

Web Privacy: an Essential Part of Electronic Commerce

Igor Margasiński, Krzysztof Szczypiorski

Warsaw University of Technology, Institute of Telecommunications,
ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
e-mail: {I.Margasinski, K.Szczypiorski}@tele.pw.edu.pl

Abstract. Web users' concerns about their privacy result in reduction of online shopping activity. Widely spread Web marketing is not possible without an appropriate protection of customers' privacy. The significance of privacy in electronic commerce is illustrated in this article. We present a classification of Web privacy risks and describe an evolution of privacy enhancing technologies. It includes a review of Client-side Utilities, new protocol – P3P, simple Third Party Proxy Servers and more sophisticated Chaining with Encryption technique. The possibilities of mentioned solutions and their disadvantages are specified. Effective solutions are characterized by low performance and high costs of their realization. Electronic society's unrealized demands are highlighted against this background. Our proposal of a new method of providing versatile anonymity for Web users is included in the summary of this discussion. The solution overcomes mentioned weaknesses.

1 Introduction

In the broad spectrum of Web security, aside from such problems as server or confidential data security, the subject of end users privacy abuses is getting more pronounced. Getting and collecting personal information constitute major attacks on Web browser's users. The intrusion of privacy is intentional and can be seen in dollar signs. Personally identifiable information is a hot commodity.

According to new survey [9] most customers do not trust companies to handle their personal information properly. Users are looking for solutions that can assure their privacy protection. They require tools allowing a control of information about them on the Internet. The survey reveals, for example, that having a company's privacy practices verified by a third party would lead 91% customers to say they would do more business with such firm. More than half of customers say that if they were confident that a company really follows its privacy policies, they would be likely to recommend that company to friends and family.

Canadian report [5] is even more pronounced. It found that: 83% of consumers who have not shopped online cited that their reluctance is due to not knowing what was being done with their information and who was watching their surfing habits. 69% of frequent Internet purchasers say they have concerns about handing out personal information like credit card numbers online.

2 Igor Margasiński, Krzysztof Szczypiorski

This contribution is mostly a review paper. Our goal was to describe a spectrum of present anonymity solutions as tools for Web privacy, help to realize their deficiency and to point directions of require development. The last part of paper includes a proposal of original solution which overcomes described weaknesses.

2 Privacy Risks

The World Wide Web (WWW) – in its present shape – does not provide adequate privacy protection. We are now witnessing growing privacy's endangerment. The increasing number of commerce applications makes personal data a marketable good. We can see a progress both in Internet technology and electronic commerce, but a gap in protection of our right to privacy remains. Today, big specialized companies are the major drivers in utilization of this technology gap to their own advantage. However these practices effectually spoil electronic market and scare customers away from online shopping. The privacy risks should be examined from few different angles. Depending on placement, they can be divided into: internal, communication link originated and Web server originated.

2.1 Internal Risks

Here, the main problem is an access to users' personal data and Web activity information (for example visited websites) for local network administrators, employers or other third parties like Internet service providers. What is more important, the URL (*Uniform Resource Locator*) addresses can inform not only about which sites were visited by a user but also about the way he filled out HTML (*HyperText Markup Language*) forms. This takes place when using method GET of HTTP (*HyperText Transfer Protocol*) protocol [1], [6].

2.2 Communication Link Originated Risks

The second point is the risks from the communication link, with the major attack – sniffing. The main protocol in WWW system is HTTP. Its security is based only on reliability of TCP/IP (*Transmission Control Protocol / Internet Protocol*) protocol suite and does not forecast possibility of attacks. User should than take into account the fact, that information about visited sites and data from completed forms can be easily accessed also by other Internet users. According to Harris Interactive report [9]: 70% of customers claim that their major concern about online shopping is that Web transactions may not be secure and 69% that hackers could steal their personal data.

2.3 Web Server Originated Risks

The last issue is the risks related to the Web server. The danger lays in the fact that the website can obtain a wide range of information about a client. Server gets client IP

while establishing connection. The origin of each individual request and its association with each host are known to a server. A standard practice for most websites is to log HTTP requests to their Web server. This means that owners of website know the originating IP address of a user agent requesting a URL. They also have access to: date of request, realization time, user's name – HTTP identification, information about errors of HTTP transaction, referrer link, user-agent information.

Sending a URL address of previously visited page (referrer link) to a server constitutes a major privacy violation. In spite HTTP specification, which says that the availability of information about referrer link should be optional, not one existing user agent did in fact incorporate the possibility of turning off this mechanism. The significance of this violation is more pronounced because the URI (*Uniform Resource Identifier*) address can often contain data from HTML forms. An URI string may in particular contain keywords introduced into Web search engines.

