Privacy on the Web: Facts, Challenges, and Solutions

Despite important regulatory and technical efforts aimed at tackling aspects of the problem, privacy violation incidents on the Web continue to hit the headlines. The authors outline the salient issues and proposed solutions, focusing on generic Web users' Web privacy.

Abdelmounaam Rezgui, Athman Bouguettaya, and Mohamed Y. Eltoweissy
Virginia Tech

The Web has spurred an information revolution, even reaching sectors left untouched by the personal computing boom of the 80s. It made information ubiquity a reality for sizeable segments of the world population, transcending all socioeconomic levels. The ease of information access, coupled with the ready availability of personal data, also made it easier and more tempting for interested parties (individuals, businesses, and governments) to intrude on people's privacy in unprecedented ways. In this context, researchers have proposed a range of techniques to preserve Web users' privacy.[1–3] (See the "Defining privacy" sidebar for details on the evolving definitions of privacy.)

However, despite considerable attention, Web privacy continues to pose significant challenges. Regulatory and self-regulatory measures addressing one or more aspects of this problem have achieved limited success. Differences and incompatibilities in privacy regulations and standards have significant impact on e-business. For example, US Web-based businesses might be unable to trade with millions of European consumers because their practices do not conform with the European Union's Data Protection Directive.[4]

Clearly, to address these issues, we must start by synthesizing ideas from various sources. We tackle this problem by surveying the issue of Web privacy and investigating the main sources of privacy violations on the Web. With a taxonomy of several current technical and regulatory approaches aimed at enforcing Web users' privacy, we hope to form a comprehensive picture of the Web privacy problem and its solutions.

In this article, we focus on Web privacy from users' perspectives. Although we recognize that different levels of privacy violations exist, our discussion on privacy focuses on its preservation or loss. This lets us use a lowest-common-denominator approach to provide a meaningful discussion about the various privacy issues and solutions.

The privacy problem

Two major factors contribute to the privacy problem on the Web:

- the inherently open, nondeterministic nature of the Web and
- the complex, leakage-prone information flow of many Web-based transactions that involve the transfer of sensitive, personal information.

To comprehend the first factor, we can contrast the Web with traditional, closed, deterministic multiuser systems, such as enterprise networks. In these systems, only known users with a set of predefined privileges can access data sources. On the contrary, the Web is an open environment in which numerous and a priori unknown users can access information. Examples of the second factor include applications involving citizen–government, customer–business, business–business, and business–government interactions. In some of these applications, personal information that a Web user submits to a given party might, as a result of the application's intrinsic workflow, be disclosed to one or more other parties.

Preserving privacy on the Web has an important impact on many Web activities and Web applications. Of these, e-business and digital government are two of the

Defining privacy

Individual privacy is an important dimension of human life. The need for privacy is almost as old as the human species. Definitions of privacy vary according to context, culture, and environment. In an 1890 paper, Samuel Warren and Louis Brandeis defined privacy as "the right to be let alone."[1] In a seminal paper published in 1967, Alan Westin defined privacy as "the desire of people to choose freely under what circumstances and to what extent they will expose themselves, their attitude and their behavior to others."[2] More recently, Ferdinand Schoeman defined privacy as the "right to determine what (personal) information is communicated to others" or "the control an individual has over information about himself or herself."[3] One of the earliest legal references to privacy was made in the Universal Declaration of Human Rights (1948). Its Article 17 states, "No one shall be subjected to arbitrary or unlawful interference with his privacy, family, home, or correspondence, nor to unlawful attacks on his honor and reputation." It also states, "Everyone has the right to the protection of the law against such interference or attacks."

Generally, privacy is viewed as a social and cultural concept. With the ubiquity of computers and the emergence of the Web, privacy has also become a digital problem. In particular, with the Web revolution, privacy has come to the fore as a problem that poses a set of challenges fundamentally different from those of the pre-Web era. This problem is commonly referred to as Web privacy. In general, the phrase Web privacy refers to the right of Web users to conceal their personal information and have some degree of control over the use of any personal information disclosed to others.

References

1. S.D. Warren and L.D. Brandeis, "The Right to Privacy," *Harvard Law Review*, vol. 4, no. 5, 1890, pp. 193–220.
2. A.F. Westin, *The Right to Privacy*, Atheneum, 1967.
3. F.D. Schoeman, *Philosophical Dimensions of Privacy*, Cambridge Univ. Press, 1984.

best examples. In the context of e-business, privacy violations tend to be associated mostly with marketing practices. Typical cases occur when businesses capture, store, process, and exchange their customers' preferences to provide customized products and services. In many cases, these customers do not explicitly authorize businesses to use their personal information. In addition, a legitimate fear exists that companies will be forced to disclose their customer's personal data in court. For example, in the *Recording Industry Association of America (RIAA)* v. *Verizon* (summer 2002), the music recording industry forced ISPs to disclose IP information about users who allegedly illegally downloaded music.

