BILATERAL ANONYMITY AND PREVENTION OF ABUSING LOGGED WEB ADDRESSES

Thomas Demuth
Department of Communication Systems
Universität Hagen
Germany
thomas.demuth@fernuni-hagen.de

Andreas Rieke
ISL Internet Sicherheitslösungen GmbH
Germany
andreas.rieke@isl-online.de

ABSTRACT

A lot of effort has been taken in recent years to hide the content of a message from eavesdroppers. However, often not only the content, but also the address and identity of sender and/or receiver of the message are of interest for attackers. For that reason, several approaches were developed to guarantee anonymity in the case of email.

But nowadays the World Wide Web (WWW) is an established service like email some years ago and it is used by people all over the world. Most of them do not recognise the fact that they reveal plenty of information about themselves or their affiliation and computer equipment to the providers of the web pages they connect to. As a result, a lot of services offer users to access web pages unrecognised or without the risk of being backtracked, respectively. This kind of anonymity is called user or "client anonymity".

But on the other hand, there are only a few offers that provide an equivalent protection for content providers, although this feature is desirable for many situations in which the identity of a publisher or content provider is to be hidden. This property is called server anonymity.

The term "server anonymity" will be explained in detail with the help of an existing system fulfilling some hundreds of thousand user requests per day. We will also describe our experiences in providing such a system with respect to misuse.

Furthermore there is another sensitive fact: While browsing web pages, the used URLs are logged both by the web client (web browser) which is used and the Internet service provider (ISP), or any other instance or organisation that is involved in the communication. Hence the ISP can investigate the content a user is interested in afterwards simply by reusing the logged URLs. The same problem results from the behaviour of regular web browsers to build an address history and local copies (browser cache) of the visited web pages.

We will demonstrate a way of preventing the reuse of logged web addresses by introducing the concept of temporarily valid web addresses.

INTRODUCTION

In recent years the usage of the WWW increased proportionally concerning its meaning and the number of users. It developed from a service focused on academic areas offering scientific content into a medium for providers of information of very different kinds and sometimes doubtful seriousness.

So the demand for anonymity in the WWW is immense; based on the fact that a user in the WWW gives away plenty of personal information while navigating through the WWW, it is possible to build a very personal profile of this user. The accumulation and connection of this information contradicts the usual idea of data security, violates personal rights, and offers many illegal possibilities like insertion into unwanted address lists which often results in undesired advertisement or spying out personal tendencies.

Because of this situation, services have already been built offering web users to stay anonymous. The development of appropriate mechanisms to protect anonymity is part of the research work in cryptology.

Besides the protection of users, the protection of the service provider becomes neccessary. The reasons for establishing this protection are manifold, some example situations are given below:

- Some years ago the owners of bookstores who offered the "Satanic Verses" by Salman Rushie were threatened by islamic fundamentalists, while it was not illegal to offer this book. Since then many bookstores have gone online.

- In the Yugoslav civil war the Internet was one of the few possibilities for civil right groups to communicate with the world. They were able to report about the injustice of their government by email and newsgroups but not by WWW because of the danger of backtracking. There and in other totalitarian countries it would be possible to realise the anonymity of speech.

In general the anonymity of the sender, here the server, is of interest in case the relation shall be hidden for the client, or for a third party.

MIXES

In his article from 1981, David L. Chaum described among other things a general purpose mail system that conceals the relation between senders and receivers of messages in a system from a global aggressor [Cha81]. Messages are transferred through several intermediate stations called mixes. Each mix is able to hide the content of a message by securing it with a virtual envelope consisting of public key encryption (often RSA [RSA78], but this kind of encryption is slowing down the whole processing). Furthermore, a mix can delay the delivery, change the sequence of arriving messages, or change their length to conceal the origin. In addition a mix is able to generate and transfer dummy messages in case too few messages are available. This guarantees that messages leave the mix at a fixed time.

Further, Chaum describes a technique to hide the identity of the receiver of a message. Therefore the address is structured in a way that the receiver cannot be recognised by inspecting the address. Chaum uses the term "untraceable return addresses"; but these addresses can also be designated as "anonymous return addresses" within the meaning of unlinkability of such an untraceable address and the clear text address of the belonging receiver. Mixes on a route through a mix compound are able to analyze an untraceable address step by step.

The system described in this paper uses untraceable return addresses to make references in WWW documents unrecognisable and to make the abusing of logged URLs difficult. In the following section, we will deal with the aspect of server anonymity in a more detailed form.

SERVER ANONYMITY

Existing projects and implementations investigate anonymity in various degrees, but only for the client ([LPWA], [CROWDS], [RR97], [GRS97], [GGMM97]). We will describe the nature of the problem from the view of an information provider and his anonymity (server anonymity). There is both the desire and the demand for participants in the WWW to publish information without revealing the server's address.

The WWW presents itself as a giant hypertext document consisting of separate globally distributed documents. Most of these documents are composed of textually and/or graphically oriented content and address information; from now on we will assume this as a standard case. But this assumption and our research results can be adapted to other document structures used in the WWW.

