

Analyzing Characteristic Host Access Patterns for Re-Identification of Web User Sessions

Dominik Herrmann, Christoph Gerber,
Christian Banse, Hannes Federrath

Management of Information Security
University of Regensburg, Germany



Nordsec 27. – 29. October 2010
Aalto University, Espoo, Finland

SECBIT

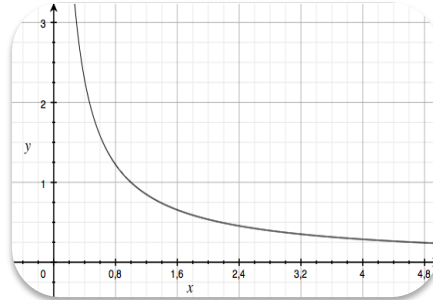
EFRE

Universität Regensburg

agenda



problem
description



relation to
text-mining



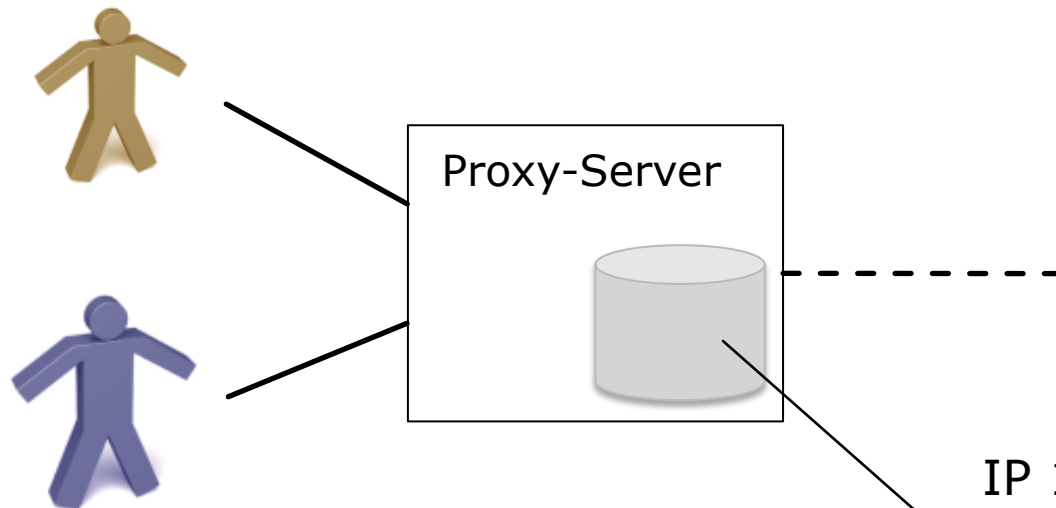
case study
and
test setting



re-
identification

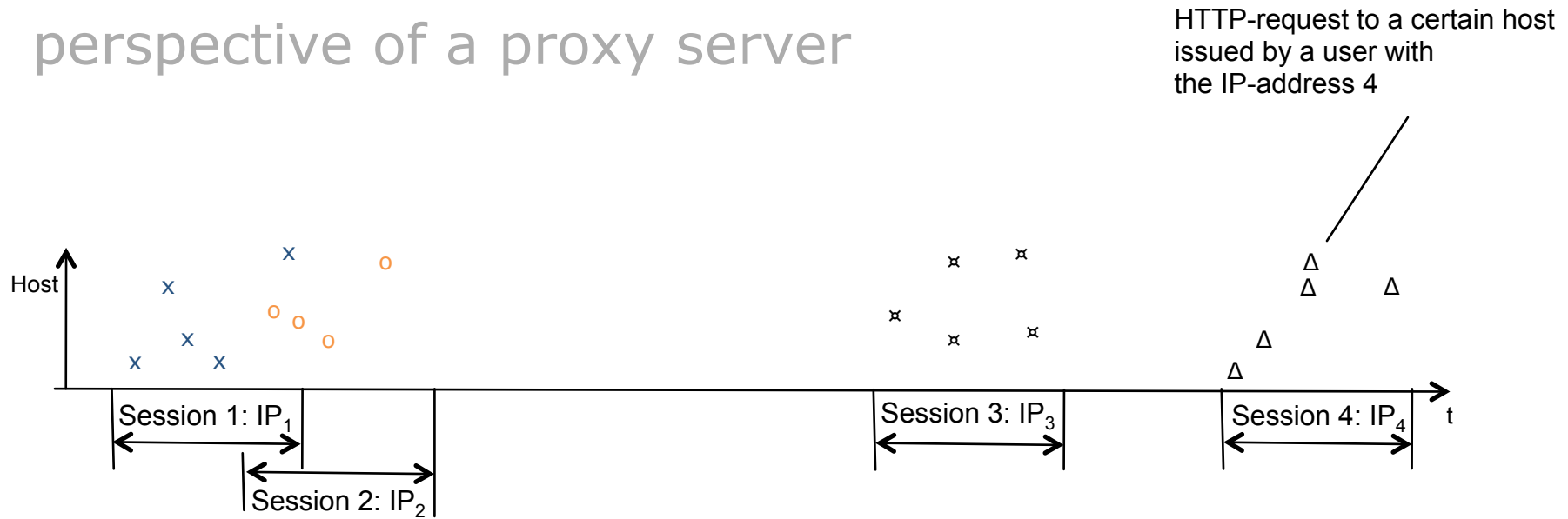
problem description

- small user group (e.g. users of a proxy-server)
- all HTTP-requests are recorded
- changing IP-addresses / different surfing sessions

