Masks: Bringing Anonymity and Personalization Together

Unlike most privacy tools, the Masks framework gives Web sites general information to personalize services without compromising the user's anonymity.

Lucila Ishitani, Virgilio Almeida, and Wagner Meira Jr.
Federal University of Minas Gerais

Most Web users realize that sites are collecting information about them, though few realize how much data is gathered or how that gathering occurs. Although some companies publish privacy policies to inform users about their practices, most policy statements are full of technical and legal jargon and are difficult to understand. Further, the Pew Internet & American Life Project's privacy survey[1] revealed that, although US users are anxious about having their activities monitored, only 5 percent use tools to "anonymize" their requests, and only 10 percent reject cookies. Still other surveys (see, for example, www.pandab.org/doubleclicksummary.html) have found that users have a strong desire for personalization, which enhances services by customizing sites to users needs—but also requires sites to gather data to that end.[2]

To bridge these conflicting needs, we developed Managing Anonymity while Sharing Knowledge to Servers,[3] a Web-based framework that balances users' privacy concerns during Web browsing activities with their desire for personalized Web services. Masks uses a selective revelation scheme that erects an anonymity barrier between the user's private data and Web services, and controls the information that flows across that barrier to the service. This kind of filtering minimizes user information exposure while still permitting some form of service personalization. Also, because it addresses privacy at the data-collection level, Masks prevents third parties from building user profiles based on links to sites and information that might reveal personal information, such as religion, travel preferences, sexual orientation, and so on.

Privacy: Issues and threats

On the Web, privacy invasion can take several forms, including hackers who gather data by attacking users' email, user groups, and computers, and online service providers who monitor user activity and habits.[4] Service providers also systematically collect personal information to personalize Web sites, which raises questions that users rarely know how to answer. For example,

- How much privacy do you give up when you make information about yourself public?
- How much information do you reveal when you interact with a Web service?
- How much information are you willing to make available to obtain better services from businesses?

As observed in *The Economist*, privacy is a residual value, hard to define or protect in the abstract.[5] In fact, people have often characterized privacy in the information age as individually determined.[6] Alan Westin's well-known definition of information privacy—"the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others"[7]—illustrates this direct dependency on individual consent.

Given this, privacy definitions will vary among individuals. What one person considers privacy invasion, another might consider a normal and acceptable exposure. How users perceive privacy risk—that is, how much information users feel comfortable disclosing or having col-

lected about them without consent—is a product of how they view privacy.

Privacy-protection layers

To implement personalized privacy protection, we divide privacy protection into layers. By adding protective layers, users can enhance their privacy level. This layering mimics what happens in the real world. How much people expose about themselves in real life depends both on their own actions and the situation. That is, people vary their information disclosure according to place, time, other people, and the various entities involved in the interaction.

Like geological layers, each privacy layer is independent of the others, and the existence of one layer does not entail that the previous one also exists. However, when more than one layer does exist, they are always organized in the same order (see Figure 1).

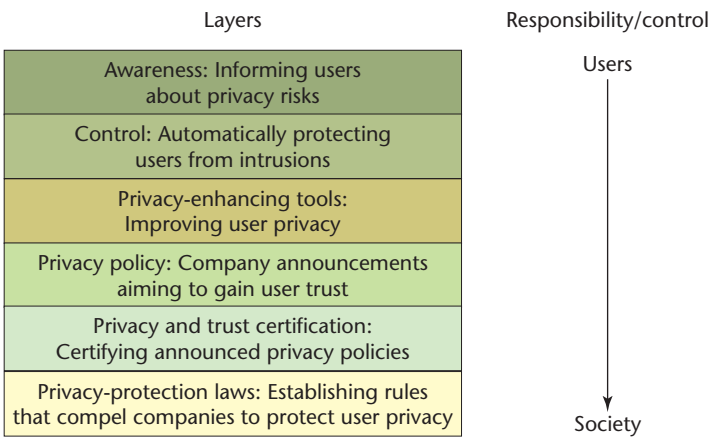

Figure 1. Privacy-protection layers. Adding subsequent layers mimics real life in that people disclose more or less information about themselves according to the situation.

