Center for Advanced Studies in Science and Technology Policy
www.advancedstudies.org

• Information • Environment • National Security •

K. A. Taipale

Final Pre-publication Draft • December 2003 v.3.0b
<http://www.taipale.org/papers/datamining.pdf>

Data mining and Domestic Security: Connecting the Dots to Make Sense of Data

To Be Published as:
5 Colum. Sci. & Tech. L. Rev. (December 2003),
<http://www.stlr.org/cite.cgi?volume=5&article=2>

Executive Summary

Article: 45,000 words

Center for Advanced Studies Preprint
• New York •

About the Center for Advanced Studies

The Center For Advanced Studies in Science and Technology Policy is a private, non-partisan research institute dedicated to developing and advocating advanced information, environment and national security policies that are pro-technology and pro-economic development but progressive, sustainable and humane.

The Center was founded on the premise that information policy, that is, how we develop, manage and regulate the creation and use of information in the emergent digital world, and environmental policy, that is, how we develop, manage and regulate the use and consumption of natural resources in the physical world, will be among the prime determinants of much of the quality of our future public and private lives. In turn, resolving fundamental conflicts within these policy areas on a national and international scale will have great direct affect on national and global security.

The Center advocates information and communication policies that promote freedom, democracy and civil liberties while encouraging and protecting intellectual property and national security. The Center advocates environment and energy policies that are sustainable and conserve resources while creating new alternative investment and growth opportunities and, where possible, enabling market mechanisms.

The Center seeks to influence on every level national and international decision makers in both the public and private sectors by providing sound, objective analysis, in particular by identifying and articulating issues that lie at the intersection of technologically enabled change and existing practice in law, policy and industry.

For more information visit the Center's web site at http://www.advancedstudies.org/.

This is a Center for Advanced Studies Preprint of:

Executive Summary, "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data" (December 2003 v3.0B) Final Pre-Publication Draft

Previous Releases: Version 1.0B (April 2003), Version 2.0B (September 2003)

Citation Format for Published Version

K. A. Taipale, "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data," 5 Colum. Sci. & Tech. L. Rev. (December 2003), available at http://www.stlr.org/cite.cgi?volume=5&article=2

Table of Contents

K. A. Taipale
Final Pre-publication Draft • December 2003 v.3.0b
<http://www.taipale.org/papers/datamining.pdf>

Data mining and Domestic Security: Connecting the Dots to Make Sense of Data

Executive Summary

Article: 45,000 words

Article Abstract

Official U.S. Government policy calls for the research, development and implementation of advanced information technologies for aggregating and analyzing data, including data mining, in the effort to protect domestic security. Civil libertarians and libertarians alike have decried and opposed these efforts as an unprecedented invasion of privacy and a threat to our freedoms.

This article examines data aggregation and automated analysis, particularly data mining, and related privacy concerns in the context of employing these techniques in domestic security. The purpose of this article is not to critique or endorse any particular proposed use of these technologies but, rather, to inform the debate by elucidating the intersection of technology potential and development with legitimate privacy concerns. It is a premise of this article that security and privacy are dual obligations, not dichotomous rivals to be traded one for the other. "In a liberal republic, liberty presupposes security; the point of security is liberty." [For citation, see FN 45 in the article]

Thus, this article argues that security with privacy can be achieved by employing value sensitive technology development strategies that take privacy concerns into account during development, in particular, by building in rule-based processing, selective revelation, and strong credential and audit features. This article does not argue that these technical features alone can eliminate privacy concerns but, rather, that these features can enable familiar, existing privacy protecting oversight and control mechanisms, procedures and doctrines (or their analogues) to be applied in order to control the use of these new technologies.

Further, this article argues that *not* proceeding with government funded research and development of these technologies (in which political oversight can incorporate privacy protecting features into the design of these technologies) will ultimately lead to a diminution in privacy protection as alternative technologies developed without oversight (in classified government programs or proprietary commercial programs) are employed in the future since those technologies may lack the technical features to protect privacy through legal and procedural mechanisms. Thus, the recent defunding of DARPA's Information Awareness Office and its Terrorism Information Awareness program and related projects is likely turn to out to be a pyrrhic 'victory' for civil liberties as this program provided a focused opportunity around which to publicly debate the rules and procedures for the future use of these technologies and, importantly, to oversee the development of the appropriate technical features required to support any concurred upon implementation or oversight policies to protect privacy.