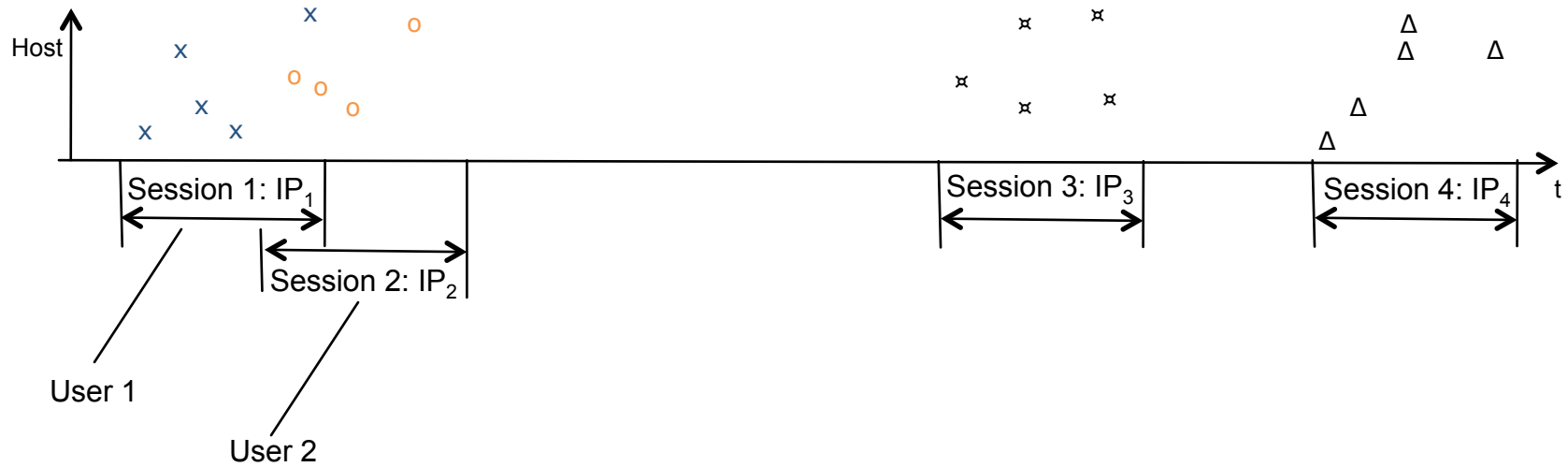


IP 1, User 1: www.wikipedia.de
IP 2, User 2: www-sec.uni-r.de
IP 2, User 2: www.cse.tkk.fi
IP 1, User 1: www.google.de