These mishaps have negatively affected businesses and, consequently, the Web-based economy. Consumers' mistrust naturally translates into a significant reluctance to engage in online business transactions. A Jupiter Communications' study estimated that, in 2002, the loss that resulted from consumers' concerns over their privacy might have reached $18 billion. This confirms the Gartner Group's view that, through 2006, information privacy will be the greatest inhibitor for consumer-based e-business.

Digital government is another class of Web applications in which Web privacy is a crucial issue. Government agencies collect, store, process, and share personal data about millions of individuals. A citizen's privacy is typically protected through regulations that government agencies and any business that interacts with them must implement. Users tend to trust government agencies more than businesses. However, law enforcement agencies are at odds with civil libertarians over collecting personal information. Law enforcement agencies have a vested interest in collecting information about unsuspecting citizens for intelligence gathering and investigations. Although anonymity is still an option for many people,[5] most Web transactions require information that can uniquely identify them.

Additionally, governments' foray in developing techniques for gathering and mining citizens' personal data has stirred controversy. One example is the US Central Intelligent Agency's investment in In-Q-tel, a semiprivate company that specializes in mining digital data for intelligence purposes. Therefore, concerns about privacy are a major factor that still prevents large segments of users from interacting with digital government infrastructures.

Understanding Web privacy

The Web is often viewed as a huge repository of information. This perception of a passive Web ignores its inherently active nature, which is the result of the intense volume of Web transactions. A *Web transaction* is any process that induces a transfer of information among two or more Web hosts. Examples include online purchases, Web sites browsers, and Web search engine use. We refer to the information exchanged as a result of a Web transaction as *Web information*. The Web information type determines the extent and consequences of a privacy violation related to that information.

Access to personal or sensitive information through Web transactions is generally subject to *privacy policies* associated with that information. These policies refer to the set of implicit and explicit rules that determine whether and how any Web transaction can manipulate that information. A Web transaction is said to be *privacy preserving* if

Table 1. Dimensions of Web privacy.

DIMENSION	REQUIREMENTS/GUARANTEES
Information collection	The privacy requirement on information collection consists of ensuring that users' private information is not collected via the Web without their knowledge and explicit consent. For example, a health insurance company can guarantee its Web customers that it will never attempt to scan their computers to determine whether they have visited Web sites of companies that sell specific medicines.
Information usage	Information usage defines the collected information's usage purposes. For example, consider a citizen using a Web-based government service such as Medicaid, which provides health care coverage for low-income citizens. The service's privacy policy might have an information usage component that limits the use of personal information to purposes directly related to providing health services.
Information storage	The storage requirement determines whether and for how long a party (such as a business) that collects private information can store the collected information. For example, Medicaid might state that collected customer information will remain in the underlying databases for one year after they leave the service.
Information disclosure	The Web privacy's information-disclosure component determines if and to whom the company can reveal collected user information. For example, a company's Web site's privacy policy might state that no information collected from customers can be transferred to a third party without their explicit approval.
Information security	This describes the security policies and mechanisms used to guarantee the security (and thus, privacy) of information (for example, firewalls, encryption, authentication).
Access control	A privacy policy must state who may access what. For example, an online business's privacy policy might state that only customer service employees are allowed to access personal information of customers. The access policy must also specify the access granularity—that is, how specific entities can get when disclosing a user's information to a third party. For example, a Web site's privacy policy might state that while it will not disclose information about specific individuals, it will disclose aggregated information about large populations (statistics).
Monitoring	Systems that collect and give access to personal information must encompass a monitoring component that builds and maintains traces of all operations that input or output sensitive information. Often, these traces are the only means to settle conflicting claims regarding privacy violation.
Policy changes	Privacy policies might evolve as a result of regulatory or internal business practice changes. However, these policies must not be retroactive. For example, if a Web site that typically collects users' personal information changes its privacy policy, the new changes must not be systematically applicable to information collected before the changes occur.

it does not violate any privacy rule before, while, and after it occurs. Privacy policies applicable to Web information could specify requirements relevant to one or multiple dimensions for Web privacy. Table 1 enumerates some of the most important dimensions.

We can classify Web users' personal information as one of three types:

- *Personal data* include information such as a person's name, marital status, mailing and email addresses, phone numbers, financial information, and health information.
- *Digital behavior* refers to Web users' activities while using the Web, including the sites they visit, frequency and duration of these visits, and online shopping patterns.
- *Communication* includes Web users' electronic messages, postings to electronic boards, and votes submitted to online polls and surveys.

Understanding Web privacy requires understanding how privacy can be violated and the possible means for preventing privacy violation.

Sources of privacy violation

Web users' privacy can be violated in different ways and with different intentions. The four major sources we identified are unauthorized information transfer, weak security, data magnets, and indirect forms of information collection.