The next tool, which can be used to take privacy away from users, is state management mechanism – Cookies [10]. This technique was introduced to allow, in stateless protocol (i.e. HTTP), a differentiation between persons visiting a server. Cookies mechanism has realized its new uses, not thought of by its creators. Today we can observe the utilization of this mechanism to create and enlarge databases with detail users' profiles. Let's follow a typical process of creating a profile. We can find banner ads on many websites. Mostly in the form of bitmap images or Shockwave animations, which come from profile creating third party (Ad Server). When visiting such page, in reality we are receiving banner ad from Ad Server, although the user does not know this. Our IP address is being automatically sent to the third party (Ad Server). Next, a Cookie information is placed. Ad Server writes information about us, such as the date, time and address of the visited page. We are identified by an ID number. There are many Web pages, which require registration before accessing some services. In the registration process we release personal identity information like name, e-mail address and so on. The registration page sends all the information to Ad Server, which creates more detailed information database, updated every time we visit pages in the banner ad network. Information gathered this way is used by banner ad network pages to target specific commercial recipients, to create personal e-mail offers and *spam* or to breach privacy even further. In user agent default settings cookie mechanism is invisible to users. The statistics [9] showed that the main (75%) concern expressed by customers is that companies they patronize will provide their information to other companies without their permission. We should also mention risks created by programs executed on client machines. They may be so called *Trojan Horses*, because they, in a way unnoticeable to users, send information about them.

All the presented growing risks make new requirements necessary to establish. Web users should have a choice to provide or not to provide their personal data.

3 Current Solutions

Two types of solutions are dominant in ordinary use. They are: Client-side Applications and Proxy Servers.

3.1 Client-side Utilities

These types of solutions can play only a secondary role. They are not able to provide full anonymity. The functions of those applications are: monitoring and control of all connections to and from user's computer (i.e. personal firewall), management of Cookies mechanism, system cleaning (removal of history or cookies files), blocking banner ads and detecting *Trojan Horses*. Today a growing number of the discussed applications allow not only for single but multiple functions. The software installed on the user's computer does not provide full protection of privacy. It is not able to conceal such data which may identify a host like IP address.

3.2 Third Party Proxy Servers

Proxy is a third party – like a mirror reflecting destination data. It acts as the middleman in the process of Web browsing.

Adopting the above kind of structure allows for hiding all kinds of information about user of HTTP client, from destination Web server. The Web server can only access information about proxy. In addition, there is a possibility to secure connection between user agent and proxy by SSL/TLS protocol (*Secure Socket Layer / Transport Layer Security*) [4]. Thanks to this other parties such as ISP (*Internet Service Provider*), LAN (*Local Area Network*) administrator or just eavesdroppers cannot access the transferred information. Another advantage to the proxy is the ease of control and filtration of transferred content. It then allows for management of Cookies mechanism and blocking of unasked for, annoying and dangerous extras (popup windows, banner ads and etc), and also deleting client-side scripts or programs.

The solution of anonymizing service based on proxy is widely used. The advantages of proxies, responsible for their high popularity are: filling the technical loop-hole in WWW system related to lack of user's privacy, high efficiency of hiding user's identity data, high capability of hiding user id data, easy access to the service, supported by no additional requirements from users (only an Internet access and standard Web browser needed), simple architecture, insignificant delays of Web navigation, relatively low costs required for system realization, no demands to modify existing network nodes and protocols – system implementation based on current well known standards, simple user interface.

Therefore, this type of solution needs to be observed more closely. Let's analyze the way it works, concentrate on problems arising from its implementation and its security limitations.

Unfortunately, there are also serious disadvantages to Third Party Proxy Servers. Currently employed proxy servers have access to information about user's Web activity. Anonymity service providers induce the belief that this data is not collected, used or shared. The user has to face the risk of concentration of personal Web activity in one place. If the anonymity service provider attempts to exercise this possibility, the user will be exposed to an even greater risk than in traditional Web browsing, because information collected by different websites is much more difficult to put together. Furthermore, proxy servers do not protect against tracking by traffic analysis. An eavesdropper can observe the volume of transmitted data and correlate

inputs and outputs (proxy server request). A very serious risk also exists here, because it is possible for third parties to track and profile users. The next disadvantage of proxy servers is the limitation of sets of elements which can be downloaded. Some of HTML standard enhancements (like JavaScript) can create high risks for the whole system. It is possible and easy to perform powerful attacks using these technologies. The mentioned attacks can completely compromise proxy sever systems. Presented security gaps should provoke further conceptual and design studies.

3.3 New Protocol – P3P

The P3P – Platform for Privacy Preferences Protocol [3] is promoted to be a new standard. Its main task is privacy protection of people surfing in WWW system. P3P introduces uniform and machine-readable format for websites privacy policies and for user's private data collected by his Web browser. Thanks to this user can easily familiarize himself with visited websites privacy statements. User's browser is able to read this statement and automatically decide, for example, if to send users id information or if to allow for Cookies.

