Technological Solutions for Protecting Privacy

Roberto J. Bayardo and Ramakrishnan Srikant
IBM Almaden Research Center

The Web is commonly viewed as an information access tool for end users. But as much as it simplifies access to stock quotes, medical libraries, or reference manuals, the Web also makes it easier for individuals and organizations to obtain and infer—with surprising detail—personal information about us.

Use of such information ranges from beneficial to criminal. On one hand, corporations that understand our preferences can customize our Web experience to save us time and increase their efficiency. On the other hand, this information can be misused in harmful ways, such as identity theft or denial of insurance on the basis of personal health details.

Alan Westin, professor emeritus of public law and government at Columbia University, defines *privacy* as the right of individuals to determine for themselves when, how, and to what extent information about them is communicated to others. However, information gathering and information management tools are not typically designed to support the right to privacy. This omission, coupled with the increasing sophistication and deployment of information-gathering systems, contributes to an ever-growing volume of misused or inappropriately shared personal information gathered from Web users.

Emerging technologies can protect privacy without restricting the information flow crucial to efficient organizations.

MISUSE, LOSS, AND LAW

Sometimes the misuse is intentional, but accidents and poor security practices also cause problems. Consider the following cases:

- GlobalHealthtrax, which sells health products online, inadvertently revealed customer names, home phone numbers, bank accounts, and credit card information for thousands of customers on its Web site (MSNBC, 19 Jan. 2000).
- Hackers recently used Google to search for vulnerable systems, which allowed them to infiltrate a database containing personal and medical information on more than 5,000 neurosurgery patients (*Wired*, Mar. 2003).

Cases of improper disclosure and outright misuse of personal information affect both individual and collective behavior. In 2001, Forrester Research, a market research firm, reported that consumer privacy apprehensions about the Web "will hold back roughly $15 billion in e-commerce revenue."

Legislative action, though essential to any comprehensive privacy strategy, is not necessarily guided by the current capabilities and limitations of information technology infrastructures. For instance, in the US, the 1996 Health Insurance Portability and Accountability Act (www.hhs.gov/ocr/hipaa/), which gives patients control over how their personal medical information is used and disclosed, required substantial IT overhauls.

Privacy legislation that impacts the IT infrastructure is not unique to the US. Sweden recently passed legislation that restricts how Web sites can use cookies, a technology that enables tracking of users across multiple visits. But cookies are also widely used in e-commerce applications, such as implementing online store shopping carts.

TECHNOLOGICAL SOLUTIONS

Information gathering on the Web is pervasive in large part because usage-tracking and data-mining technology are deeply integrated into most Web software systems, such as tools for building online storefronts. In contrast, tools for managing data privacy are uncommon. This makes addressing user and legislative privacy concerns difficult and costly.

Nevertheless, many technologies offer ways to help protect personal privacy on the Web and beyond. We focus here on emerging technologies that—by protecting privacy without restricting the information flow crucial to efficient organizations—may become core features of future information systems and Web infrastructures.

Further Reading in Privacy-Preserving Technologies

Hippocratic databases

- R. Agrawal et al., "Hippocratic Databases," *Proc. 28th Int'l Conf. Very Large Databases*, Morgan Kaufmann, 2002, pp. 143-154.

Privacy-preserving data mining

- R. Agrawal and R. Srikant, "Privacy Preserving Data Mining," *Proc. 2000 ACM SIGMOD Int'l Conf. Management of Data*, ACM Press, 2000, pp. 439-450.
- A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting Privacy Breaches in Privacy Preserving Data Mining," *Proc. 22nd ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems*, ACM Press, 2003, pp. 211-222.

Information sharing across private repositories

- Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," *Proc. Crypto 2000*, Springer-Verlag, 2000, pp. 37-55.
- R. Agrawal, A. Evfimievski, and R. Srikant, "Information Sharing across Private Databases," *Proc. 2003 ACM SIGMOD Int'l Conf. Management of Data*, ACM Press, 2003, pp. 86-97.
- J. Dyer et al., "Building the IBM 4758 Secure Coprocessor," *Computer*, Oct. 2001, pp. 57-66.

