Protection of Privacy on the Web

Thomas M. Chen
Dept. of Electrical Engineering
Southern Methodist University
PO Box 750338, Dallas, Texas 75275
Tel: +1 214-768-8541
Email: tchen@engr.smu.edu

Zhi (Judy) Fu
Network and Infrastructure Research Lab
Motorola Labs
1301 E Algonquin Rd., Schaumburg, IL 60196
Tel: +1 847-576-6656
Email: judy.fu@motorola.com

Abstract

Most people are concerned about online privacy but may not be aware of the various ways that personal information about them is collected during routine Web browsing. We review the types of personal information that may be collected voluntarily or involuntarily through the Web browser or disclosed by a Web server. We present a taxonomy of regulatory and technological approaches to protect privacy. All approaches to date have only been partial solutions. By its nature, the Web was designed to be an open system to facilitate data sharing, and hence Web privacy continues to be a challenging problem.

Introduction

The main appeal of the World Wide Web is convenient and instant access to a wealth of information and services. Many people will start research on a topic with a Google search. The number of Web sites has grown exponentially and reached more than 149 million in November 2007 according to Netcraft (http://news.netcraft.com/archives/web_server_survey.html).

In their search for services, users may not keep in mind that the Web is capable of collecting data as well as displaying data. The most obvious means of data collection are Web forms for registrations, logins, and messaging. These forms are voluntary disclosures of personal information that most people understand to be necessary for shopping, online banking, and other personalized services. However, users may not fully appreciate that Web sites collect information about them routinely without their consent or even notification. Web sites keep track of clients' IP (Internet protocol) addresses at a minimum and often additional information such as browser version, operating system, viewed resources, and clicked links. Moreover, this collected information may be shared among organizations in the background without the public's knowledge.

Some users may have unrealistic expectations about online privacy because they ignore the fact that the Web is an open system. By design, just about anyone can easily put up a Web site and make its contents globally accessible. This means that sites should not be assumed to be trustworthy. Contrary to natural inclinations, it would be more reasonable to assume sites are untrustworthy, until a trust relationship is established (e.g., through prior experience, reputation, or third-party validation).

Web privacy is certainly not a new issue. However, if anything, concerns have escalated rather than decreased due to increasing prevalence of phishing and malware (malicious software) attacks. In phishing attacks, innocent users are lured to malicious sites designed to deceive them into revealing valuable personal information. Common types of malware include spyware, bots, and keyloggers, which can steal personal data. They can be downloaded in various ways, often without a user's awareness.

The consequences of privacy loss will be growing distrust of the Web and diminishing usage of online services. Thus, protection of privacy is an important practical problem with economic ramifications. This chapter examines regulatory and technological approaches to protect privacy on the Web. First, we survey the various threats to online privacy. Then we offer a taxonomy of approaches to provide and protect privacy.

Types of Private Information

Clearly there are different types of personal information, with varying degrees of sensitivity. As shown in Figure 1, personal information on the Web might be classified into three types (Rezgui, Bouguettaya, and Eltoweissy, 2003):

- personal data such as name, address, and history;
- surfing behavior consisting of visited sites, online transactions, and searches;
- communications such as bulletin boards, messages, and feedback forms.

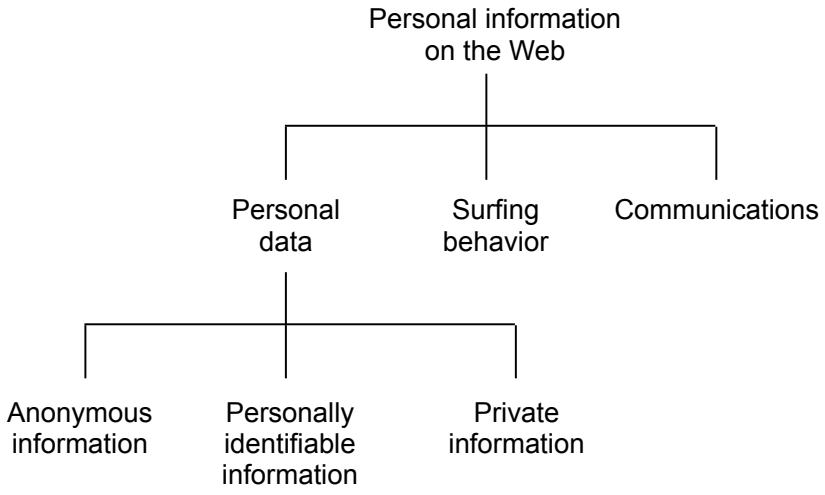

Fig. 1. Types of personal information on the Web

Personal data can be classified further into anonymous information (which can not be traceable to a specific person); personally identifiable information; or private information (Garfinkel, 2002). Information can be anonymized by "scrubbing" any identifying aspects or by aggregating multiple records into a single record. Personally identifiable information can be traced to an individual, such as name, address, e-mail address, or phone number. Although this information is personal, it is often published and can be found with effort. Disclosure of personal information may be undesirable but generally not harmful. On the other hand, disclosure of private information (such as bank records or passwords) may be considered harmful or at least embarrassing. Private information has an obvious value to criminals. Phishing and malware attacks usually have the goal to steal private information to make profits by identity theft or selling the information to other criminals.

Surfing behavior has a value primarily for advertisers who wish to understand the interests of users in order to target advertisements. Advertisements are far more effective if the audience is likely to be interested.

Another concern has been government monitoring of surfing behavior. The U.S. government has argued that Web monitoring is useful in anti-terrorism programs, which has been a federal priority since 9/11. The argument essentially seeks to barter a loss of privacy for increased security.

Threats to Web Privacy

Threats to privacy can arise at the Web client or server side. Naturally, privacy might also be lost within the network but network security is not addressed here as it is not specifically related to the Web. Eavesdropping in the network is a broad problem involving all IP-based communications. IP packets are normally not protected cryptographically against eavesdropping. Web transactions can be protected by SSL (secure sockets layer) or TLS (transport layer security). TLS is the successor standardized by the Internet Engineering Task Force (IETF) to SSL version 3.0 but they are essentially similar (Dierks and Rescorla, 2006). SSL/TLS uses encryption to protect Web transactions from eavesdropping. In addition, SSL/TLS uses cryptography to authenticate the identities of clients and servers to each other, and to protect messages from tampering in transit in the network.

