Title
Data Mining and Internet Profiling: Emerging Regulatory and Technological Approaches

Authors
Schwartz, Paul M.
Lee, Ronald D.
Rubinstein, Ira

Publication Date
2008-09-19

Data Mining and Internet Profiling: Emerging Regulatory and Technological Approaches

Ira S. Rubinstein, Ronald D. Lee, & Paul M. Schwartz †

INTRODUCTION

The 9/11 terrorists, before their deadly attacks, sought invisibility through integration into the society they hoped to destroy. In a similar fashion, the terrorists who carried out subsequent attacks in Madrid and London attempted to blend into their host lands. This strategy has forced governments, including the United States, to rethink counterterrorism strategies and tools.

One of the current favored strategies involves data mining. In its pattern-based variant, data mining searches select individuals for scrutiny by analyzing large data sets for suspicious data linkages and patterns. Because terrorists do not "stand out," intelligence and law enforcement agents want to do more than rely exclusively on investigations of known suspects. The new goal is to search "based on the premise that the planning of terrorist activity creates a pattern or 'signature' that can be found in the ocean of transaction data created in the course of everyday life."[1] Accordingly, to identify and preempt terrorist activity, intelligence agencies have begun collecting, retaining, and analyzing voluminous and largely banal transactional information about the daily activities of hundreds of millions of people.

Private organizations have their own reasons for gathering widespread information about individuals. With the expansion of internet-based services, companies can track and document a broad range of people's online activities and can develop comprehensive profiles of these people. Advertisers and marketing firms likewise have strong incentives to identify and reach internet users whose profiles have certain demographic, purchasing behavior, or other characteristics. The construction, storage, and mining of these digital dossiers by internet companies pose privacy risks. Additional privacy issues arise when

† The authors are respectively Associate General Counsel, Microsoft Corporation (ret); Partner, Arnold & Porter LLP; and Professor of Law, UC Berkeley School of Law. The views expressed in this article are those of the authors alone. All three authors received their JD degrees from Yale Law School in 1985.

1 James X. Dempsey and Lara M. Flint, *Commercial Data and National Security*, 72 Geo Wash L Rev 1459, 1464 (2004).

the government obtains this information, which it currently can without much legal process.[2]

This essay begins by examining governmental data mining; its particular focus is on pattern-based searches of databases according to a model of linkages and data patterns that are thought to indicate suspicious behavior. In Part I, this essay reviews widely held views about the necessary safeguards for the use of data mining. In Part II, this essay considers "dataveillance" by private corporations and how they have compiled rich collections of information gathered online in the absence of a robust legal framework that might help preserve online privacy.[3]

This essay then discusses some of the techniques that individuals can employ to mask their online activity as well as existing and emerging technological approaches to preventing the private sector or government from linking their personal information and tracing their activities. These technologies permit users to move about the world wide web pseudonymously and to adopt privacy-enhancing identity management systems. This essay concludes by briefly considering three topics: (1) whether and how to regulate the potential impact of identity management systems on counterterrorism efforts; (2) the requirements of transparency and understanding of the underlying models used in either data mining or identity management systems as a necessary prelude to the creation of rules on appropriate access and use; and (3) the need for research in several further areas.

I. DATA MINING

Data mining refers to a series of techniques used to extract intelligence from vast stores of digital information. One kind of data mining simply accelerates the process by which law enforcement or intelligence agents gather relevant information about subjects they already suspect of wrongdoing. This approach is termed subject-based searches. In pattern-based data mining, in contrast, the government investigator develops a model of assumptions about the activities and underlying characteristics of culpable individuals or the indicators of terrorist plans. The government official then searches databases con-

---

2 See Part I.B. See also Jon D. Michaels, *All the President's Spies: Private-Public Intelligence Gathering in the War on Terror*, 96 Cal L Rev (forthcoming 2008) (noting how the government can obtain information from private companies it might not be able to acquire itself and explaining how the Bush Administration has used informal agreements with private companies to gain private information, circumventing traditional congressional oversight).

3 See Roger A. Clarke, *Information Technology and Dataveillance*, in Charles Dunlop and Rob Kling, eds, *Computerization and Controversy: Value Conflicts and Social Choices* 496, 498 (Academic 1991) (defining dataveillance as "the systematic use of personal data systems in the investigation or monitoring of the actions or communications of one or more persons").

taining transactional and personal information for "hits" that indicate a match between the model and patterns left by potential evidence of terrorist plans or by potentially culpable individuals. The hope is that this approach will help to identify terrorists who seek to blend into the host population and its economic and social structures.

Thus, subject-based searches start from the usual predicate of reasonable suspicion.[4] In contrast, pattern-based searches depend on a theory or theories about the predictive power of data linkages to identify suspicious individuals. As a consequence, this approach may intrude in known and unknown ways into the lives of innocent people.

Privacy policy experts have raised concerns about pattern-based data mining. James Dempsey and Lara Flint note that this technique is in tension with "the constitutional presumption of innocence and the Fourth Amendment principle that the government must have individual suspicion before it can conduct a search."[5] Others have wondered about psychic harms caused by government scrutiny of innocent people.[6] Finally, there is the danger of false positives. Data analysis can lead to an innocent person being placed on a watch list, investigated, or detained.

A. Total Information Awareness, Terrorism Information Awareness, and Progeny

In the wake of the September 11 attacks, the Defense Advanced Research Projects Agency (DARPA) in the Pentagon began funding basic research in connection with the Total Information Awareness (TIA) project. TIA sought to support research into a prototypical data mining program aimed at discovering and tracking terrorists through the digital paths of their routine transactions. These interactions provide data points about communications, education, financial affairs, travel, medical history, immigration, transportation, and housing.[7]

---

4 See Paul Rosenzweig, *Proposals for Implementing the Terrorism Information Awareness System*, Legal Memorandum No 8 (Heritage Foundation, Aug 7, 2003), online at http://www.heritage.org/Research/HomelandDefense/lm8.cfm (visited Jan 12, 2008) (recommending a reasonable suspicion standard for searches conducted by the Terrorism Information Awareness program that break "the anonymity barrier" because "[r]equiring more would impose a burden of proof . . . that is more substantial than in any other investigative context" but "[r]equiring less would effectively eliminate any constraint on the technology's use").

5 Dempsey and Flint, 72 Geo Wash L Rev at 1466–67 (cited in note 1) (contending that "[p]attern analysis raises the most serious privacy and civil liberties concerns because it involves examination of the lawful daily activities of millions of people").

6 See, for example, Technology and Privacy Advisory Committee, Report, *Safeguarding Privacy in the Fight against Terrorism* ("TAPAC Report") 35–36 (Mar 1, 2004), online at http://www.cdt.org/security/usapatriot/20040300tapac.pdf (visited Jan 12, 2008).

7 William Safire, *You Are a Suspect*, NY Times A35 (Nov 14, 2002).

DARPA mismanaged external communications regarding TIA in several ways, such as creating an ominous logo utilizing the symbol for the "all seeing eye" and failing to inform Congress or the public of its funding of privacy research. As to the latter issue, TIA funded a study by an external research advisory board, the Information Science and Technology Study Group (ISAT), which was entitled "Security with Privacy."[8] This study investigated the development and extension of technologies for employing exploratory data mining techniques to pursue terrorists while ensuring privacy to individuals.[9]

DARPA's funding of TIA first garnered significant criticism in November 2002 when William Safire criticized the research program as a "supersnoop's dream." Safire termed TIA a "virtual, centralized grand database" that contained commercial and governmental dossiers on every US citizen.[10] Faced with a storm of public and congressional objections, DARPA quickly renamed the program Terrorism Information Awareness, which preserved the identical acronym as the original name for the project.