Layer 1: Awareness

Individuals can generally provide information to Web services in two ways: voluntarily or involuntarily. In the first case, users fill out forms or send emails with specific information. In the second case, Web services collect users' data and monitor their activities without notice or consent.

We define privacy risk as the information that is disclosed or derived without consent when users interact with a site. Generally, users are unaware of privacy risks. For example, many users still don't know what cookies are, and of those who do, many are unaware that blocking cookies can sacrifice desirable services, such as personalization. In general, users are unaware that Web servers can use each mouse click to create detailed personal profiles about them and find out which sites they previously visited. Clearly, we need to better inform users about such privacy risks.

Privacy Critics,[8] for example, is a first-layer privacy-protection tool that issues warnings and suggestions in response to user interactions. Its goal is to improve users' knowledge of privacy risks so they can understand their exposure level during interactions with Web sites. Simply informing users about privacy risks, however, is not the same as acting to protect privacy.

Layer 2: Control

Some technologies undeniably create conditions for privacy invasion. History files, third-party cookies, and Web bugs all support user behavior analysis without users' knowledge or consent. Layer 2 includes mechanisms that let users take control by choosing mechanisms or tools to fight such explicit attempts to violate their privacy.

The key technology here is the Web browser and its extensions (such as plug-ins), which must let users easily reject or filter undesirable data gathering methods. Malicious code can easily retrieve history files and disclose to third parties all the pages that users visit during a given session. Browsers should either automatically delete these files or make it easy for users to delete them.

Cookies record information in a user's machine, which, for example, can let sites identify the user on future visits. Web browsers like Microsoft Internet Explorer and Netscape let users reject cookies. However, for many users, selecting that option is a cumbersome task. Also, only users who are already conscious of privacy risks will opt to stop undesirable services.

Users can also install a filter, such as the one offered by Anonymizer (www.anonymizer.com). Filters are software programs that block cookies, banner advertisements, and Web bugs. The disadvantage of filters is that they fail to consider consent; they block all cookies, and thus users lose access to all personalized services, even those from the most trustworthy of sites. Also, filters make privacy invasion difficult, but not impossible. A site can still identify users by IP address, interaction time, and geographical location, for example. Given this, users might need additional levels of privacy protection.

Layer 3: Privacy-enhancing tools

This layer includes most existing privacy-protection tools. The main difference between this layer and the previous one is in the location of privacy-protection mechanisms. In Layer 2, the user controls privacy protection; in Layer 3, the privacy-preserving mechanisms operate elsewhere. Some privacy advocates argue that users should control their privacy rather than rely completely on a

Web site's privacy policies. Usually, the mechanisms they advocate use anonymity or pseudonymity: users adopt a virtual name, individually or as part of a collective, and use the name to interact with Web sites. The difference between anonymity and pseudonymity is that pseudonyms are unique and persistent—users each own a distinct identification, which they can use throughout their interactions with a site.

Although it's not easy to associate pseudonyms with real users, it is possible to associate a group of messages with a user. Anonymity makes this impossible, however. The Anonymizer, for example, acts as a proxy, submitting Web requests on users' behalf. As a result, sites only know the proxy's IP address. Anonymizer does have drawbacks, however. Because a Web server can't discern users, it can't identify their preferences and provide desirable personalization and customization services.

The Lucent Personalized Web Assistant (LPWA; www.bell-labs.com/projects/lpwa) is a pseudonym tool that lets Web sites offer identification-based services without linking to users' actual identities. This technique also has problems, however. As in real life, if someone discovers the real identity of a pseudonymous user, all of the user's past actions are automatically exposed.