Unauthorized information transfer

Personal information is increasingly viewed as an important financial asset. Businesses frequently sell individuals' private information to other businesses and organizations. Often, information is transferred without an individual's explicit consent. For example, in 2002, medical information Web site DrKoop.com announced that, as a result of its bankruptcy, it was selling customers' data to vitacost.com.[6]

Weak security

The Web's inherently open nature has led to situations in which individuals and organizations exploit the vulnerability of Web-based services and applications to access classified or private information. In general, unauthorized access is the result of weak security. A common form of these accesses occurs when foreign entities penetrate (for

example, through hacking) Web users' computers. Consequences generally include exposure of sensitive and private information to unauthorized viewers. The consequences are even more important when the attack's target is a system containing sensitive information about groups of people. For example, in 2000, a hacker penetrated a Seattle hospital's computer network, extracting files containing information on more than 5,000 patients.[7]

Data magnets

Data magnets are techniques and tools that any party can use to collect personal data.[8] Users might or might not be aware that their information is being collected or do not know how that information is collected. Various data-magnet techniques exist:

Explicitly collecting information through online registration. Online registration entails that users provide personal information such as name, address, telephone number, email address, and so on. More importantly, in the registration process, users might have to disclose other sensitive information such as their credit card or checking account numbers to make online payments.

Identifying users through IP addresses. Generally, each time a person accesses a Web server, several things about that person are revealed to that server. In particular, a user's request to access a given Web page contains the user's machine's IP address. Web servers can use that to track the user's online behavior. In many situations, the address can uniquely identify the actual user "behind" it.

Software downloads. Companies that let their customers download their software via the Internet typically require a unique identifier from each user. In some cases, companies use these identifiers to track users' online activity. For example, in 1999, RealNetworks came under fire for its alleged use of unique identifiers to track the music CDs or MP3 files that users played with its RealPlayer software.

Cookies. A cookie is a piece of information that a server and a client pass back and forth.[9] In a typical scenario, a server sends a cookie to a client that stores it locally. The client then sends it back to the server when the server subsequently requests it. Cookies are generally used to overcome the HTTP protocol's stateless nature; they let a server remember a client's state at the time of their most recent interaction. They also let Web servers track Web users' online activities—for example, the Web pages they visit, items accessed, and duration of their access to every Web page. In many situations, this monitoring constitutes a violation of users' privacy.

Trojan horses. These applications might seem benign but can have destructive effects when they run on a user's computer. Examples of Trojan horses include programs that users install as antviruses but that actually introduce viruses to their computers. For example, a Trojan attack might start when a user downloads and installs free software from a Web site. The installation procedure might then launch a process that sends back to the attack initiator sensitive personal information stored on the local computer.

Web beacons. A Web beacon—also known as a Web bug, pixel tag, or clear gif—is a small transparent graphic image that is used in conjunction with cookies to monitor users' actions.[8] A Web beacon is placed in the code of a Web site or a commercial email to let the provider monitor the behavior of Web site visitors or those sending an email. When the HTML code associated with a Web beacon is invoked (to retrieve the image), it can simultaneously transfer information such as the IP address of the computer that retrieved the image, when the Web beacon was viewed, for how long, and so forth.

Screen scraping. Screen scraping is a process that uses programs to capture valuable information from Web pages. The basic idea is to parse the Web pages' HTML content with programs designed to recognize particular patterns of content, such as personal email addresses. A case that illustrates how screen scraping can violate privacy is the one in which the US Federal Trade Commission alleged that ReverseAuction.com had illegally harvested data from the online auction site eBay.com to gain access to eBay's customers.[8]

Federated identity. A Web user's federated identity is a form of identity (for example, a user name and password pair) that lets a user access several Web resources. Microsoft's .Net My Services is an example of one architecture that provides a federated identity mechanism, with which a user can create an identity at one Web site and use it to access another Web site's services. This extensive sharing of users' private information raises concerns about the misuse of that information.

Indirectly collecting information. Users can authorize organizations or businesses to collect some of their private information. However, their privacy can be implicitly violated if their information undergoes analysis processes that produce new knowledge about their personality, wealth, behavior, and so on. This deductive analysis might, for example, use data mining techniques to draw conclusions and produce new facts about the users' shopping patterns, hobbies, or preferences. These facts might be used in recommender systems through a process called *personalization*, in which the systems use personalized information (collected and derived from customers' past activity) to predict or affect their future shopping patterns. Undeniably, personalization makes users' shopping