Information Collected at Clients

Information can be collected at the client side either voluntarily, automatically, or involuntarily.

Voluntary disclosure: In many routine cases, users voluntarily submit many types of information in order to access online services such as banking, shopping, downloads, and searches. Common examples of voluntarily disclosed information include names, addresses, e-mail addresses, telephone numbers, credit card numbers, login IDs, passwords, and search queries. Voluntarily disclosed information is typically sent to a Web site and stored in a database.

Unfortunately, it can be difficult to determine if a site is legitimate, particularly if a criminal has carefully created a malicious site to masquerade as a legitimate site. Phishing attacks depend on social engineering techniques to trick users into believing that a phishing site is legitimate and submitting their personal information voluntarily. For example, a user may receive an e-mail appearing to come from a bank requesting the customer to update his/her account information. Following a link in the e-mail, the user is directed to a phishing Web site, looking exactly like the real bank Web site. When the user enters his/her login information, it is sent to the malicious site's owner.

For the year from October 2006 to September 2007, the Anti-Phishing Working Group counted an average of 31,987 new phishing sites per month (http://www.antiphishing.org). It has been estimated that millions of U.S. consumers are victimized annually by phishing.

Automatic disclosure: Most users may not realize that information about their computer is automatically disclosed in the normal HTTP (hypertext transfer protocol) messages used in Web communications. HTTP is an application layer protocol working over TCP (transmission control protocol) and IP. IP packets reveal the client's IP address to the server. Moreover, HTTP request messages contain a so-called user agent string that identifies the browser name, version, client operating system, and language. An example user agent string might be: "Mozilla/5.0 (Macintosh; U; Intel Mac OS X; en; rv:1.8.1.11) Gecko/20071128 Camino/1.5.4" which reveals that the client is running Mac OS X on an Intel-based Mac computer, and the browser is Camino version 1.5.4 in English (containing the Mozilla-based Gecko engine). The client's IP address and user agent string may be stored in a Web server's logs, which usually records all transactions.

HTTP requests also contain a "referer URL" which is the URL for the page viewed by the user before making the HTTP request. That is, it identifies the URL where the user is coming from. An example of unintentional information disclosure is the referer link: "http://www.google.com/search?hl=en&q=privacy." Unfortunately, it reveals that the user just viewed a Google search for "privacy." The problem is that the search query terms are encoded in the URL itself. The URL probably reveals more than the user expects.

Cookies are another means of automatic information disclosure (Kristol, 2001). Cookies were first developed and implemented by Netscape Communications as a way to keep "state" in Web transactions, particularly online shopping. HTTP is a stateless protocol meaning that a server does not remember a client's previous visits. This makes it difficult to keep track of items in an online shopping cart. Cookies solve the problem by caching text-based information in the client browser, in the form of "name=value" strings. A cookie can also include an expiration time, domain name, path, and whether SSL/TLS should be used. A server first passes a cookie to the browser, and the browser returns the cookie upon revisiting the same site.

Third-party cookies have been controversial. An HTML document often contains images or elements provided by other sites. When these elements are requested, the other sites may set a third-party cookie (called third party because the other sites may be outside the domain of the requested page). A controversial practice is Web banners or advertisements setting third-party cookies to track users across different sites. Once a cookie is set, the cookie is returned whenever the user visits a site with the same ads. A well known example is DoubleClick which serves ads on many different sites. An example cookie for the domain "doubleclick.net" might be:

"id=800009fb4447d5a; expires Oct. 14, 2009; secure=no." The cookie contains an identifier that is unique for each user, which allows DoubleClick to track individuals across multiple sites. Browser such as Mozilla and Opera can be configured to block third-party cookies. However, DoubleClick can also track the same client IP address across sites which serve DoubleClick ads.

Involuntary disclosure: Clients can disclose information involuntarily by means of malware. On the Web, malware can be downloaded to a client through social engineering (human deception) or exploiting a vulnerability in the client. A so-called drive-by download is triggered by a client visiting a malicious site which contains, for instance, a zero-dimension iframe (inline frame). The iframe would not be displayed but its HTML contents would be interpreted by the browser. The contents might include exploit code to compromise the client through a vulnerability and download malware without the user's knowledge.

Many forms of malware are known, including:

- viruses: self-replicating pieces of code attached to normal programs or files;
- worms: self-replicating standalone programs that spread to vulnerable hosts through the network;
- downloaders: small programs for downloading other malware;
- Trojan horses: programs hiding their malicious functions by appearing to be useful;
- bots: programs to remotely control a compromised client as a member of a bot net following instructions from a bot herder;
- spyware: programs to covertly steal client data.

In the context of the Web privacy, spyware has become a particularly worrisome problem affecting a significant fraction of PC users (Shukla and Nah, 2005). Thousands of spyware variants have been found, creating a new commercial market for anti-spyware programs. Common types of spyware include: adware tracking browsing behavior to target ads more effectively; browser changers that modify start pages and other browser settings; browser plug-ins adding functions to browsers; bundleware installed with other software; keyloggers recording keyboard inputs; and dialers changing dial-up connection settings.

The Web is known to be a common avenue for spyware, installed knowingly by consent (perhaps bundled with useful software) or unknowingly by drive-by downloads. Studies have found that spyware can be encountered on a wide variety of sites, even well known popular sites and sites rated as trustworthy.

Information Collected at Servers

As already mentioned, Web servers collect data in logs and databases. Concerns about server side privacy are related to how servers collect data, control access to data, share data, and make use of data (Rezgui, Bouguettaya, and Eltoweissy, 2003).

Web bugs: One of the covert means for tracking users is Web bugs, also called Web beacons or clear gifs. One study found that 58% of popular sites and 36% of randomly chosen sites had Web bugs close to their home pages (Martin, Wu, and Alsaid, 2003). A Web bug is a transparent GIF image with dimensions of one pixel by one pixel hosted on a third party server. When the Web bug is requested as an element in a Web page, the client's IP address is recorded by the hosting server. By matching IP addresses, a specific user might be tracked across multiple

Web sites. Tracking could be done with any HTML element but a small transparent GIF image will not be perceptible to users.