Even if it were possible, controlling technology through law alone, for example, by outlawing the use of certain technologies or shutting down any particular research project, is likely to provide little or no security and only brittle privacy protection.

Article Overview

Vast Data Volumes Exceed Analytic Resources

Recent reports by the U.S. Congress, the National Research Council, the Markle Foundation and others have highlighted that the amount of available data to be analyzed for domestic security purposes exceeds the capacity to analyze it. Further, these reports identify a failure to use information technology to effectively address this problem.

> "While technology remains one of this nation's greatest advantages, it has not been fully and effectively applied in support of U. S. counter-terrorism efforts." [FN 47]

Among the recommendations put forth in these reports are the increased use of data aggregation (information sharing) and automated analysis (in particular data mining) technologies.

Data Aggregation and Automated Analysis

Data aggregation (including data integration and data sharing) is intended to overcome the "stovepipe" nature of existing datasets. Research here is focused on making information available to analysts regardless of where it is located or how it is structured. A threshold issue that has technical, security and privacy implications is whether to aggregate data in a centralized data warehouse or to access information directly in distributed databases.

Automated data analysis (including data-mining) is intended to turn low-level data, usually too voluminous to understand, into higher forms (information or knowledge) that might be more compact (for example, a summary), more abstract (for example, a descriptive model), or more useful (for example, a predictive model).

> "A key problem [for using data mining for counter-terrorism] is to identify high-level things – organizations and activities – based on low-level data – people, places, things and events." [FN 67]

The application of data aggregation and automated analysis technologies to domestic security is the attempt to "make sense of data" by automating certain analytic tasks to allow for better and more timely analysis of existing datasets in order to prevent terrorist acts by identifying and cataloging various threads and pieces of information that may already exist but remain unnoticed using traditional means, and to develop predictive models based on known or unknown patterns to identify additional people, objects or actions that are deserving of further resource commitment or law enforcement attention.

Compounding the problem in domestic security applications is that relevant data (that is, information about terrorist organizations and activities) is hidden within vast amounts of irrelevant data and appears innocuous (or at least ambivalent) when viewed in isolation. Individual data items – relating to people, places and events, even if identified as relevant – are essentially meaningless unless viewed in context of their relation to other data points. It is the network or pattern itself that must be identified, analyzed and acted upon.

Thus, there are three discrete applications for automated analysis in the context of domestic security:

> first, *subject-oriented link analysis*, that is, automated analysis to learn more about a particular data subject, their relationships, associations and actions;
>
> second, *pattern-analysis* (or data mining in the narrow sense), that is, automated analysis to develop a descriptive or predictive model based on discovered patterns; and,
>
> third, *pattern-matching*, that is, automated analysis using a descriptive or predictive model (whether itself developed through automated analysis or not) against additional datasets to identify other related (or "like") data subjects (people, places, things, relationships, etc.).

Because spectacular terrorist events may be too rare or infrequent for automated analysis to extract useful patterns, the focus of these techniques in counter terrorism is to identify lower level, frequently repeated events (for example, illegal immigration, money transfers, front businesses and recruiting activity) that together may warrant further attention or resource commitment.

Thus, data aggregation and automated analysis are not substitutes for human analytic decision-making, rather, they are tools that can help manage vast data volumes and potentially identify relational networks that may remain hidden to traditional analysis. If successful, these technologies can help allocate available domestic security resources to more likely targets.

Privacy Concerns

Because data aggregation and automated analysis technologies can cast suspicion based on recognizing relationships between individually innocuous data, they raise legitimate privacy concerns. However, much of the public debate regarding the potential use of these technologies is overshadowed by simplifications, misunderstandings and misrepresentations about what the technologies can do, how they are likely to be employed and what actual affects their employ may have on privacy and security.

The significant privacy concerns relating to these technologies are primarily of two kinds: those that arise from the aggregation (or integration) of data itself and those that arise from the automated analysis of data that may not be based on any individualized suspicion – the former might be called the *database problem* and the latter the *mining problem*.

