Effectivity of Various Data Retention Schemes for Single-Hop Proxy Servers

Dominik Herrmann
University of Regensburg, Germany

Rolf Wendolsky
JonDos GmbH, Regensburg, Germany

Abstract

Recently, member states of the European Union have legislated new data retention policies. Anonymisation services and proxy servers undermine such data retention efforts, as they allow users to masquerade their IP addresses. Providers of such services have to implement effective data retention mechanisms allowing for traceability while at the same time preserving users' privacy as far as possible. In this paper we analyse the effectivity of four data retention schemes for single-hop proxy servers which use information already stored in logs today. We assess their effectivity by applying them to the historic logs of a mid-range proxy server. According to our evaluation it is insufficient to record data on session-level. Users can only be unambiguously identified with high probability if access time and source address of each request are stored together with the destination address. This result indicates that effective data retention based on currently available identifiers comes at a high cost for users' privacy.

1 Introduction

In 2006, the European Union issued the Data Retention Directive [4]. For the purpose of law enforcement, Internet Service Providers (ISPs) may thereby be required to uncover the identity of a user, given an IP address and a timestamp. The Directive has to be implemented by member states until March 15th, 2009. The German implementation for Internet access has gone into force on January 1st, 2009. Since then, providers of telecommunication services have to retain transformation data for a period of six months.

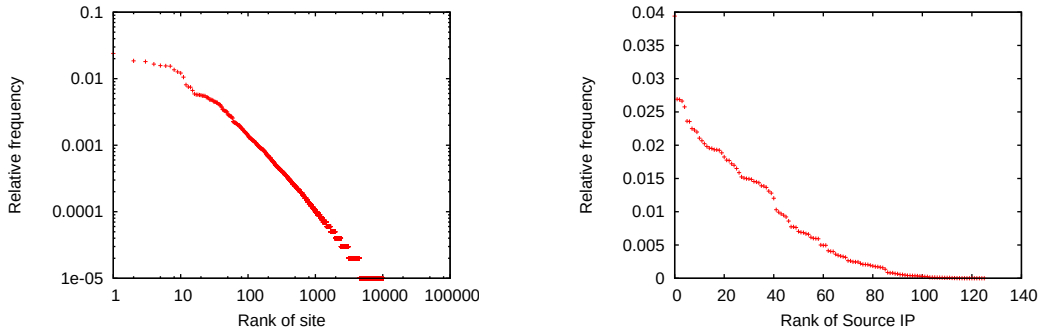
While the implementation of data retention measures is rather straightforward for ISPs, interesting questions arise when data retention is applied to proxy servers and anonymisation services. There is still considerable uncertainty about which types of

services are affected by the new data retention regulations. Although those legal discussions are of high practical relevance, they often neglect the implications regarding the involved technologies and the impact on users' privacy. The factual evidence presented in this paper is intended to foster a more technology-aware discussion.

In the context of anonymisation services and proxies, data retention measures have to allow for *traceability*, i. e., uncovering the IP address of a user from whom a suspicious connection originated (cf. Richard Clayton's PhD thesis for more information on that topic [3]). While some people consider traceability of Internet users fundamentally necessary to enable crime detection and prevention, it is criticised by others for unduly infringing users' privacy. Moreover, ISPs complain that implementing and operating a data retention infrastructure is a costly undertaking. Law enforcement agencies (LEAs) or related governmental organisations have not specified technical requirements regarding data retention on proxy servers and anonymisation services so far. Devising effective data retention mechanisms allowing for traceability while at the same time preserving users' privacy is the challenge at hand.

In this context Kesdogan et al. [7] have researched the effectivity of various intersection attacks from the literature using the log files of a proxy server. Berthold et al. [1] have evaluated the effectivity of intersection attacks on the AN.ON/JonDonym anonymisation service, i. e., whether the provider of the anonymisation service can unambiguously reconstruct the source IP address of an offender, given a number of events when the designated offender was using the service. The authors find that the size of the anonymity group decreases rapidly with an increasing number of events available for building the intersection. According to their results another means to improve traceability is increasing the accuracy of the timestamps used by LEAs. Intersection attacks have one drawback, though: they rely on the fact that LEAs are able to identify multiple requests from the same offender, all of them coming from the source IP address. Köpsell et al. [8] propose a request-level data retention scheme specifically designed for distributed anonymisation services. It is based on threshold group signatures to allow for the revocation of the anonymity of offending users while preserving the privacy of all other users. Köpsell et al. do not define which kind of information is stored to identify offending users, though. The schemes in this paper are possible realisations for their proposal.