Search queries: Many users may not realize that their search queries are logged and could be used for profiling. Google has been the main focus of privacy concerns due to the enormous amount of data it sees as the predominant search engine, and the common belief that Google retains permanent records of every search query (although Google claims that its data is not personally identifiable).

Concerns about Google were heightened by its acquisition of DoubleClick in April 2007. Google has extensive data about searching behavior, while DoubleClick specializes in monitoring the viewing of ads across many different sites. Their combination places an unprecedented amount of personal data in one organization.

Server vulnerabilities: Various attacks on servers are possible. Like other computers, servers control access mainly through passwords which are subject to cracking attacks. Hence, the security of server data depends mainly on password security. In addition, servers can have vulnerabilities like other computers. The vast majority of Web servers use Apache or Microsoft IIS (Internet Information Server), and both have records of vulnerabilities (easily found on various Web sites).

Database vulnerabilities: Web servers often have a SQL (structured query language) database backend. SQL is an ANSI/ISO standard language for querying and manipulating databases. SQL injection attacks take advantage of user data input through Web forms. If a SQL injection vulnerability exists, the input data is passed to the SQL database without proper filtering for string literal escape characters, such as quotes. These characters in the input data will cause the database to interpret the data as SQL statements. By carefully crafting the input data, an attacker could learn data or manipulate data stored in the database.

Cross-site user tracking: In November 2007, it was discovered that Facebook's Beacon system tracks the activities of Facebook users on more than 40 affiliate Beacon sites, which report those activities to the user's Facebook friends. The reporting is done even when users are logged out of Facebook or deactivated their Facebook account. Furthermore, the reporting is done by affiliate Beacon sites without explicit notification to users.

Assurances of Privacy

An individual's legal right to privacy was argued by Samuel Warren and Louis Brandeis in 1890 (Warren and Brandeis, 1890). The right protects anyone from having to unwillingly disclose information that is not of "legitimate" public concern. Furthermore, the right is violated if a person's information is disclosed against their will, regardless of whether the information is accurate or not, and regardless of the motive. The notion of an inherent right to privacy against involuntary disclosure was reaffirmed in the United Nations' 1948 Universal Declaration of Human Rights.

More recently, Alan Westin described "informational privacy" as the right of people to "determine for themselves when, how, and to what extent information about them is communicated to others" (Westin, 1967). Thus, privacy involves an individual's control over how personal information is shared. The issue of control matters, for example, when a user

submits personal information to a trusted Web site. The user may be trusting the site to keep the information internal from other organizations. However, the user does not have much control in reality over how the site uses the submitted information.

Approaches to protect privacy can be classified as regulatory or technological (Rezgui, Bouguettaya, and Eltoweissy, 2003). A taxonomy is shown in Figure 2.

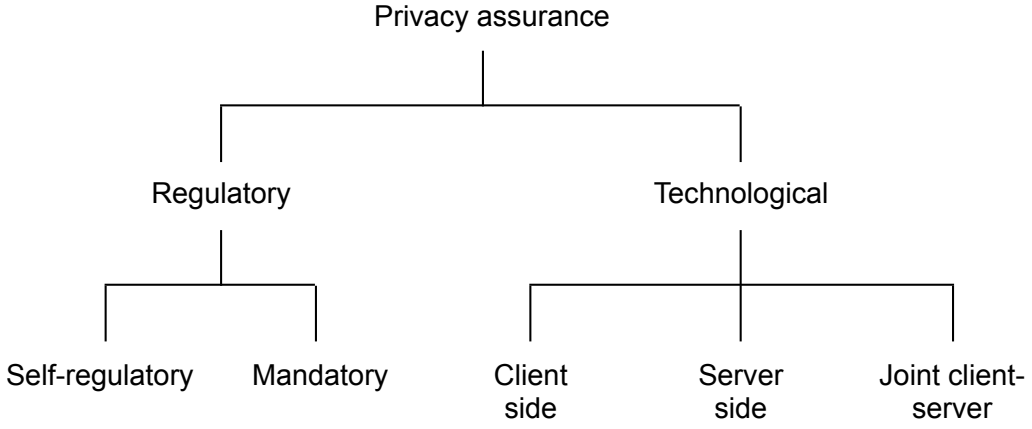

Fig. 2. Approaches to privacy assurance

Regulatory Protection of Privacy

As we discussed before, Web sites may collect consumers' personal data and share the data with third parties without the consumers' consent. Regulations provide some governance on how businesses should collect and process consumers personal data. Regulatory protection of privacy includes voluntary and mandatory approaches, and both approaches are being used today (Rezgui, Bouguettaya, and Eltoweissy, 2003).

Voluntary self-regulations: The U.S. government has historically preferred not to use an overarching federal law to protect consumers' online privacy but instead recommended self-regulation for industries to implement. In general, technology industries have also preferred self-regulation over the alternative of government regulation.

In 1998, the Online Privacy Alliance (OPA) was formed as a group of global companies and associations to "lead and support self-regulatory initiatives to promote business-wide actions that create an environment of trust and that foster the protection of individuals' privacy online and in electronic commerce" (http://www.privacyalliance.org). The OPA advocates online businesses to post privacy policies following at least five principles:

- adoption of a privacy policy;
- identification of what information is being collected and how it is shared and used;
- a choice for individuals to give consent to (or opt out of) other uses of their data;
- adequate protection of personal data from misuse, loss, or tampering;
- assurance that data is kept accurate, complete, and up to date.

Additionally OPA offers guidelines effective self regulation and principles for protecting the privacy of children's online activities. These guidelines are recognized as the industry standard for online privacy.

Privacy seal programs are one of the means for effective self regulation. The major privacy seal programs are TRUSTe, CPA WebTrust, and BBBOnline. TRUSTe was founded by the Electronic Frontier Foundation (EFF) and the CommerceNet Consortium to act as an independent, nonprofit organization dedicated to building consumers' trust and confidence on the Internet (Benassi, 1999). Web sites conforming to TRUSTe's privacy standard can display the TRUSTe "trustmark" shown in Figure 3.