The Safire column marked the start of the public's opposition to TIA.[11] Soon afterward, Congress passed an amendment prohibiting the expenditure of funds on TIA unless the attorney general, the director of central intelligence, and the secretary of defense jointly reported on the development of the TIA and its effect on civil liberties. The amendment also prohibited use of the TIA to conduct searches on US persons without specific congressional authorization.[12] Ultimately, in its September 2003 defense appropriations bill, Congress terminated the bulk of funding for TIA and directed that the TIA office be closed.[13]

---

8 The Pentagon released the study in response to a FOIA request submitted by the Electronic Privacy Information Center. See Information Science and Technology Study Group, Report, *Security with Privacy* ("ISAT Report") (Dec 13, 2002), online at http://www.epic.org/privacy/profiling/tia/isat_study.pdf (visited Jan 12, 2008) (urging DARPA to adopt selective revelation, tamper resistance technologies, and rule processing technologies).

9 Email from Eric Horvitz, a member of the ISAT board who proposed and helped organize the *Security with Privacy* study, to Ira Rubinstein (June 1, 2007).

10 Safire, *You Are a Suspect*, NY Times at A35 (cited in note 7).

11 TAPAC Report at 16 (cited in note 6) ("In the seven months between the initial disclosure of TIA and Safire's column, only 12 press reports had appeared about the program. In the next 30 days, the press carried 285 stories.").

12 See Consolidated Appropriations Resolution, 2003, Pub L No 108-7, 117 Stat 11, 534 ("[T]he Total Information Awareness program should not be used to develop technologies for use in conducting intelligence activities or law enforcement activities against United States persons without appropriate consultation with Congress or without clear adherence to principles to protect civil liberties and privacy.").

13 Department of Defense Appropriations Act of 2004 § 8131(a), Pub L No 108-87, 117 Stat 1054, 1102 (2003). See also Making Appropriations for the Department of Defense for the

There are two important postscripts to TIA. First, the Technology and Privacy Advisory Committee (TAPAC), convened for the purpose of studying TIA, provided an important review of the program. TAPAC noted that the current laws regulating electronic surveillance were inadequate to address the range of privacy concerns introduced by a project of TIA's novelty, scale, and ambition.[14] In highlighting the dangers of data mining and recommending enhanced privacy protections, TAPAC set the agenda for development of a consensus view on data mining. TAPAC viewed data mining as a powerful and, at times, necessary tool in the fight against terrorism, but one that must be regulated effectively in order to protect civil liberties and to limit the number of disruptive and potentially devastating false positives.[15] In a similar fashion, the Markle Foundation's Task Force on National Security in the Information Age considered data mining as a useful tool, but one that should only be used "where there is a focused and demonstrable need to know, balanced against the dangers to civil liberties."[16]

Second, data mining of the kind that TIA contemplated in 2002 has quickly gone from theory to practice. The Department of Homeland Security (DHS) stated in a 2006 report to Congress, "Several components of DHS engage or plan to engage in data mining activities."[17] The DHS report did not distinguish between pattern-based and subject-based searches, but presumably the DHS does or will engage in both kinds of data mining. In August 2005, the GAO examined five data mining efforts by the federal government.[18] Finally, according to some media reports, funding for some aspects of TIA activities continues in classified parts of the federal budget.[19] In short, the consensus policy view regarding data mining is of great relevance.

Fiscal Year Ending September 30, 2004, and for Other Purposes, HR Rep No 108-283, 108th Cong, 1st Sess (2003), reprinted in 2003 USCCAN 1168, 1189 (conference report).

14 See TAPAC Report at 6 (cited in note 6). See also Dempsey and Flint, 72 Geo Wash L Rev at 1472–73, 1488–89 (cited in note 1) (describing the landscape of potential sources of regulation of privacy concerns relating to electronic data surveillance, suggesting that even those laws that seem like they might apply "are riddled with exceptions," and concluding that no constitutional or statutory law really addresses those privacy concerns).

15 See TAPAC Report at 39, 48 (cited in note 6).

16 Task Force on National Security in the Information Age, *Protecting America's Freedom in the Information Age: A Report of the Markle Foundation Task Force* 27 (Markle Foundation, Oct 2002), online at http://www.markle.org/downloadable_assets/nstf_full.pdf (visited Jan 12, 2008).

17 DHS Privacy Office, *Data Mining Report: DHS Privacy Office Response to House Report 108-774* ("DHS Privacy Office Report") 2 (July 6, 2006), online at http://www.dhs.gov/xlibrary/assets/privacy/privacy_data_%20mining_%20report.pdf (visited Jan 12, 2008).

18 See GAO, *Data Mining: Agencies Have Taken Key Steps to Protect Privacy in Selected Efforts, But Significant Compliance Issues Remain* preface (Aug 2005), online at http://www.gao.gov/new.items/d05866.pdf (visited Jan 12, 2008).

19 See, for example, Shane Harris, *TIA Lives On*, Natl J 66, 66 (Feb 25, 2006) ("It is no secret that some parts of TIA lived on behind the veil of the classified intelligence budget. How-

B. The Consensus View

The policy community has, in general, arrived at certain shared views of the technological systems and legal safeguards needed to protect civil liberties when pattern-based data mining is employed against terrorism. Within the consensus view, a basic premise is that data mining has substantial potential to protect against terrorism. But policy experts also insist that technological and legal safeguards are needed.[20] As TAPAC stated, "Data mining is a vital tool in the fight against terrorism, but when used in connection with personal data concerning U.S. persons, data mining can present significant privacy issues."[21]

Within this consensus, opinions diverge concerning how much regulation is necessary and the extent to which these safeguards should be allowed to modify the data mining processes. There is also a debate regarding acceptable levels of imprecision and the tolerable range of false positives. Finally, there is concern about the threshold for a privacy injury. Is one harmed by a false identification as a person of interest? Or, does a privacy harm materialize only when one is wrongly interrogated or detained?[22]

There are also two important experts who strongly disagree with the consensus view. First, Bruce Schneier is fundamentally skeptical about the underlying worth of data mining. He likens the search for a terrorist to looking for the proverbial needle in a haystack and views data mining based on pattern-based searches as only enlarging the haystack.[23] Second, Judge Richard Posner does not accept that data mining per se affects privacy interests. When a computer "sifts" through data, this is merely activity by a machine. Until human scrutiny occurs, no privacy has been invaded and no potential harm incurred.[24]

---

ever, the projects that moved, their new code names, and the agencies that took them over haven't previously been disclosed.").

20 See, for example, Daniel J. Solove, *The Digital Person: Technology and Privacy in the Information Age* 175–85 (NYU 2004) (highlighting "Orwellian" dangers to democratic values and "Kafkaesque" dangers from bureaucratic decisions with insufficient accountability).

21 TAPAC Report at viii (cited in note 6). See also ISAT Report at 4–7 (cited in note 8) (arguing against a dichotomy between security and privacy).

22 See Robert Popp and John Poindexter, *Countering Terrorism through Information and Privacy Protection Technologies*, IEEE Sec & Privacy 18, 24 (Nov/Dec 2006), online at http://csdl.computer.org/dl/mags/sp/2006/06/j6018.pdf (visited Jan 12, 2008) (describing how data mining may be used to reduce the incidence of unmerited interrogations and detentions).

