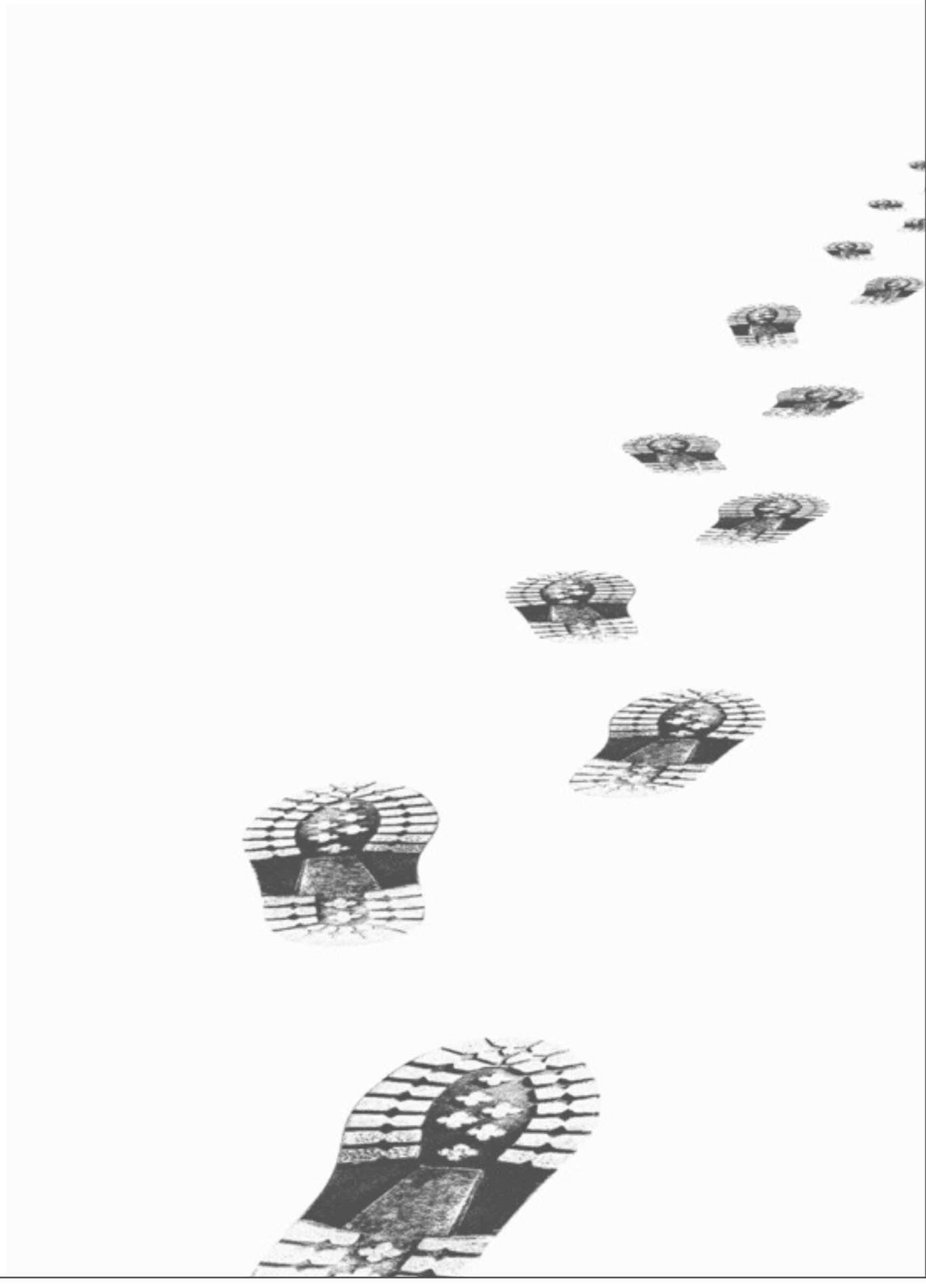


Donnerstag, 26. März 2009

Trace Me If You Can

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

2009-03-24

Dominik Herrmann
University of Regensburg

Rolf Wendolsky
JonDos GmbH

Donnerstag, 26. März 2009

joint work; Preliminary study in order to bringconcrete numbers into discussion on data retention

DATA RETENTION DIRECTIVE

2006

Donnerstag, 26. März 2009

EU issued Data Retention Directive. Member states had to implement it within 18 months, Germany and Austria postponed implementation until 2009.