Privacy-preserving search

- M. Bawa, R.J. Bayardo, and R. Agrawal, "Privacy Preserving Indexing of Documents on the Network," to appear in *Proc. 29th Int'l Conf. Very Large Databases*, Morgan Kaufmann, 2003.
- B. Chor, etal., "Private Information Retrieval," *IEEE Symp. Foundations of Computer Science*, IEEE CS Press, 1995, pp. 41-50.

Privacy policy encoding

One of the most well-known Web privacy technologies is the Platform for Privacy Preferences developed by the World Wide Web Consortium (W3C). With P3P, an organization with a Web presence can encode its data-collection and data-use practices in a machine-readable XML format known as a P3P policy. Browsers such as Microsoft Internet Explorer and Mozilla can programmatically compare site policies against a user's privacy preferences and take actions based on the comparison. For example, the browser can block the site altogether or limit the types of cookies it will accept.

The current P3P standard only provides a mechanism for Web sites to state their intentions regarding use of the personal information that they collect. Mechanisms for enforcing that sites act according to their stated policies are beyond its scope.

IBM developed the Enterprise Privacy Authorization Language for encoding an enterprise's internal privacy-related data-handling policies and practices. EPAL and P3P have different goals. While P3P enables automated matching between privacy policies and user preferences, EPAL allows privacy-enforcement systems such as IBM's Tivoli Privacy Manager to import and enforce the enterprise's privacy policy.

Hippocratic databases

Inspired by the privacy tenet of the Hippocratic oath, Hippocratic databases include responsibility for the privacy of data they manage as a fundamental tenet, and are thus a natural solution for the problem of enforcing privacy policies. Hippocratic databases incorporate 10 fundamental privacy principles. For example, the "purpose specification" principle states that the purposes for which information has been collected should be associated with any personal information stored in the database; the "limited use" principle states that the database will run only queries that are consistent with the purposes for which the information has been collected.

To illustrate how Hippocratic databases can automatically enforce these principles, consider what happens when queries, tagged with purpose, are submitted to the database. The database first checks whether the user issuing the query is among the users authorized by the privacy policy for that purpose. Next, the database analyzes the query to check whether it accesses any fields not explicitly listed for the query's purpose in the privacy policy. Finally, the database ensures that only records having a purpose attribute that includes the query's purpose will be visible to the query, thereby enforcing any opt-in or opt-out preferences.

Anonymization

While Hippocratic databases can help organizations appropriately manage and use the information they collect, some customers may prefer to prevent organizations from collecting information about them in the first place. Various anonymization technologies let Web users prevent data collection by hiding or blocking potentially identifying information such as cookies and IP addresses. These technologies range from centralized privacy proxies such as anonymizer.com to decentralized Web-browsing networks such as Crowds from AT&T. In fact, companies like iPrivacy.com even allow users to anonymously purchase items (by creating special arrangements with credit-card companies).

Privacy-preserving data mining

Despite their advantages, anonymization methods may prevent sites from understanding their customers and

improving their products and services accordingly. Privacy-preserving data mining lets businesses derive the understanding they need without collecting accurate personal information. By randomizing customer data, this approach precludes the recovery of anything meaningful at the individual level but still supports algorithms that can recover aggregate information, build mining models, and deliver actionable insights to businesses.

To make this idea concrete, consider a scenario in which an online merchant asks a Web site visitor for demographic information such as age. Client-side software scrambles or "randomizes" the data entered by the visitor before sending it to the merchant. The scrambling involves taking the entered number and adding or subtracting a random value. The software performs this randomization step independently for every user who opts to enter an age. So, an entry of 30 might become 42, while an entry of 34 might become 28.

The software that online merchants use cannot determine the true age value of visitors. It has access only to the randomized values and the randomization parameters (for example, that the randomization values ranged from -30 to +30). Solely on the basis of this information, the software can reconstruct a close approximation of the true distribution. This reconstruction will only be accurate over thousands of people—not for single users—thereby preserving privacy.

The merchant can then use this reconstructed distribution to build an accurate data-mining model and, for example, understand the demographics of the people who buy something versus those who don't. Or, if the goal is to give the user personalized recommendations, the merchant can ship the data-mining model to the visitor who then applies it locally.