Fig. 3. TRUSTe privacy seal

A displayed TRUSTe trustmark gives assurance to users that the visited site has agreed to disclose their information collection and dissemination practices, and that their disclosure and practice have been verified by a credible third party. Specifically, the site has met at least five requirements:

- a posted privacy statement disclosing its personal data collection and dissemination practices;
- a choice for users to opt out of having their personal information used for non-primary purposes;
- a means for users to correct inaccurate information;
- reasonable procedures to protect personal information from loss, misuse, or tampering;
- verification of compliance by TRUSTe.

It should be noted that a privacy seal does not mean that data will be kept private. A privacy seal only implies that policy statements have been made and verified, but policies are not required to protect privacy. The adequacy of a site's policies must be judged by each user. Although privacy seals are intended to ultimately encourage more e-commerce, studies suggest that most users are unaware of the requirements for a site to obtain a privacy seal, and can not differentiate between a genuine or fake privacy seal (Moores, 2005).

Mandatory regulations: In the 1960s and 1970s, the U.S. government investigated the implications of computers on consumer credit reporting and privacy. One of the important outcomes was a 1973 commission report that created the Code of Fair Information Practices. This report would have a later impact on European regulations on data protection. The Code of Fair Information Practices prescribed five principles:

- systems for recording personal data should not be kept secret;

- individuals are entitled to discover what personal data about them is recorded and how it is used;
- personal data about a person obtained for one purpose must have the person's consent to be used for another purpose;
- individuals are entitled to dispute and correct a personal record;
- organizations involved in keeping personal records must ensure that data is used only as intended.

Regulations for data protection in the U.S. have been piecemeal, covered by various laws over the years, such as the Fair Credit Reporting Act, Privacy Act of 1974, Freedom of Information Act, and the Children's Online Privacy Protection Act (COPPA).

For health care industries, the U.S. Congress recognized the need to protect the privacy of patients' personal data and enacted the Health Insurance Portability and Accountability Act of 1996 (HIPAA), also known as Public Law 104-191 (Anton, Earp, Vail, Jain, Gheen, and Frink, 2007). It included administrative simplification to make healthcare services more efficient, portability of medical coverage for pre-existing conditions, and standards for electronic billing and claims transmission. The privacy part of HIPAA requires that access to patient information be limited to only authorized individuals and only the information necessary to perform the assigned tasks. All personal health information must be protected and kept confidential. The final version of the HIPAA privacy regulations were issued in December 2000 and went into effect in April 2001 with a two-year grace period for compliance.

Recently, the collection and sharing of consumers' personal information by financial institutions was addressed at the federal level by the Gramm-Leach-Bliley Act, also known as the Financial Modernization Act of 1999. There are three parts of the legislation related to data privacy:

- The Financial Privacy Rule addresses the collection and disclosure of customers' personal financial information by financial institutions. Provisions require financial institutions to give consumers privacy notices that clearly explain the organizations' information-sharing practices. In turn, consumers have the right to opt out of sharing their information with third parties. However, opting out is not allowed in certain circumstances, such as when data sharing is required by law.
- The Safeguards Rule requires all financial institutions, including credit reporting agencies, to implement safeguards to protect customer information.
- The Pretexting provision protects consumers from individuals and companies that obtain their personal financial information under false pretenses (pretexting).

The federal law allows states to pass stronger privacy laws. For example, California passed the Online Privacy Protection Act of 2003 (OPPA) which requires all commercial sites or online services that collect personal information from California residents to:

- post their privacy policies and dates on their Web sites and comply with those posted policies;
- describe the types of personal data collected and how the data is shared with third parties;
- describe the process for notifying users of policy changes;
- describe the process for consumers to request changes to any of their information (if allowed).

Violators are given 30 days to comply after notification, under threat of civil suit for unfair business practices.

While the U.S. has traditionally preferred an approach combining self-regulation and local legislation, the European Union (EU) has been more consistent in broadly recognizing privacy rights. A recognition of the fundamental right to privacy was included in the 1950 European Convention on Human Rights (Article 8) and the Council of Europe's 1981 Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. Concerned with maintaining uniformity, the EU issued an overarching Directive on Protection of Personal Data in 1995, formally known as Directive 95/46/EC, to regulate the collection and processing of consumers' personal information within the EU member countries. Based on the Code of Fair Information Practices from the U.S. and 1980 recommendations issued by the Organization for Economic Cooperation and Development (OECD), the directive aims for a balance "between a high level of protection for the privacy of individuals and the free movement of personal data within the EU." The directive sets certain limits on the collection and use of personal data and requires each EU member to set up an independent national body (supervisory authority) responsible for the protection of personal data. The directive sets conditions related to three categories: transparency, legitimate purpose and proportionality.

- Transparency means that personal data may be processed when that person has given his consent or when the data processing is necessary (e.g., for compliance with law or for contracted services). In addition, the person has the right to access his/her personal data and correct or delete inaccurate data.
- Legitimate purpose means that personal data can be processed only for specified and legitimate purposes.
- Proportionality means that personal data may be processed only as it is adequate, relevant and not excessive in relation to the purposes for which they are collected. Personal data must be accurate and kept up to date.

The directive also regulates the transfer of personal data to countries that do not belong to the EU and may not have adequate privacy protection. The U.S. has an arrangement with the EU called the Safe Harbor Program to streamline the process for US companies to comply with Directive 95/46/EC. US companies can opt into the program if they adhere to the seven basic principles outlined in the directive:

- individuals should be informed that their data is being collected and about how it will be used;
- individuals must be given the choice to opt out of data collection or data sharing.
- data may be transferred only to third parties with adequate data protection;
- reasonable efforts must be made to protect collected information from loss;
- data must be relevant and reliable for the purpose it was collected for;
- individuals must be able to access their personal information, and correct or delete inaccurate data.
- there must be effective means of enforcing these rules.

U.S. companies can demonstrate compliance by self-assessment or third-party assessement.