Donnerstag, 26. März 2009

DR is all about Traceability. concentrate on Web traffic. traceability means linking offending HTTP requests to originating user via his IP. Traceability is easy for direct requests as ISPs are now required by law to store the IP-user mapping for at least 6 months. LEAs take source IP of observed packet and request contact info from the ISP.

Proxy Servers and Anonymisers

Src IP #1 | Src IP #2

Src IP #1 | Src IP #3

Proxy's IP | Proxy's IP

Proxy's IP | Proxy's IP

make traceability difficult

(some) are subject to
data retention obligations

It is much more difficult when the user uses a proxy server. like a funnel, proxies substitute the origin source IP of all requests with their own; are required by law to support backtracking users

How to do Data Retention on Proxies?

The law does not tell us!

There are lots of ideas and we can borrow from research on anonymisation services, key words are log file pseudonymisation, intersection attacks, and so on.

Long-Term Research Question

find a data retention scheme for proxy servers

which honours privacy of users
+
allows for optimum traceability of offenders

Goal of Preliminary Empirical Study

assess effectivity of 4 data retention schemes

(which utilise data already available today to proxy providers)

no new technology required
(i.e., cheaply+easily implemented)

intersection attacks neglected
(for now)

Donnerstag, 26. März 2009

wanted to find out whether information already available to providers of proxies or anonymisers is sufficient to do effective data retention.

Simplistic Effectivity Metric

$$\frac{\text{\# of successfully traceable requests}}{\text{\# of all requests}}$$

ratio of requests which
could have been attributed to their
true source IP unambiguously

For this preliminary study we concentrated on one simplistic metric to measure the effectivity of a data retention scheme. Curious to learn about your ideas regarding metrics.

Characteristics of Sample

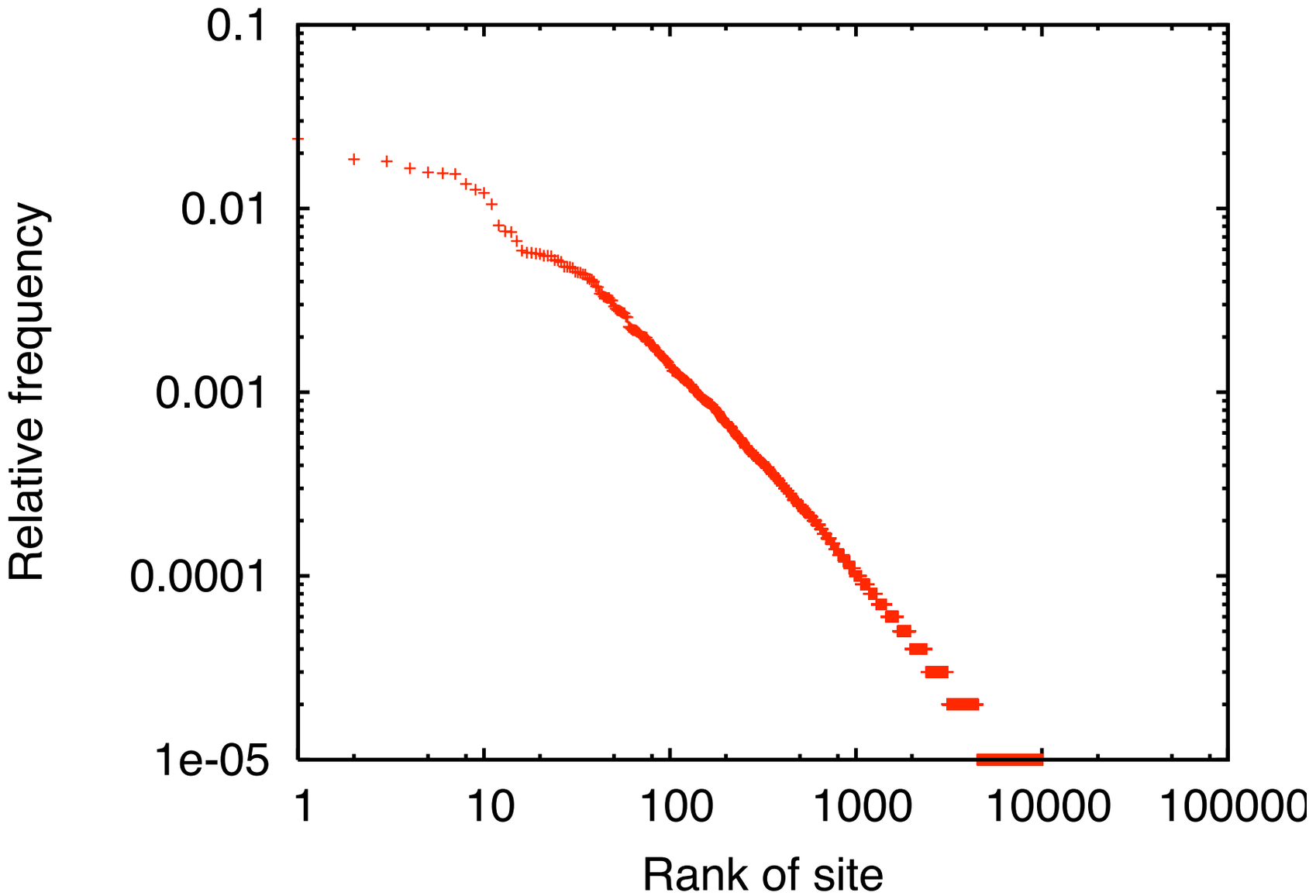
Squid Log of a Local School
1,100 unique users