The database problem is implicated in subject-based inquiries that access distributed databases to find more information about a particular subject. To the extent that maintaining certain government inefficiencies helps protect individual rights from centralized state power, the primary privacy question involved in aggregation is one of increased government efficiency.

The mining problem is implicated in the use of pattern-matching inquiries, in which profiles or models are run against data to identify unknown individuals. To some, pattern-matching raises privacy issues relating to non-particularized suspicion in violation of the Fourth Amendment.

Additional concerns are that the technology will not work for the intended purpose (providing either a false sense of security by generating false negatives or imposing civil liberties costs on too many innocent people by generating false positives), and that the technology is subject to potential abuse or that it will be vulnerable to attack.

The issue of false positives and false negatives is not insignificant but is an issue of efficacy and requires further research to determine whether an appropriate *confidence interval* for counter terrorism applications can be achieved. The point of the research is to find out if the technologies can work – if they cannot, other privacy concerns are moot since the technologies will not be employed. If they can, then appropriate policies and procedures to manage and compensate for error rates can be developed before implementation.

Building in Technical Constraints

Assuming some acceptable baseline efficacy, it is the premise of this article that privacy concerns relating to data aggregation and data mining in the context of domestic security can be significantly mitigated by developing technologies that enable the application of existing legal doctrines and related procedures to their use:

> First, that *rule-based processing* and a *distributed database architecture* can significantly ameliorate the general data aggregation problem by limiting the scope of inquiry and the subsequent processing and use of data within policy guidelines;
>
> Second, that *selective revelation* can reduce the non-particularized suspicion problem, by requiring an articulated particularized suspicion and intervention of a judicial procedure before identity is revealed; and

> Finally, *strong credential and audit features* and *diversifying authorization and oversight* can make misuse and abuse "difficult to achieve and easy to uncover". [FN 62]

Further, this article contends that developing these features for use in domestic security applications will lead to significant opportunities to enhance overall privacy protection more broadly in the U.S. (and elsewhere) by making these technical procedures and supporting features available for voluntary or legislated adoption in the private sector. In addition, the development of these technologies will have significant beneficial "spill-over" uses for commercial and scientific applications, including improved information infrastructure security (better user authentication, encryption, and network security), protection of intellectual property (through rule-based processing) and the reduction or elimination of spam (through improved analytic filtering).

Overriding Principles

This article proffers certain guiding principles for the development and implementation of these technologies:

> First, that these technologies only be used as investigative, not evidentiary, tools (that is, used as a predicate for further investigation not proof of guilt) and only for investigations of activities about which there is a political consensus that aggressive preventative strategies are appropriate (for example, counter-terrorism and national security).
>
> Second, that specific implementations be subject to strict congressional oversight and review, be subject to appropriate administrative procedures within executive agencies where they are to be employed, and be subject to appropriate judicial review in accordance with existing due process doctrines.
>
> And, third, that specific technical features that protect privacy by providing opportunities for existing doctrines of due process and reinforcing procedures to function effectively, including rule-based processing, selective revelation and secure credentialing and tamper-proof audit functions, are developed and built into the technologies.

Article Structure

The Prelude and Introduction to this article contextualize the debate about the need for and potential use of these technologies. Part I then provides a more detailed introduction to data aggregation and analysis technologies, in particular, data mining. Part II examines certain government initiatives, including TIA and CAPPS II, as paradigmatic examples of development efforts in these areas. Part III outlines the primary privacy concerns and the related legal framework. Part IV suggests certain technology development strategies that could help ameliorate some of the privacy concerns. And, Part V concludes by restating the overlying principles that should guide development in these technologies.

About the Author

K. A. Taipale
b.a., j.d. (New York University), m.a., ed.m., ll.m. (Columbia University)
Executive Director, the Center for Advanced Studies in Science and Technology Policy

Bio: <http://www.taipale.org/>
Email: <datamining@advancedstudies.org>

Acknowledgements

The author would like to thank the editorial board of the Columbia Science and Technology Law Review and Eben Moglen, Daniel Solove, Paul Rosenzweig, Daniel Gallington, Usama Fayyad and David Jensen whose insights, comments or work helped inform this article.

The views and any errors are solely those of the author.