Unlike Anonymizer and LPWA, which require a unique third party to forward requests, Onion (www.onion-router.net) and Crowds[9] basically hide the request's real originator in a group. In Onion, a message's path through the group members is predetermined; in Crowds, the path is configured during request transmission. As with all anonymity tools, these tools do not make data available for personalization.

It's difficult to reach consensus on Web privacy because the privacy concept is heavily dependent on cultural and political issues.

Layer 4: Privacy policies

The idea at this level is to give users information about a site's privacy policies, and then let them negotiate how the site gathers and uses their information. One proposal to this end is the World Wide Web Consortium's Platform for Privacy Preference Project (P3P; www.w3.org/p3p), a proposal for controlling Web sites' personal information use. P3P offers a way for Web sites to disclose how they handle user information, and for users to describe their privacy preferences. P3P-enabled Web sites make this information available in a standard, machine-readable format. P3P-enabled browsers then "read" this snapshot automatically and compare it to users' own privacy preferences. If the policy matches the user's security configuration, the browser continues the requisition of pages from the site. If not, the user can resolve disparities by interacting with an agent that notifies them of the disparity and presents alternatives to resolve the conflict.

P3P neither sets a minimum privacy standard nor monitors whether sites adhere to their own stated procedures. Users thus have to trust entirely in the Web site. Also, companies might change their privacy policy in undetectable ways. To protect users from such risks, the next two privacy-protection layers are crucial.

Layer 5: Privacy and trust certification

This layer is concerned with ensuring that Web sites observe their announced privacy policies. To do this, privacy groups and organizations could periodically verify Web sites' announced privacy policies and assign each site a grade that is explicitly available to users. These grades could be based on Huaiqing Wang and colleagues' taxonomy: improper access, improper collection, improper monitoring, improper analysis, improper transfer, unwanted solicitation, and improper storage.[4]

A recent survey emphasized the importance of privacy policies, finding that the vast majority of consumers and businesses surveyed expected to both see and understand privacy policies when they visit an e-commerce site.[10] However, we must approach privacy certification with caution. It has been reported in the news (http://abcnews.go.com/sections/tech/DailyNews/toysmartftc000711.html) for example, that some bankrupt companies transferred assets, including private user information, to other companies. The companies purchasing this information did not necessarily feel obligated to uphold the privacy policies that were in place when the data was collected.

Layer 6: Privacy-protection laws

In many countries, governments have discussed and proposed laws to regulate privacy protection and mechanisms to punish people and organizations that break the rules. Until privacy laws are really enforced, however, companies will find few incentives to protect and respect user privacy, mainly because most users don't even realize that their privacy can be violated. A central problem is that behavior on the Web can't be controlled. To regulate the Web, governments would have to regulate code writing or how Web applications (browsers, Java, email systems, and so on) function.[11]

Also, it's difficult to reach international consensus on Web privacy because the privacy concept is heavily dependent on widely variable cultural and political issues. Despite this, however, there is a set of common activities

that are undoubtedly privacy invasion:

- collecting and analyzing user data without the user's notice or authorization
- employing user data in a way other than was authorized, and
- disclosing or sending user data to others without the user's knowledge and authorization.

Even if international privacy laws existed, some countries and companies would still be likely to operate in an opprobrious way. Consequently, users can't rely only on laws to protect their privacy. Mechanisms must exist to let users improve the protection of their data.

Masks

To the best of our knowledge, no mechanism implements all six privacy-protection layers—mainly because Layers 5 and 6 rely on the laws of different countries, rather than on individual initiatives. However, it is possible to design and implement integrated tools that aim to cover most protection layers. To that end, we've designed Masks, a distributed, consent-based privacy architecture that implements the first three layers of the privacy-protection framework.

In our framework, a *mask* is a temporary identification that a user adopts while interacting with a Web site (see Figure 2). This identification is associated with the user's interest in a given topic or specific site. Whenever Masks users visit a Web site that identifies its visitors, they grab masks so that they're not identified when they interact with the site. Users can use different masks during a site interaction, depending on their interest in any given moment.