Technological Protection of Privacy

While regulations are obviously important in establishing trust between online businesses and consumers, they are not likely to be sufficient by themselves. Fortunately, a variety of technological solutions exist (Linn, 2005; Rezgui, Bouguettaya, and Eltoweissy, 2003). Technological approaches to privacy protection might be divided into client side, server side, or joint client-server (again, network-based protection such as IPSEC is not addressed here because it is not specific to the Web). Protection done at the client includes encryption, anonymizers, personal firewalls, cookie disabling, ad blocking, anti-spyware, and anti-phishing. On the server side, data protection consists of preventing unauthorized intrusions by means of strong passwords, firewalls, vulnerability testing, and intrusion detection systems. Joint client-server approaches involve cooperation between clients and servers, and the main example today is the Platform for Privacy Preferences Project (P3P).

Encryption: Encryption uses mathematical algorithms to change plaintext (the original data) before transmission into ciphertext (encrypted data) that would be not understandable to an eavesdropper. Many encryption algorithms, such as RSA (Rivest-Shamir-Adleman) and the U.S. standardized AES (advanced encryption standard), are known and used in practical applications. Typically, the encryption algorithm used for communications is known, but not the encryption key. In symmetric or secret key encryption, the key is known only by the sender and receiver. The key at the receiver allows decryption of the ciphertext into plaintext, exactly reversing the encryption process, as shown in Figure 4.

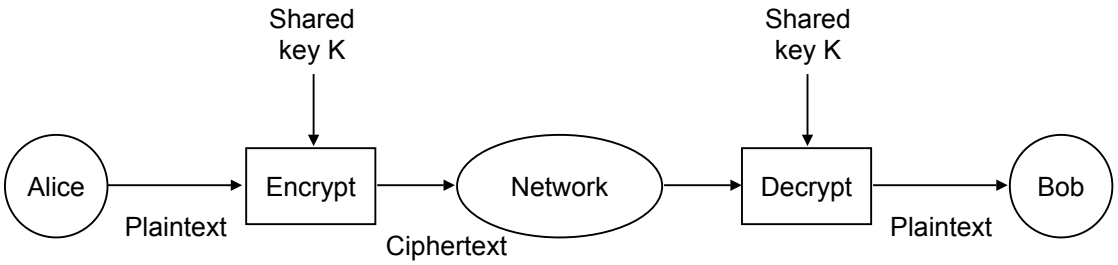

Fig. 4. Secret key cryptography

Protection against eavesdropping can be accomplished similarly by asymmetric or public key cryptography, where the sender uses a public key for encryption and the receiver uses the corresponding private key for decryption as shown in Figure 5. The public key is known to everyone while the private key is known only to its owner. Although the public and private keys are mathematically related to each other, it should be very difficult to deduce the private key from the public key. Compared to secret key encryption, public key encryption offers the great advantage that the sender and receiver do not have to share a secret before they can start communicating with each other.

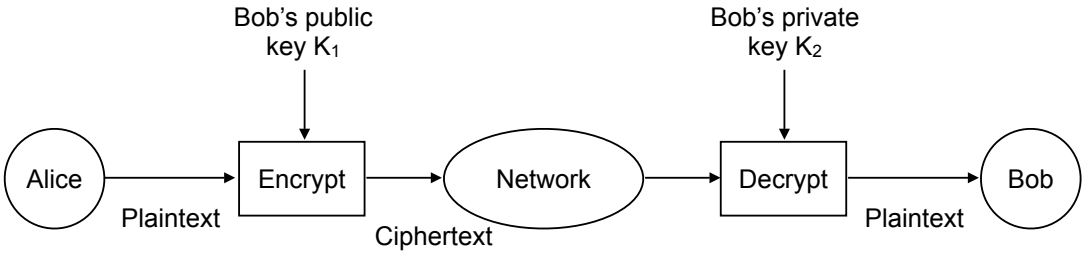

Fig. 5. Public key cryptography

A Web server can present its public key to clients in a verifiable public key certificate. To obtain a certificate, the site owner registers with a trusted third-party certificate authority. A public-private key pair is arranged by the certificate authority. The public key is recorded with the key owner's identity and expiration time in a certificate, along with the certificate authority's digital signature. The digital signature is cryptographic evidence that the certificate was created by the certificate authority and has not been altered. The certificate authority essentially vouches for the public key owned by the server. With the certificate presented by the server, a client can send data encrypted with the public key; presumably the server is the only one who owns the private key and can decrypt the encrypted data.

TLS/SSL is the protocol used in practice for ensuring private communications between a Web client and server. First, the server presents its public key certificate to the client. The client may present a certificate to the server but usually this is not done. Web browsers are pre-programmed with a set of certificate authorities that it will recognize. Then the client and server perform a TLS/SSL handshaking procedure to negotiate a secret key encryption algorithm, parameters, and a "master secret" to use for the duration of the session. Two encryption keys, one for each direction between the client and server, are derived from the master secret. The actual data exchanged between the client and server will use secret key encryption with the derived keys. The secret key encryption protects the transmitted data from eavesdropping.

It should be noted that encryption provides confidentiality in messages but does not hide the messages themselves. In particular, it is impossible to encrypt the source and destination IP addresses in IP packets because routers must be able to read the addresses to forward the packets. Hence, encryption does not prevent the source and destination IP addresses from being observed by eavesdroppers. An eavesdropper would be able to see that a client has contacted a server, but not see the contents of the request or reply.

Anonymizing agents: The purpose of anonymizing agents is to prevent Web requests from being traceable to the original IP address. Anonymizing agents can be single point agents such as Anonymizer or networked agents such as onion routers.

The basic idea of Anonymizer is to submit Web requests on behalf of its users through secure servers. The original version in 1996 was simply a proxy server sitting between clients and servers. A user submitted a URL into the "www.anonymizer.com" site, which fetched and forwarded the requested Web page. The server then deleted all history of the transaction.

A popular example of a Web proxy is Privoxy (privacy enhancing proxy) based on an earlier commercial program called Internet Junkbuster (http://www.privoxy.org). Sitting between

the client and server, Privoxy works by filtering ads, banners, Web bugs, animated GIFs, Javascript annoyances, and other unwanted contents from fetched Web pages. Ads are recognized by examining the image's size and URL reference.

The original Anonymizer service was vulnerable to eavesdropping because URL requests were sent without protection. The next version of Anonymizer added SSL encryption between the client and the anonymizer proxy. Additional features include cookie caching, filtering out viruses and spyware, cleaning up malicious Javascript or other scripts, and blocking the client from reaching known malicious Web sites.

