The Laboratory for International Data Privacy

LIDAP WP 1

Latanya Sweeney, Director
January 2001

Abstract

This document describes the research and educational activities of the Laboratory for International Data Privacy (LIDAP) at Carnegie Mellon University, the role of industrial participants, the role of government agencies, and the intended impact of the LIDAP on society. The overall objective of the LIDAP is to provide intellectual leadership to society in shaping the evolving relationship between technology, public policy and the legal right to or public expectation of privacy in the collection and sharing of data. The LIDAP is unique and necessary because its interdisciplinary perspective within an academic setting provides a vantage point for examining data privacy problems across different kinds of applications and data, including financial, economic, criminal, medical and genetic information, to name a few, in various socio-political settings.

The primary goal of LIDAP research is to create architectural, algorithmic and technological foundations for the maintenance of the privacy of individuals, the confidentiality of organizations, and the protection of sensitive information, given the requirement that information be released publicly or semi-publicly. We seek to invent balanced approaches that integrate technology and policy together for the purpose of satisfying society's need for data while protecting society's need for privacy. We will develop these new approaches by studying problems and specific data collections contributed by our industrial and government partners.

1 Laboratory Mission

The overall objective of the Laboratory for International Data Privacy (LIDAP) at Carnegie Mellon University is to provide intellectual leadership to society in shaping the evolving relationship between technology and the legal right to or public expectation of privacy in the collection and sharing of data.

The LIDAP is inspired by the astonishing proliferation of public information made available on the Internet and by recent access to inexpensive, fast computers with large storage capacities. Most data holders do not even realize the jeopardy at which they place financial, medical, or national security information when they erroneously rely on existing or past security practices. Technology has eroded common protections, leaving the information vulnerable. In the past, a person seeking to reconstruct private information was limited to visiting disparate file rooms and engaging in the labor-intensive review of printed material in geographically distributed locations. Today, one can access voluminous worldwide public information using a standard handheld computer and ubiquitous network resources. Thus, from seemingly innocuous anonymous data and available public and semi-public information, one can draw an electronic image of a person or organization that is as identifying and personal as a fingerprint, even when the information contains no explicit identifiers, such as a name or phone number. However, one cannot seriously propose that all information with any links to sensitive information be suppressed. Society has developed an insatiable appetite for all kinds of detailed information for many worthy purposes, and modern systems tend to distribute information widely. A goal of the LIDAP is to inform on-going discussions and to assess and propose balanced approaches in which data can be shared but in which inferences about the identities of people and organizations contained in the released data cannot reliably be made. In this way, information that is practically useful can be shared freely with guarantees that it is sufficiently anonymous and declassified.

The LIDAP is unique and necessary because its interdisciplinary perspective within an academic setting provides a vantage point for examining data privacy problems across different kinds of applications and data, including financial, economic, criminal, medical and genetic information, to name a few, in various socio-political settings. The LIDAP constructs integrated solutions across boundaries, weaving technology and policy together.

The specific mission of the LIDAP is:

1. To conduct research on computational techniques and integrated policies for sharing information in such a way that privacy and confidentiality are maintained while the data remain practically useful;

2. To conduct research that characterizes the nature and extent of data privacy problems as society becomes increasingly technically-empowered;

3. To assess the impact of proposed practices, policies and regulations on data privacy problems;

4. To work closely with industrial and government partners to explore the evolving space of data privacy problems and solutions; and,

5. To transfer resulting technology and learned information into real-world applications and practice.

2 Who is involved

The LIDAP is a newly established interdisciplinary research facility at Carnegie Mellon University. The LIDAP is located within the H. John Heinz III School of Public Policy and Management because of the school's historical interdisciplinary leadership in assessing and informing policy discussions worldwide. Members of the LIDAP are also drawn from the School of Computer Science. U.S. News and World Reports rated the Heinz School's program in information technology as number one in the country. U.S. News and World Reports rated the computer science department in the School of Computer Science as number one in the country. In creating the LIDAP, we have assembled a team of eminent faculty and graduate students from public policy, computer science, information systems management, computer security, medical informatics, biology and other fields across the university, and of renowned legal scholars beyond the university.

3 Laboratory Organization

This section describes the research and educational activities of the LIDAP, the role of industrial participants, the role of government agencies, and the intended impact of the LIDAP on society.

3.1 Research Activities

The primary goal of LIDAP research is to create architectural, algorithmic and technological foundations for the maintenance of the privacy of individuals, the confidentiality of organizations, and the protection of sensitive information, given the requirement that information be released publicly or semi-publicly. We seek to invent balanced approaches that integrate technology and policy together for the purpose of satisfying society's need for data while protecting society's need for privacy. We will develop these new approaches by studying problems and specific data collections contributed by our industrial and government partners, and we will make our results immediately available to LIDAP partners. Thus, our partners will have access to new methods and findings long before they become commercially or publicly available.