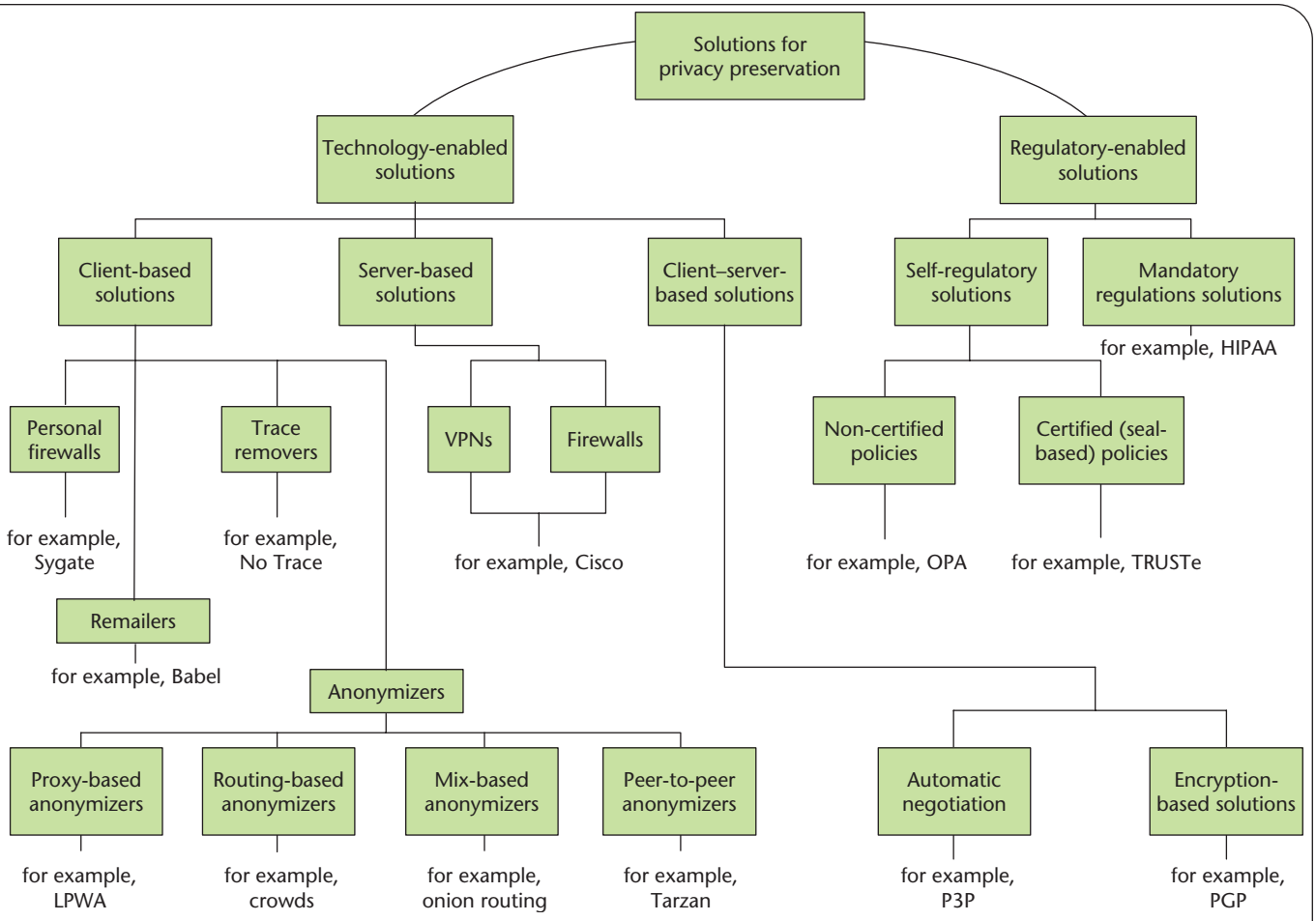

Figure 1. A taxonomy of technology- and regulation-enabled solutions for privacy preservation in the Web.

experience more convenient. However, in more aggressive marketing practices (such as advertising phone calls) it can negatively affect customers' privacy.

Privacy can also be violated through the misuse of statistical databases, which contain information about numerous individuals. Examples include databases that provide general information about the health, education, or employment of groups of individuals living in a city, state, or country. Typical queries to statistical databases provide aggregated information such as sums, averages, *p*th percentiles, and so on. A privacy-related challenge is to provide statistical information without disclosing sensitive information about the individuals whose information is part of the database.

A taxonomy of privacy preserving solutions

We categorize solutions to the Web privacy problem based on the main enablers of privacy preservation (Figure 1). The two main categories are technology- and regulation-enabled solutions. The implementation approach further refines this taxonomy.

Technology-enabled solutions

A typical Web transaction involves a Web client and a Web server. We classify technology-enabled solutions according to the type of Web entities that are responsible for their implementation: clients, servers, or clients/servers.

Client-based solutions. These solutions target privacy aspects relevant to individual users. Examples include protecting personal data stored on a personal computer, protecting email addresses, deleting any trace of Web access, and hiding Web surfers' real identities. We discuss four types of solutions: personal firewalls, remailers, trace removers, and anonymizers (see Figure 1).