The URL (Uniform Resource Locator) is the central and connecting element in documents. The URL makes it possible to identify and localise a document in a unique way. However, this URL also reveals information about the origin of a document, because it is often built according to the following syntax:

[scheme]://[server].[domain]/[path]/[document]

Each of these components reveals more or less information about the author of a document; usually the domain section represents the most sensitive element. Even if there is no clue from the domain to the geographical locality or the membership of an institution or organisation of the server, there are many more mechanisms to extract detailed facts from this restricted information. Examples are global or national institutions like the NIC whose tasks are to administrate the relation between domain name and the organisation to which it belongs.

The HTTP (HyperText Transfer Protocol) describes the syntax of an URL, the mechanism of communication between web server and browser, but also the structure of messages exchanged between them. The specification of HTTP version 1.1, which is supported by our system can be found online in the WWW [HTTP].

The communication between web browser and web server works in a bilateral way. The browser initiates a request consisting of two parts:

1. Header: The header contains meta information and data fields specifying the address to be contacted. It may contain the client's email address, the type of browser the client uses, and other information.

2. Body: The actual content (e. g. parameters for forms) is transported in this part.

After receiving a request the web server reacts with a response which is constructed similarly, but contains information different to the request; now, the content may be either the requested data or an error message of the server. Therefore, one has to remember that there is data in the header giving information about the server. The body of the message, representing web pages or other web objects, contains similar information to that in the header. It often includes web pages built with the language HTML (HyperText Markup Language [HTML]. HTML is the most often used language to construct hypertext documents in the WWW. HTML consists of instructions (tags) and the actual information. Some of these tags contain references to other objects in the WWW. Because of the references' nature of being address information they have to be treated specially by the described system.

The system "Rewebber" [REWEBBER] described in the following represents a network of stations working with mix methods. It is the successor of a research project at the FernUniversität Hagen, Germany [DERI99]. The partners of a communication stay anonymous. That means that the identity of the content provider stays hidden to the user in the situation described above. The transport of the messages takes a route over one or more instances to increase security.

METHOD OF WORKING

Because the communication is not restricted to one instance of a "Rewebber"; it is possible to use cascaded "Rewebber" systems to improve security and immunize the system against attacks in the form of eavesdropping of incoming and outgoing messages.

An approved method for the encryption of data, in this case of web addresses, are public key algorithms. To encrypt an URL, the provider of a web page uses the actual function of "Rewebber", a special web page or he encrypts the URL with the public key, published by the operators of the "Rewebber" server.

With a symmetric algorithm it would not be possible to publish the secret (and only) key, because everybody using our system would be able to decrypt URLs encrypted by another user.

The URL `http://www.milcom2000.org/`, anonymised by "Rewebber" becomes:

```
http://www.rewebber.com/surf_encrypted/MTBul
n5EFN8u$EFEnK68VF898GjCmhmsjYvIwvAndRGHFBF$K
pVEVXUXzHC5ezz0hAmzSSEH2gRh4N6Iy4ifTXxs4lmbW
K94ERRUUuauT8C6RyF+sN8KyQUR0BT1Vv9UX5s=
```

These encrypted URLs are submitted to users who want to use them to contact the web page via "Rewebber". When this encrypted URL is sent from a web browser to "Rewebber" it decrypts it with the secret key. After that "Rewebber" acts like a web browser itself and requests the web page the URL is pointing to. The received page is analysed for URLs to be encrypted, header information is anonymised or filtered, respectively; the resulting page is sent to the client.

We have to pay attention to the fact that there is not only content in a request or response. Administrative data is transmitted in the header of the message, too, and this data has to be treated in the same way to achieve anonymity. Our method is a kind of partial realisation of Chaum's concept. But our encryption relates to possible revealing references instead of the whole content.

This means, that a content provider who wants to be anonymous, has to publish the URL of a web page in encrypted form. He is responsible himself not to reveal his identity by compromising content.

MISUSE

A service that is publicly available making it possible for a content provider to stay anonymous may cause misuse. "Rewebber" is not available for providers of web pages containing information that offends against national or international law or violates common moral standards. Those providers will be inserted in a restriction list. Entries in this list have the effect that a user will get a page with a notification from "Rewebber" instead of the real addressed web page, if he wants to access this page. The URL of offending web pages will be added to the list if the authors notice a case of "Rewebber"'s misuse stated above. In addition, the decrypted URL of this provider can be notified to prosecuting authorities on demand after the proof of misuse.

Very few cases of misuse have been reported to the providers of "Rewebber". Most of them concerned harassment in chat groups and in one case a globally agitating pseudo church requested to block the access to documents illegally published via anonymous URLs and therefore violating the copyright.

LOGGED WEB ADDRESSES

While browsing web pages, a surfer is implicitly presenting the addresses of the visited pages to various instances. This fact and its consequence is mostly not known to web users.