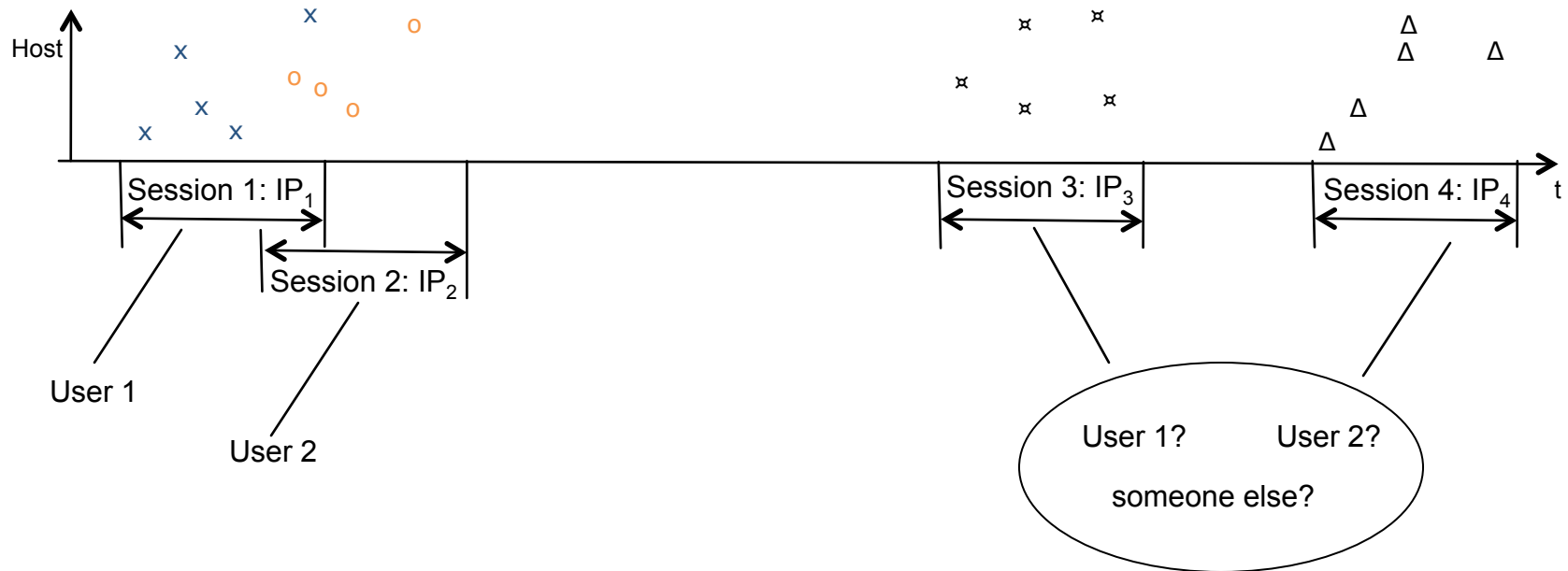
perspective of a proxy server



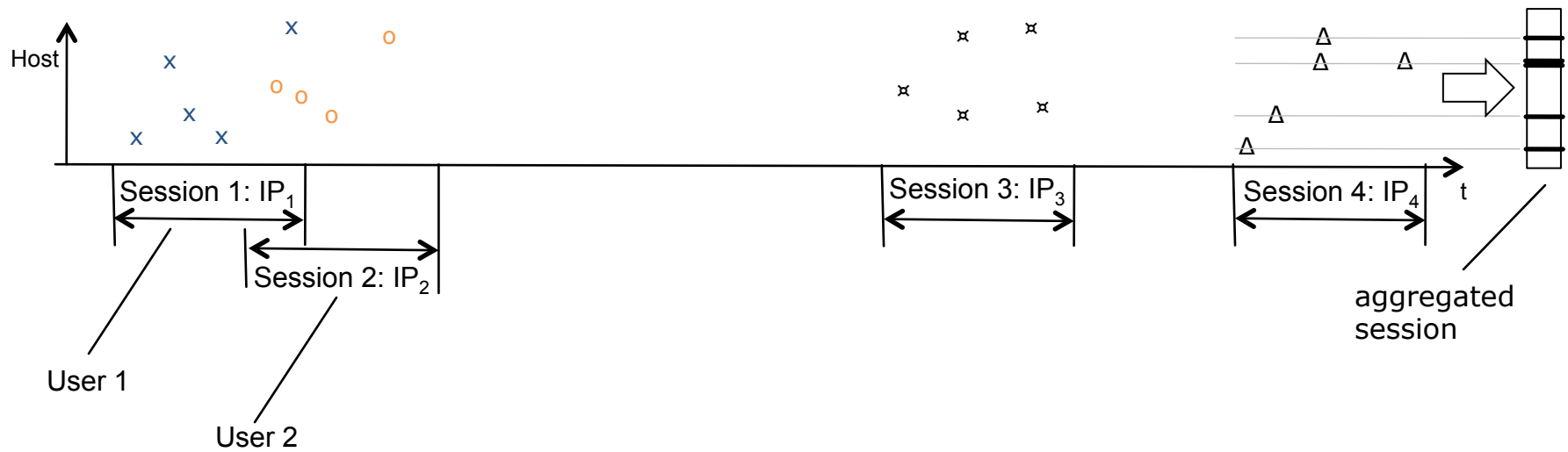
perspective of a proxy server



perspective of a proxy server

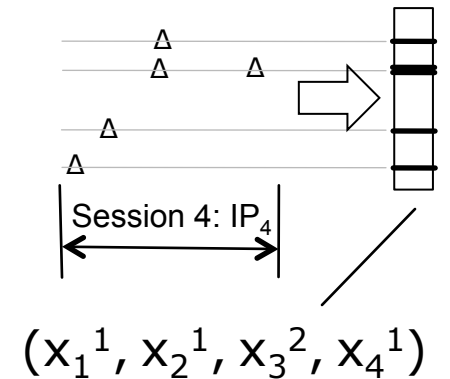


perspective of a proxy server



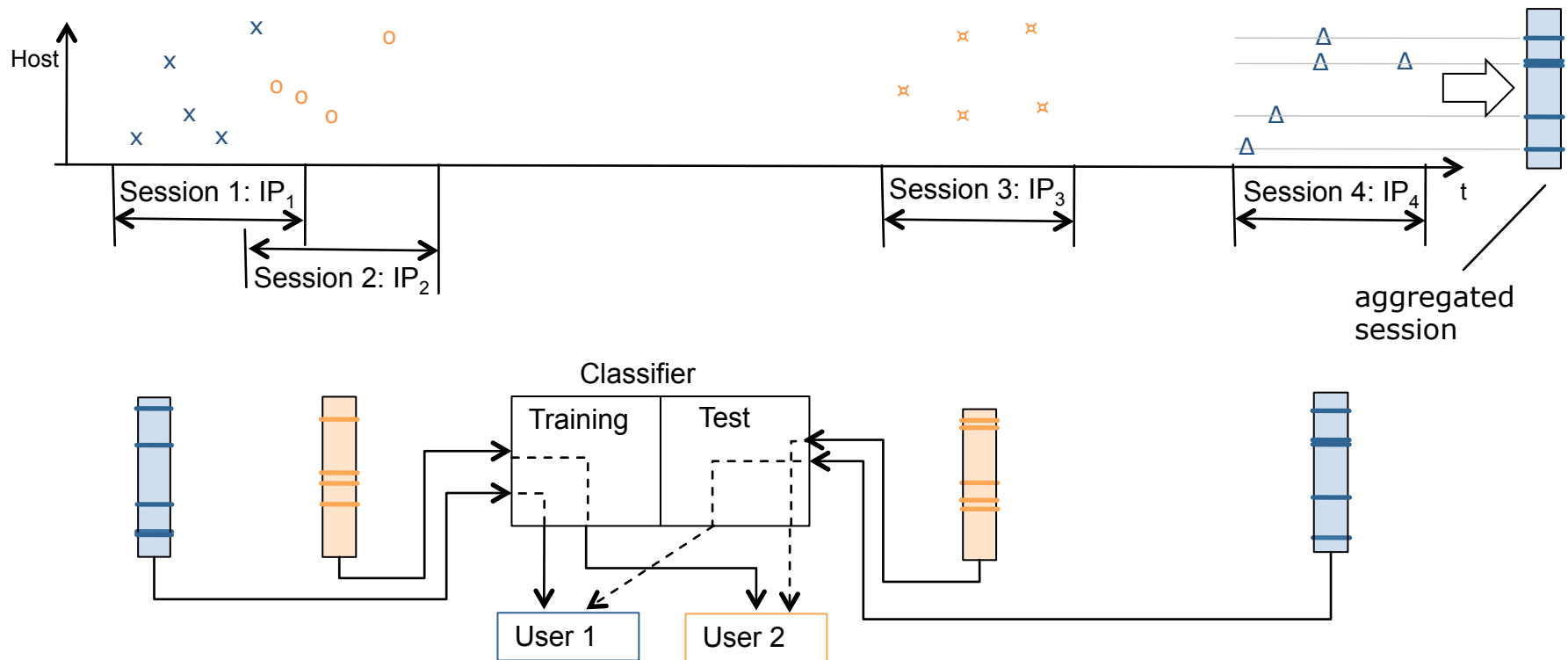
modeling the classification problem

- each session (s) consists of a multiset $(x_1^{f_{x_1}}, x_2^{f_{x_2}}, \dots, x_m^{f_{x_m}})$
- each surfing session (s) is an instance of a class $c_i \in C$