A *firewall* is a software and/or hardware system that provides a private network with bidirectional protection from external entities gaining unauthorized access. Generally, firewalls protect medium-to-large networks (such as an enterprise's intranet). A *personal firewall* is a software firewall that protects a single user's system (typically, a single machine). It runs in the background on a PC or a server and watches for any malicious behavior. A user

might even configure the firewall to detect specific types of unwanted events—for example, access from a specific IP address or a given port.

Personal firewalls have recently become a significant market. Many software firms propose personal firewalls with different capabilities. Examples include ZoneAlarm, NetBoz, and Outpost. In addition, general Web users can also use network address translation (NAT) devices to help preserve network privacy. Developers have initially proposed NATs to provide one IP for a set of home machines, thus providing a single point of entry for that network. While providing relative anonymity, its strength is on providing a firewall to provide reasonable security against external attacks.

A *remailer* is an application that receives emails from senders and forwards them to their respective recipients after it alters them so that the recipients cannot identify the actual senders. If necessary, a recipient can send a reply to the remailer, which then forwards it to the sender of the original message. Babel and Mixminion are examples of remailers.

When users navigate through the Web, their browsers or any other external code (such as a downloaded script) can store different types of information on their computers. This *navigation trace* provides details of users' surfing behavior, including the sites they visit, the time and duration of each visit, what files they download, and so on. *Trace removers* are available as a conservative measure to prevent disclosure of users' Web navigation history. They simply erase users' navigation histories from their computers. Examples of trace removers include Bullet Proof Soft and No Trace.

For many reasons, Web users would like to visit a Web site with the guarantee that neither that site nor any other party can identify them. Researchers have proposed several techniques to provide this anonymous Web surfing. These solutions' basic principle is preventing requests to a Web site from being linked to specific IP addresses. We can classify *anonymizing* techniques into four types:

- *Proxy-based anonymizers*. A proxy-based anonymizer uses a proxy host to which users' HTTP requests are first submitted. The proxy then transforms the requests in such a way that the final destination cannot identify its source. Requests received at the destination contain only the anonymizer's IP address. Examples of proxy-based anonymizers include Anonymizer, Lucent Personal Web Assistant (LPWA), iPrivacy, and WebSecure. Some proxy-based anonymizers can also be used to access registration-based Web sites. For example, LPWA uses alias generators, giving users consistent access to registration-based systems without revealing potentially sensitive personal data. More effective proxy-based anonymizers such as iPrivacy can conceal users' identity even while making online purchases that, normally, would require them to disclose their actual identities.
- *Routing-based anonymizers*. This class of anonymizers has Web requests traverse several hosts before delivering them to their final destination so that the destination cannot determine the requests' sources. An example of a tool that uses this technique is Crowds.[10] Its philosophy is that a good way to become invisible is to get lost in a crowd. The solution is to group Web users geographically into different groups, or crowds. A crowd performs Web transactions on behalf of its members. When users join a crowd, a process called *jondo* starts running on their local machines. This process represents the users in the crowd. It engages in a protocol to join the crowd, during which it is informed of the current crowd members. Once users' jondos have been admitted to the crowd, they can use the crowd to anonymously issue requests to Web servers. Users' requests are routed through a random sequence of jondos before they are finally delivered to their destinations. Neither the Web servers nor any other crowd members can determine who initiated a specific request.
- *Mix-based anonymizers*. Mix-based anonymizers are typically used to protect communication privacy. In particular, they protect against *traffic-analysis attacks*, which aim to identify who is talking to whom but not necessarily to directly identify that conversation's content. One technique that protects against traffic-analysis attacks is *onion routing*.[2] It is based on the idea that mingling connections from different users and applications makes them difficult to distinguish. The technique operates by dynamically building anonymous connections within a network of real-time Chaum mixes.[1] A *Chaum mix* is a store-and-forward device that accepts fixed-length messages from numerous sources, performs cryptographic transformations on the messages, and then forwards the messages to the next destination in a random order.
- *Peer-to-peer anonymizers*. Mix-based anonymizers generally use static sets of mixes to route traffic. This obviously poses three major problems: scalability, performance, and reliability. One way to overcome these drawbacks is to use peer-to-peer (P2P) anonymizers, which distribute the anonymizing tasks uniformly on a set of hosts. Examples of P2P anonymizers include Tarzan, MorphMix, and P5 (Peer-to-Peer Personal Privacy Protocol). For example, Tarzan uses a pool of voluntary nodes that form mix relays. It operates transparently at the IP level and, therefore, works for any Internet application.

Server-based solutions. Server-based solutions target aspects of Web privacy relevant to large organizations such as enterprises and government agencies. For example, an online business might deploy a server-based privacy-

preserving solution to protect hospital patients' records or a customers database. Privacy preservation in these solutions is a side effect of strong security mechanisms typically employed in large organizations. Virtual private networks (VPNs) and firewalls are two mechanisms that have been particularly effective in protecting security and privacy at an enterprise scale. VPNs are secure virtual networks built on top of public networks such as the Internet. They generally use several security mechanisms (such as encryption, authentication, and digital certificates) and are often used in conjunction with firewalls to provide more stringent levels of security and privacy enforcement.