6 months
9 mn requests
126 Source IPs
33k Destination IPs
51k Destination Host Names

Donnerstag, 26. März 2009

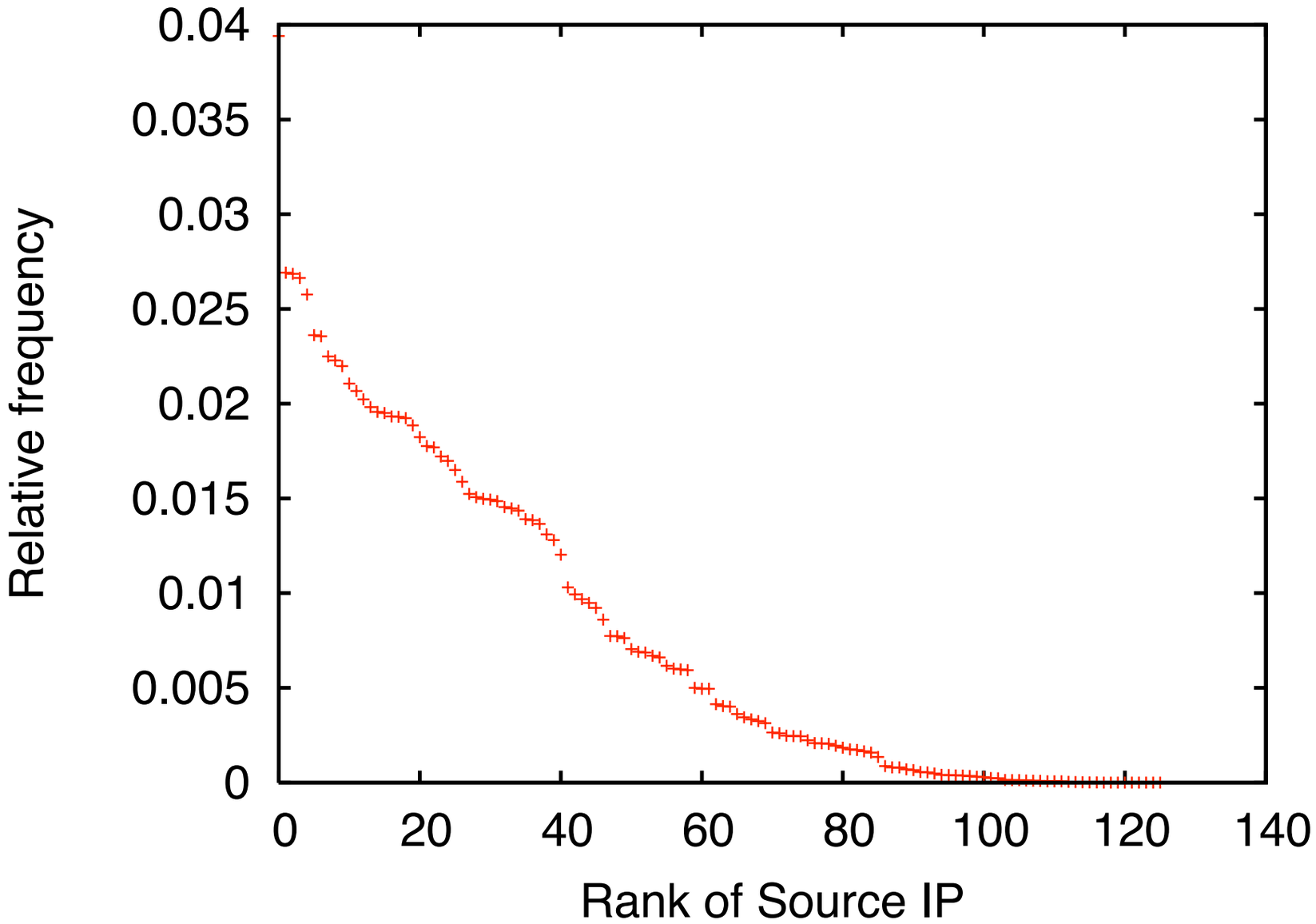
For the study we pulled the proxy log files of a local school. In order to make sure that the sample was not biased too badly we analysed it descriptively...

