

Transparency and End-to-End Accountability: Requirements for Web Privacy Policy Languages

¹Daniel J. Weitzner, ¹Harold Abelson, ¹Tim Berners-Lee, ¹Chris Hanson, ²James Hendler, ¹Lalana Kagal, ¹Gerald Jay Sussman, ¹

¹Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, 32 Vassar St., Cambridge, MA USA

²University of Maryland, MIND Lab, 8400 Baltimore Ave., College Park, MD USA

djweitzner@csail.mit.edu, hal@mit.edu, timbl@csail.mit.edu, cph@csail.mit.edu, hendler@cs.umd.edu, lkagal@csail.mit.edu
gjs@mit.edu

Abstract

Attempts to address issues of personal privacy in a world of computerized databases and information networks -- from security technology to data protection regulation to United States Fourth Amendment law jurisprudence -- typically proceed from the perspective of controlling or preventing access to information. We argue that this perspective has become inadequate and obsolete, overtaken by the ease of sharing and copying data and of aggregating and searching across multiple data bases, to reveal private information from public sources. To replace this obsolete framework, we propose that issues of privacy protection currently viewed in terms of data *access* be re-conceptualized in terms of data *use*. From a technology perspective, this requires supplementing legal and technical mechanisms for access control with new mechanisms for transparency and end-to-end accountability of data use.

I. Introduction

Information systems upon which we depend are becoming ever more complex and decentralized. While this makes their power and flexibility grow, it also raises substantial concern about the potential for privacy intrusion and other abuses. Understanding how to incorporate transparency and accountability into decentralized information systems will be critical in helping society to manage the privacy risks that accrue from the explosive progress in communications, storage, and search technology. A prime example of a growing, decentralized information system is the World Wide Web, recently augmented with structured data capabilities and enhanced reasoning power. As the Web gets better and better at storing and manipulating structured data it will become more like a vast global spreadsheet or database, than merely a medium for easy exchange and discovery of documents. Technologies such as XML, Web Services, grids, and the Semantic Web all contribute to this transformation of the Web. While this added structure increases inferencing power, it also leads to the need for far greater transparency and accountability of the inferencing process. By *transparency* we mean that

the history of data manipulations and inferences is maintained and can be examined by authorized parties (who may be the general public). By *end-to-end accountability* we mean that one can check whether the policies that govern data manipulations and inferences were in fact adhered to [FW06]. Furthermore, this accountability assessment must be possible across all parts of the Web over which the transaction in question reaches. Transparency in inferencing systems enables users to have a clear view into the logical and factual bases for the inferences presented by the system. Accountability in inferencing enables users or third parties to assess whether or not the inferences presented comply with the rules and policies applicable to the legal, regulatory or other context in which the inference is relied upon.

Today, when an individual or an enterprise uses a single, self-contained set of data and applications, the controls necessary to assure accuracy and contextualize the results of queries or other analyses are available and generally well understood. But as we leave the well-bounded world of enterprise databases and enter the open, unbounded world of the Web, data users need a new class of tools to verify that the results they see are based on data that is from trustworthy sources and is used according to agreed upon institutional and legal requirements. Hence, we must develop technical, legal and policy foundations for transparency and end-to-end accountability of large-scale aggregation and inferencing across heterogeneous data sources. We can expect a wide range of legal and regulatory requirements on inferencing systems, and some requirements may well overlap or contradict others. This expected diversity of rulesets makes it all the more important to have one common technical framework for managing accountability to rules.

Such transparency and accountability will be important in a variety of cases: for compliance with financial regulations [SOX] and new security and privacy rules for health care data [HIPAA]. Finance and health are just two areas in which the higher quality data management practices are seen as important in connect with greater

reliance on complex information systems. In the most general case, we will trust inferences only when we have a transparent view into their antecedents and will use them appropriately only when we know that we may be held accountable for misuse. A wide range of public and private sector data mining and inferencing applications will benefit from the transparency and accountability mechanisms described here [JoCrPa04].

Transparency and accountability are important features of a larger architectural project to make Web more 'policy aware'. *Policy awareness* is a property of the Web that will provide users with accessible and understandable views of the policies associated with resources, enable agents to act in response to rules on a user's behalf, thereby making compliance with stated rules easier, and afford a greater opportunity for accountability when rules are intentionally or accidentally broken. [WHBC05]

The fundamental technical challenge that must be addressed in order to provide transparency and accountability for reasoning on the Semantic Web is rooted in the open, decentralized architecture of the Web itself. The Semantic Web [BLHL01] is an enhancement of the current Web to allow machine-processable data to span application boundaries in the same way that human-readable documents do currently. The goal of the Semantic Web is as broad as that of the Web: to be a universal medium for data. It is envisaged eventually to smoothly interconnect personal information management, enterprise application integration, and the global sharing of commercial, scientific and cultural data. Introducing transparency into the reasoning occurring over the Semantic Web requires innovative techniques that account for the open, decentralized architecture of the Web.