- each class represents an user



- X_4 : www.google.de
- X_3 : www.cse.tkk.fi
- X_2 : www-sec.uni-r.de
- X_1 : www.wikipedia.de

classification of user sessions



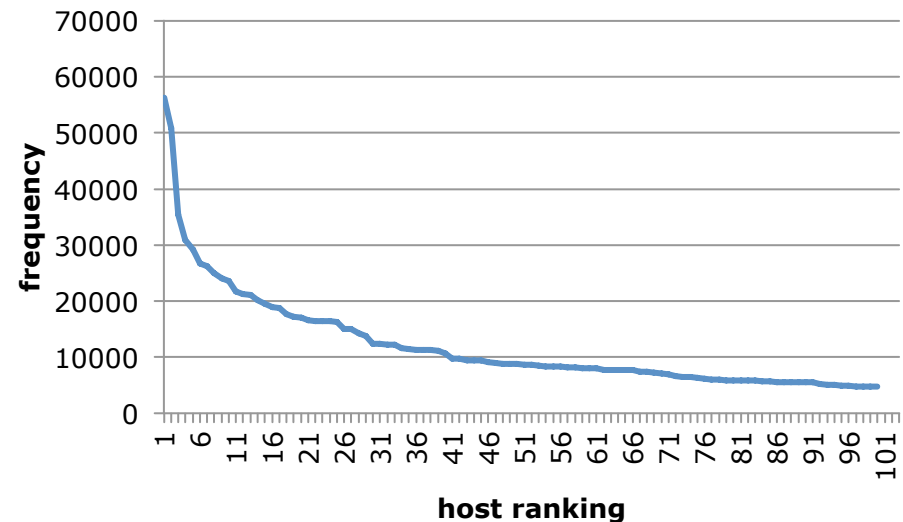
similarity to text-mining-problems

- word frequency and host frequency following a power-law

text-retrieval



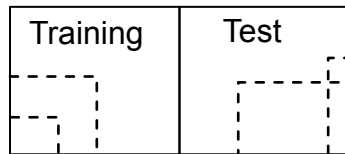
user re-identification



<http://www.cs.princeton.edu/introcs/data/bible.txt>

text-mining toolbox

- multinomial naive bayes (MNB)