23 Bruce Schneier, *Why Data Mining Won't Stop Terror*, Wired (Mar 9, 2006), online at http://www.wired.com/politics/security/commentary/securitymatters/2006/03/70357 (visited Jan 12, 2008). For his criticisms of TIA, see Bruce Schneier, *Beyond Fear: Thinking Sensibly about Security in an Uncertain World* 253–54 (Copernicus 2003) (discussing the probability of large numbers of false positives and concluding that "TIA is not worth it").

24 Richard A. Posner, *Not a Suicide Pact: The Constitution in a Time of National Emergency* 96–97 (Oxford 2006).

We wish now to examine the elements of the generally accepted (and in our view useful) framework for government data mining.

1. Legal authorization.

Data mining should occur only with legal authorization in place. In some instances, this authorization should be statutory. In others, the head of the agency must establish a regulatory framework governing searches of information relating to US persons. The agency head should make a written finding authorizing the project, specifying its purpose, and describing how the information will be used. In certain circumstances, there should be a required finding that less intrusive means of achieving the same purpose are not practically available or are less effective. Finally, the written authorization statement should require an additional layer of scrutiny before a warrant will be sought, establish an "acceptable" false positive rate, and develop a means of responding to the false positives.[25]

2. Access controls and authentication of users.

Existing technologies may be incorporated into the data mining process to restrict unauthorized use of data mining tools. These tools ensure that only authorized analysts gain access and that designated users do not misappropriate the information, either for personal ends or for an unrelated and unauthorized investigation, such as using the data to locate parents who owe child support.[26]

---

25 See TAPAC Report at 49–50 (cited in note 6) (proposing measures that would require agency heads to specify in writing, among other things, the existence of a satisfactorily low rate for false positives and of a system in place for dealing with false positives before using data mining); Dempsey and Flint, 72 Geo Wash L Rev at 1501–02 (cited in note 1) (proposing that one way "to structure a judicial role that would provide that check without unduly burdening executive branch efficiency" would be to "require a court order approving the use of pattern-based analysis in the first instance").

26 For discussions of data mining "mission creep," see Mary DeRosa, *Data Mining and Data Analysis for Counterterrorism* 16 (Center for Strategic and International Studies, Mar 2004), online at http://www.cdt.org/security/usapatriot/20040300csis.pdf (visited Jan 12, 2008) ("At any time, another type of illegal behavior could take on a high profile, and authorities will be under pressure to expand the use of these techniques, for example, to help investigate other violent criminals, immigration law violators, or even 'deadbeat dads.'"); TAPAC Report at 39–40 (cited in note 6) (suggesting that mission creep "is a particularly acute risk when the use of personal data about U.S. persons is justified by an extraordinary need such as protecting against terrorist threats"); Rosenzweig, *Proposals for Implementing the Terrorism Information Awareness System* (cited in note 4) (suggesting that "initial scanning must be automated and structured to prevent unauthorized access" and must also meet other requirements in order for data mining technology to conform with the "idea of preserving anonymity unless and until a good reason for breaching the anonymity barrier arises").

3. Rule-based processing.

Rules should be built into data search queries to ensure that results are tailored to the analyst's authorization. For example, search queries could carry information about the type of permission that the analyst has been granted, or the system could ask an analyst for additional proof or authorization before sharing certain kinds of results. Thus, the analyst might be asked to specify whether she has a search warrant, and if she does not, the system might not allow her to retrieve certain kinds of information.[27]

Additionally, "data labeling" may be used to describe how data should be accessed. Metadata may be included that summarizes the information, its source, and even the reliability and age of the source.[28] Information might be accessed differently if an analyst is advised that the data relate to a US citizen, rather than a foreign person,[29] and it might be treated as more or less reliable depending on where the information came from and how recently it has been verified. This final point raises the important issue of data quality. As the DHS Privacy Office has argued, strong data quality standards should be adopted for all information used in data mining.[30]

4. Anonymization and selective revelation.

With the goal of minimizing the amount of personal information revealed in the course of running pattern-based searches, the anonymization of data (such as names, addresses, and social security numbers) is essential.[31] The disclosure of personal information would occur only after the "sanitized" pattern-query results establish a reason to pursue further investigations of a subset of the original pool of individuals. Even then, identifying information would only be selectively revealed.[32] Access to additional personal details would require

---

27 K.A. Taipale, *Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data*, 5 Colum Sci & Tech L Rev 2, 75–76 (2003).

28 See ISAT Report at 16–17 (cited in note 8) (explaining that "[s]ince computers in general cannot understand the underlying representation of private information, it is necessary to label data with information that will allow it to be properly processed, both with respect to privacy constraints but also with respect to general constraints").

29 DeRosa, *Data Mining and Data Analysis* at 19 (cited in note 26); Taipale, 5 Colum Sci & Tech L Rev at 75–76 (cited in note 27).

30 See DHS Privacy Office Report at 16–17 (cited in note 17).

31 See DeRosa, *Data Mining and Data Analysis* at 17–18 (cited in note 26) (describing some of the possible techniques for anonymization).

32 See ISAT Report at 9–11 (cited in note 8) (recommending a system of "selective revelation" in which data in the form of general statistics and categories is revealed first and more specific information is revealed as justified by prior general results).

even further narrowing of the searches, independent authorization,[33] or a combination of the two. The Markle Foundation, for example, proposes that "personally identifiable data can be anonymized so that personal data is not seen unless and until the requisite showing (specified in guidelines) is made."[34]

5. Audit function.

Given the vast amounts of personal information made available to intelligence analysts, a means must exist to "watch the watchers."[35] Thus, an audit system is needed to provide a complete and tamper-proof record of the searches that have been conducted and the identity of the analysts involved. An audit system will also permit the investigation of any data security breaches as well as any misappropriation of information.[36] TAPAC advocated annual audits of any data mining programs involving personal information of US citizens.[37] In addition to favoring "strong, automatic audit capabilities" for data mining programs, the DHS Privacy Office requested the use of "random audits at regular intervals" as well as notice to government employees that their activities are subject to these audits.[38]

6. Addressing false positives.

A false positive occurs when a data relationship identifies an innocent individual. Mary DeRosa has worried that "there will be great temptation for the government to . . . take action based on the results of data-analysis queries alone."[39] One way to handle false positives is

---

33 See TAPAC Report at 51–52 (cited in note 6) (making specific recommendations for regulation of data mining that would involve "personally identifiable information" including requiring a written order from the Foreign Intelligence Surveillance Court); Rosenzweig, *Proposals for Implementing the Terrorism Awareness Information System* (cited in note 4) (suggesting that judicial authorization should be required at the point where sufficient personal details are available to break "the anonymity barrier" and begin identifying specific individuals).

34 Task Force on National Security in the Information Age, *Creating a Trusted Network for Homeland Security* 35 (Markle Foundation, 2003), online at http://www.markle.org/downloadable_assets/nstf_report2_full_report.pdf (visited Jan 12, 2008).

35 See ISAT Report at 13 (cited in note 8) (suggesting that strong audit mechanisms are "[p]erhaps the strongest protection against abuse of information systems").

36 See TAPAC Report at 50 (cited in note 6); Popp and Poindexter, *Countering Terrorism*, IEEE Sec & Privacy at 25 (cited in note 22) (proposing a "privacy appliance concept" that would "create an immutable audit log that captures the user's activity and transmits it to an appropriate trusted third-party oversight authority to ensure that abuses are detected, stopped, and reported"); DeRosa, *Data Mining and Data Analysis* at 19 (cited in note 26) (discussing the need for auditing generally and the challenges that an effective auditing program would face).

