invoking rhetorical terms such as "spam" and "DoS" critics seek to cast doubt on our efforts and intentions by associating TMN with activities generally deemed reprehensible; we see these accusations, however, as question begging. After consulting numerous sources, we are confident that TMN fits no reasonable, commonly accepted definition of either DoS or spamware. In Wikipedia, for example, a "denial of service" attack is defined as "an attempt to make a computer resource unavailable to its intended users. Typically the targets are high-profile Web servers, and the attack attempts to make the hosted Web pages unavailable on the Internet."[22] And spam is defined as "the abuse of electronic messaging systems to send unsolicited bulk messages, which are generally undesired."[23] Neither is applicable to TMN.

Behind the rhetoric of DoS and spam, however, lies a question that deserves attention: the extent of TMN's impact on servers and bandwidth. This concerns us as well, as we had set forth with the principle of minimizing resource consumption as a design constraint, as stated earlier. The relevant facts are that TMN's resource usage is relatively low—tiny, in fact, compared with common components of Web traffic such as animations, music, and video—and consequently it is unlikely to have any appreciable effect on network bandwidth. It is conceivable, however, depending on the number of users and their use patterns (e.g., the mode and frequency settings they select), that TMN might have an impact on search engine performance by placing additional demands on server processing and bandwidth. Our intention and our expectation, based on the current usage and trajectory, is that the impact on search engines will be minimal. Universal deployment is not the goal of the project; our intention is to offer a degree of protection to individuals who may feel threatened and to afford such users a voice in the evolving debate over Web search privacy. We are confident that search companies will take steps to address user dissatisfaction long before TMN usage reaches significant proportions.

We still have not addressed a key driver of this critique that TMN "wastes" bandwidth and server resources. We know, both from anecdotes and through search statistics aggregated by Google Zeitgeist, the Lycos 50, and other such services,[24] that people search the Web for a vast range of items of information.

22. Wikipedia, http://www.wikipedia.org (accessed March 29, 2007).

23. Wikipedia, http://www.wikipedia.org (accessed March 29, 2007).

24. Google Zeitgeist, http://www.google.com/press/zeitgeist.html; Lycos 50, http://50.lycos.com/.

Judging by perennial favorites—the likes of "Britney Spears," "Paris Hilton," and "Pokemon"—we conclude that most search subjects are not terribly weighty. Further, people and enterprises download and distribute large video, music, and image files with no apparent socially redeeming value, and search companies constantly seek out new customers and markets with the hope of enticing to their services millions of new users from all over the world. All these activities use bandwidth and place heavier and heavier burdens on servers and services, but they are generally not criticized for wasting network and server resources. Why is this? These patterns reveal an underlying presumption about what constitutes proper use of the network; the aforementioned uses, however trivial, are assumed to be legitimate, whereas TMN-generated traffic is not. We challenge this assumption. Because adequate privacy protection is incorporated into neither the technology of search engines nor the policies governing it, steps taken by individuals to protect themselves do constitute a legitimate draw on resources, certainly no less legitimate than the myriad of others drawing on these resources. In this regard, we place TMN in a category along with uses of encryption technologies for securing transactions and proxies for anonymization (e.g., Tor). Despite the latter's incremental draw on resources, these additional burdens are generally understood as warranted and at times necessary. The same goes for TMN.

Another critique charges that TMN is morally indefensible because it violates search engines' terms of service (ToS) forbidding access by automated means such as scripts and Web crawlers (e.g. http://www.google.com/accounts/TOS). Although the legal enforceability of Web site ToS is a broader question than concerns us here, the active debates surrounding it provide valuable input.[25] Since the very beginnings of the Web, a complex and constantly evolving system of social norms, derived from a combination of law, morality, and affordances of architecture, has formed the background against which online actions and transactions are evaluated. Terms of service can be controversial because they unilaterally impose obligations on users that go beyond those implied by the background norms—in particular, a ToS designed to control users' experiences of a Web site or service. To be sure, Web site owners' preferences in setting the terms of engagement deserve consideration, but these expressed preferences do not automatically imply moral obligations, particularly ones that society needs to honor and defend. Owners' preferences need to be weighed against a range of other considerations.

25. The authors acknowledge the limitations of their perspective on legal issues, formed exclusively with reference to the United States legal system; Dan L. Burk, "The Trouble with Trespass," *Journal of Small and Emerging Business Law* (1998): 3; Maureen A. O'Rourke, "Is Virtual Trespass an Apt Analogy?" *Communications of the ACM* 44, no. 2 (2001): 98–103; Orin S. Kerr, "Cybercrime's Scope: Interpreting 'Access' and 'Authorization' in Computer Misuse Statutes," *NYU Law Review* 78, no. 5 (2003): 1596–1668.