Beyond the basic architecture of the Web, four more general trends in the use of information should encourage privacy-sensitive system designers to rethink their approach to privacy protection: first, the gradual demise of stove-pipe applications in favor of enterprise-wide data integration; second, the rapidly declining cost of web-scale query; and third, the rapid spread of sensor networks in both public and private settings. Fourth, the cost of data storage is becoming cheaper and cheaper to the point that is often less expensive to just keep all data rather than figure out which information to discard and which to retain. No doubt, there is a fixed cost associated with operation of data storage facilities, but with the rapidly declining cost of disk storage, the cost per data element is approaching zero.

II. Inadequacy of current privacy protection approaches

Current technical investigations of the impact of data mining on privacy have generally focused on limiting access to data at the point of collection or storage. Much effort has been put into the application of cryptographic

and statistical techniques to construct finely tuned access-limiting mechanisms.

Our proposal to rely on transparency and accountability as privacy protection mechanisms stands in contrast to other efforts to engineer privacy protection into information systems. Recently, much work has been done on distributed database systems with secure private computation algorithms (SPCA) [GoMi82] as a means of protecting privacy [BFSW04]. Privacy-preserving data mining algorithms [LiPi02] have shown that it is possible to constrain query power based on some predefined measure of how much information the requestor is entitled to have and some quantified notion of privacy [EGS03]. While such systems may well have their place in some privacy applications, it has not yet been demonstrated that they can be successfully deployed at the scale required to meet privacy requirements for either large scale private sector or government data mining. What's more, the ability to constrain queries in this manner depends on a mathematically-expressible definition of privacy describing the quantitative limits on how much information the government can have [AgSr00].

Compliance with privacy rules can often depend on factual circumstances only manifest *after* a given query has been made, so it is simply impossible to rely on control over query (data collection rules) alone to protect privacy. Furthermore, it will not always be possible to articulate a computable definition of privacy. In many cases, privacy laws rely on some judgment of whether one set of facts 'reasonably' justifies access to some larger set of information, as is the case with a "probable cause" requirement for electronic surveillance. Finally, while SPCA can enable control of the scope of queries within the bounds of a given information system, data may leak out of systems instrumented with SPCA through a variety of channels, not subject to control of the query control mechanisms.

We believe that exclusive reliance on secure, private computation algorithms both under-emphasize the vital need for transparency into the use of data mining, and also may result in over-constraining the use of data mining capability to the detriment of law enforcement needs. Even if such privacy-preserving data mining techniques prove to be practical, they are unlikely to provide sufficient public assurance that government inferences conform to legal restrictions. They also do not address the need to provide citizens the certainty that adverse government action is based on factually accurate data. In sum, *while privacy-preserving data mining techniques are certainly necessary in some contexts, they are not sufficient privacy protection without the transparency and accountability.*

Yet for all this emphasis on access restriction, the reality is that the Web is making it increasingly difficult to limit access to data, while at the same time making it increasingly easy to aggregate data from multiple

information sources, and to do searching and inferencing based on these aggregations. In the long run, access restriction alone cannot suffice neither to protect privacy nor to ensure reliable conclusions. It must be augmented by attention to increased transparency and accountability for the inferencing and aggregation process itself.

III. Transparency and Accountability in the Current Privacy Policy Debate

We have argued that large scale inferencing capabilities pose novel privacy challenges which require a novel response. However, our efforts to structure laws and develop technologies with sensitivity for privacy values should seek guidance from the nearly century-long interplay between ever-growing surveillance capabilities of new technologies and fundamental privacy principles. Historically, we learn that as electronic communications have become more sophisticated and more ubiquitous, communications privacy law has responded to the advance in law enforcement needs *and* privacy threats by tying the growth in surveillance capabilities to gradually expanding privacy protections that kept pace with new intrusion powers. Over the last hundred years in the United States and elsewhere around the world, privacy protections were extended to voice telephone calls, then email, then transactional records, and other communications-related information [De97]. Web-scale inferencing that powers data mining is only the latest in the series of technology advances that demands new privacy protection alongside intrusive surveillance powers [Hsrpt86].

The inherent complexity of web-scale inferencing and data mining dictates that privacy values will not be protected merely by controlled access to personal information in the way that wiretapping laws could simply grant or deny access to a telephone conversation. We will have to supplement *a priori* access control with *a posteriori* accountability to rules. Privacy protection will require both the ability to assure that adverse actions are premised on factually correct antecedents, and that the adverse conclusions are logically grounded in permissible uses of personal information. As the conclusions are reached and acted upon long after the information supporting those conclusions were collected, we obviously cannot rely upon *a priori* control mechanisms operating only at the time of collection. Rather, full accountability to privacy rules cannot be achieved without the *a posteriori* proof techniques we have described here.