Information sharing across private repositories

In February 2000, DoubleClick announced plans to combine consumer information it collected from Web users with information in the databases of an acquired subsidiary, Abacus Direct, raising the ire of privacy advocates and consumers alike. The message from this uproar was clear: While consumers might in some cases choose to disclose personal information, they do not want the information they disclose combined into massively detailed consumer dossiers.

Solutions to privacy concerns must combine laws, societal norms, markets, and technology.

Once again, though, businesses have a legitimate desire to understand their customers. When the information necessary for an accurate understanding is scattered across multiple databases created for disparate purposes, the problem is to allow businesses to compile aggregate models without having to merge—and hence disclose—the individual data on which the models are built. This problem belongs in the general framework of secure multiparty computation: Given two parties with inputs x and y, secure multiparty computation computes a function f such that the two parties learn only $f(x,y)$ and nothing else.

Cryptographic protocols. In 1986, Andrew Chi-Chih Yao showed that for any function computable by a circuit of AND, OR, and NOT gates, a cryptographic protocol exists that can perform the computation in an encrypted space and reveal only the function's output ("How to Generate and Exchange Secrets," *Proc. 27th IEEE Symp. Foundations of Computer Science*, IEEE, 1986, pp. 162-167).

However, such circuit-based protocols do not scale to computations over millions of records. There are two broad strategies for improving scalability in the context of computing models or aggregate statistics. The first breaks down the function in such a way that each party can perform the bulk of the computation locally on their unencrypted data, leaving only a small portion for secure multiparty computation protocols. The second strategy involves finding specialized protocols that can solve specific problems much faster than general solutions.

Secure coprocessors. Another approach is to use a secure coprocessor—a tamper-resistant device designed so that any physical tampering will clear its memory. Participants in a group computation can verify that the secure coprocessor is running an agreed-upon program—for example, one that outputs a customer model from its input and nothing else—even if the device is in a remote location. Participants can communicate securely with the device to deliver their share of the input, and the secure coprocessor performs the computation.

Privacy-preserving search

Both data owners and people searching for information might have privacy concerns.

Data owner's privacy. To avoid the privacy concerns raised by merging private information sources, institutions often manage their private information databases with their own incompatible authentication and access-control mechanisms. This approach has privacy advantages over aggregating such information at a central host, but it is inconvenient at best for users. Users searching for access-controlled information that is legitimately available to them must independently search each relevant repository, assuming they know the entire set of relevant providers.

Efficient and uniform search of multiple access-controlled repositories would seem to require a central trusted-index host. But a typical search index almost perfectly represents the indexed files and databases, so a central host removes any privacy benefits associated with distributed maintenance of private data.

Methods from the peer-to-peer domain, however, can uniformly search

distributed content without relying on centralized resources. For example, developers could extend query-flooding methods such as Gnutella with decentralized authentication and access-control policy-enforcement mechanisms to support uniform searching of access-controlled content. While query flooding does not scale well, recent results show promise for addressing scalability in decentralized search with stronger privacy properties.

Searcher's privacy. The detailed personal information that Web site or data repository owners can infer from a list of a user's searches raises another privacy concern. Anonymization methods can protect a user's privacy in public Web searches by preventing the results from being associated with a user's identity. But authentication requirements keep anonymization methods from protecting privacy in searches of access-controlled data.

Techniques from the private information retrieval (PIR) domain may potentially apply to this particular problem. PIR techniques let authenticated users retrieve information from remote databases while preventing the database owner from identifying the specific information accessed. Significant work remains, however, to extend the current theoretical formulations of the problem to the real-world scenarios that arise on the Web.

Technology alone cannot address all the concerns surrounding a complex issue like privacy. The total solution must combine laws, societal norms, markets, and technology. However, by advancing what is technically feasible, we can influence the ingredient mix and improve the overall quality of the solution. ■

Roberto J. Bayardo is a research staff member at IBM Almaden Research Center. Contact him at bayardo@alum.mit.edu.

Ramakrishnan Srikant is a research staff member at IBM Almaden Research Center. Contact him at srikant@almaden.ibm.com.

Editor: Sumi Helal, Computer and Information Science and Engineering Dept., University of Florida, P.O. Box 116125, Gainesville, FL, 32611-6120; helal@cise.ufl.edu