As Figure 2 shows, the Masks framework has two major components: a Masks server, which acts as an intermediary between users and Web sites, and a privacy and security agent (PSA), which acts as an intermediary between users and the Masks server. The PSA runs in conjunction with a user's browser and has several functions, including

- ciphering user requests to prevent eavesdropping
- informing users about both potential privacy intrusions and their assigned masks, and
- providing mechanisms that let users configure the masks.

The PSA also lets users turn off the masking process if they want to interact directly with sites without anonymity. Finally, the PSA blocks and filters known methods of privacy violations, such as cookies and Web bugs. Given these functions, PSA offers users privacy protection that corresponds to the first two protection layers: awareness and control.

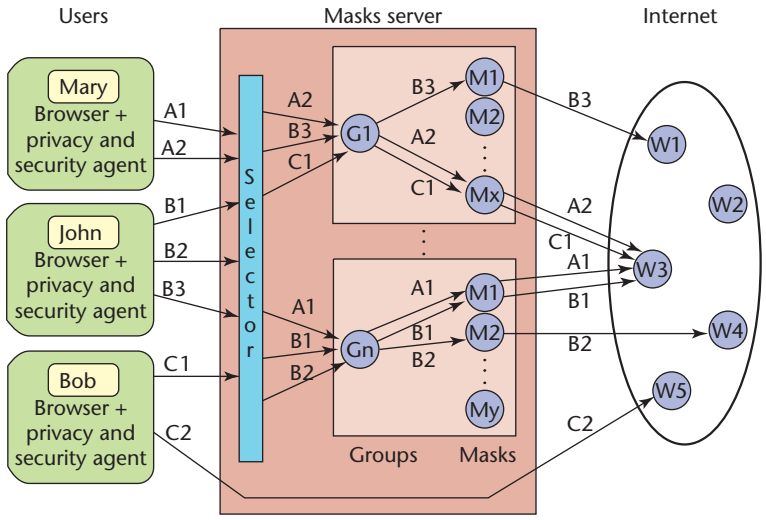

Figure 2. A simplified view of the Masks architecture. The architecture implements the first three privacy-protection layers and provides several user benefits, including making it possible for sites to offer partial personalization services.

The Masks server

The Masks server works as a proxy that can be deployed over special network locations, such as in an intranet or in a service provider's proxy. The Masks server manages masks and assigns them to users according to groups. A group represents a topic of interest, and a user's request is assigned to a group according to the requested object's semantics. Because a Masks server makes requests on behalf of a group rather than an individual user, it can offer Web sites relevant data about users' interests without disclosing their identity. Web servers can then use the data to personalize the interaction.

The Masks server's *selector* chooses which group to assign users to on a per-request basis. We work at the request level for two reasons. First, it's difficult to predict overall user behavior. Second, a single user might express diverse interests during a single session. Thus, it is more appropriate to characterize user interests based on each request. Because the requested object's nature is one of the main sources of information for assigning users to groups, users need not disclose any personal information. The Masks server thus keeps no private data, and it's virtually impossible for Web sites to determine whether a sequence of accesses assigned to the same mask belongs to a single user or a group of users with similar interests.

For example, in Figure 2, Web site W3 believes that requests A2 (Mary) and C1 (Bob) came from the same user. With Masks, Web sites can access the group's navigation pattern and offer the group personalized services,