Popularity of Requested Sites
follows Power Function (Zipf-like)



and found that access frequencies of sites ranked by popularity show the expected Zipf-like distribution. I.e., users in our group have - to a certain degree - shared interests.

Heterogenuous Usage Intensity



In order to get a better feeling for our users we looked at the number of requests from the various SRC IPs. we found that activity varies wildly.

Biased Sample!

YMMV

(calls for follow-up study)

We are well aware that our sample is biased due to the environment we pulled it from. Therefore, results are only valid for our user base.

Evaluation Methodology

For each request in the sample
simulate a typical Law Enforcement Agency query and
calculate the *Simplistic Effectivity Metric*

Query: *From which source IP address originated the request at <TIMESTAMP> to <URL> using your IP address <IP>?*

Evaluation of
4 Data Retention Schemes

Sessions

Requests

+ DST IP

+DST Host

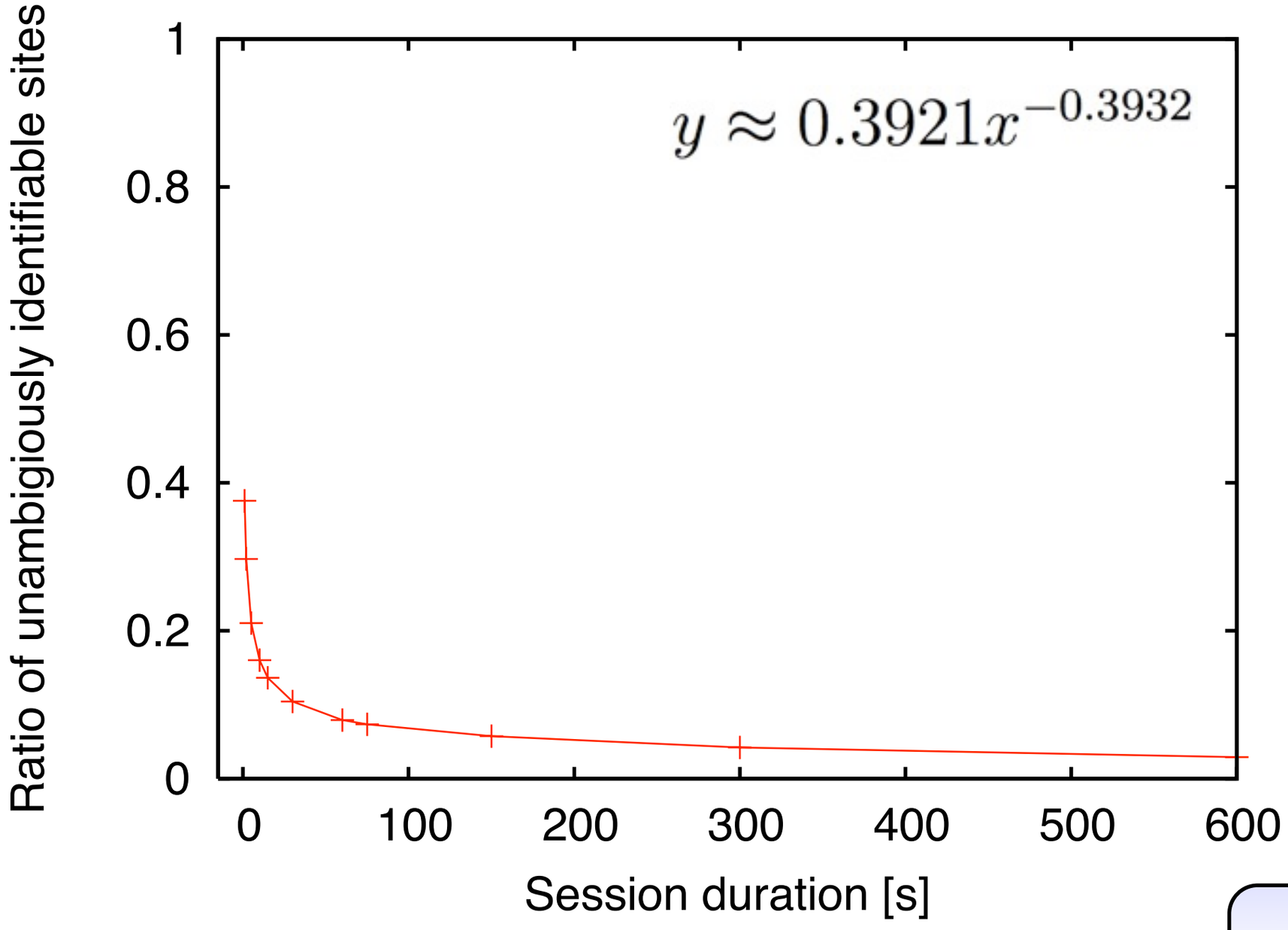
Session-Based Logging

available for VPNs, anonymisation services, etc.

Timestamp of start of user session
Timestamp of end of user session
Source IP
Proxy IP