Client–server-based solutions. In these solutions, clients and servers cooperate to achieve a given set of privacy requirements. Two examples illustrate this: negotiation- and encryption-based solutions.

Negotiation-based solutions use a protocol on which both the Web client and server agree. Enforcing privacy through a negotiated privacy policy is a viable and practical option only if the negotiation process is automated. Automatic negotiation of privacy requirements is generally enabled through software agents that users configure to implement specific privacy preferences. Client–server negotiation of privacy requirements is the driving design paradigm behind the platform for privacy preferences project (P3P), the World Wide Web Consortium's standard for privacy preservation. P3P lets users automatically manage the use of their personal information on Web sites they visit. A site implementing P3P expresses its privacy policy in a machine-readable format. Its users can configure their browsers to automatically determine whether the Web site's privacy policy reflects their personal privacy needs .

Typically, negotiation-based Web interactions use XML to specify and exchange policies.[11] In P3P, Web sites' privacy policies and users' privacy preferences are encoded using XML. On a P3P-enabled Web site, a *policy reference file* provides the P3P privacy policy file's location for the different parts of the Web site. A user's agent first sends an HTTP request to get the policy reference file. It then fetches the file, interprets it, and makes decisions according to which user instructed it through privacy preferences. Developers can build user agents into Web browsers, browser plug-ins, or proxy servers as Java applets or scripts.

Encryption-based solutions encrypt the information exchanged between two or more Web hosts so that only legitimate recipients can decrypt it. Web users might use encryption in different Web activities and to enforce several privacy requirements. One of these requirements is the privacy of personal communication, or email. Typically, Internet-based communication is exchanged in clear text. An encryption-based protocol that has particularly addressed protecting email is Pretty Good Privacy. PGP has become the de facto standard for email encryption. It enables people to securely exchange messages and to secure files, disk volumes, and network connections with both privacy and strong authentication. It ensures privacy by encrypting emails or documents so that only the intended person can read them.

Regulation-enabled solutions

Regulation-enabled solutions encompass two types: self- and mandatory-regulation solutions. *Self regulation* refers to the information keepers' ability to voluntarily guarantee data privacy. *Mandatory regulation* refers to legislation aimed at protecting citizens' privacy while they transact on the Web.

Self regulation. In the absence of comprehensive regulations addressing the Web privacy problem, self-discipline has been an alternative approach adopted by many Web-based businesses. This typically manifests in the form of privacy statements that businesses post on their Web sites. An important problem with self-regulation is that it is also self-defined—that is, different organizations generally adopt different privacy rules in handling their customers' information. Businesses tend to advocate self-regulation to avoid government involvement. Examples of industry groups that push for self-regulating privacy policies include the Online Privacy Alliance, NetCoalition, and the Personalization Consortium.

Self-regulated privacy policies can be *certified* or *noncertified*. This certification is the process of formally asserting to users that a party's claimed policy is actually implemented. A trusted third party is usually responsible of certifying privacy policies. Upon request, the trusted party checks a given Web site's practices with regard to its privacy policy. If the trusted party deems that the Web site does respect its privacy policy, it delivers a certificate of good conduct that the site can display, typically in the form of a trust seal. Major trust seals include TRUSTe, BBBOnline, WebTrust, and SecureBiz.

Different third parties might have different requirements to approve a given site. For example, to approve a Web site's privacy policy, TRUSTe requires information about what type of information is collected, who collects it, how it is used, whether it is shared, a minimum of an opt-out provision for consumer choice, security measures, and how to correct information.

Mandatory regulation. Several countries and political entities have adopted laws and legal measures to address the Web privacy problem. A notable example of privacy-preserving regulations is the European Union's Data Protection Directive, adopted in October 1995. The directive limits access to electronic data contained in the EU member nations. According to the directive, certain personal information (such as an individual's race, creed, sex-

Table 2. Dimension-solution summary.

DIMENSION	TECHNOLOGY-ENABLED SOLUTIONS	REGULATION-ENABLED SOLUTIONS
Information collection	Yes	Mostly no
Information usage	Mostly yes	Yes
Information storage	No	Mostly no
Information disclosure	Mostly yes	Mostly no
Information security	Mostly yes	Mostly no
Access control	Mostly yes	Mostly no
Monitoring	Mostly no	Mostly no
Policy changes	No	Mostly no

ual orientation, or medical records) cannot leave the EU unless it is going to a nation with laws offering privacy levels that the EU has deemed adequate .

In the US, the regulatory approach to preserving privacy is reactive and not based on a national privacy policy. In fact, most privacy-related laws were enacted in response to particular events or needs for a specific industry. Examples include the 1978 Financial Services Privacy Act (FSPA), the 1986 Electronic Communications Privacy Act (ECPA), the 1996 Health Insurance Portability and Accountability Act (HIPAA), and the 1998 Child Online Privacy Protection Act (COPPA).