$$P(\mathbf{f}|c_i) \sim \prod_{j=1}^m P(X = x_j | c_i)^{f_{x_j}}$$

- vector transformations

- TF transformation
- IDF transformation
- cosine normalisation (N)

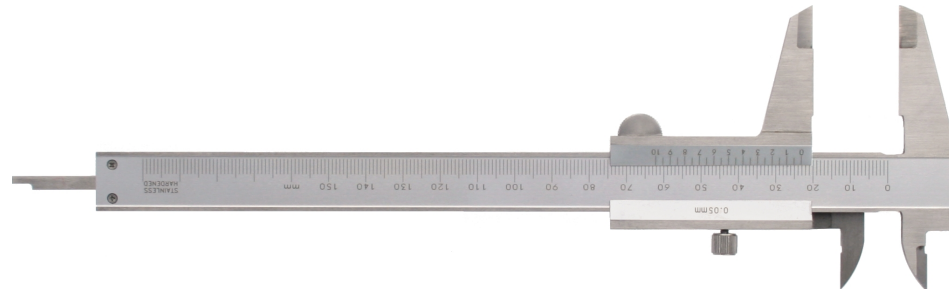
$$f_{x_j}^* = \log(1 + f_{x_j})$$

$$f_{x_j}^* = f_{x_j} \cdot \log \frac{n}{df_{x_j}}$$

$$f_{x_j}^{\text{norm}} = \frac{f_{x_j}^*}{\|(f_{x_1}^*, \dots, f_{x_m}^*)\|}$$

related work

- Pang et al. (2007)
 - re-identification of users in 802.11 wireless networks
- Yang (2008)
 - focus on fraud detection
- Kumpost (2009)
 - focus on re-identification of web users



test setting and case study

- test users
- local proxy server
- host obfuscation
- client/server architecture

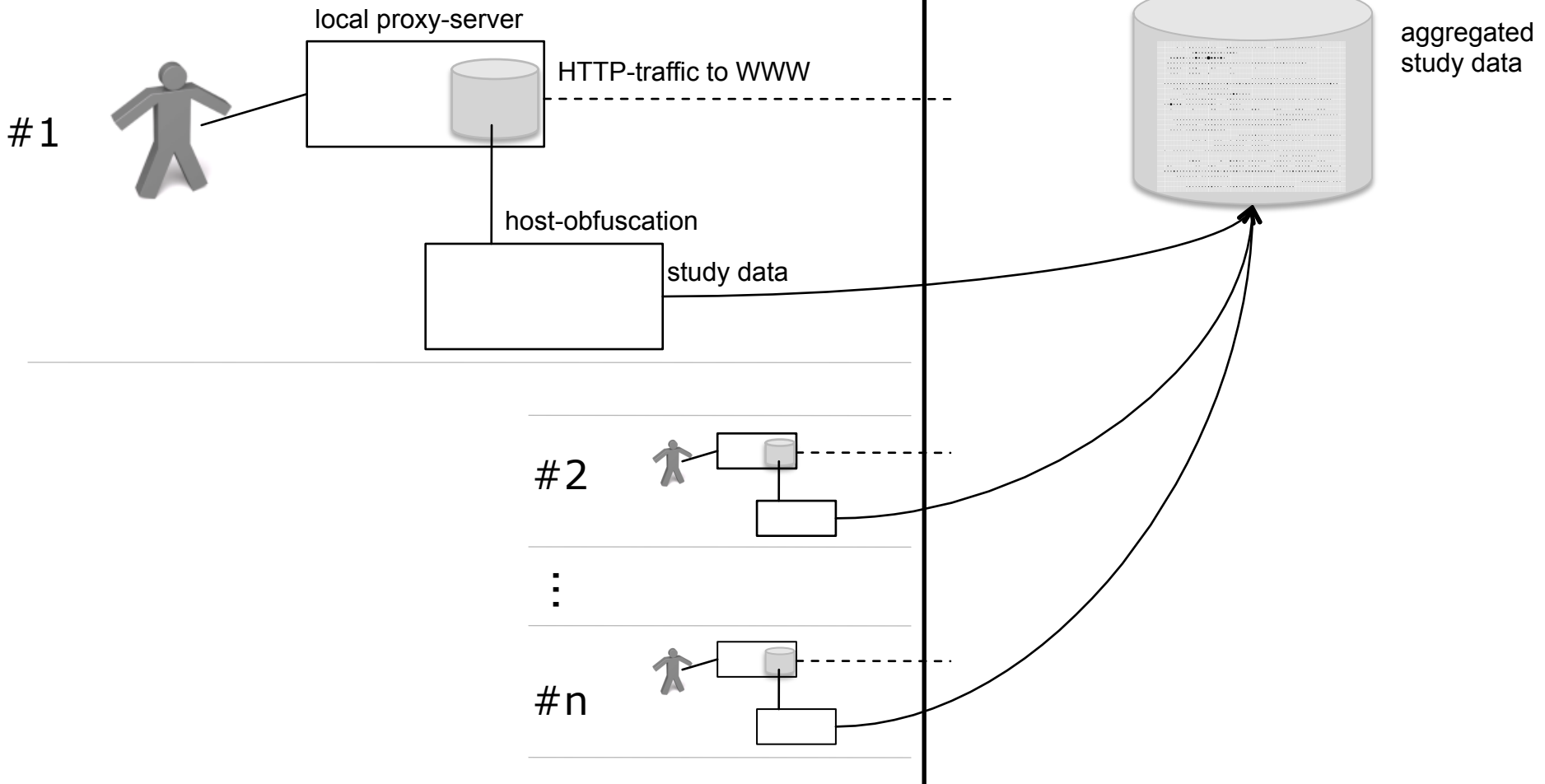
key	value
participants	28
duration of study in days	57
number of HTTP requests	2,684,736
number of unique hosts	25,124