- Web browsers
  Common web browsers are logging used web addresses in a so called "URL history", offered to the user to simplify the access to the most recently visited web pages. This history log can be used by everyone with access to the user's computer. Especially in office environments computers are accessible if the owner is not present. Colleagues and, even more, a superior may be interested in this sensitive information.
  Another scenario is the visit of an internet cafe. There the computers are used by many web surfers that are not known to each other in general. Therefore the URLs of visted web pages can be seen by succeeding computer users.

- Internet Service Providers
  Most of the ISPs are logging the URLs their clients have used by default. Even if an ISP does not check the content of a requested web page simultaneously, it can request it at any time afterwards.
  This ability is given to every instance located between the user (his web browser) and the web server.
  A solution would be end to end encryption of the communication via SSL (Secure Socket Layer). But the use of SSL is disabled by some ISPs.

These facts are more disturbing if one recognises that an URL can carry additional information. Password, context, or personal information is coded in the URL to realise online shops, databases, etc. with the originally stateless HTTP protocol.
Furthermore it is possible that copies of requested pages are available at all instances between the web server and the client. ISPs, but also other network nodes, are caching web pages to be able to deliver often requested pages in a shorter time. As a result, the ISP is able to see user requested web pages afterwards even if it does not log web addresses.

PREVENTION OF ABUSE

The requirements to an URL that cannot be used by another instance than the original user are:

- The user wants to use the URL, maybe several times, in a predefined time interval.

- After that time no one shall be able to use the URL to access the corresponding web page. One has to be aware of the method by which ISPs are evaluating their log files. Normally this is done once per day or even less often. But even if an ISP is able to check the web pages "simultaneously" there is a minimal delay because of the huge number of web addresses clients of an ISP are using.

The solution is to restrict the time of validity. Thus, an additional field is introduced to the anonymised URL; this field works as a time stamp. While building an anonymous and time restricted URL via the corresponding "Rewebber" web pages, the user can determine the (relative) time the URL will be valid. This time should be as short as possible to offer prevention against abuse. Therefore the user can request a web page via this URL and can even reload it, if needed. But an ISP (or any other person utilising the user's computer some time after) will receive an error message from "Rewebber", signalling that the used URL is not longer valid.
The time stamp consists of two elements: One absolute and one relative date. With the combination of these, "Rewebber" is able to determine if a requested time valid URL is allowed to be accessed. As explained above, the corresponding web page normally contains further references which have to be converted to time valid URLs before the page can be delivered to the user. The time stamps for these URLs are based again on the known relative time, but now the absolute element is replaced with the current time. As a result, the URLs in each successively requested web page are valid for the time the user has determined originally.

To disable the creation of copies by caching web pages, a feature of the HyperText Transfer Protocol can be used. The instruction "*pragma : no-cache*" is a command to all transporting instances not to cache the corresponding page. In particular the web browser of the user does not cache the page into a file

on the computer's hard disk.
Our experiments have shown, that the most well known web browsers "Netscape Navigator" and "Internet Explorer" obey the pragma instruction.
A mentionable fact is the handling of the newer instructions "*Cache-Control : no-cache*" and "*Cache-Control : no-store*". These have been introduced into the current standard HTTP/1.1 and are ignored by browsers like Netscape and Opera.
Thus, "Rewebber" modifies requested web pages and inserts the HTTP/1.0 instruction *pragma : no-cache* into the header fields of the page delivered to the user to disable caching of web pages.

The presented mechanisms offer bilateral anonymity and untraceability with respect to web server and the user of web pages and prevent the abuse of logged web addresses by intermediate instances (i.e. ISPs).

We wish to thank the department of Communication Systems at the Universität Hagen under supervision of Prof. Dr.-Ing. Firoz Kaderali for his support in our research.

REFERENCES

Cha81 D. L. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms", Communications of the ACM 24, pp. 84–88, February 1981.

CROWDS http://www.research.att.com/ projects/crowds

DERI99 T. Demuth, A. Rieke, "On Securing the Anonymity of Content Providers in the World Wide Web", Proceedings of SPIE '99, Vol. 3657, San Jose, CA, USA, January 1999, pp. 494-502

GGMM97 E. Gabber, P. B. Gibbons, Y. Matias, and A. Mayer, "How to make personalized web browsing simple, secure, and anonymous", Bell Labs Technical Memorandum, Bell Laboratories, Lucent Technologies, Murray Hill, NJ, USA, May 1997.

GRS97 D. M. Goldschlag, M. G. Reed, and P. F. Syverson, "Privacy on the internet", tech. rep., http://www.itd.nrl.navy.mil/ITD/5540/ projects/onion-routing/inet97/ index.htm, 1997.

HTML http://www.w3.org/markup/wilbur/

HTTP http://www.ietf.org/rfc/rfc1945.txt

LPWA http://www.bell-labs.com/project/lpwa

REWEBBER http://www.rewebber.com/

RR97 M. K. Reiter and A. D. Rubin, "Crowds: Anonymity for web transactions", DIMACS Technical Report 97-15, AT&T Labs, Murray Hill, NJ, USA, August 1997.

RSA78 R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems", Communications of the ACM 21, pp. 120–126, February 1978.