Governments might also impose privacy-related regulations on their own agencies. The US has passed statutes and laws to regulate its federal agencies' data collection. In fact, some of these laws were passed even before the Web era. One example is the Privacy Act passed in 1974. The act aimed at regulating activities of all agencies that collect and maintain personal information.

Assessing solutions

It is useful to provide an assessment on the adequacy of the proposed Web privacy solutions. However, this could not be totally objective because of the various perceptions on privacy violations. Therefore, our assessment (see Table 2) contains a subjective element that reflects our perceptions of privacy violations. We use the taxonomy of issues in Table 1 for the rows. For brevity's sake, we use technology- and regulation-enabled solutions as the two main categories of solutions. The values we used are "Yes," "No," "Mostly yes," and "Mostly no." "Yes" indicates that all approaches in that category address part of or the whole corresponding issue. "No" indicates that no approach in that category addresses the corresponding issue in a meaningful way. "Mostly yes" indicates that the majority of approaches in the category address the corresponding issue in some meaningful way. "Mostly no" indicates that only a minority of approaches in that category address the corresponding issue in some meaningful way.

Privacy in the Semantic Web

In the vision of the Semantic Web, the Web evolves into an environment in which "machines become much better able to process and 'understand' the data that they merely display at present."[12] In this environment, Web services and Web agents interact. *Web services* are applications that expose interfaces through which Web clients can automatically invoke them. *Web agents* are intelligent software modules that are responsible for some specific tasks—for example, searching for an appropriate doctor for a user.

Web services and Web agents interact to carry out sophisticated tasks on users' behalf. In the course of this interaction, they might automatically exchange sensitive, private information about these users. A natural result of this increasing trend toward less human involvement and more automation is that users will have less control over how Web agents and Web services manipulate their personal information. The issues of privacy preservation must therefore be appropriately tackled before the Semantic Web vision fully materializes.

Two key concepts are essential in solving the privacy problem in the Semantic Web, namely, *ontologies* and *reputation*. Artificial intelligence researchers first introduced the ontologies concept to facilitate knowledge sharing and reuse. An ontology is a "set of knowledge terms, including the vocabulary, the semantic interconnections, and some simple rules of inference and logic for some particular topic."[13] Researchers have widely recognized the importance of ontologies in building the Semantic Web. In particular, ontologies are a central building block in making Web services computer interpretable.[14] This, in turn, lets us automate the tasks of discovering, invoking, composing, validating, and monitoring the execution of Web services.[15]

Ontologies will also play a central role in solving the Semantic Web's privacy problem. In fact, building a *privacy ontology* for the Semantic Web is one of several recent propositions to let Web agents carry out users' tasks while preserving their privacy. In a recent paper on ontologies,[16] researchers presented a privacy framework for Web services that lets user agents automatically negotiate

with Web services on the amount of personal information they will disclosed. In this framework, users specify their privacy preferences in different permission levels on the basis of a domain-specific ontology based on DAML-S, the DARPA Agent Markup Language set of ontologies to describe the functionalities of Web services.

Another important research direction in solving the Semantic Web's privacy problem is based on the *reputation* concept . Researchers suggest that using this concept lets Web agents and Web services interact with better assurances about their mutual conduct. In the highly dynamic Semantic Web environment, a service or agent will often be required to disclose sensitive information to Web-based entities (such as government agencies or businesses) that are unknown and/or whose trustworthiness is uncertain. The reputation-based approach consists of deploying mechanisms through which agents can accurately predict services' "conduct" with regard to preserving the privacy of personal information that they exchange with other services and agents. In another work,[15] we proposed a Web reputation management system that monitors Web services and collects, evaluates, updates, and disseminates information related to their reputation for the purpose of privacy preservation.

Most of the technology-based solutions target network privacy. These solutions typically use a combination of encryption or request rerouting to provide data privacy and some anonymity. These systems have several limitations. Installing, configuring, and using these tools might be complicated. Systems requiring modification of network protocols or access to proxy servers might be behind firewalls or inaccessible to users of custom Internet access software. Privacy-enhancing technologies have not met the challenge of safeguarding people's data on the Web mostly due to the underlying assumption that third-party providers can implement privacy preservation. As the P3P effort shows, providers have no vested interest in insuring Web privacy. Therefore, the design of privacy-enhancing techniques must focus on how to make the privacy-preservation part of the data it is supposed to protect.