37 TAPAC Report at 52 (cited in note 6).

38 DHS Privacy Office Report at 4 (cited in note 17).

39 DeRosa, *Data Mining and Data Analysis* at 15 (cited in note 26).

to require an intermediate step in which analysts investigate and independently corroborate computerized search results before further action is permitted.[40] Should it be determined that a false positive has been made, agencies not only have an obligation to redress the immediate harm (for example, removing the innocent individual's name from a no-fly list), but also to use this result to improve the underlying model of the data mining program.[41] False results must be fed into a periodic, mandatory, and ongoing revalidation of the data mining program.

7. Accountability measures.

Internal controls need to be developed and adhered to, as well as steps taken to promote a culture of professionalism among the analysts.[42] The inspectors general of agencies that engage in data mining could play an important role in oversight of internal controls. Externally, review should be undertaken on a periodic basis by Congress, perhaps through the GAO. In addition, there should be regular public reports describing the nonclassified aspects of any data mining program. Finally, the government should develop standards for the validation of models used in data modeling and of the results of these programs. As the DHS Privacy Office observes, "[J]ust because a pattern exists in the data does not mean that the pattern is meaningful or valid."[43] The need is for independent validation of the model's predictive accuracy.

II. DATAVEILLANCE

In this Part, we examine how private companies are now amassing and analyzing rich databases of personal information collected online. We contrast the weak legal regulation of these practices with the privacy-protective abilities of different technologies, most notably identity management systems. These technologies, if widely adopted, might pose a "front end" challenge to the government's ability, on the "back end," to analyze information through data mining.

---

40 See id (suggesting that if data mining is used to aid investigation rather than as the sole basis for government action, then "[d]ata-mining results will [ ] lead only to more analysis or investigation, and false positives can be discovered before there are significant negative consequences for the individual").

41 Rosenzweig, *Proposals for Implementing the Terrorism Awareness System* (cited in note 4).

42 Id.

43 DHS Privacy Office Report at 10 (cited in note 17).

A. From Cookies to the "Database of Intentions"

Cookies are small text files placed on a user's computer to store information about the user and her preferences. Websites use cookies both to offer a personalized experience to users and to track online behavior and usage patterns in order to tailor online ads to groups of users based on demographics or likely purchasing behavior. Cookies are often placed without users' express knowledge or consent; they raise additional privacy concerns to the extent that they capture and transmit data about individual users. This information can include the searches that users have run, the identifying information that they have disclosed (for example, to register for and log onto a given service), their browsing patterns while visiting a site, and their "clickstream" behavior (that is, what links they clicked on while browsing the web).[44] Finally, third-party ad-serving companies use cookies to compile information about users' online behavior as they visit multiple sites that rely on the same ad network to deliver targeted ads.

The use of cookies for advertising purposes prompted significant privacy complaints beginning in the late 1990s. Most of the focus then was on the activities of DoubleClick, an ad-serving company that had compiled information from cookies to develop profiles on more than 100 million internet users.[45] When DoubleClick announced plans in 2000 to acquire Abacus Direct, a leading marketing firm, and to integrate its own online profile caches with Abacus's offline data, privacy advocates sounded the alarm.[46] Their concern was with the connection of largely pseudonymous online profiles with personally identifiable offline profiles.

The initial result was several state and federal investigations and class action consumer lawsuits. The controversy subsided a few months later when DoubleClick entered into a series of settlements, with the FTC in a lead role. DoubleClick agreed to preserve the anonymity of user profiles, to enhance opportunities for users to "opt out" of direct-marketing profiling, and to give users access to their online profiles.

---

44 See, for example, *In re DoubleClick Inc Privacy Litigation*, 154 F Supp 2d 497, 504–05 (SDNY 2001) (describing how DoubleClick employs cookies to record a user's browsing history while visiting DoubleClick-affiliated websites).

45 Heather Green, *Privacy: Outrage on the Web*, Bus Wk 38, 38 (Feb 14, 2000).

46 Id (describing how DoubleClick "quietly reversed" its policy of providing only anonymous data to marketers when combining data from direct mailers, Abacus, and other web services); Greg Miller, *Ad Firm's Practice Seen as Threat to Net Anonymity*, LA Times A1 (Feb 3, 2000) ("For years, DoubleClick has insisted that its cookie files were never connected to a person's name or address. That is changing in part because DoubleClick aims to take advantage of a giant offline database, Abacus, that it acquired last year.").

Today, concerns over cookies seem almost quaint. More of our lives take place online, and the online/offline distinction has less practical resonance. Recent privacy concerns now center on web services—and especially search engines. Like many other web services, companies offer free searching to consumers in exchange for targeted internet advertising based on the queries that individuals submit and other diverse information.

Internet searches raise significant privacy concerns because they can represent the most intimate and spontaneous of one's online activities. An internet search reflects unvarnished thoughts and ponderings rather than one's more considered communications or transactions. For John Battelle, the collection of such searches is no less than "the Database of Intentions." Battelle has pointed to the acute privacy implications of these new databases involving

> the aggregate results of every search ever entered, every result list ever tendered, and every path taken as a result. It lives in many places, but . . . AOL, Google, MSN, Yahoo . . . hold a massive amount of this data. Taken together, this information represents . . . a massive clickstream database of desires, needs, wants, and preferences that can be discovered, subpoenaed, archived, tracked, and exploited for all sorts of ends.[47]

These search queries can be subpoenaed and used against litigants to show motive or preparation for certain behaviors of interest.[48] They can be accessed by hackers, disclosed by wayward insiders, or subpoenaed by the DOJ in support of government efforts to enforce an online child pornography law.[49]

Web searches only mark the first level of services offered online. Calendars, web-based emails, and a coming generation of new services mean that internet companies will collect, organize, and store ever more information. As Eric Schmidt, the CEO of Google, has explained, gathering more personal data is the key to Google's future. Schmidt states, "We are very early in the total information we have

---

47 John Battelle, *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture* 6 (Penguin 2005).

48 See, for example, Elise Ackerman, *What Do They Know about You?: MN Survey of Big 4 Firms Shows Your Wanderings Online May Not Be as Secret as You Would Like*, San Jose Mercury News A1 (Aug 20, 2006) (describing the successful murder prosecution of Robert Petrick, which relied in part on evidence that Petrick had googled the words "neck," "snap," and "break," and that he had researched information pertaining to the depth, currents, and accessibility of the lake in which his victim's body was found).

49 See Tom Zeller, Jr., *Privacy vs. Viewing the Internet User as a Commodity*, NY Times C1 (Aug 12, 2006) (mentioning attempts to obtain private data in the fight against online child pornography and other possible security risks to private data).

within Google. The algorithms will get better and we will get better at personalization." He also stated, "The goal is to enable Google users to be able to ask the question such as 'What shall I do tomorrow?' and 'What job shall I take?'"[50]

B. US Information Privacy Law

These privacy concerns are exacerbated because there is no comprehensive information privacy law in the US regulating private sector collection and use of personal data.[51] Despite a patchwork of sector-specific privacy regulations, neither the Constitution nor a general set of laws regulates commercial companies' overall data practices as they affect privacy. Moreover, the government faces few hurdles in gaining access to any information that the private sector collects.[52]

Once a person discloses information to a third party, as she does when requesting a URL or when running search queries, she relinquishes any reasonable expectation of constitutional privacy she has in that information.[53] As one of the authors of this essay has noted, information privacy law in the US contains a strand that considers privacy merely as an interest in "data seclusion."[54] Individuals have a right to keep their information secluded, but once they share it with others, privacy rights end. The Supreme Court relies on this paradigm and interprets the Fourth Amendment as protecting only information that has not been shared with others.[55] Thus, the Fourth Amendment

---

50 Caroline Daniel and Maija Palmer, *Google's Goal: To Organise Your Daily Life*, Fin Times (May 22, 2007) ("Mr Schmidt [also] told journalists in London: 'We cannot even answer the most basic questions because we don't know enough about you. That is the most important aspect of Google's expansion.'").