LPWA (Lucent Personalized Web Assistant) was a research project adding a pseudonym agent ("persona generator") to a Web proxy (Gabber, Gibbons, Kristol, Matias, and Mayer, 1999). The purpose of the persona generator is to maintain a persistent session with an alias on behalf of a user. The user connects to the Web proxy. For each requested Web site, the persona generator creates an alias consisting of an alias username, alias password, and alias e-mail address. A Web site sees the alias instead of the user's real information, and can even send e-mail to the user (through the alias e-mail address). The main advantage of LPWA over an anonymizing service such as Anonymizer is the capability of LPWA to maintain personalized services for a user.

Numerous Web sites are available to act as proxies for anonymous Web surfing and searching. Proxies offering anonymous search often use Google for searching but will block Google from setting their cookie (an unique identifier to track a user's searches).

A disadvantage of single anonymizing agents is a requirement that the agent is trusted. The agent is an attractive target for attacks. It is also tempting for attackers to observe the inputs and outputs, and attempt to analyze traffic by correlating the inputs and outputs. Logically, a network of anonymizing agents might offer more resilience against attacks and reveal less to traffic analysis.

A widely influential idea for a network of anonymizing agents was David Chaum's "cascade of mixes" (Chaum, 1981). A mix is a computer that sits between a set of senders and receivers, as shown simplified in Figure 6. The goal is to confuse any observer from analyzing the traffic to learn both the source and destination of a message. Here K_m is the public key of the mix; K_b is the public key of receiver Bob; and M is the message. First, the message M is encrypted with Bob's public key K_b. Alice attaches Bob's address to the encrypted message, and encrypts the entire thing with the mix's public key K_m. One can view the message as double layers of encryption. Only the mix can decrypt the outer layer and read Bob's address as the recipient. The mix will forward (output) messages in a different order than it receives (inputs) messages. The mix does not have to be entirely trusted. If the mix delivers Bob's message to a different receiver, only Bob can decrypt the message with his private key.

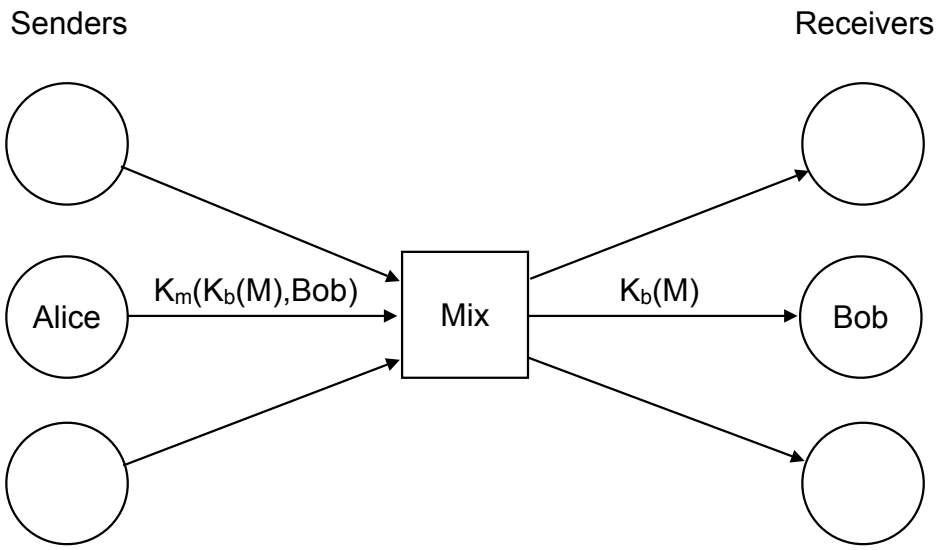

Fig. 6. Simplified operation of a mix

The process is slightly complicated by cascading or placing multiple mixes in series, but the idea is the same. The message must be encrypted with an additional layer of encryption for each mix in the cascade. Each mix successively removes a layer of encryption until the innermost encrypted message is delivered to the recipient. The advantage of a cascade is that any single mix can provide the secrecy of the correspondence between the senders and receivers.

Onion routing is conceptually similar to cascades of mixes. Onion routing works by forwarding encrypted HTTP requests through a series of onion routers (Goldschlag, Reed, and Syverson, 1999). Each onion router successively removes a "layer" of encryption until the HTTP request is ultimately delivered to the Web server with the IP address of the last onion router, making it impossible to trace back to the original requester. Each onion router knows its predecessor and successor in a message's route but not the entire route. A proof of concept was demonstrated, but a practical system was not built.

TOR (The Onion Router) is a second-generation onion router (Dingledine, Mathewson, and Syverson, 2004). It adds numerous improvements to the original onion routing design, including low latency and better anonymity. Perhaps most importantly, TOR has been released as free software and deployed with several hundred nodes with support from the Electronic Frontier Foundation. Consequently, various other software projects are being designed to work with TOR. For instance, XeroBank Browser, formerly called Torpark, is a variant of Firefox with built-in TOR access through secure and encrypted connections. OperaTOR is a bundle combining the Opera browser, TOR, and Privoxy.

As the name suggests, Crowds works by blending individuals into a big group of Crowds users (Reiter and Rubin, 1999). Each Crowds user runs a local "jondo" process that serves as a Web proxy and an HTTP request forwarding agent. When the user wants to visit a Web site, the HTTP request is handled by his/her local jondo. The jondo will randomly choose to forward the request to another jondo (Crowds member) with probability P or to the requested Web site with probability 1-P. If forwarded to another jondo, the next jondo will make another random choice to forward the request to another jondo with probability P or the requested Web site with

probability 1-P. The system-wide parameter P is configurable and affects the method's effectiveness. HTTP requests are thus forwarded along a random path through the Crowds. None of the group knows the identity of the original requester. Each jondo remembers its predecessor and successor, because replies from the server are returned along the same path in the backward direction. For additional privacy, all links are encrypted using a private key shared by the nodes on each end of the link.

Another variation of Chaum's cascade of mixes can be found in Freenet (Clarke, Miller, Wong, Sandberg, and Wiley, 2002). Freenet is a self-organizing peer-to-peer network designed to pool unused disk space on member PCs into a vast collaborative virtual file system. Privacy is enforced by forwarding messages from node to node, re-encrypted on each link. Each node is aware of its predecessor and successor, but not the original sender or ultimate recipient. Tarzan is another peer-to-peer network realization of mixes (Freedman and Morris, 2002).