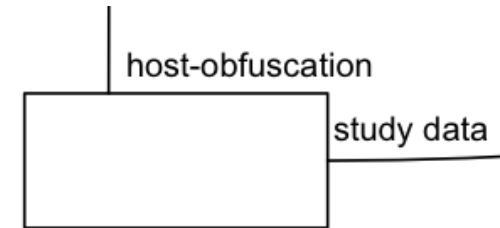
data acquisition

users scope of protection

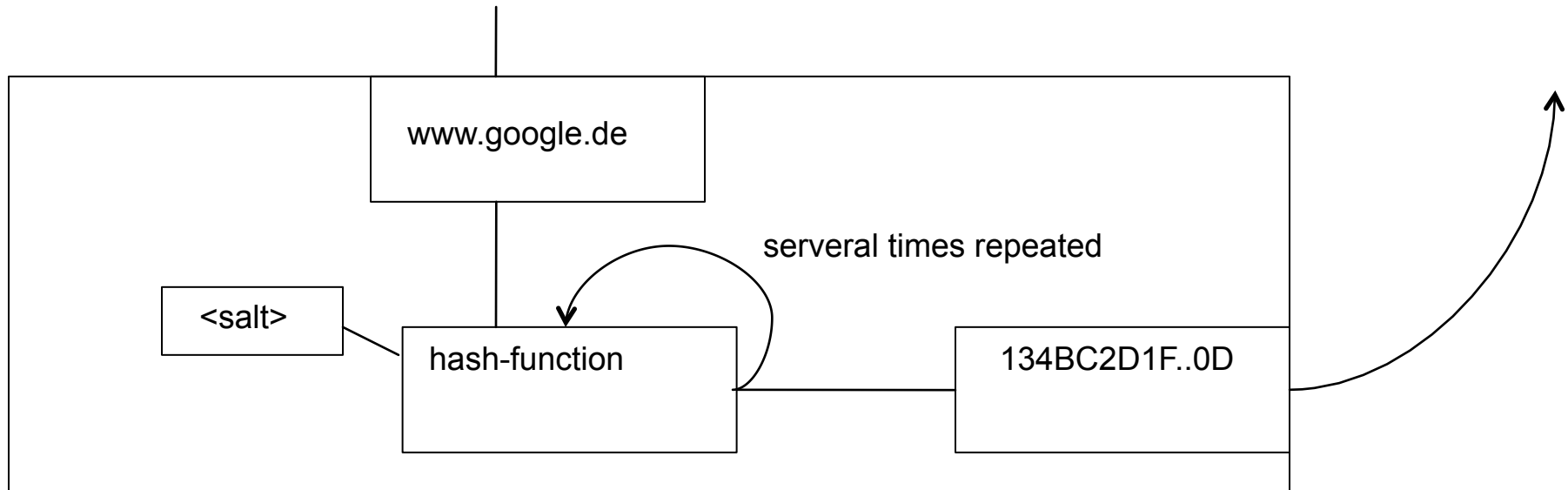
WWW

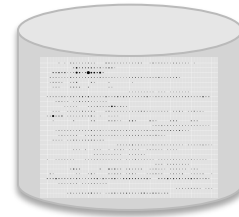
aggregated
study data

host obfuscation

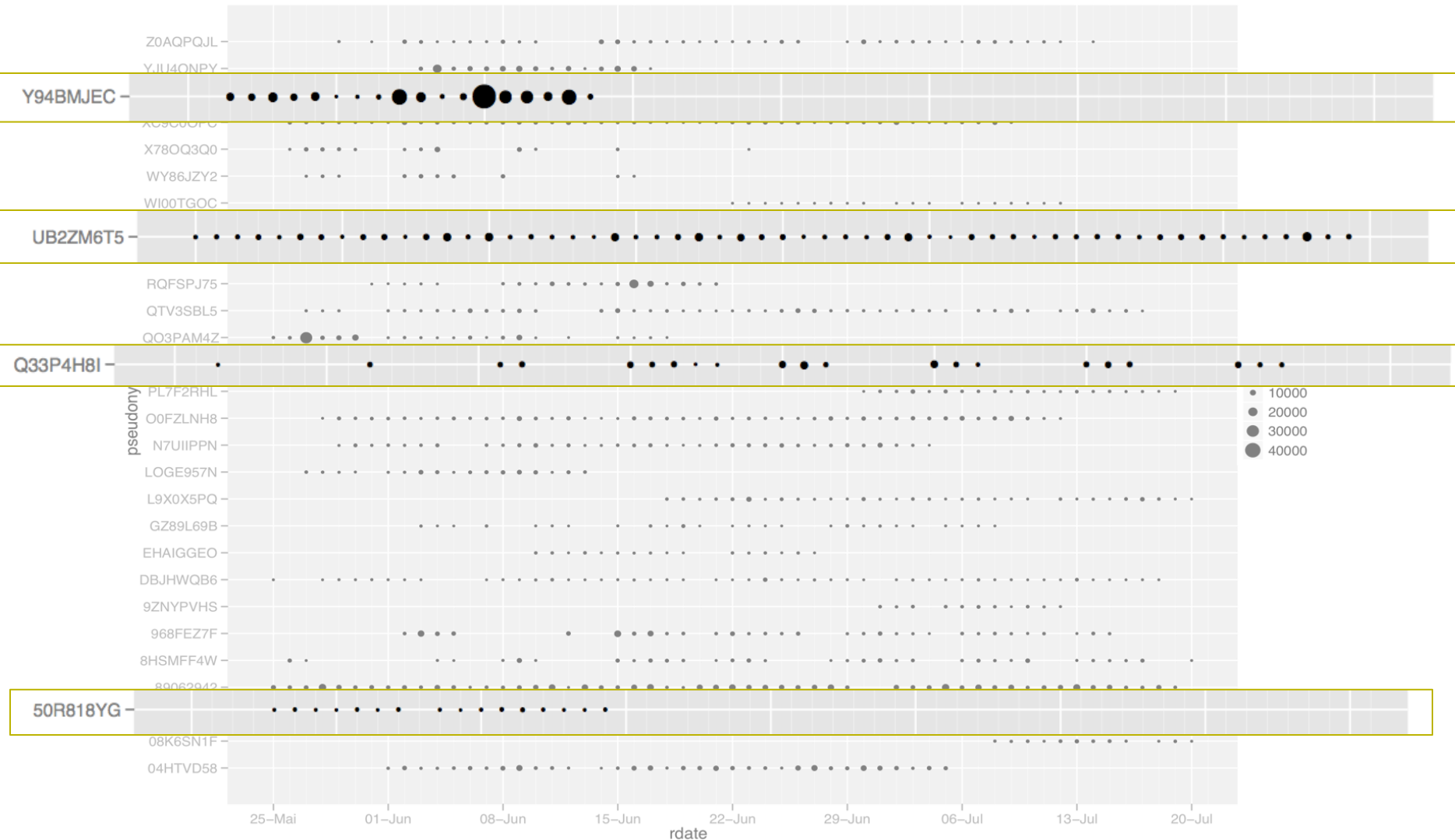


- hashing of hostnames
- + salt to prevent dictionary attacks
- + iterations to prevent building of own dictionary





user contribution on a daily basis



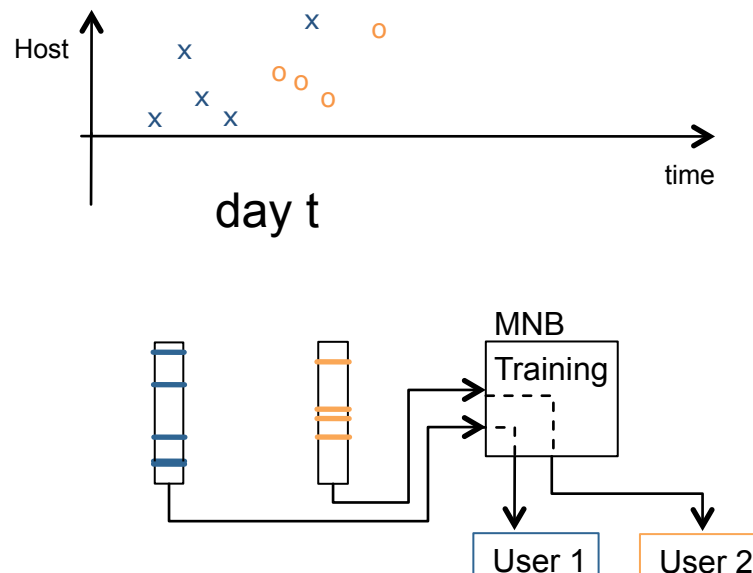
re-identification attack

- attacker's view
 - limited knowledge
 - practical relevance
- simulations
 - for evaluating the driving factors
- countermeasures



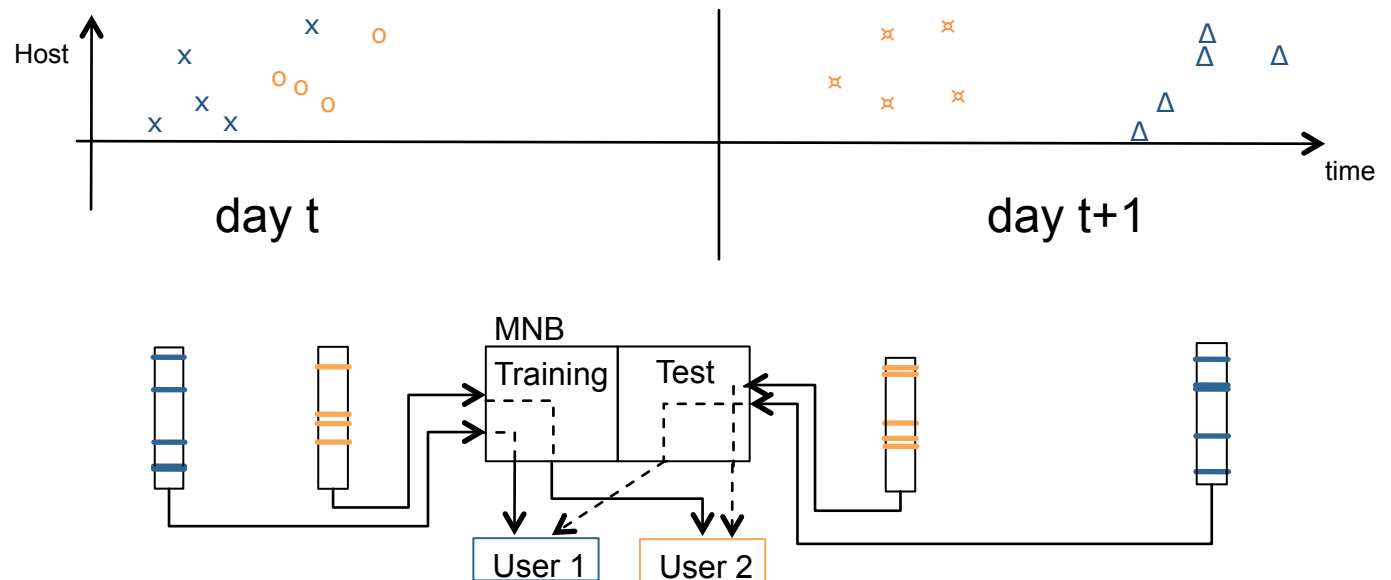
attacker's view (training)

- $\Delta t = 24h$
- decision to track a specific user u^t on day t
- training with U^t classes on day t with S^t sessions



attacker's view (attack)