51 Solove, *The Digital Person* at 67–72 (cited in note 20) (mentioning that Congress has enacted twenty laws dealing with privacy, all of which are narrow in scope). But see *Tech Giants Plan to Push for Privacy Law*, AP (Dec 11, 2006) (noting that Microsoft, HP, eBay, and other high-tech firms are advocating comprehensive federal privacy legislation).

52 See Part II.A.

53 Daniel J. Solove, *Digital Dossiers and the Dissipation of Fourth Amendment Privacy*, 75 S Cal L Rev 1083, 1135 (2002) ("[I]f information is in the hands of third parties, then an individual can have no reasonable expectation of privacy in that information, which means that the Fourth Amendment does not apply. Individuals thus probably do not have a reasonable expectation of privacy in communications and records maintained by ISPs or [ ] network system administrators.").

54 See Paul M. Schwartz, *Privacy and Democracy in Cyberspace*, 52 Vand L Rev 1609, 1662–64 (1999).

55 See Solove, 75 S Cal L Rev at 1134–37 (cited in note 53) (describing the Supreme Court's Fourth Amendment jurisprudence with respect to protections afforded to personal information shared with others). See also *Smith v Maryland*, 442 US 735, 742 (1979) (holding that there was no reasonable expectation that phone numbers individuals dial will be kept private because "[a]ll telephone users realize that they must 'convey' phone numbers to the telephone company, since it is through telephone company switching equipment that their calls are completed"); *United States v Miller*, 425 US 435, 443 (1976) (holding that there was not a reasonable expectation that

protects neither bank records nor information shared with one's accountant. And whatever vestige of choice an individual had before the internet to keep data isolated quickly vanishes once she keeps her daily calendar or conducts her personal communications online.

In response to the absence of constitutional protections, Congress has enacted an incomplete patchwork of information privacy statutes. For a variety of reasons, the current statutory framework is, by and large, inadequate to protect privacy against the growing availability of personal information. As a consequence, information possessed by third parties, such as search engines and ISPs, can be acquired by the government through subpoenas or court orders that do not carry with them the same judicial oversight, or require the same level of particularized suspicion, that the law prescribes for warrants.[56] Numerous scholars have proposed that more restrictions be placed on the government's gathering of information from third parties.[57]

C. Technologies That Preserve Anonymity or Pseudonymity

In the absence of new legal protections, individuals are largely on their own in avoiding commercial dataveillance. Techniques and technologies do exist, however, to assist a user in concealing her true identity or in resisting online profiling.[58] This concealment can, in turn, hinder law enforcement or commercial surveillance, the other side of a two-edged sword that we briefly discuss in Part II.D. For example, a user may be able to shield or distort her "digital persona" by adopting

bank information would remain private in spite of the narrow purpose for which documents were provided).

[56] See Solove, 75 S Cal L Rev at 1085–86, 1138–51 (cited in note 53) (detailing modest statutory restrictions on government access to information possessed by third parties).

[57] See, for example, Center for Democracy & Technology, *Digital Search and Seizure: Updating Privacy Protections to Keep Pace with Technology* 30 (Feb 2006), online at http://www.cdt.org/publications/digital-search-and-seizure.pdf (visited Jan 12, 2008) ("[W]e believe that a probable cause standard should . . . be applied to stored location information obtained from third party providers, such as cellular phone companies and car navigation system companies."); Solove, 75 S Cal L Rev at 1084–87 (cited in note 53) ("Since information maintained by third parties is exposed to others, it is not private, and therefore not protected by the Fourth Amendment. This conception of privacy is not responsive to life in the modern Information Age, where most personal information exists in the record systems of hundreds of entities."); Schwartz, 52 Vand L Rev at 1667–70 (cited in note 54) (citing Robert Post's "pessimism about the creation of privacy rules under the conditions of contemporary life" but concluding that "in the context of online privacy we can reject Post's negative conclusions about the inability to create meaningful privacy rules in the age of organizations").

[58] See, for example, Kevin M. Martin, *Internet Anonymizing Techniques*, online at http://www.usenix.org/publications/login/1998-5/martin.html (visited Jan 12, 2008). See also Center for Democracy & Technology, *CDT's Guide to Online Privacy*, online at http://www.cdt.org/privacy/guide/basic/topten.html (visited Jan 12, 2008) (enumerating a commonsensical "top ten" list for protecting online privacy).

multiple online identities,[59] selectively providing inaccurate identifying information,[60] employing cookie-blocking browser settings, using cookie-managing technologies that prevent websites from tracking her internet habits,[61] or installing commercially available tools (such as the Anonymizer) that facilitate anonymous web surfing.[62]

In this Part, we limit our discussion to a small number of techniques invented by David Chaum and modified and enhanced by other cryptographers. In particular, we look at "onion routing" (which is based on Chaum's "mix" networks) and the "Tor" implementation of second-generation onion routing.[63] We also review a pair of pseudonymity techniques (unlinkable pseudonyms and anonymous credentials) that are finding their way into a new generation of sophisticated solutions for managing digital identities.[64] We assess how well these techniques meet two conditions for success: (1) full or partial concealing of identity to protect privacy; and (2) a combination of ease of use and ubiquity that will make such tools widely available to ordinary users.

Tor is one of the better known applications of so-called "onion routing" technology that enables a user to communicate anonymously on the internet.[65] Onion routing is a technique that involves the direct-

---

59 See Roger Clarke, *Privacy on the Internet—Threats* (Oct 19, 1997), online at http://www.anu.edu.au/people/Roger.Clarke/DV/InternetThreats.html (visited Jan 12, 2008) (advocating the adoption of a false identity or "digital persona" as one of "many ways in which little people can do significant harm to the interests of large organisations, and force them to change their behaviour"); Roger Clarke, *The Digital Persona and Its Application to Data Surveillance*, 10 Info Socy No 2, 77 (June 1994), online at http://www.anu.edu.au/people/Roger.Clarke/DV/DigPersona.html (visited Jan 12, 2008).

60 Clarke also suggests that at times one might consider providing explicitly false information, particularly in instances where the personally identifying information sought is not necessary. See Clarke, *Privacy on the Internet—Threats* (cited in note 59) (advocating even subtle changes to identifying information given out to different sites).

61 See Dennis O'Reilly, *Utilities Clean Cookie Crumbs from Your Hard Drive*, PC World (Apr 3, 2001), online at http://www.pcworld.com/article/id,44901-page,1/article.html (visited Jan 12, 2008) (reviewing different cookie-management products).

62 See, for example, Anonymizer, *Anonymous Surfing*, online at http://www.anonymizer.com/consumer/products/anonymous_surfing (visited Jan 12, 2008) ("Anonymous Surfing hides your IP address so online snoops are unable to track the sites you visit and build profiles on your Internet activities.").

63 See David Chaum, *Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms*, 24 Commun of the ACM 84, 84 (1981) (theoretically describing "[a] technique based on public key cryptography . . . that allows an electronic mail system to hide who a participant communicates with as well as the content of that communication—in spite of an unsecured underlying telecommunication system" or the lack of "a universally trusted authority").

64 See David Chaum, *Security without Identification: Transaction Systems to Make Big Brother Obsolete*, 28 Commun of the ACM 1030, 1030–31 (1985).