One such consideration is efficiency. Legal discourse cautions that enforcement of the arbitrary preferences of Web site owners "for this or that type of usage,"[26] subjecting users to exclusions and exceptions, would result in the need for users to pick their way cautiously through the Web. Such a requirement would degrade the efficiency and positive externalities of the Web.[27] A Web requiring such cautious engagement represents a sadly diminished alternative to the Web extolled for its provision of freewheeling access to vast repositories of information goods and services.

Fairness is another consideration. Since search engines are able to generate value by skimming information off the Web by means of crawlers, thereby benefiting from the willingness of others to place informational resources online with no strings attached, it is unfair to prevent others from doing the same. Fairness precludes making an exception of oneself, imposing arbitrary restrictions on how others may use the resources one has placed on the open Web, while taking full advantage of the norms of open access embraced by others. Legal scholar Dan Burk argues that granting Web site owners overly strong exclusionary rights would make it possible for them "to free-ride upon the benefits of the network, while at will avoiding contribution of such benefits to others."[28]

Although efficiency and fairness are important considerations, they do not necessarily trump the claims of search engines (expressed in ToS) against any and all automated access. Essential to TMN's defense is its role in promoting the morally legitimate ends of user privacy and autonomy in Web searching and, ultimately, freedom of expression, association, and inquiry. Also relevant is TMN's relatively low imposition on resources. In other words, even if a case could be made for favoring the preferences of search engine developers, as expressed in ToS, in opposing *some* forms of automated query (e.g., ones that are frivolous and seriously undermine performance), fairness and efficiency considerations should place the burden of proof on the search companies; in the case of TMN, we believe, they ought not prevail.

VII. FUTURE WORK AND CONCLUSION

TMN provides individuals a means of both expressing and asserting a commitment to privacy in Web searches without having to depend on the largesse or intervention of third parties. Although it is fully functional, TMN is best

26. Burk, "The Trouble with Trespass," 3 (n. 24).

27. Some, like Orin Kerr, have argued that only those ToS expressible in "code" should be enforceable. We imagine such protocols as robots.txt would qualify, but the question is a larger, more general one than can be adequately addressed here.

28. Burk, "The Trouble with Trespass," 3 (n. 24).

considered a prototype, a proof of concept for a particular approach to privacy, that is, privacy through obfuscation. As discussed, TMN's greatest potential lies in its capacity to protect individuals against profiling, and its greatest challenge is to stay abreast of evolving search services themselves. Beyond the challenges of simply keeping up, there are challenges of providing rigorous, scientific assessments of performance as well as of improving the system in several ways.

A scientific means of evaluating TMN's performance, or the performance of any system adopting this approach, needs to address at least one key question, which we are not equipped to answer: what measure one employs to confirm that user-generated searches have been successfully obfuscated by TMN-generated searches. To be efficacious, TMN needs to introduce into the set of user search queries not only sufficient noise, and not only noise in the correct format, but noise of the right kind in relation to the type of protection being sought and the information being mined. Such needs are likely to turn not only on the statistical analysis of signal-to-noise ratios, but also on a practical understanding of how search query data is actually mined and how users are profiled.

Future work on TMN will focus on making various improvements to the search query terms. One alternative is to incorporate into TMN further mechanisms that effectively generate hot-button and identifier queries. As mentioned in an earlier section, after going some distance along this path, we chose not to follow it. A second avenue for future work is to explore P2P approaches to generating both search queries and timing patterns as a possible alternative to current mechanisms. A central challenge is to develop a system that meets functional criteria as well as the design constraints discussed early in this paper, such as usability and independence from third parties (i.e., central servers or potentially untrustworthy third parties.) We are not sure this is practically achievable.

We conclude with a philosophical point. TrackMeNot operates in an environment that is not only technologically complex, as we have tried briefly to demonstrate, but is also socially complex. Search engines provide an important service in a volatile and competitive marketplace in which search query logs are a valuable resource and source of revenue. For individuals, however, whether or not they view their discrete acts of search and retrieval as sensitive, patterns recorded over time potentially open a window into their lives, interests, and ambitions. Thus, such faculties are not only a source of individual vulnerability, but they could interfere with the rights to free and autonomous inquiry, association, and expression that are essential to sustaining a healthy democratic society. Consequently, there remains a tension in the relationship between individual users, important political values, and search service providers. In a better world, this tension would be resolved in a transparent, trust-based mutual accommodation of respective interests.

Instead, users who are concerned with privacy in searching perceive little transparency and few credible assurances in the policies of search engine companies that privacy will ever trump pursuit of direct profit as a priority. In

light of this, trust-based mutual accommodation of necessity gives way to an adversarial relationship. TMN, a tool for *this* world and *this* relationship, gives users a say in shaping the terms of engagement with search companies. Although obvious measures of TrackMeNot's success include impenetrable camouflage and 100 percent adoption, we would prefer a world in which TMN is not needed.

PART III

ANONYMITY