LIDAP research can be viewed either in terms of basic scientific issues to be addressed, or in terms of specific data and applications. The exact list of applications and specific data collections examined will be determined in great part by the needs of our industrial and government partners. The list of basic research topics is therefore based on those needs and on faculty research interests and expertise. Clearly, the most important scientific and policy issues will have significant impact across many different application areas. This allows LIDAP to spread the cost of this basic research over multiple problem domains and multiple funding sources. Below is a sample of current research activities in the LIDAP.

Distributed Privacy. The goal of this work is to provide the architectural means to electronically coordinate information from vast numbers of distributed, autonomous data

holders so that intended disclosure and declassification policies can be collectively enforced, even when related inferences may not have been explicitly stated. An example includes automated surveillance of data in order to detect bioterrorist attacks and naturally occurring outbreaks. Another example concerns the automated construction of meta-level data systems.

Automated policy enforcement. The goal of this work is to automatically transfer textually stated policy statements into enforceable software actions even when related inferences are not necessarily stated. Examples include: sharing information internationally while respecting the European Union directive; collecting personal information over the Internet in one country (e.g. the United States) on citizens in another country (e.g. Canada); and, sharing patient-specific medical information under the new HHS HIPAA privacy regulation.

Anonymity certification. The goal of this work is to automatically assess and report the identifiability of information contained within a given data set. An example includes an automated system that determines the number of people who could be identified in a publicly available data set.

Privacy metrics. The goal of this work is to define and assess useful metrics for measuring the extent and character of privacy problems, risks and liabilities. Having such allows stated practices and proposed policies to be compared. While the LIDAP already has some effective metrics for computing the amount of information collected on individuals and for measuring privacy risk and liability, demonstrating the effectiveness of these metrics and exploring supporting metrics remains an ongoing research activity.

Privacy policy frameworks. The goal of this work is to explore and assess existing and proposed policies concerning data privacy in comprehensive frameworks. Examples include: examining the intent and usefulness of informed consent in the secondary sharing of medical data; devising dynamic on-line consent systems; and, designing holistic models of data collection and sharing.

Anonymous linking. The goal of this work is to develop computational tools and practices to electronically link data sets and render the results sufficiently anonymous even though the original data sets are themselves rendered sufficiently anonymous and originate from data holders that do not share information with each other. The idea is that data are linked on cryptographic equivalents of the fully identified data, which are hidden within the sufficiently anonymous data. Results appear as visible values in a regular text-based flat file.

Disclosure control techniques and systems. The LIDAP already has the leading systems and algorithms for rendering data sufficiently anonymous. However, exploring new techniques, protection models, and anonymous data systems, as well as applying known systems and models to different kinds of data, such as genetic, GIS and visual images, remains an ongoing research activity.

LIDAP faculty research interests include many additional topics as well, such as visualization of data collections and privacy problems, re-identification experiments, linking and

profiling techniques, genetic discrimination, information warfare and privacy issues specific to the Internet.

3.2 Educational Activities

There are three components to LIDAP's educational plans: an ongoing seminar series, continuing professional education and a curriculum for students in the Masters of Information Systems Management specific to data privacy.

The LIDAP hosts a regular seminar series entitled, "Privacy at Lunch" that brings together local LIDAP partners with faculty, students and others in the community to discuss current issues in privacy. Guest speakers are drawn from around the world. Topics cover all areas of privacy.

During the summer, the Center for Automated Learning and Discovery in the School of Computer Science offers a week-long series of continuing professional education courses related to data mining. Among these courses is typically a LIDAP lecture and lab course on data privacy.

The LIDAP provides an interdisciplinary course on data privacy to full-time CMU students in the Masters for Automated Learning and Discovery program and to full and part-time students in the Masters of Information Systems Management. CMU students in the Masters of Information Systems Management program can also use the LIDAP courses combined with a concentration in computer security or in medical informatics to prepare for a position as a Chief Privacy Officer.

3.3 Corporate, Non-profit and Government Participation

Participation by corporate, non-profit and government partners in the LIDAP is crucial to our success. Because of the nature of the problems we wish to study, it is essential that we ground our research in real-world problems and data sets. We look to our corporate and government partners for sharing problems and data sets with which to experiment and for funding to support a portion of these efforts. We offer two levels of participation, which we term Partners and Members.

Corporate and government partnership is intended as a low-cost mechanism for partners to learn more about data privacy problems and solutions, to gain access to LIDAP faculty and graduating students and to obtain a rapid turnaround study by LIDAP researchers specific to the partner's needs. Corporate and government membership, on the other hand, is a way in which organizations can be affiliated with the LIDAP and receive internal publications, but members do not work directly with faculty and students.

Turnaround Project

One benefit to which partners are entitled is a quick turnaround study on a problem of interest to the partner. This is intended to provide corporate and government partners with useful short-term input on problems they consider important, while enabling both the corporate partner and LIDAP faculty to evaluate the potential of this topic for a longer-term targeted research effort.