65 See generally Roger Dingledine, Nick Matthewson, and Paul Syverson, *Tor: The Second-Generation Onion Router* (Aug 2004), online at http://www.torproject.org/svn/trunk/doc/design-paper/tor-design.pdf (visited Jan 12, 2008) (describing a second-generation onion-routing system with "perfect forward secrecy, congestion control, directory servers, integrity checking, configurable exit policies, and a practical design for location-hidden services via rendezvous points").

ing of messages (including web traffic and email) from their source to their destination via a sequence of proxies (called onion routers) that reroute messages in an unpredictable path. By routing a sender's data through a number of separately encrypted servers, each of which can read only where the data immediately came from and where the data are immediately going, it allows the sender to conceal her identity.[66] Tor readily meets our first condition: it allows anonymous web browsing and offers additional privacy features when used in combination with Privoxy.[67] But Tor fails to meet our second condition. By this failure, it shares the fate of a long list of anonymity tools, many of which remain underutilized for a variety of reasons.[68]

First, average users have not embraced anonymization tools. This reluctance is due to a combination of poor ease of use[69] and ignorance of and, perhaps, apathy towards invasions of privacy.[70] Second, the lack of commercial success of such tools has deterred entrepreneurs from investing in tools that might be more accessible to a broader audience.[71] Third, as long as ISPs have financial incentives to collect data, they are not particularly eager to take the lead in promoting anonymi-

---

66 Id.

67 Id. See also *Privoxy—Homepage*, online at http://www.privoxy.org (visited Jan 12, 2008) ("Privoxy is a web proxy with advanced filtering capabilities for protecting privacy, modifying web page data, managing cookies, controlling access, and removing ads, banners, pop-ups and other obnoxious Internet junk.").

68 For extensive listings of anonymization programs, see http://www.freeproxy.ru/en/free_proxy/cgi-proxy.htm (visited Jan 12, 2008); http://www.hsinlin.com/tips/anonymous_surfing.html (visited Jan 12, 2008).

69 Incorporating these technologies requires users to invest time and energy in finding the right tool and then installing, configuring, using, and maintaining it. Moreover, to the extent these tools slow down internet use, even sophisticated users may discount their utility. See Roger Dingledine and Nick Mathewson, *Anonymity Loves Company: Usability and the Network Effect*, in Lorrie Faith Cranor and Simson Garfinkel, eds, *Security and Usability: Designing Secure Systems That People Can Use* 547, 548–49 (O'Reilly 2005) (describing a variety of reasons why users disable security measures and explaining why usability is important for privacy software).

70 *Search Engines Are at the Center of Privacy Debate*, Info Wk (Mar 1, 2006), online at http://www.informationweek.com/news/showArticle.jhtml?articleID=181401639 (visited Jan 12, 2008) (quoting Ramez Naam, group program manager for MSN Search, as saying, "Privacy is not something that people are saying this is priority one, you have to have this [protection;] . . . few people are so motivated [to make themselves anonymous] that they would install this").

71 See Jonathan D. Glater, *Online, But Out of Sight: Anonymity Gets More Popular*, Intl Herald Trib 17 (Jan 26, 2006) (noting that companies have "moved away from marketing products that protect identity" because of the failure of private companies in the industry and describing how commercial efforts are now focused on selling security to ISPs, not privacy protection to consumers); Ian Goldberg, *Privacy-Enhancing Technologies for the Internet, II: Five Years Later* 6 (2002), online at http://www.cypherpunks.ca/~iang/pubs/pet2.pdf (visited Jan 12, 2008) (noting that "every commercial privacy technology venture [but the Anonymizer] has failed"); Anick Jesdanun, *Privacy Service Gives Less Secrecy*, Chi Trib C3 (Mar 18, 2002) (describing the failure of Zero Knowledge's Freedom Network).

zation tools and making them pervasive and easy to use (in a way that would unburden the individual user).[72]

The next two reasons for underutilization of these technologies may be based on the understanding that worse than a lack of anonymization on the internet is flawed, weak, or incomplete anonymization. Thus, fourth, some users may recognize that many anonymization tools may be vulnerable to attacks.[73] Moreover, fifth, users may also be concerned that their very use of these technologies, especially during a period of low overall adoption, may draw unwanted attention from government security agencies and other parties. As some have argued, "anonymity loves company."[74] Or, to express the same thought through a related metaphor, anonymity systems function best in a crowd. Finally, to the extent social or legal sanctions play a role in controlling behavior, anonymity will lessen inhibitions. Heightened anonymity may lead to reckless behavior that violates civil and criminal laws, as well as standards of decency and propriety. In summary, there are many reasons why industry and government have been reluctant to facilitate the widespread use of anonymization tools, and why users have not adopted them.

Unlinkable pseudonyms and anonymous credentials also satisfy our first condition, which concerns protecting privacy by concealing identity. With unlinkable pseudonyms, "you can only have one pseudonym per organization, but *no-one* can link your pseudonyms to each other or to your real identity, even if all the organizations in the system conspire against you."[75] Unlinkable pseudonyms also protect against dataveillance by allowing a user to register with multiple websites by using a different pseudonym with each one, thereby avoiding profiling based on use of identity credentials. In addition to a pseudonym, a website may need verification of one or more claims or privileges relating to that pseudonym. An anonymous credential supplies this verification; it "is a proof about some fact about one of your pseudonyms which does not reveal either this pseudonym or [ ] your identity."[76] An unlinkable pseudonym, in conjunction with an anonymous credential, makes linking of identities more difficult by enabling a user to prove some single fact without having to reveal extraneous personal data. Stefan Brands refers to this capability to decide how much data to disclose in a given transaction as "selective disclosure"

---

72 See *Search Engines Are at the Center of Privacy Debate* (cited in note 70).

73 Dingledine, Mathewson, and Syverson, *Tor* §§ 7, 9 (cited in note 65).

74 Dingledine and Mathewson, *Anonymity Loves Company* at 549 & n 4 (cited in note 69).

75 Miranda Mowbray, *Implementing Pseudonymity*, 3 SCRIPT-ed 34, 36 (Mar 2006), online at http://www.law.ed.ac.uk/ahrc/script-ed/vol3-1/mowbray.pdf (visited Jan 12, 2008).

76 Id.

(which is the inverse of "selective revelation" in the back-end data mining context).[77] Thus, a user can use this technology to prove only that she meets an age requirement, or is authorized to access a restricted website, or even that she is not on a no-fly list. She need reveal nothing more about herself. Indeed, two of the new breed of identity management tools discussed below implement these or similar features.[78]

This pair of pseudonymity tools also meets our second condition. This positive judgment requires some additional background as well as some speculation about future demands that users will place upon the internet. Over the past several years, software developers have begun to develop sophisticated, user-centric solutions to managing digital identity.[79] As Kim Cameron has noted, there are many existing identity technologies, and no single identity system is likely to replace them.[80] Yet these identity solutions share a number of common traits, beginning with a recognition that the internet was built without an identity layer but needs one to overcome identity theft and other forms of fraud. Moreover, identity solutions view identities as serving different purposes within different contexts and individuals as relying on multiple identities with the goal of controlling how much personal information to reveal in any given situation. Finally, digital identity systems must meet three core privacy requirements. The systems must (1) make data flows explicit and subject to data owners' control; (2) support data minimization by disclosing no more data than is needed in a given context; and (3) impose limits on linkability.[81]

---

77 Stefan A. Brands, *Rethinking Public Key Infrastructures and Digital Certificates: Building in Privacy* § 1.2.3 at 31 (MIT 2000) (noting that the selective disclosure paradigm accommodates a diversity of privacy preferences).