Transparency and accountability mechanisms are a vital part of privacy protection going forward because we expect continued expansion in the depth and breadth of data available both to the government and the private sector. The great power of data mining to reveal intimate details about individuals has yet to be matched with either legal or technical measures that balance its impact with privacy

requirements [CDT03]. What's more, there are proposals to expand law enforcement data analysis powers even further. Taking an example from the United States: in calling for the creation of a nationwide network to respond to threat of terrorism, a Markle Foundation Task Force explains that an open, decentralized Web-like architecture is really the only design strategy that could possibly succeed in linking that many disparate entities in law enforcement, homeland security, intelligence, and defense with a role to play. In addition to the twenty-two federal agencies now under the DHS umbrella, the following organizations must be integrated into a single, coordinated information sharing environment:

- 18 federal agencies in the US cabinet
- 17,784 State & Local law enforcement agencies
- 30,020 Fire departments
- 5,801 Hospitals
- 1,700 Private critical infrastructure

[BJS2000][Pa2004]

In such a far-flung and heterogeneous environment, both collection and analysis of data must "occur at multiple nodes, rather than only in a few centralized locations" [Mark03]. Reliance on Web architecture as a model for sharing, analyzing, and managing this data is appropriate not because of any desire to make all of this data public (as much of the Web is) but because institutions have learned that the decentralized addressing model of the Web has been uniquely successful in enabling large-scale coordination of data both inside and outside enterprise boundaries.

How much larger that universe of data grows and how quickly this happens is a matter for public policy makers to decide in an open, democratic process. As technology designers, however, we can provide information infrastructure that help society be more certain that data mining power is used only in legally-approved ways, and that the data which may give rise to adverse consequences for individuals is based on inferences that are derived from accurate data. We can meet these goals by making sure that the architecture of new Web technologies provides transparency into the inferencing mechanisms and creates technical means for assuring that government data mining efforts are accountable for improper use of data.

IV. Toward a public policy agenda based on transparency and accountability

Transparency and accountability technologies are necessary, but certainly not sufficient for privacy protection in an age of large scale public and private sector data mining. Our Policy Aware Web infrastructure can provide meaningful privacy protection through transparency and accountability *only if social conventions and legal requirements make such mechanisms available*

and effective. While it is beyond the scope of this paper to develop detailed public proposals, we believe that policy aware systems bring added focus to policy questions regarding data mining privacy. In order to realize the promise of transparency and accountability in support of privacy values, the legal system will have to address questions such as these:

- What degree of transparency rights (also known as ‘access rights’ in privacy law) should those subject to data mining have?
- What will be the mechanism for correction of data found to be incorrect?
- Will there be legal recourse in the event agencies rely on incorrect information after the error has been pointed out by the data subject?

Accountability mechanisms hold significant promise, but only meaningful if the legal rules against which data miners are held accountable are properly reflective of privacy values. Rules are needed to address questions such as:

- Under what circumstances, if ever, can inferences generated in one type of profiling system (anti-terrorism passenger screening, for example) be used to further criminal investigations?
- If data mining results can be shared across the national security/domestic criminal investigation "wall", is this true in all cases or only for certain classes of crimes?
- If data mining is used in a criminal investigation, can those results be applied to any other type of crime? For example, should someone under suspicion of late tax payment also be subject to checks for unpaid parking tickets or expired drivers license.

The Policy Aware systems we have described have the ability to deal with a wide range of rules in the above categories, but the rules, whatever they are, must be specific enough provide real transparency and accountability.

V. Conclusion

Our goal is to develop technical and legal design strategies for increasing the transparency of complex inferences across the Semantic Web and data mining environments. We believe that transparent reasoning will be important for a variety of applications on the Web of the future, including compliance with laws and assessing the trustworthiness of conclusions presented in a wide variety of applications. We also expect that this technical approach

will provide important guidance to policy makers who are considering how to fashion laws to address privacy challenges raised by data mining in both private sector and homeland security contexts.

Acknowledgements

This paper is based on going work with colleagues at MIT, University of Maryland and Stanford University, described in more detail in [WABH06]. This work is supported by US National Science Foundation grants: the Transparent Accountable Data Mining Initiative (award #0524481) and Policy Aware Web project (award #0427275).