The debate regarding to what extent providers of proxy servers and anonymisation services will have to implement data retention has not settled yet. In this paper we will analyse various conceivable retention schemes which only utilise data already available today to the providers of such services. The evaluated schemes do not rely on intersection attacks and could be implemented easily. Based on an empircial study using the log files of a medium-range proxy server we find that data retention schemes utilising currently available data is only effective if information about the requested destination addresses is stored, which is not satisfactory from a user's perspective. Therefore, our paper motivates further research in this field in order to find better data retention schemes which address the security and privacy requirements of all involved parties.



(a) Popularity of web sites

(b) Activity of source IP addresses

Figure 1: Distribution of request numbers for the evaluated sample

2 Evaluation Methodology

In the interest of conciseness we limit our analysis to HTTP traffic which is relayed by single-hop web proxies. In order to get comparable results we implemented various data retention schemes and applied all of them to a common log file of a proxy server. As providers of anonymisation services refrain from keeping log files containing information to the necessary extent, for this preliminary study we used Squid log files of a local school with about 1,000 students and about 100 staff members. The log files contained the pseudonymised requests of six months (August 2008 to February 2009).

The combined log file contains 9,074,962 requests in total originating from 126 distinct (local) source IP addresses. The users requested objects from 33,258 destination IP addresses which have been accessed via 51,746 different host names. The plot in Figure 1a shows the relative access frequencies of the host names ordered by their popularity (based on the number of total requests per host name, most active first), which indicates that in our sample the retrieved web sites follow a Zipf-like or power law distribution [10]. This feature has been observed in several earlier studies for web requests from a homogenous community of users (cf. [2, 5]). According to the histogram in Figure 1b the user group consists of both, power users and less active ones.

These characteristics have to be kept in mind when interpreting the results of our study, i. e., they only apply to systems which serve a rather small and homogenuous user group and probably cannot be easily generalised to large-scale anonymisation services. The absolute values of the results are certainly affected by the specific composition of our user group and its behaviour in a school setting.[1] Nevertheless, we believe our proposed methodology may be used to assess the effectivity of data retention schemes on such systems.

For the evaluation we created stripped-down versions of the Squid log file containing only the information which would be available for the examined data retention schemes. We then analysed the effectivity as expressed by the ratio of requests which could have been unambiguously attributed to the correct source IP address for the var-

[1] Some pages containing unsuitable content for students are filtered at the proxy level. This may add to the bias in our sample.

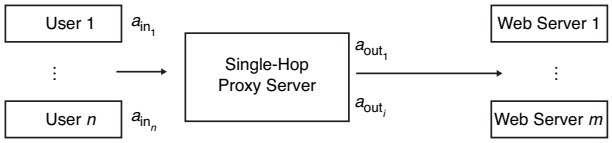


Figure 2: Model of the evaluated single-hop proxy scenario

ious schemes. The ratio was calculated by complete enumeration, i. e., we created LEA queries for each request contained in the log file, every time recording the number of potentially matching requests. For maximum effectivity the result set would have to contain only one request for each query.