78 See, for example, Credentica, *U-Prove SDK Overview* 4 (Apr 16, 2007), online at http://www.credentica.com/files/U-ProveSDKWhitepaper.pdf (visited Jan 12, 2008) (describing Credentica's ID Token technology as supporting the full privacy spectrum from anonymity to pseudonymity to full identification). ID Tokens implement privacy-protective cryptographic protocols invented by Stefan Brands and described in his book. See Brands, *Rethinking Public Key Infrastructures* § 2.6.2–.3 at 87–90 (cited in note 77) (explaining how cryptographic actions can be used with secret-key certificates to protect anonymity while allowing for a "non-trivial part" of a secret communication to be sent). See also Jan Camenisch and Els Van Herreweghen, *Design and Implementation of the* idemix *Anonymous Credential System* § 1 (Nov 18–22, 2002), online at http://www.zurich.ibm.com/security/publications/2002/camher02b.pdf (visited Jan 12, 2008) (describing an anonymous credential system developed by an IBM researcher).

79 See Mike Neuenschwander, *User-centric Identity Management and the Enterprise: Why Empowering Users Is Good Business* (Burton Group, Dec 2005), online at http://www.tbg.com/Research/PublicDocument.aspx?cid=736 (password protected) (describing efforts by Microsoft, Sxip, and Credentica).

80 See Kim Cameron, *The Laws of Identity* (May 2005), online at http://msdn2.microsoft.com/en-us/library/ms996456.aspx (visited Jan 12, 2008) (noting that it is difficult if not impossible to add a single identity layer for the entire internet because of different contexts of use by different players).

81 Such identity solutions are frequently described as privacy-enhancing. See, for example, Marit Hansen, et al, *Privacy-enhancing Identity Management*, 9 Info Sec Technical Rep 35, 35–44

With this background, the pseudonymity techniques implemented in new identity management tools such as Credentica and *idemix* should be easy to use, and, in our judgment, stand a reasonable chance of gaining widespread acceptance over time. First, these new tools are designed to be user-centric. They eliminate cumbersome username/password credentials (and the temptation to store these credentials in insecure places for convenience) and enable users to store identity "tokens" from a variety of service providers in an easy-to-use digital "wallet." Second, they offer greater protection against phishing attacks and identity theft by improving user interfaces and authenticating sites to users. Third, they also enable organizations to meet key security and privacy requirements in a number of scenarios where the absence of sophisticated identity tools has impeded successful deployments. These situations include e-health, where patients and medical professionals need tightly controlled but ready access to patient health records, and national defense, which presents unique access control, security clearance, and audit requirements.[82] In addition, identity management tools can satisfy the interests of commercial companies in knowing and reaching their customers.[83] Finally, the new generation of identity tools is already in the hands of the general public and beginning to gain acceptance.[84] In sum, it seems clear that the identity tools described above meet both our conditions: they are de-

(2004) (discussing required components of "pervasive privacy-enhancing identity management"). Of course, poorly designed identity systems have also been criticized for creating serious privacy concerns to the extent that they involve the use, transfer, and retention of personal information. This is especially true in the case of large-scale systems such as a national ID system, which may rely on a centralized authentication service, thereby increasing the risk of improper information sharing, data mining, and profiling by government agencies and even private enterprises connected to the centralized services. See Brands, *Rethinking Public Key Infrastructures* § 1.2.3 at 31 (cited in note 77) (describing the privacy concerns of authentication systems based on public-key infrastructures and digital certificates). See also Stephen T. Kent and Lynette I. Millett, eds, *Who Goes There?: Authentication through the Lens of Privacy* 177 (National Academies 2003) (highlighting the risk that a nationwide identity system "could easily result in inappropriate linkages among nominally independent databases").

82 For a discussion of these and other scenarios, see Credentica, *Target Markets*, online at http://www.credentica.com/target_markets.html (visited Jan 12, 2008) (noting how Credentica addresses data protection and scalability requirements in various markets, including government and healthcare, and linking to further information about those contexts).

83 Wide deployment of privacy-enhanced identity tools may result in a greater willingness by consumers to identify themselves and even to receive more personalized ads provided they remain in control. But mere use of CardSpaces or ID Tokens has no immediate impact on privacy concerns posed by the use of cookies for tracking purposes because cookies use the HTTP protocol, which is independent of the protocols on which privacy-enhanced identity tools rely.

84 For example, Microsoft's identity management tool, Windows CardSpace, shipped with Windows Vista and will also work with Windows XP. IBM and Novell have announced their support for an open-source identity framework called Higgins. See IBM, *Open Source Initiative to Give People More Control over Their Personal Online Information* (Feb 27, 2006), online at http://www-03.ibm.com/press/us/en/pressrelease/19280.wss (visited Jan 12, 2008).

signed, at least in part, to protect privacy by allowing users greater control over their online identities and they are reasonably likely to be widely deployed in a user-friendly manner.

D. Tradeoffs and Similarities: A Shared Consideration of Data Mining and Identity Management Systems

Thus far, we have considered the consensus view concerning how the law should regulate data mining to further counterterrorism goals and also to protect privacy interests. Data mining can be viewed as a "back end" use of personal data that is already collected and resident in public and private sector databases. This essay has also discussed emerging technologies that complement safeguards on "back end" use by limiting the "front end" identification of users and collection of personal data about them. We next turn to a brief look at issues in common shared by the consensus view's proposed limits on data mining and the new identity management systems.

Two issues are of special interest. The first concerns the risk that these systems will limit at least some data collection or use, and thereby make counterterrorism and other public safety efforts more difficult. The second concerns the need for transparency as these complex systems are developed and controlled and as security tradeoffs emerge.

Regarding the first issue, the consensus view calls for checks and balances on government data mining that are not only justified in their own right but may also contribute to accuracy. However, privacy-enhancing identity management systems might hamper the government's ability to identify online users or associate a digital identity with a "real" person and therefore with that person's potentially threatening conduct offline.[85] In response, some countries have sought to overcome anonymity and pseudonymity by imposing mandatory user registration systems. For example, China has tried to require bloggers to use their real names and official identification, but it seems recently to have backed away from this requirement.[86]

One policy model for regulating these kinds of emerging technologies appears in the Communications Assistance for Law En-

---

[85] The government's capacity to conduct surveillance or to comply with legal limits on surveillance and protect individual privacy depends on its ability to identify a particular online user. This is because judicial authority to conduct electronic surveillance is in large part based on an individual's identity as well as on other factors, including her geographical location.

[86] Steven Schwankert, *China Drops Real-name Blogger Plan*, Infoworld (May 23, 2007), online at http://www.infoworld.com/article/07/05/23/China-drops-real-name-blogs_1.html (visited Jan 12, 2008).

forcement Act[87] (CALEA), but we question the extension of this conceptual approach to the identity management context. Congress enacted CALEA to preserve the ability of law enforcement officials to conduct electronic surveillance involving digital telephony. This law requires telecommunications carriers and manufacturers of telecommunications equipment to design their equipment, facilities, and services to ensure that a required level of surveillance capabilities will be built in.[88] In August 2005, the FCC interpreted CALEA as covering internet broadband providers and Voice over Internet Protocol (VoIP) providers.[89] This rulemaking, upheld by the DC Circuit in June 2006, established that broadband and VoIP are hybrid telecommunications-information services that fall under CALEA to the extent that they qualify as "telecommunications carriers."[90] The CALEA model assimilates new technologies to a status quo at a given date, roughly analog telephony as it existed in 1994. New technologies for telephony are shaped to allow at least as much surveillance capacity as at that time.