- $\Delta t = 24h$
- decision to track a specific user u^t on day t
- training with U^t classes on day t with S^t sessions
- on day $t+1$ assigning each session s to a class u_t
- evaluating the classification result for class c_u



prediction scheme of attacker's view

correctly classified by proxy-server

- attacker successfully recognizes the user
- attacker successfully recognizes the absence of the user

wrong classification – error is detectable for proxy-server

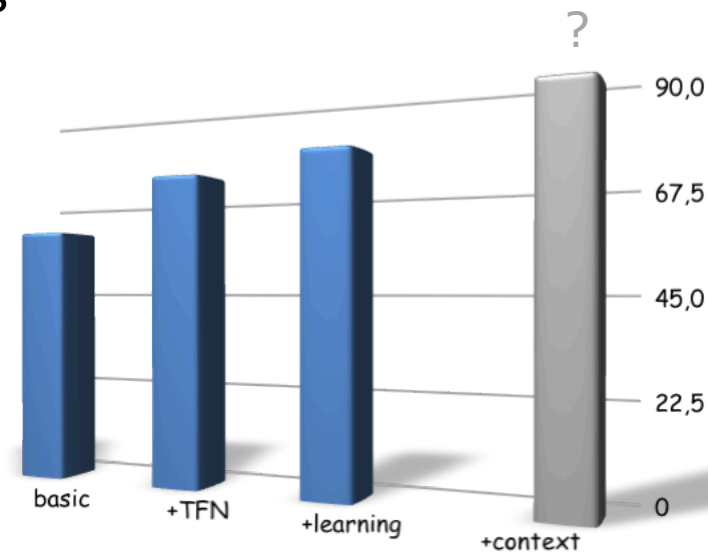
- more than one user was predicted to belong to class c_u

wrong classification – error not detectable for proxy-server

- attacker detects absence of user; but user was online
- attacker wrongly recognizes the user

results from the attacker's view

- user re-identification works
 - 60.5% correctly classified sessions
- and can be improved by vector transformations
 - 73.1% by applying TF-N transformation
- further improvements are possible
 - 77.6% by 'learning' the user habits
- more improvements conceivable
 - timing-information
 - filenames
 - GET-parameters
 - destination-ports
 - ...

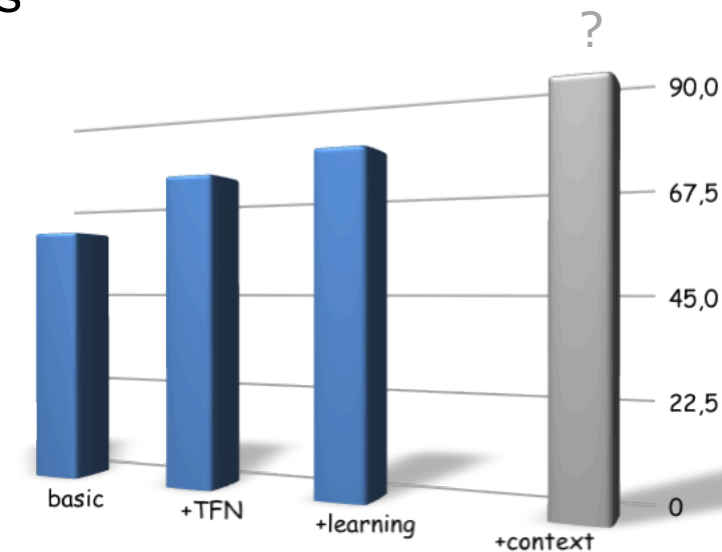


results from the attacker's view

- user re-identification works
 - 60.5% correctly classified sessions

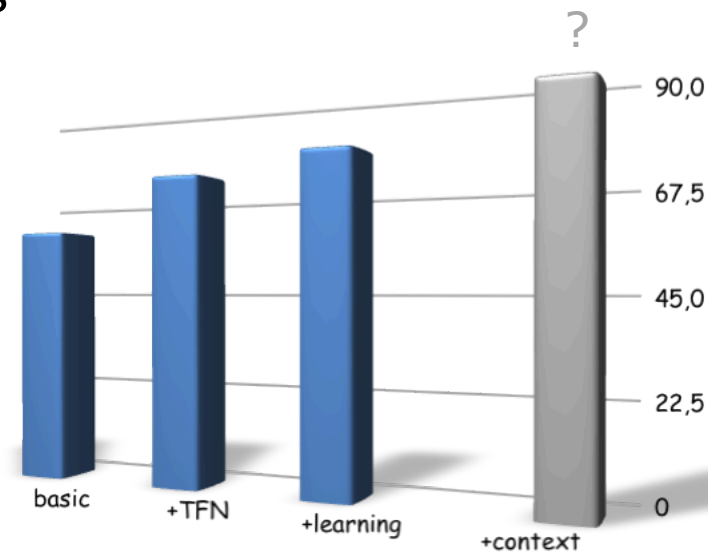
none	N	IDF	IDF-N	TF	TF-N	TF-IDF	TF-IDF-N
60.5%	62.9%	65.0%	62.8%	56.0%	73.1%	66.1%	72.8%

- further improvements are possible
 - 77.6% by 'learning' the user habits
- more improvements conceivable
 - timing-information
 - filenames
 - GET-parameters
 - destination-ports
 - ...



results from the attacker's view

- user re-identification works
 - 60.5% correctly classified sessions
- and can be improved by vector transformations
 - 73.1% by applying TF-N transformation
- further improvements are possible
 - 77.6% by 'learning' the user habits
- more improvements conceivable
 - timing-information
 - filenames
 - GET-parameters
 - destination-ports
 - ...