3 Data Retention Schemes for Single-Hop Systems

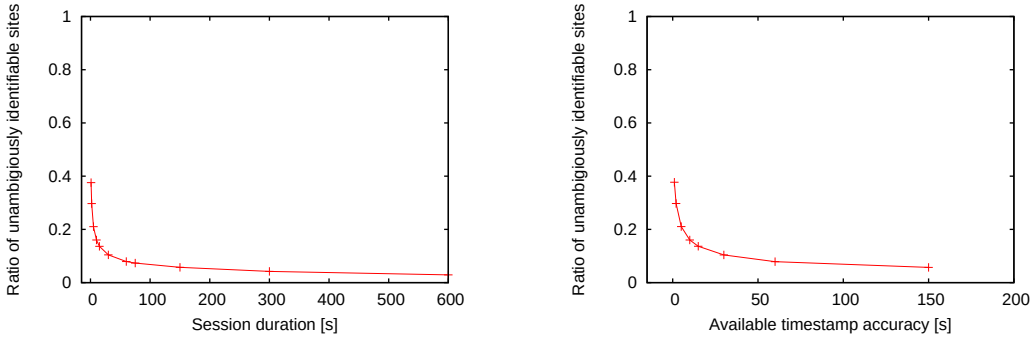
Figure 2 illustrates the single-hop setup. The proxy is used by n users with IP addresses $a_{\text{in}_i} \in A_{\text{in}}$. From the viewpoint of the destination server, the request originates from an IP address $a_{\text{out}_j} \in A_{\text{out}}$. Note that $|A_{\text{in}}| > |A_{\text{out}}|$ in most cases, i. e., the number of unique input addresses exceeds the number of IP addresses of the proxy. For our proxy $|A_{\text{out}}| = 1$. We will present four different retention schemes in the following sections.

3.1 Recording Input Addresses on Session-Level

Session-oriented services like VPN-based anonymisation services could record the relevant session-level information. If t_{start} and t_{end} denote begin and end timestamps of a user's session with the anonymisation service, the provider would store the tuple $(t_{\text{start}}, t_{\text{end}}, a_{\text{in}}, a_{\text{out}})$ for each session. Note that individual HTTP requests, which are relayed during a session, are not considered. From a privacy point of view this solution is the most desirable form of data retention. Only a bare minimum of information is recorded. Personal information – apart from the usage time – is not stored.

Traceability cannot be guaranteed at all times with this approach. Faced with a LEA query $q = (t^{(q)}, a^{(q)}_{\text{out}}, a^{(q)}_{\text{dest}})$ for some timestamp $t^{(q)}$, one of the proxy's output addresses $a^{(q)}_{\text{out}} \in A_{\text{out}}$ and the destination address $a^{(q)}_{\text{dest}}$, e. g., q=(2008-10-10 9:43am GMT, 132.199.2.111, 66.249.93.104), the service provider may not be able to uniquely identify one of its users as requested. He can only provide all source IP addresses a_{in_i} of all sessions established at $t^{(q)}$ and relayed over a_{out}. Note that the destination address $a^{(q)}_{\text{dest}}$ does not help to reduce the anonymity group because the service provider is not storing any destination addresses in this scheme.

With this scheme even inactive users contribute to the anonymity group. Intuitively, tracing a request back to its originator is only possible if there is only a single session at $t^{(q)}$, which is very unlikely for popular proxies. If multiple requests from different sessions could be attributed to the same user, LEAs could intersect the result sets to decrease the size of the anonymity group.



(a) Varying session durations

(b) Varying timestamp accuracy

Figure 3: Data retention effectivity for session- and request-based services

Obviously, traceability largely depends on the duration of the individual user sessions. We analysed the influence of the session length on the effectivity by grouping consecutive requests from an individual IP within the simulated session duration into one contiguous session. As shown in Figure 3a the effectivity of this scheme is dropping extremely fast with increasing session durations. Even for rather short sessions of only 300 seconds, less than 5 % of requests can be identified unambiguously. For busier proxies with thousands of users this figure is expected to approach zero.

Using a regression analysis we found that the plotted data closely fits a power function ($y \approx 0.3921x^{-0.3932}$ with a residual sum of squares $\mathrm{rss} \approx 3.018 \cdot 10^{-4}$).