For example, a company with a database of customer records may wish to define a turnaround project to apply existing LIDAP techniques to sufficiently render the data anonymous so that copies can be sold to marketing companies without revealing the identities of the customers. In this case, the company will provide the information in a single file and LIDAP researchers would perform analyses and experiments. Within three to six months, a brief written report will be provided and a face-to-face meeting will be held to discuss the results.

As another example, a government agency may historically release data each year on a certain population and may wish to define a turnaround project to apply existing LIDAP re-identification techniques to determine whether the data are sufficiently anonymous; and if not, what vulnerabilities may exist. Within three to six months, a brief written report will be provided and a face-to-face meeting will be held to discuss the results.

As another example, when somewhat aged replicated information is declassified differently by one government agency than by another, the overall declassification effort suffers; by using two partial releases, the original may be reconstructed in its entirety. One of the government agencies may wish to define a turnaround project to apply existing LIDAP knowledge of data sources and automated techniques to refine declassification rules so information can be released with assurances of confidentiality. Within three to six months, a brief written report will be provided and a face-to-face meeting will be held to discuss the results.

As another example, when a government agency is faced with a Freedom of Information Act request for sensitive person-specific information, the agency may wish to define a turnaround project to apply existing LIDAP techniques to demonstrate which data values may be released or not to ensure privacy. Within three to six months, a brief written report will be provided and a face-to-face meeting will be held to discuss the results.

As a final example, a non-profit organization may be exploring different privacy policies or proposals for legislation and may wish to define a turnaround project to apply existing LIDAP metrics to determine the comparative impact and effectiveness of these proposals. Within three to six months, a brief written report will be provided and a face-to-face meeting will be held to discuss the results.

These are just a few examples of the kinds of turnaround projects possible through the LIDAP. If the results of any of these appear promising, the corporate or government partner may choose to fund a more targeted research effort on this topic.

Summary of Benefits

Below is a summary of the benefits available to LIDAP members and to LIDAP partners. LIDAP partners receive all the benefits of LIDAP members and more, including a turnaround project.

Membership includes:

- Access to LIDAP internal publications.
- Access to all LIDAP reports (not including reports specific to a partner) and seminars.
- Access to faculty for separate consulting or targeted research projects.
- Dues: $25k/year

Partnership includes:

- Access to LIDAP internal publications.
- Access to all LIDAP reports (not including reports specific to a partner) and seminars.
- Access to faculty for separate consulting or targeted research projects.
- Access to LIDAP non-restricted software
- Access to LIDAP non-restricted data
- Short-term visitors to LIDAP by Partner's personnel
- Turnaround Project
- Dues: $50k/year

The LIDAP and Carnegie Mellon University reserve the right to refuse to accept any membership, partnership or project.

Targeted Research Projects

Corporate and government members and partners are invited to arrange additional targeted research projects with individual LIDAP faculty or groups of faculty. Such specific projects will be performed at cost. A prototypical two-year project involving a single faculty member and a single graduate student will typically cost in the range of $90-150k/year. We expect successful turnaround projects may lead to such long-term targeted projects.

Specialized LIDAP software

The LIDAP has the leading systems and algorithms for rendering data sufficiently anonymous. They include: (1) the Scrub System, which locates and replaces personally identifying information in textual documents; (2) the Datafly System, which generalizes and suppresses values in field-structured data sets; and, (3) the *k*-Similar algorithm, which finds optimal solutions such that data are minimally distorted while still providing adequate protection. These are some of the tools in the restricted LIDAP collection. Licenses for the private or commercial use of restricted LIDAP software and data are available through Carnegie Mellon University's Technology Transfer Office.

3.4 Additional Government Participation

The scientific and policy problems addressed by the LIDAP are also of considerable interest to government agencies such as NSF, DARPA, NIST, DOE, NIH and the U.S. Bureau of the Census. The majority of current research funding for LIDAP faculty comes from such agencies, and we expect this to continue in the future.

4 Benefits to Society

Because results from the LIDAP can facilitate a responsible means for providing detailed medical data to researchers, financial information to economists, and military intelligence information to analysts, to name a few, society can reap tremendous benefits in allocation of resources, financial efficiencies, and protection of national information interests. Of course, this is only possible because the abstracted data does not compromise individuals, organizations or national interests.

5 Benefits to CMU

The intended benefits of the LIDAP to the University include:

- Establishment of a world-class interdisciplinary research facility addressing these important problems
- Gaining access to significant real-world research problems in data privacy
- Producing integrated course offerings for students
- Integrating work currently done at CMU that combines the strengths of researchers found within the University
- Assisting CMU in contributing to an important societal issue

6 For Further Information

- See our World Wide Web page at http://sos.heinz.cmu.edu/dataprivacy/index.html
- Contact LIDAP@sos.heinz.cmu.edu
- Contact Latanya Sweeney, LIDAP Director, at latanya@andrew.cmu.edu.
- Contact Debra Dennison, Administrative Assistant, at (412) 268-4456
- Contact Casey Porto of the Technology Transfer Office at Carnegie Mellon University, at (412) 268-1241, to acquire a license for restricted software or data, such as the Scrub System, the Datafly System and the *k*-Similar algorithm.