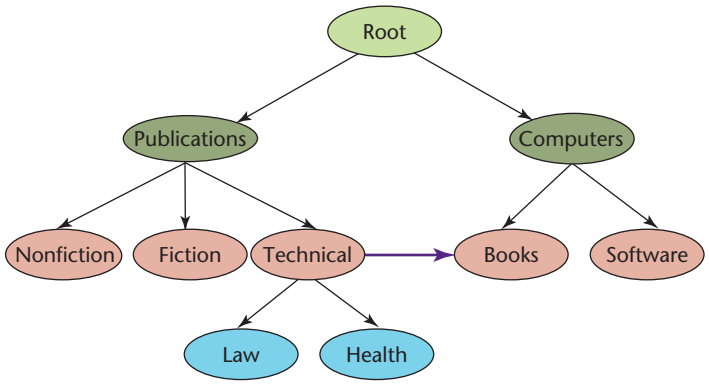

Figure 3. Example of a semantic tree. The tree defines increasingly specialized concepts, with the root group being the most generic. The specializations of some nodes lead to nodes in other subtrees, as represented by the purple line between Root/Publications/Technical/Books and Root/Computers/Books.

but they can't profile users because they can't determine the user associated with a request. Moreover, because a Masks server works as a proxy, the sites don't even realize that the requests are from a group of users. Also, as Figure 2 shows, each group can have several associated masks (one for each site that offers relevant information). Many sites offer travel information, for example, and the group interested in travel will have one mask for each of the previously visited travel-related sites.

Users can associate with different groups and masks. In Figure 2, John is associated with two different groups and three different masks. The same user can also be assigned to different masks while browsing one site. Suppose that Web site W3 in Figure 2 is a portal that offers different classes of information, such as travel and finance. If Mary requests travel-related services (A1) and later requests finance information (A2), W3 would view the two requests as coming from two distinct users, because Mary is in two distinct groups. Finally, Masks lets users relinquish their anonymity by turning off the PSA agent and interacting directly with the site (as with request C2 in Figure 2).

Because the masks are divided into groups, our server's effectiveness depends on how we define groups and how we assign objects to them. One strategy we evaluated is the use of a semantic tree as defined by the Open Directory Project (http://dmoz.org). This tree organizes millions of Web sites according to their semantics and is maintained by volunteer editors all over the world. The tree is free and readily available, and we use it as our starting point for defining groups and their relationships to each other. As Figure 3 shows, each tree node represents a semantic category, or, in our approach, a group, and the categories are organized hierarchically.

Benefits

Masks offers its users several benefits that not only make them aware of privacy risks, but also help them control the amount of information they transparently provide. We categorize Masks benefits according to six major features:

- *Privacy protection*. Masks uses anonymity to preserve privacy.
- *Partial personalization*. Unlike other privacy tools, Masks discloses some data, which lets Web sites personalize services without profiling individual users.
- *Safety*. Masks stores only the last user request. This enhances safety: the less information Masks maintains, the less likely it is to be the target of an attack.
- *Transparency*. Users can choose their exposure level. Masks tunes its selective revelation scheme accordingly, providing better service to users.
- *Efficiency*. To associate masks with users, Masks uses simple data structures, such as lists, trees, and hashing tables, and a simple graph traversal algorithm rather than complex data mining or clustering algorithms. Thus, it is unlikely to increase user-perceived latency.
- *Flexibility*. Masks offers adaptive services that adjust dynamically to its users' behaviors.

Finally, Masks offers both interoperability and ease of use. It uses standard HTTP and TCP protocols and works with the usual identification mechanisms, such as cookies. No special protocol or proprietary technology is required. Masks uses a simple interface, requiring no information from users beyond their requests to a Web site.

We tested the Masks algorithms using the logs of an actual electronic bookstore. We based the groups on a five-level semantic tree extracted from the Open Directory Project. We evaluated Masks by assigning groups to the requests submitted to the bookstore, in particular verifying whether the Masks server was able to find a specialized topic that matched the request parameters. Our test showed that Masks assigned 80 percent of the requests to a specialized mask (that is, to a mask different from the root).[3] We're now developing a fully operational version of Masks.