This protocol is a useful tool in electronic commerce both for merchant and client. P3P can help achieve harmony between companies' economical needs for information required to provide services and customers' rights to privacy and control over personal information. P3P technology can also decrease practices of profiling. However it is well known, that this mechanism cannot provide full privacy. It is not possible to hide all user agent id data – for example IP address.

The newest Web browsers are equipped in Platform for Privacy Preferences Protocol client. For now this is mostly experimental – there are few Web pages with P3P technology support.

3.4 Adaptation of Chaining with Encryption Technique

More or less adequate adaptation of Chaumian concept to WWW system was already implemented by few organizations. Examples of systems based on *MIXNET* include: *Onion Routing* [14], [8], *Crowds* [13], *Freedom* [7] (first profitable system of this kind). They are all based on a network of proxy nodes. Data packets are divided into uniform length frames (for example 128 bytes). Every frame is encrypted repeatedly and then send on its winding way.

To accomplish encryption and decryption of exchanged packets these systems use additional software executed on user's computer. This software takes over all communication with Internet and rerouting packets to service servers. Usually the systems also provide anonymous access to another services such as email, news, file sharing.

The popularity of this type of systems is limited. Systems based on chaining with encryption do not eliminate all risks of traffic analysis attacks. Anonymity is still dependent on a third party – information about user's Web activity was only dispersed between many proxies. However, there is no certainty that proxies do not collaborate with each other. What is more, anonymity service provider cannot submit a proof that

system proxies do not collaborate. Originators usually omit this fact and only mention curtly that each of proxies should belong to a different infrastructure provider. Is this a reason enough to support a belief that proxies do not collaborate? Nowadays we can see many examples of cooperation between independent companies in the process of tracking Web users and profiling them. Why would it be any different in this case? Implementation of this class of methods is not widely spread. We can only presume that in case of their popularization and implementation of proxies by many different companies, this possibility could be misused.

Systems based on a network of many proxies require construction of expensive infrastructure. This brings in a discouraging factor for potential investors. The use of network of proxies is a great technique for providing anonymity during e-mail correspondence. However, serious delays, which are not inconvenient in sending emails, are a serious obstacle in Web browsing. To increase system's performance speed, it is necessary to employ powerful and fast computers as proxies. Yet, they are expensive.

4 VAST System

Today more than ever the WWW service increases its multimedia character. The users expect texts and graphics to be available immediately. Today's solutions, thought, still require very high financial investments.

We have selected the following principles in our solution design: preservation of all advantages of popular anonymous third party proxy servers, providing versatile anonymity (including service provider and also preventing traffic analysis attack), speed, no additional requirements from users – general access, in particular, no applications installed on user's host, easy to implement outside laboratories – low costs.

4.1 VAST Concept

VAST – Versatile Anonymous System for Web Users [11] (see Figure 1) contains only one proxy node. To achieve anonymity from proxy server's point of view and also to disable traffic analysis attack we are introducing a technique of *dummy traffic* generation. More pages than actually requested by user are transferred from proxy to client. Information about which content is the object of interest to the user stays with him. An agent (JAVA applet) cooperates with user's browser. At the time when the user is familiarizing himself with the page content, the agent simulates

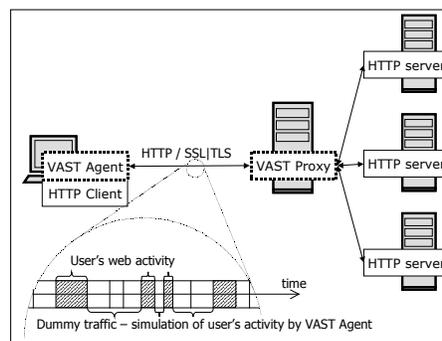


Fig. 1. VAST system scheme

users Web activity – by requesting random websites from proxy. The basis of this solution is the observation of a typical Web navigation. User does not request websites at all times. Requests get sent in various time intervals and in between them, the user reads the content. VAST utilizes this fact. Additionally, VAST takes advantage of the simplicity of generating *dummy traffic* in WWW system – we mean that a wide range of Internet resources is indexed in Web search engines. We can simulate user’s activity using it. The VAST system, unlike other existing solutions, does not conduct additional major activity while the transaction takes place, but it utilizes free time to take actions providing versatile anonymity. A source code of agent applet would be available to all. Then users would be able to check if it is not a *Trojan Horse*.