Firewalls: Organizations typically deploy firewalls at the perimeters of their enterprise network to filter out unwanted traffic from the Internet. Firewalls are an essential component for protecting servers from attackers. However, the effectiveness of firewalls depends greatly on its filtering rules. Effective firewall rules generally require a high level of technical expertise. Also, although other ports can be closed, port 80 (default for HTTP) and perhaps port 443 (for secure HTTP) must be kept open for a Web server. Attacks can still be carried out over these ports.

PCs typically contain personal firewalls which can block incoming malicious traffic. Again, their effectiveness depends on proper configuration of filtering rules.

Cookie disabling: Fortunately, modern browsers offer users a good degree of control of cookies. Browsers can be set to reject cookies, selectively accept cookies, or ask every time before accepting each cookie. Cookies can be deleted manually, or automatically after each session or a specific expiration time.

Ad blocking: Modern browsers can be configured to block images, although this is mostly to improve rendering speed. It would be desirable to selectively block images of zero or one pixel dimensions that are probably Web bugs, but this is complicated by the practice of some sites to use small images for composing the site's appearance. Hence, blocking these small images could change the appearance of a site.

Anti-spyware: Spyware has become a considerable threat for identity theft. Many anti-virus programs will look for spyware, and numerous specialized anti-spyware software is available as well. The main function of anti-spyware is recognition of spyware, which is more difficult than might be expected because spyware authors are constantly developing new spyware and trying new tactics to evade detection. The most accurate approach to detection is signatures (unique characteristics) developed from careful examination of spyware samples. Unfortunately, signatures can take considerable time to develop, test, and distribute. Also, anti-spyware must be continually updated with the latest signatures. If signatures are inaccurate, anti-spyware might miss detection of spyware (a false negative) or mistake a legitimate program as spyware (a false positive). A zero rate of false negatives and false positives would obviously be ideal, but the complexity and stealthiness of spyware make false negatives and positives unavoidable. If detection is possible, anti-spyware should ideally be able to disinfect a computer (remove the spyware) and protect it from future infections. Unfortunately, these tasks may also be

complicated by the damage done by spyware. In many cases, the best approach may be erasing a computer completely and installing a clean operating system.

Anti-phishing tools: Phishing has become a widespread threat due to the ease of spam, ease of setting up phishing sites, and mostly, the ease of deceiving enough consumers into disclosing their valuable personal information to make phishing profitable. A variety of anti-phishing methods exist today. The main non-technical method is user education. If users are alert to phishing attempts, the response rate might be reduced sufficiently to make it unprofitable for criminals to continue. Technical methods include spam filtering to block phishing lures; phishing site detection (by various methods) and blacklisting or take down; and browser aids such as toolbar add-ons. Browser toolbars usually work by checking URLs against blacklists of known phishing sites and whitelists of legitimate sites, combined with heuristic tests of a site's HTML contents.

Strong passwords: Web servers are typically protected by passwords. Most computers today enforce policies for choosing strong passwords. Unfortunately, passwords are still vulnerable to a multitude of password cracking tools. Also, system administrators sometimes neglect to change the default accounts and passwords shipped with many computers.

Vulnerability testing and patching: Like all complex software programs, browsers and servers are susceptible to having vulnerabilities - security weaknesses that may be targeted by exploits, pieces of software written specifically to take advantage of a security weakness. Historically, common exploits have been buffer overflow exploits which attempt to compromise a target by running arbitrary malicious code.

The computer industry publishes known vulnerabilities to enable system administrators to patch their computers with the latest software updates. In addition, numerous vulnerability scanners are available for testing systems to identify their vulnerabilities. Vulnerabilities should be eliminated by patching, or if patches are not available (because patches take time to create and distribute), by other means such as firewalls.

Intrusion detection: Host-based intrusion detection systems run on computers and monitor their activities for signs of malicious intent, while network-based intrusion detection systems monitor network traffic for malicious signs. Both host-based and network-based intrusion detection systems are widely deployed in combination with firewalls and other security measures.

Similar to anti-spyware, intrusion detection systems depend mostly on attack signatures, called misuse detection. Signatures allow accurate identification of attacks, but signature-based detection is limited to known attacks. It is possible for new unknown attacks to evade detection if they are significantly different from signatures. A new signature can be developed after a new attack is detected.

A different approach is anomaly detection, which characterizes normal activities or traffic. Anything different from the normal profile is detected as an anomaly. Anomaly detection has the advantage of possibly catching new unknown attacks, under the assumption that attacks will be different from the normal profile. However, there are significant drawbacks to anomaly detection that continue to challenge researchers. First, normal behavior or traffic is complex and always changing. The normal profile must be continually updated. Second, anomaly detection is imprecise. Anomalies are unusual activities or traffic but not necessarily malicious. By itself,

anomaly detection does not identify the exact nature of the anomaly, only that it is not normal. Indeed, only a small fraction of anomalies may be malicious, but all of them will require investigation, possibly wasting a great amount of effort.

P3P: The World Wide Web Consortium's Platform for Privacy Preferences Project (P3P) aims to provide a standardized method for Web sites to communicate their policies for data collection and use. Descriptions of these policies are sent from a P3P-enabled server to a P3P-enabled browser in machine-readable XML (extensible markup language) format. The site's policies are automatically compared to the user's preferences, saving the user from having to read and understand the site's entire policy. The browser can warn the user about differences or disable certain functionalities. However, P3P does not specify what policies should be, nor does it ensure that a site will actually follow its stated policy. It simply describes an XML-based vocabulary for representing privacy policies.

The number of commercial sites with privacy policies satisfying customers with strict P3P privacy preferences is an issue. The concern is that the lack of acceptable sites will cause users to lower their privacy requirements, defeating the purpose of P3P. Hence, the future of P3P is in doubt.

Few sites appear to be motivated to use P3P due to the time and money required to prepare their privacy policy statements in support of P3P. In fact, the total number of P3P-enabled Web sites appears to be small. A survey done in 2003 found only 9.4% sites adopting P3P, and a later survey completed in late 2005 found about the same unchanged adoption level (Reay, Beatty, Dick, and Miller, 2007). Deployment of P3P appears to be stagnant.