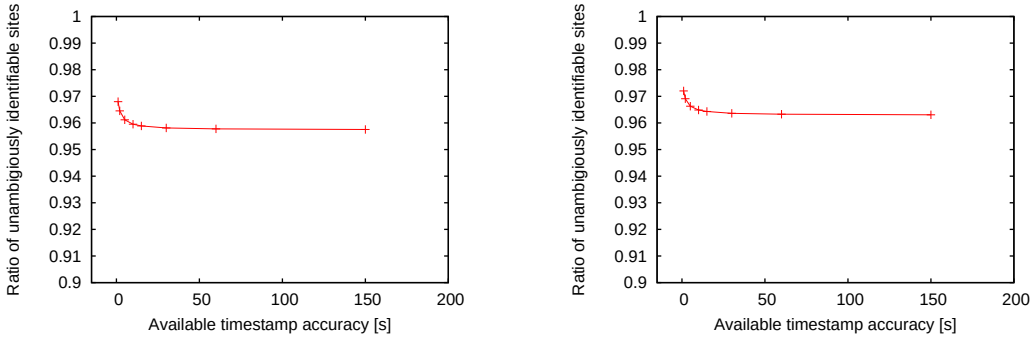
3.2 Recording Input Addresses on Request-Level

Common web proxy servers, e. g., the Squid cache proxy or many form-based CGI proxies, operate on individual HTTP requests. They could store the tuple $(t_{\mathrm{transform}}, a_{\mathrm{in}}, a_{\mathrm{out}})$, where $t_{\mathrm{transform}}$ is the point in time when the input address was transformed into the output address.[2] Anonymity groups become considerably smaller as inactive users are not included in the result set any more. Traceability cannot be guaranteed when multiple users issue requests at the same time, though.

Figure 3b depicts the effectivity of this scheme. Although the plot looks similar to the session-based case, request-based data retention is more effective: the effectivity depends only on the accuracy of the timestamps used in the log files and the LEA query. The accuracy will be degraded if the clocks of the service provider and the destination site are not synchronized or if non-deterministic network latencies cause unforeseen delays.

In comparison to the session-based data retention scheme, logging data on the request level offers potentially higher effectivity because of a much more precise time resolution. Given a hypothetical timestamp accuracy of 60 seconds, all requests within a time window of 30 seconds around the point in time specified in the LEA query are part of

[2]Of course, this scheme is not limited to services operating on a request level, i. e., session-based services like VPNs could store request-level data, too.



(a) Recording IP addresses

(b) Recording host names

Figure 4: Impact of storing destination addresses on data retention effectivity

the result set. For a hypothetical timestamp accuracy of 60 seconds about 7.9 % of requests can be unambiguously identified in our sample. This ratio climbs up to 39 % if timestamp accuracy would be increased to one second. Again, we expect these figures to decrease tremendously on busy proxies.

We presume that realistic values of the timestamp accuracy for Internet hosts lie in the range between one and 60 seconds. To the best of our knowledge the available timestamp accuracy has not been analysed so far. Further research is necessary.

3.3 Recording Destination IP Addresses

In the previous section we have excluded destination addresses from data retention. For increased traceability, anonymisation services might be forced to store the IP addresses of the destination servers. In this case they would store $(t_{\text{transform}}, a_{\text{in}}, a_{\text{out}}, a_{\text{dest}})$ for each request. This approach reduces the size of the anonymity group considerably. Now, only IP addresses of users requesting an object from $a_{\text{dest}}^{(q)}$ at time $t^{(q)}$ are included. So, again, the effectivity of this scheme depends on the available timestamp accuracy (cf. Figure 4a) . Given a timestamp accuracy of 60 seconds 95.8 % of the requests in our sample can be unambiguously attributed to a single user with this scheme (96.8 % given an accuracy of one second). Effectivity is still not perfect, though, as there is still a (relatively small) possibility that several users are accessing different objects on the same destination server within the requested time window, which may happen for example when various web sites are (virtually) hosted on the same physical server.

From a privacy viewpoint storing destination IP addresses is not desirable, though, as they may reveal information about the interests of users to the service provider for the whole retention time span.

3.4 Recording Destination Host Names

The last scheme we present in this paper is based on the previous one. Instead of recording destination IP addresses, DNS host names are stored in order to further re-

duce the size of the result set. The result set will then only contain source IP addresses of users who have accessed the same (virtual) host at a given point in time, thus allowing for an exact match in most cases.