Sessions

Session-Based Logging



Sessions

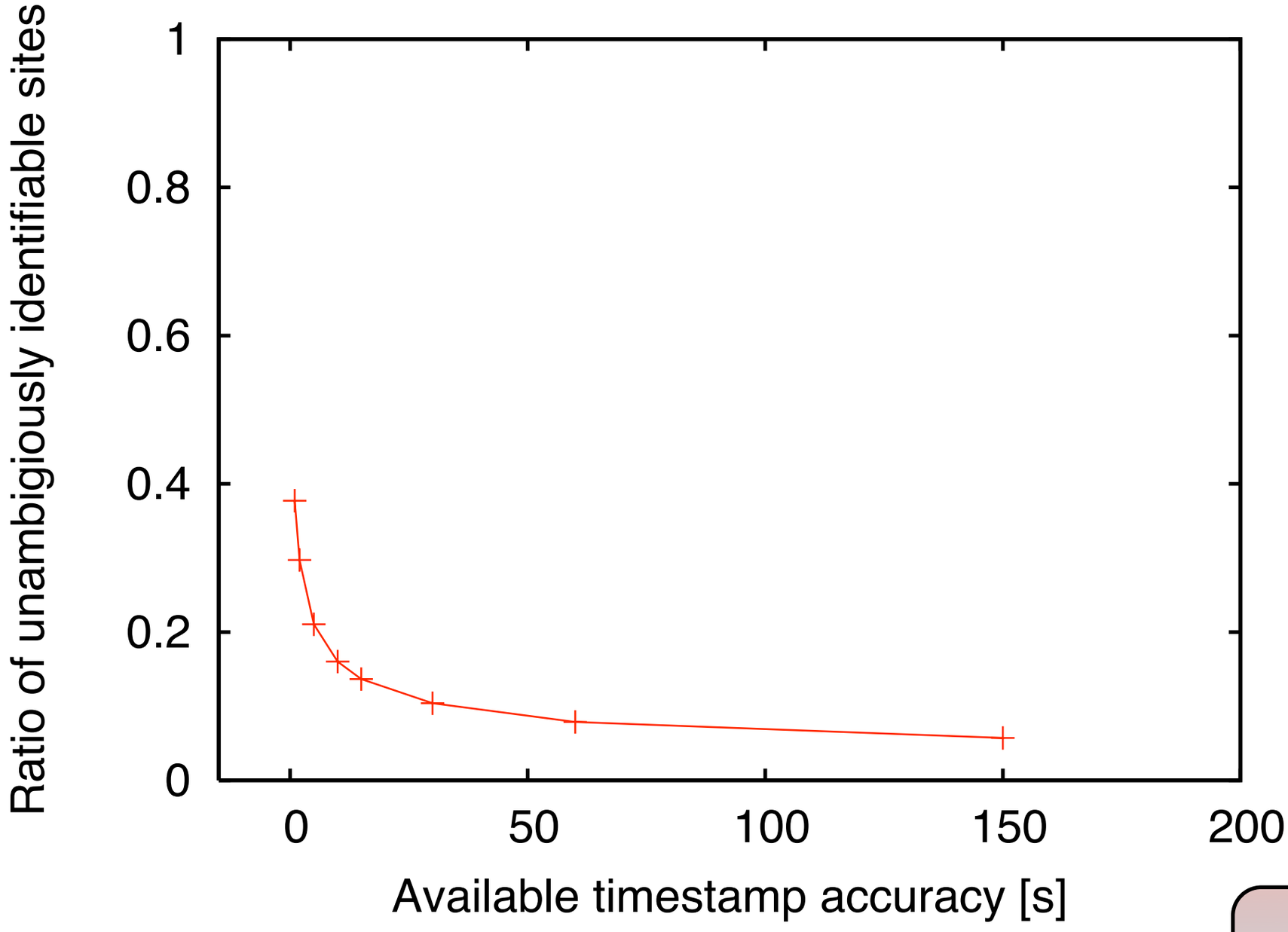
Request-Based Logging

available for HTTP proxy servers, etc.