Privacy is emerging as a central concern for both commercial and governmental spheres. Although we initially designed Masks for electronic business environments, its tenets can be extended to the other security and privacy contexts as well. □

References

1. S. Fox et al., *Trust and Privacy Online: Why Americans Want to Rewrite the Rules*, The Pew Internet & American Life Pro-

ject, Washington, D.C., Aug. 2000; www.pewinternet.org/reports/toc.asp?Report=19.
2. D. Riecken, "Personalized Views of Personalization," *Comm. ACM*, vol. 43, no. 8, Aug. 2000, pp. 27–28.
3. B. Gusmão et al., "Disclosing Users' Information in an Environment that Preserves Privacy," *Proc. ACM Workshop on Privacy in Electronic Society* (WPES 2002), ACM Press, Nov. 2002.
4. H. Wang, M.K.O. Lee, and C. Wang, "Consumer Privacy Concerns about Internet Marketing," *Comm. ACM*, vol. 41, no. 3, Mar. 1998, pp. 63–70.
5. "The End of Privacy" editorial, *The Economist*, vol. 351, no. 8117, 1 May 1999, p. 15.
6. J.R. Rao and P. Rohatgi, "Can Pseudonymity Really Guarantee Privacy?" *Proc. 9th Usenix Security Symp.*, Usenix Assoc., 2000, pp. 85-96.
7. A. Westin, *Privacy and Freedom*, Bodley Head, 1987, p. 7.
8. M.S. Ackerman and L.F. Cranor, "Privacy Critics—Safeguarding Users' Personal Data," *Web Techniques*, Sept. 1999; www.newarchitectmag.com/archives/1999/09/ackerman.
9. M.K. Reiter and A.D. Rubin, "Crowds: Anonymity for Web Transactions," *ACM Trans. Information and Systems Security*, vol. 1, no. 1, Jan. 1998, pp. 66–92.
10. J.B. Earp and D. Baumer, "Innovative Web Use to Learn About Consumer Behavior and Online Privacy," *Comm. ACM*, vol. 46, no. 4, Apr. 2003, pp. 81-83.
11. L. Lessig, *Code and other Laws of Cyberspace*, Basic Books, 1999.

Lucila Ishitani *is a PhD candidate in computer science at the Federal University of Minas Gerais, Brazil. Her current interests include Web privacy and data and knowledge mining. She has an MS in computer science from UFMG. Contact her at Computer Science Department, Federal University of Minas Gerais, Av. Antônio Carlos, 6627, Belo Horizonte, MG, Brazil, 31270-010; lucila@dcc.ufmg.br.*

Virgilio Almeida *is a professor and chair of the Computer Science Department at the Federal University of Minas Gerais, Brazil. His research interests include performance evaluation and large-scale distributed systems modeling. He has been a visiting professor at Boston University and Polytechnic University of Catalunya in Barcelona, and held visiting appointments at Xerox PARC and Hewlett-Packard Research Laboratory. He has a PhD in computer science from Vanderbilt University. Contact him at Computer Science Department, Federal University of Minas Gerais, Av. Antônio Carlos, 6627, Belo Horizonte, MG, Brazil, 31270-010; virgilio@dcc.ufmg.br.*

Wagner Meira Jr. *is an associate professor of computer science at the Federal University of Minas Gerais, Brazil. His research interests include performance analysis, data mining, and large-scale distributed system modeling. He has a PhD in computer science from the University of Rochester. Contact him at Computer Science Department, Federal University of Minas Gerais, Av. Antônio Carlos, 6627, Belo Horizonte, MG, Brazil, 31270-010; meira@dcc.ufmg.br.*

Now Available

IEEE Pervasive Computing
Special Issue on
The Human Experience

The April–June issue of *IEEE Pervasive Computing* features articles on designing for and having a better understanding of human interactions in a world of increased technological accompaniment.

Articles explore quantitative and qualitative evaluation techniques for investigating formative design challenges for ubiquitous computing, evaluate pervasive technologies in a European retail environment, review a sensor-rich assisted-living community, and discuss the development of a remote, sychronous teaching experience that emphasizes coordinating existing and novel techniques.

For more information or to purchase an article, visit
http://computer.org/pervasive