With the emerging Semantic Web, services and systems will be able to automatically understand data semantics. For some Web users, this provides a more convenient Web. Unfortunately, this also provides an increased incentive to intrude in people's privacy because of the enhanced quality of information available to Web users. Therefore, more effective techniques are necessary to protect this high quality Web information from illegitimate access and use. Although legislation can work for paper-based information, it has limited effect on Web-based information. A promising research direction is to explore the concept of code shipping to develop novel mechanisms for data protection. The objective is to empower users to have better control over the access and the use of their data. This approach meshes well with the Semantic Web. The idea is to embed user agents with the data. These agents would travel with the data, setting access protection dynamically. □

Acknowledgments

The second author's research is supported by the National Science Foundation under grant 9983249-EIA and grant SE 2001-01 from the Commonwealth Technology Research Fund through the Commonwealth Information Security Center Information Security Center (CISC). We thank Brahim Medjahed and Mourad Ouzzani for their valuable comments on earlier versions of this article.

References

1. D. Chaum, "Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms," *Comm. ACM*, vol. 24, no. 2, 1981, pp. 84–88.
2. D. Goldschlag, M. Reed, and P. Syverson, "Onion Routing," *Comm. ACM*, vol. 42, no. 2, 1999, pp. 39–41.
3. A. Rezgui, A. Bouguettaya, and Z. Malik, "A Reputation-Based Approach to Preserving Privacy in Web Services," *Proc. 4th VLDB Workshop on Technologies for E-Services* (TES '03), Springer-Verlag, 2003.
4. European Union, "Final EU Data Protection Directive," *Official J. European Communities*, no L. 281, 23 Nov. 1995, p. 31.
5. M. Froomkin, "Anonymity in the Balance," *Digital Anonymity and the Law: Tensions and Dimensions*, C. Nicoll, J.E.J. Prins, and M.J.M. van Dellen, eds., Uitgeverij T.M.C. Asser Press, 2003.
6. Health Data Management, "Koop Clients' E-Mail Addresses for Sale," 3 July 2002; www.healthdata/management.com/html/PortalStory.cfm?type=trend&DID = 8775
7. Security Focus, "Hospital Records Hacked," 6 Dec. 2000; www.securityfocus.com/news/122.
8. PriceWaterhouseCoopers, "E-Privacy: Solving the On-Line Equation, 2002; www.pwcglobal.com/extweb/pwcpublications.nsf/DocID/ED95B02AC583D4E480256A380030E82F.
9. D.M. Kristol, "HTTP Cookies: Standards, Privacy, and Politics," *ACM Trans. Internet Technology*, vol. 1, no. 2, 2001, pp. 151–198.
10. M.K. Reiter and A.D. Rubin, "Anonymous Web Transactions with Crowds," *Comm. ACM*, vol. 42, no. 2, 1999, pp. 32–48.
11. E. Bertino, E. Ferrari, and A. Squicciarini, "X-TNL: An XML-Based Language for Trust Negotiations," *Proc. IEEE 4th Int'l Workshop Policies for Distributed Systems and Networks*, IEEE Press, 2003.
12. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, May 2001, vol. 284, no. 5, pp. 34–43.

13. J. Hendler, "Agents and the Semantic Web," *IEEE Intelligent Systems*, vol. 16, no. 2, 2001, pp. 30–37.
14. D. Fensel and M.A. Musen, "The Semantic Web: A Brain for Humankind," *IEEE Intelligent Systems*, vol. 16, no. 2, 2001, pp. 24–25.
15. S. McIlraith, T.C. Son, and H. Zeng, "Semantic Web Services," *IEEE Intelligent Systems*, vol. 16, no. 2, 2001, pp. 46–53.
16. A. Tumer, A. Dogac, and H. Toroslu, "A Semantic Based Privacy Framework for Web Services," *Proc. WWW '03 Workshop on e-Services and the Semantic Web* (ESSW 03), 2003.

Abdelmounaam Rezgui *is a PhD candidate in the Department of Computer Science, Virginia Tech. His current research interests include privacy, trust, and reputation in the Semantic Web. As a member of Virginia Tech's E-Government and E-Commerce Research Lab, he is also involved in other research topics related to the design and development of digital government and e-commerce infrastructres. He has an MSc in computer science from Purdue University. While at Purdue, he worked on video segmentation and object extraction, high performance video servers, and multimedia databases. He is a member of the IEEE and the ACM. Contact him at rezgui@vt.edu.*

Athman Bouguettaya *is program director of computer science at Virginia Tech. He is also director of the E-Commerce and E-Government Research Lab. He is on the editorial boards of the* Distributed and Parallel Databases Journal *and the* International Journal of Web Services Research. *He is a senior member of the IEEE and a member of the ACM. Contact him at athman@vt.edu.*

Mohamed Eltoweissy *is a visiting professor and associate professor of computer science at Virginia Tech and James Madison University, respectively. His research interests include network and information security and privacy, computer supported cooperative work, and distributed computing. He has a PhD in computer science from Old Dominion University and a BS and MS in computer science from Alexandria University, Egypt. He cofounded the Commonwealth Information Security Center in Virginia. He is a member of the ACM and the honor societies of Phi Kappa Phi and Upsilon Pi Epsilon. Contact him at eltowemy@vt.edu.*