4.2 VAST Performance

Presented idea was designed to allow high performance – similar to efficiency of traditional Web browsing. All additional processing intended to provide anonymity takes place not at the time of requesting or downloading page, but at the time when user reads/watches downloaded pages. However, there is still a question: how *dummy traffic* delays browsing? Sometimes a user familiarizes him with page content fast. How long must he wait than? VAST can block all content which comes from the third side servers. This means excluding all multimedia banner ads placed frequently on many sites. Statistic analyses show that size of ads placed on popular websites and portals often occuppies 50% of total website size or even more. Requests to third side servers are replaced by *dummy traffic* requests in VAST system.

The volume of dummy traffic should be on a proper level to perform effective masking of user’s activity. The number of transactions performed in appropriate sessions should be approximately equal. Let τ_d denotes average time of downloading of single Webpage; τ_f – average time of familiarizing with page content; τ_w – average delay of Webpage downloading introduced by VAST system in comparison to traditional proxy server; n – number of dummy sessions. Then the delay τ_w can be described as follows:

$$\tau_w = 0.5 \tau_d \quad \text{for } \tau_f \geq n \tau_d \quad (1)$$

$$\tau_w = n \tau_d - \tau_f + 0.5 \tau_d \quad \text{for } \tau_f < n \tau_d \quad (2)$$

This can be written:

$$\tau_w = \frac{|n \tau_d - \tau_f| + (n+1) \tau_d - \tau_f}{2} \quad (3)$$

The following results can be obtained for a typical downloading time $\tau_d = 8$ [s]:

Table 1. Delays introduced by VAST (τ_w) system as a function of user’s familiarizing with Web content time

n	τ_f [s]	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	τ_w [s]	12	11	10	9	8	7	6	5	4	4	4	4	4	4	4	4	4	4	4	4	4	
2	τ_w [s]	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	4	4	4	4	4

The computations shows that acceptable delays (similar to delays present in traditional anonymous proxy server systems) occur when users spend time to familiarize with Webpage content equal to multiplication of a number of sessions and a single page download time.

5 Summary

Today's popular tools are not perfect. They do not provide full anonymity. New proposals are not able to reach a wider audience. Technology based on network of Chaum's nodes is successful in electronic mail. It does not mean that it is directly applicable to WWW, which has different requirements (most of all speed). Our proposition of solving this problem is the system dedicated to WWW. The concept is based on its characteristic and utilizes it. This comprehensive technique overcomes weaknesses of existing systems such as: serious, noticeable delays, access of service provider to user's private data and high costs of service implementation. The novel idea in this system – *dummy traffic* generation – may be in some cases viewed as its weakness. For users, whose fees for the Internet access are based on the amount of downloaded data, it means higher costs. We should stress that the system can block third party servers advertisement elements. It means that the graphic files from third parties are exchanged for *dummy traffic*.

References

1. Berners-Lee, T., Fielding, R., Frystyk, H.: Hypertext Transfer Protocol – HTTP/1.0. RFC 1945 (1996)
2. Chaum, D.: Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. Communications of the ACM Vol. 24 no 2 (1981) 84 - 88
3. Cranor, L., Langheinrich, M., Marchiori, M., Presler-Marchall, M.: The Platform for Preferences 1.0 (P3P 1.0) Specification, W3C Recommendation (2002)
4. Dierks T., Allen C.: The TLS-Protocol Version 1.0. RFC 2246 (1999)
5. Ferneyhough, C.: Online Security and Privacy Concerns on the Increase in Canada, Ipsos-Reid (2001)
6. Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee T.: HyperText Transfer Protocol – HTTP/1.1. RFC 2616 (1999)
7. Goldberg, I., Shostack, A.: Freedom Network 1.0 Architecture and Protocols. Zero-Knowledge Systems. White Paper (1999)
8. Goldschlag, D. M., Reed, M. G., Syverson, P. F.: Onion Routing for Anonymous and Private Internet Connections. Communications of the ACM Vol. 42 no 2 (1999) 39-41
9. Krane, D., Light, L., Gravitch D.: Privacy On and Off the Internet: What Consumers Want. Harris Interactive (2002)
10. Kristol, R., Montulli, L.: HTTP State Management Mechanism. RFC 2965 (2000)
11. Margasiński, I., Szczypiorski, K.: VAST: Versatile Anonymous System for Web Users. The Tenth International Multi-Conference on Advanced Computer Systems ACS'2003 (2003)
12. Martin, D., Schulman, A.: Deanonymizing Users of the SafeWeb Anonymizing Service. Privacy Foundation, Boston University (2002)
13. Presler-Marshall, M.: The Platform for Privacy Preferences 1.0 Deployment Guide, W3C Note (2001)
14. Reiter, M.K., Rubin, A.D.: Crowds: Anonymity for Web Transactions. ACM Transactions on Information and System Security (1998) 66-92
15. Syverson, P. F., Goldschlag, D. M., Reed, M. G.: Anonymous Connections and Onion Routing. IEEE Symposium on Security and Privacy (1998)