References

- [AgSt00] Agrawal D. and Srikant, R. Privacy preserving datamining, Proc 2000 ACM SIGMOD Conference on Management of Data, 2000, 439-450
- [BFSW04] D. Boneh, J. Feigenbaum, A. Silberschatz, R. Wright, PORTIA: Privacy, Obligations, and Rights in Technologies of Information Assessment, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 27, pp. 10-18 (2004).
- [BLHL01] Berners-Lee, T., Hendler, J. and Lassila, O. The Semantic Web: When the Internet gets smart, Scientific American, May 2001.
- [BJS2000] <http://www.ojp.usdoj.gov/bjs/lawenf.htm> (last visited 25 October 2005)
- [BrKa03] Jeen Broekstra and Arjohn Kampman. Inferencing and Truth Maintenance in RDF Schema: Exploring a naive practical approach. In Workshop on Practical and Scalable Semantic Systems (PSSS), Sanibel Island, FL, 2003.
- [CDT03] CDT Report - Privacy's Gap: The Largely Non-Existent Legal Framework for Government Mining of Commercial Data, May 28, 2003. <http://www.cdt.org/security/usapatriot/030528cdt.pdf>
- [De97] J. Dempsey, "Communications Privacy In The Digital Age: Revitalizing The Federal Wiretap Laws To Enhance Privacy," Albany Law Journal of Science & Technology, 1997. <http://www.cdt.org/publications/lawreview/1997albany.shtml>
- [EGS03] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting Privacy Breaches in Privacy Preserving Data Mining," in Proceedings of the 22nd Symposium on Principles of Database Systems, ACM Press, New York, 2003, pp. 211–222.
- [FW06] Feigenbaum and Weitzner (ed.), "Report on the 2006 TAMI/Portia Workshop on Privacy and Accountability." <http://dig.csail.mit.edu/2006/tami-portia-accountability-wws/summary> (August 2006)
- [GoMi82] S. Goldwasser, S.Micali, Probabilistic encryption & how to play mental poker keeping secret all partial information, Proceedings of the fourteenth annual ACM symposium on Theory of computing, pp. 365-377, 1982, ACM Press.

[Hsrpt86] United States House of Representatives, Judiciary Committee Report on the Electronic Communications Privacy Act of 1986 (House Report 99-647).

[HIPAA] Health Insurance Portability and Accountability Act of 1996 (Pub. L. 104-191).

[JoCrPa04] D. Johnson, S. Crawford, J Palfry, The Accountable Internet: Peer Production of Internet Governance, 9 Virginia Journal of Law and Technology 9 (2004)

[KFJ03] L. Kagal, T. Finin, A. Joshi, "A Policy Based Approach to Security for the Semantic Web", In Proceedings, 2nd International Semantic Web Conference (ISWC2003), September 2003.

[Ka04] L. Kagal, "A Policy-Based Approach to Governing Autonomous Behavior in Distributed Environments", Phd Thesis, University of Maryland Baltimore County, November 2004.

[LiPi02] Y. Lindell and B. Pinkas, "Privacy preserving data mining," J. of Cryptology, 15:177-206, 2002.

[McFiMc03] McCool, R.; Fikes, R.; & McGuinness, D. Semantic Web Tools for Enhanced Authoring. KSL, 2003. http://www.ksl.stanford.edu/KSL_Abstracts/KSL-03-07.html

[McP04] Deborah L. McGuinness and Paulo Pinheiro da Silva. Explaining Answers from the Semantic Web: The Inference Web Approach. Journal of WebSemantics. Vol.1 No.4., pages 397-413, October 2004.

[Mark02] Protecting America's Freedom in the Information Age. Markle Foundation, 2002. http://www.markle.org/downloadable_assets/nstf_full.pdf

[Mark03] Creating a Trusted Network for Homeland Security: Second Report of the Markle Foundation Task Force, 2003. http://www.markle.org/downloadable_assets/nstf_report2_full_report.pdf

[Pa2004] Guarding America: Security Guards and U.S. Critical Infrastructure Protection. Congressional Research Service (14 November 2004) <http://www.fas.org/sgp/crs/RL32670.pdf>

[SOX] Sarbanes-Oxley Act of 2002 (Pub. L. 107-204).

[WHBC05] Weitzner, Hendler, Berners-Lee, Connolly, Creating the Policy-Aware Web: Discretionary, Rules-based Access for the World Wide Web in Elena Ferrari and Bhavani Thuraisingham, editors, Web and Information Security. IOS Press, forthcoming.

[Weit00] Weitzner, D. Testimony before the United States Senate Commerce Committee Hearing on Online Privacy. May 2000

[WABH06] Weitzner, Abelson, Berners-Lee, Hanson, Hendler, Kagal, McGuinness, Sussman, Waterman, Transparent Accountable Data Mining: New Strategies for Privacy Protection.; MIT CSAIL Technical Report MIT-CSAIL-TR-2006-007(27 January 2006).