As expected our results show only small increases in effectivity when host names are stored (cf. Figure 4b). Given the timestamp accuracy of 60 seconds, for 96.3 % of the requests the originator can be identified. Apparently, the set of simultaneously retrieved pages which are co-located on the same host is rather small in our sample. Note that effectivity could still be improved slightly if – instead of host names – the complete URLs including HTTP query parameters would be stored. Even then, traceability could not be guaranteed for encrpyted web sites (HTTPS), though, because the proxy could only log host name and port of them. And of course multiple users might still coincidentally request the same URL. As the expected benefits of this scheme are rather low for our sample, we have not implemented it so far.

The effectivity of this approach comes at a high cost. While host names may disclose the personal interests and habits of users, URLs may even contain personal or sensitive information (e. g., search engine queries, session IDs, and unencrypted credentials). Storing information of this kind on a proxy server over a period of six months causes considerable privacy and security issues and therefore seems disproportionate.

4 Conclusion

This paper examined four data retention schemes in terms of their effectivity. The presented schemes only rely on data easily available to providers of proxy and anonymisation services, i. e., they are straghtforward to implement based on already existing logging facilities. Effective data retention schemes have to offer traceability of – ideally – all requests which are handled by such services to law enforcement agencies.

According to our empirical study, none of the examined schemes can guarantee traceability for all requests. Namely, we found that storing *session-level data* is not sufficient because the anonymity groups become too large even on our little-frequented proxy for typical session lengths. Logging on a *request-level basis* seems more promising, but only if the *destination address* of each request is recorded – which infringes users' privacy. None of the evaluated data retention schemes provides effective traceability while respecting users' privacy. Although we have utilised a synthetic sample, we believe that our methodology is of general value and it could be applied to many kinds of anonymisation services, e. g., CGI-based proxies using HTML forms or VPN solutions (as provided by anonymizer.com), mix cascades (provided by JonDonym [6]) and Onion Routing (cf. the Tor project [9]).

In future work we plan to repeat the evaluation with log files from a proxy server with a higher load and a more diverse user base or even a real-world anonymisation service. This will allow us to rule out any bias caused by the data source chosen for this preliminary study. Within this future story we will also be able to examine the efficacy of intersection attacks, i. e., under which circumstances they reduce the size of the anonymity groups over time.

Furthermore, we plan to evaluate what timestamp accuracy can be achieved in a practical environment in order to quantify the actual size of the anonymity groups for the various schemes. Another promising field for future research activities is the design of more advanced data retention techniques, e. g., by introducing dedicated retention identifiers which preserve the privacy of users, while at the same time offering improved traceability.

References

[1] Stefan Berthold, Rainer Böhme, Stefan Köpsell. Data Retention and Anonymity Services. In: Proceedings of IFIP/FIDIS Summer School 2008, Brno, Czech Republic, http://www.buslab.org/SummerSchool2008/slides/Stefan_Koepsell.pdf.

[2] Lee Breslau, Pei Cao, Li Fan, Graham Phillips, and Scott Shenker. Web Caching and Zipf-like Distributions: Evidence and Implications. In: INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings, New York, USA, 1999.

[3] Richard Clayton. Anonymity and traceability in cyberspace. Technical Report, based on dissertation submitted August 2005. University of Cambridge, 2005.

[4] European Parliament & Council. Directive on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks (Directive 2006/24/EC), March 15, 2006.

[5] Steven Glassman. A Caching Relay for the World Wide Web. In First International Conference on the World-Wide Web, CERN, Geneva, Switzerland, May 1994.

[6] JonDonym. http://www.jondos.de/

[7] Dogan Kesdogan, Lexi Pimenidis, and Tobias Köllsch. Intersection Attacks on Web-Mixes: Bringing the Theory into Praxis. In: Proceedings of First Workshop on Quality of Protection, Milan, Italy, 2005, http://www.freehaven.net/anonbib/cache/KesdoganPK06.pdf.

[8] Stefan Köpsell, Rolf Wendolsky, and Hannes Federrath. Revocable Anonymity. In: Emerging Trends in Information and Communication Security. Lecture Notes in Computer Science, 3995. Springer, Berlin, pp. 206-220, 2006.

[9] The Tor Project. http://www.torproject.org/

[10] George Kingsley Zipf. Relative frequency as a determinant of phonetic change. Reprinted from the Harvard Studies in Claasical Philology, Volume XL, 1929.