Timestamp of request
Source IP
Proxy IP

Requests

Request-Based Logging

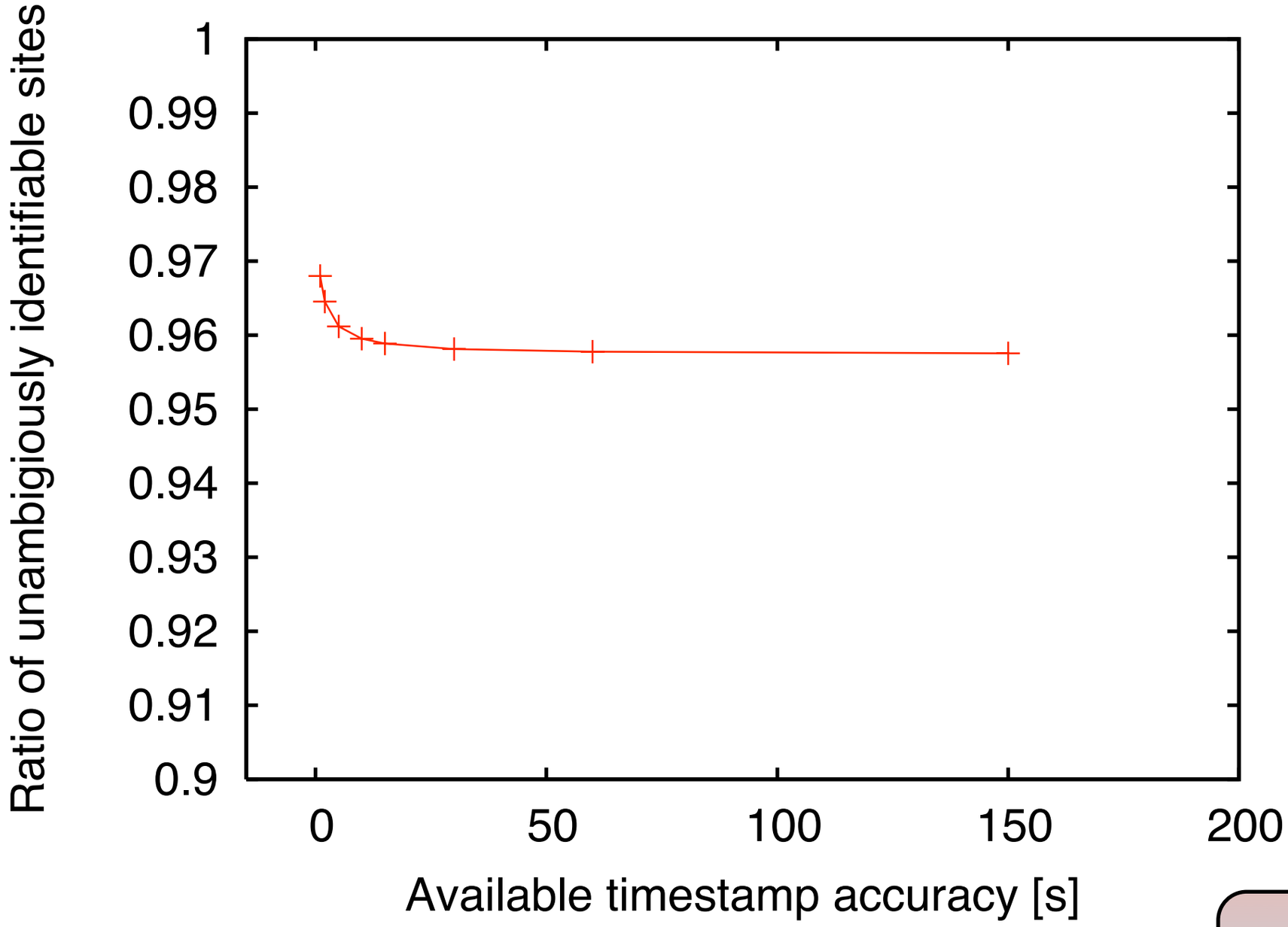


Requests

Request-Based Logging
+ Storing Destination Address

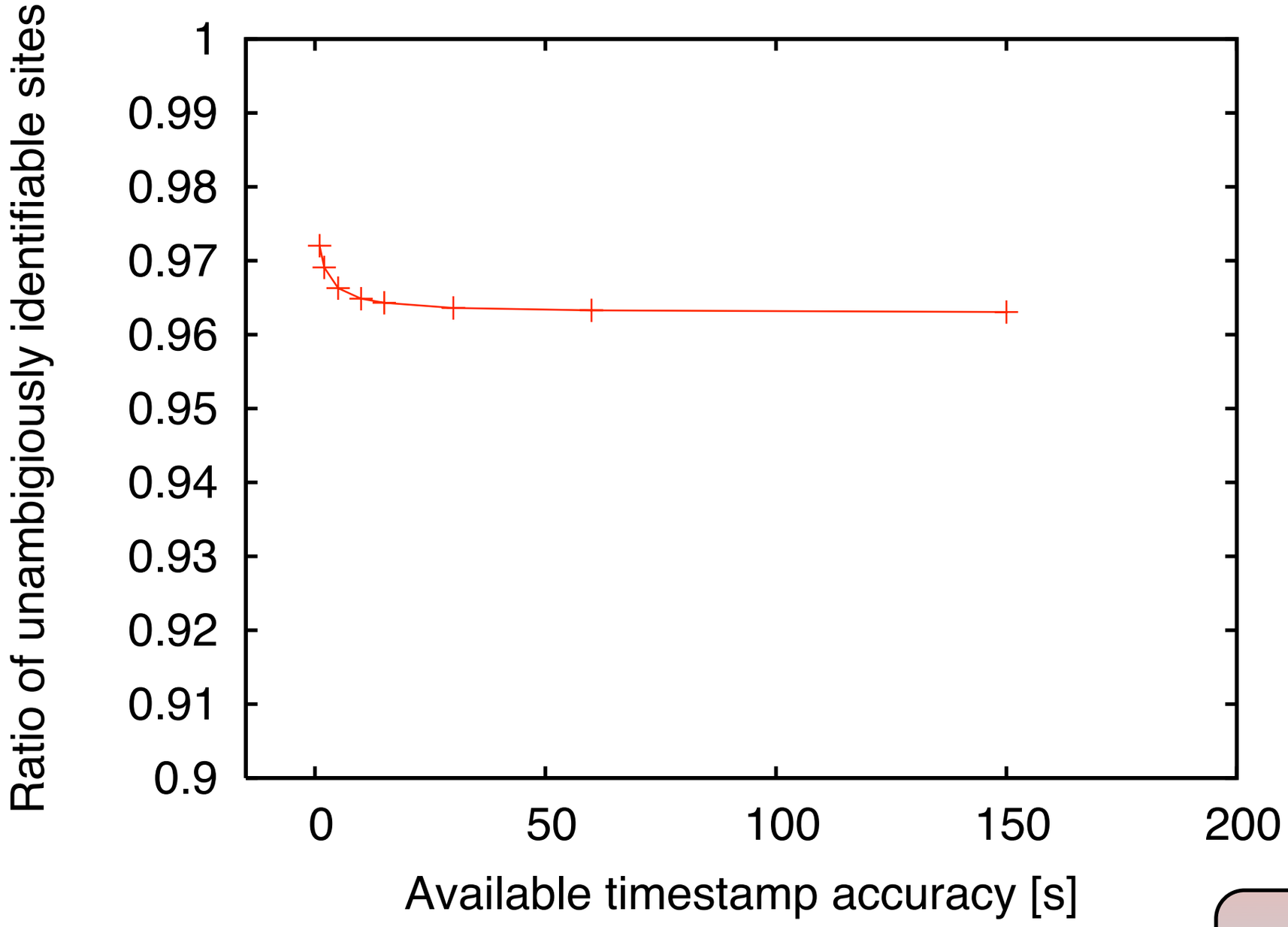
Timestamp of request
Source IP
Proxy IP
Destination IP or hostname

Request-Based Logging
+ Storing Destination IP



+ DST IP

Request-Based Logging
+ Storing Destination Hostname



+ DST Host

Results Overview

Sessions	Requests	+ DST IP	+DST Host
sessions of 300s: 5% traceable	accuracy 60s: 8%	60s: 95.8%	60s: 96.3%
	accuracy 1s: 39%		
privacy: good	privacy: okay	privacy: poor	privacy: poor

Open Questions

How does homogeneity of users influence effectivity?

What accuracy is achievable for timestamps in real world?

How effective are intersection attacks in the real world?

How would privacy benefit if proxies used huge IPv6 ranges?

What about advanced schemes, e.g., embedding dedicated data retention tags in HTTP header or using TCP source ports?

Real challenge is to find a data retention scheme that combines

cannot be solved by the 4 schemes we evaluated. Search goes on...

Trace Me If You Can

studied 4 data retention schemes based on already available data using log files from a small proxy server

results indicate that the schemes based on session-based and request-based logging offer no satisfactory traceability

traceability will improve significantly, if destination IPs are stored; which comes at the cost of privacy of users

Dominik Herrmann

dominik.herrmann@wiwi.uni-regensburg.de http://www-sec.uni-regensburg.de/herrmann/