The conceptual gap that CALEA glosses over is, of course, that new technologies for telephony go far beyond the capacities present in 1994—at that time, for example, analog systems did not make available the array of call-related information now available in digital systems, cell phones were not widely adopted, and people were not using their telephones to access the internet. CALEA requires the construction of legal fictions to bridge the world of now and then. It largely allows the FCC, FBI, and telecommunications carriers to elaborate

---

87 Communications Assistance for Law Enforcement Act, Pub L No 103-404, 108 Stat 4279, 4280–81, codified at 47 USC § 1002 (2000).

88 47 USC § 1002:

> [A] telecommunications carrier shall ensure that its equipment, facilities, or services that provide a customer or subscriber with the ability to originate, terminate, or direct communications are capable of [ ] expeditiously isolating and enabling the government, pursuant to a court order or other lawful authorization, to intercept, to the exclusion of any other communications, all wire and electronic communications carried by the carrier.

89 See *In re CALEA and Broadband Access and Services*, 20 FCCR 14989, 14989 (2005):

> In this Order, we conclude that the Communications Assistance for Law Enforcement Act (CALEA) applies to facilities-based broadband Internet access providers and providers of interconnected [VoIP] service. This Order is the first critical step to apply CALEA obligations to new technologies and services that are increasingly relied upon by the American public to meet their communications needs.

The decision was reconsidered in part. *In re CALEA and Broadband Access and Services*, 21 FCCR 5360, 5361 (2006) (providing facilities-based broadband internet access providers, VoIP, and owners of similar services a period of time in which to bring themselves into compliance with CALEA and declining to intervene in the standards development process).

90 See *In re CALEA*, 20 FCCR at 15002, 15009; *American Council on Education v FCC*, 451 F3d 226, 235 (DC Cir 2006).

these fictions and then work out their practical implementation through an administrative process.

Even greater problems would exist if this policy approach were extended to identity management systems. There is no readily ascertainable status quo to which identity management systems might be compared. Even if there were such a status quo, policymakers would need to decide the reference date for the status quo, just as CALEA chose the reference date of 1994. The extent to which public safety interests should figure in the development of identity management systems, and the safeguards for government access to investigative information needed when pseudonymous identities are managed through systems on the internet, deserve to be thought out on their own terms. Therefore, it would be premature for the government to consider unilaterally imposing any surveillance-based design mandates on identity management systems at this time.

Our second common theme is the need for transparency in both data mining and identity management systems. The emerging framework for regulating data mining represents a worthwhile first attempt to provide a social and legal context for its use. A similar discourse is needed to guide internet profiling. In turn, increased knowledge about the reliability and track record of both government and commercial data mining and the actual makeup and operation of data mining systems is important for an informed debate about how best to regulate data mining. We are unwilling, at this juncture, to join Schneier's condemnation of it. Judge Posner, in contrast, fails to consider that there is human intervention in data mining even before the first automated search is run; humans will write the software, shape the database parameters, and decide on the kinds of matches that count. And the task of data mining itself is guided by some degree of human interaction. Like data mining, identity management systems rely on theoretical models and assumptions. We note the risk that a false sense of security in identity management systems may lead to potentially greater privacy invasions compared to more guarded or less extensive internet activities without such systems.

The necessary response in both areas is for documentation and study of the design, performance, and privacy protections of the systems both before deployment and over time. One analogy is with fingerprinting and its acceptance as an investigative technique. Fingerprinting was used for a century before the law began the process of developing reasonably authoritative standards to put it on a sound

empirical footing.[91] Only recently has research been carried out regarding issues such as the number of points needed to be matched between a fingerprint under examination and a known exemplar in order to conclude that the two examples belong to the same individual.[92] Any use of data mining should occur in parallel with development of sound scientific models for its use.

E. Questions for Further Study

The preceding discussion touches on the implications for counterterrorism efforts of technologies that enable identity concealment. But this is only one of several policy issues requiring further research and analysis. This essay concludes by raising three related issues.

First, to what extent should the various safeguards and privacy protections identified in the consensus view on government data mining activities apply to similar efforts in which the private sector engages? How would the consensus safeguards be developed and administered by numerous private sector actors? Moreover, there might be additional privacy-enhancing practices that consumers might wish certain companies to adopt, and, in contrast, other circumstances in which they might accept fewer safeguards or more revelation of their identities in exchange for lower prices or more services.[93] Hence, it is an open question whether and how the law should: require private companies that data mine to implement access controls and authentication of personnel; impose data quality standards and anonymization of data mining results used for online advertising purposes; require periodic audits and validation of data mining programs; or mandate oversight of internal controls to ensure accountability.

Second, what is the likelihood that ad-funded web services will adopt the new breed of privacy-enhanced identity management systems? As noted above, websites tailor online ads to the interests of their customers by using cookies to collect data and track web behavior, often without their customers' knowledge and consent. The new identity technologies may induce consumers to share more information with web sites willingly, thereby opening up possibilities for personalized ads based on an explicit value proposition. Indeed, firms might even allow a user not only to exchange personal data for tai-

---

91 Jennifer L. Mnookin, *Fingerprint Evidence in an Age of DNA Profiling*, 67 Brooklyn L Rev 13, 16–43 (2001).

92 Id at 57–71 (pointing out issues recently raised about the statistical likelihood of a match given an incomplete print).

93 See Paul M. Schwartz, *Property, Privacy, and Personal Data*, 117 Harv L Rev 2056, 2076–84 (2004) (noting, however, the difficulties involved in achieving price discrimination in personal information markets).

lored ads, but also to keep her personal data under her own control. But web firms may be reluctant to make this transition if they perceive the new identity technologies as threatening their online advertising revenues. As an additional difficulty, even if websites adopt these new technologies, it is unclear whether they will offer users the full spectrum of privacy capabilities from anonymity to pseudonymity to full identification. These choices might also mean foregoing the opportunity to collect personal information for ad targeting and other profitable uses.

Third, and finally, if the new identity technologies with privacy features are widely adopted and succeed in preserving anonymity and pseudonymity, what will be the broader societal impact? The internet already suffers from a lack of accountability, which many commentators view as the underlying cause of some users behaving irresponsibly since they are unafraid of the consequences and will not suffer any sanctions if they violate various criminal laws or social norms.[94] Indeed, anonymity is often blamed for a variety of undesirable and prohibited behavior, ranging from defamatory or libelous statements and the distribution of offensive or pornographic materials to a host of specifically internet-related offenses. These internet offenses include launching viruses, engaging in phishing attacks, and sending spam or downloading spyware on a PC without the user's consent. The question will be whether new identity technologies help preserve privacy without exacerbating the problems associated with a lack of accountability. The assumption underlying this question, of course, is that these identity technologies gain widespread use. The hope is that they might provide more nuanced controls to help achieve the ideal balance between these sometimes competing values. Yet, and as a last question, we wonder who will turn the dial on these controls as among end-users, service providers, software designers, and the government.

CONCLUSION

Predictive data mining by the government offers promise and peril in its response to terrorism. In this essay, we have considered ways for heightening the positive and reducing the negative aspects of this technique. We have also evaluated the likely emergence of identity management systems within the private sector and some tradeoffs between and similarities with data mining. Greater transparency is

94 See, for example, David R. Johnson, Susan P. Crawford, and John G. Palfrey, Jr., *The Accountable Internet: Peer Production of Internet Governance*, 9 Va J L & Tech 1, 4–5 (2004) (arguing that accountability on the internet naturally results from the ability to accurately identify those with whom we are communicating).

needed regarding the reliability, track record, and operation of government and commercial data mining systems. In addition, questions remain about the extension of consensus safeguards around government data mining to commercial data mining, the extent and speed of ad-funded web services' adoption of identity management systems, and the broader social impact of greater online anonymity and pseudonymity.