Future Trends

The real problem is to keep the convenience of personalized online services while protecting privacy at the same time (Kobsa, 2007). By privacy protection, we mean that individuals have control over the collection, sharing, and use of their personal data. As seen in this chapter, privacy protection can be approached through regulations or technology.

Nations have a number of regulations related to privacy protection, but mandatory regulations are not uniform or very far reaching. The U.S. has been reluctant to attempt comprehensive mandatory regulations. Most likely the U.S. will continue to rely on voluntary regulations. Today self regulations consist mainly of privacy seal programs. However, they are probably not well understood by most consumers, nor do privacy seals actually ensure the privacy of data. Hence, one could argue that the regulatory approach to privacy protection will continue to be largely ineffectual.

It could be argued that technological approach are more promising. In the past few years, significant progress has been seen in the deployment of TLS/SSL, anonymizing agents (such as TOR), and privacy options in major browsers (such as ad blocking). It is relative easy today to do Web surfing or searches with anonymity. In a way, these have been the easy problems.

Identity theft by phishing or spyware continue to be open problems. Although various methods are used to fight phishing, identity theft is a growing problem because phishers invent new techniques. The same is true for spyware. These problems are more difficult because they are adversarial and continually changing.

Conclusion

By just about every measure, the Web has been phenomenally successful. Obviously the Web has become pervasive in social, business, and industrial contexts. Its ubiquity is evidence of its success but also the reason for public concern. As our dependence on the Web grows, more pieces of personal information become exposed. Not only are pieces collected, but it is possible that isolated pieces of personal information may be correlated together and mined into more threatening invasions of privacy.

As seen in this chapter, a variety of non-technical and technical approaches exist to protect online privacy. Yet all approaches to date have only been partial solutions. By its nature, the Web was designed as an open system to facilitate data sharing. There are many ways to collect personal information through the Web, and there are companies and even governments that are interested in collecting this information. Also, most people tend to be willing to lose a certain amount of privacy to gain access to online services. A universal solution does not seem likely, but even if a solution was available, it is not clear where the balance between privacy and convenience should be drawn.

References

Anton, A., Earp, J., Vail, M., Jain, N., Gheen, C., and Frink, J. (2007). HIPAA's effect on Web site privacy policies. IEEE Security and Privacy, 5(1), 45-52.

Benassi, P. (1999). TRUSTe: an online privacy seal program. Communications of the ACM, 42(2), 56-59.

Chaum, D. (1981). Untraceable electronic mail, return addresses, and digital pseudonyms. Communications of the ACM, 24(2), 84-88.

Clarke, I., Miller, S., Wong, T., Sandberg, O., and Wiley, B. (2002). Protecting free expression online with Freenet. IEEE Internet Computing, 6(1), 40-49.

Dierks, T., and Rescorla, E. (2006). The Transport Layer Security (TLS) protocol version 1.1. Internet Engineering Task Force RFC 4346.

Dingledine, R., Mathewson, N., and Syverson, P. (2004). Tor: the second-generation onion router. Presented at 13th USENIX Security Symposium, San Diego, CA.

Freedman, M., and Morris, R. (2002). Tarzan: a peer-to-peer anonymizing network layer. Proceedings of the 9th Conference on Computer and Communications Security, ACM Press, 193-206.

Gabber, E., Gibbons, P., Kristol, D., Matias, Y., and Mayer, A. (1999). Consistent, yet anonymous, Web access with LPWA. Communications of the ACM, 42(2), 42-47.

Garfinkel, S. (2002). Web Security, Privacy, and Commerce, 2nd ed. Sebastopol, CA: O'Reilly and Associates.

Goldschlag, D., Reed, M., and Syverson, P. (1999). Onion routing. Communications of the ACM, 42(2), 39-41.

Kobsa, A. (2007). Privacy-enhanced personalization. Communications of the ACM, 50(8), 24-33.

Kristol, D. (2001). HTTP cookies: standards, privacy, and politics. ACM Transactions on Internet Technology, 1(2), 151-198.

Linn, J. (2005). Technology and Web user data privacy. IEEE Security and Privacy, 3(1), 52-58.

Martin, D., Wu, H., and Alsaid, A. (2003). Hidden surveillance by Web sites: Web bugs in contemporary use. Communications of the ACM, 46(12), 258-264.

Moores, T. (2005). Do consumers understand the role of privacy seals in e-commerce? Communications of the ACM, 48(3), 86-91.

Reay, I., Beatty, P., Dick, S., and Miller, J. (2007). A survey and analysis of the P3P protocol's agents, adoption, maintenance, and future. IEEE Transactions on Dependable and Secure Computing, 4(2), 151-164.

Reiter, M, and Rubin, A. (1999). Anonymous Web transactions with Crowds. Communications of the ACM, 42(2), 32-48.

Rezgui, A., Bougeuettaya, A., and Eltoweissy, M. (2003). Privacy on the Web: facts, challenges, and solutions. IEEE Security & Privacy, 1(6), 40-49.

Shukla, S., and Nah, F. (2005). Web browsing and spyware intrusion. Communications of the ACM, 48(8), 85-90.

Warren, S., and Brandeis, L. (1890). The right to privacy. Harvard Law Review, 4(5), 193-220.

Westin, A. (1967). The Right to Privacy. Boston, MA: Atheneum Press.

Key Terms and Definitions

Anonymizing agent: a program acting as an intermediary between client and server to make Web requests untraceable to the original client.

Client: an application such as a Web browser running on a user's computer that sends a request to a server as necessary.

Cookie: a text string stored in a browser to keep state across multiple Web transactions.

Cryptography: the use of mathematical algorithms to protect transmitted data from eavesdropping.

Phishing: the use of malicious Web sites masquerading as legitimate sites to deceive users into disclosing personal information.

Privacy: protection of personal information from disclosure.

Server: an application running on a computer for responding to a client's request.

Spyware: a type of malicious software designed to covertly steal a computer user's personal information.

Web bug: a small, usually invisible image embedded in a Web page used to detect the IP addresses of users viewing the page.