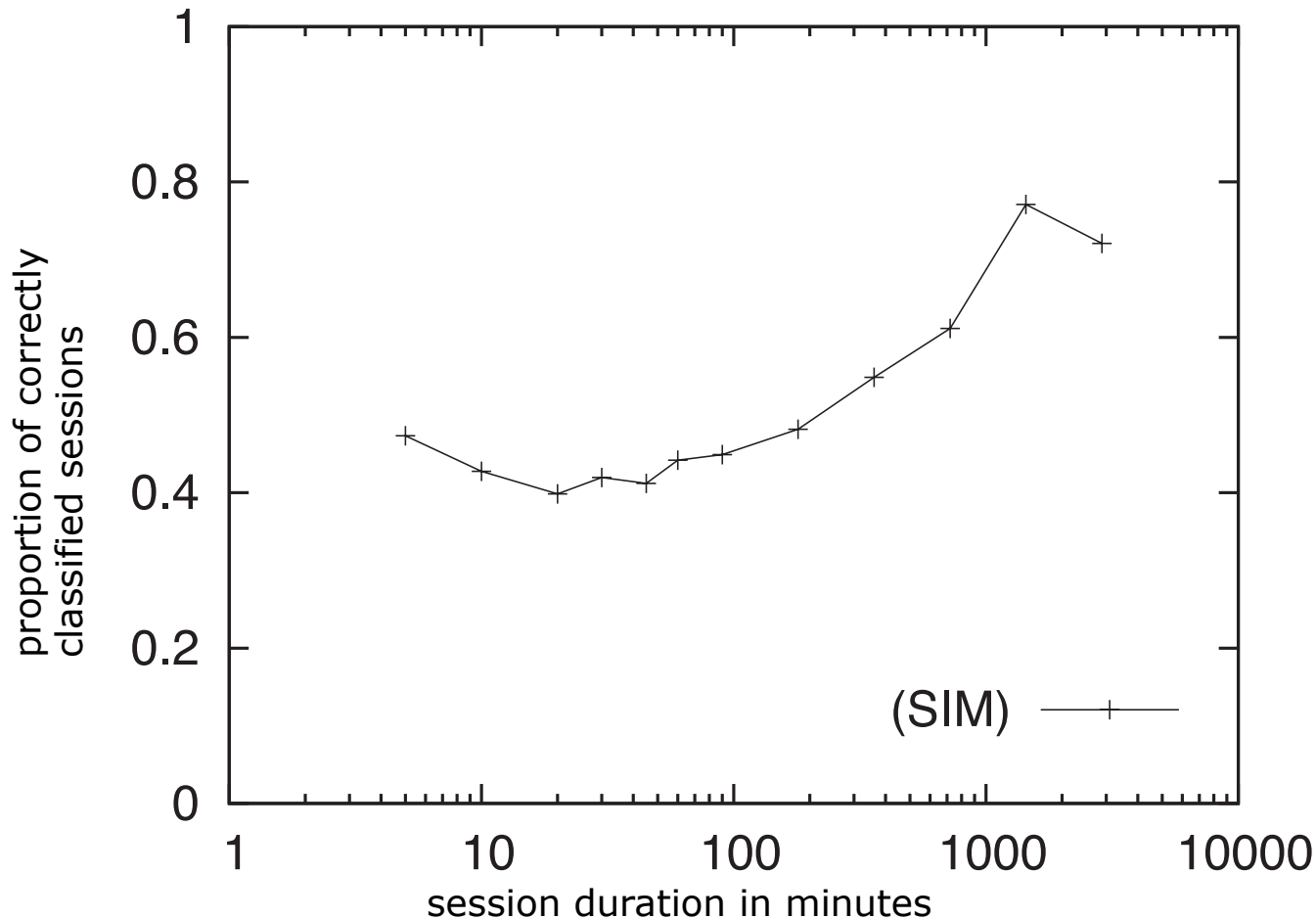
simulations

- simulation of simultaneously surfing sessions
 - putting together the cronologically succeeding sessions
 - always 28 users / session
- in each experiment one parameter was modified
 - session duration
 - number of simultaneous users
 - offset between last training and first test session
 - number of consecutive training instances
- each experiment was repeated 25 times



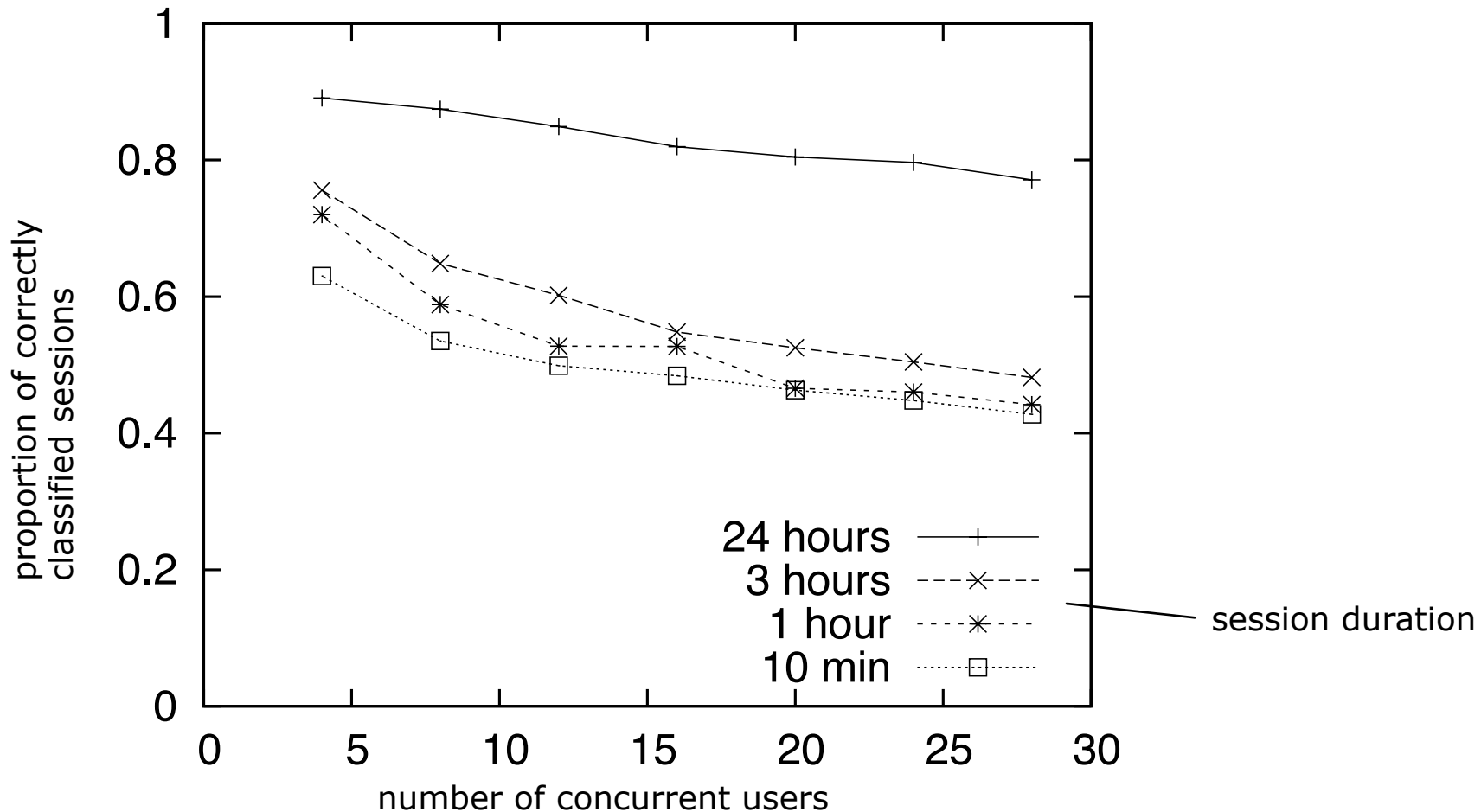
session duration

- longer session times support re-identification



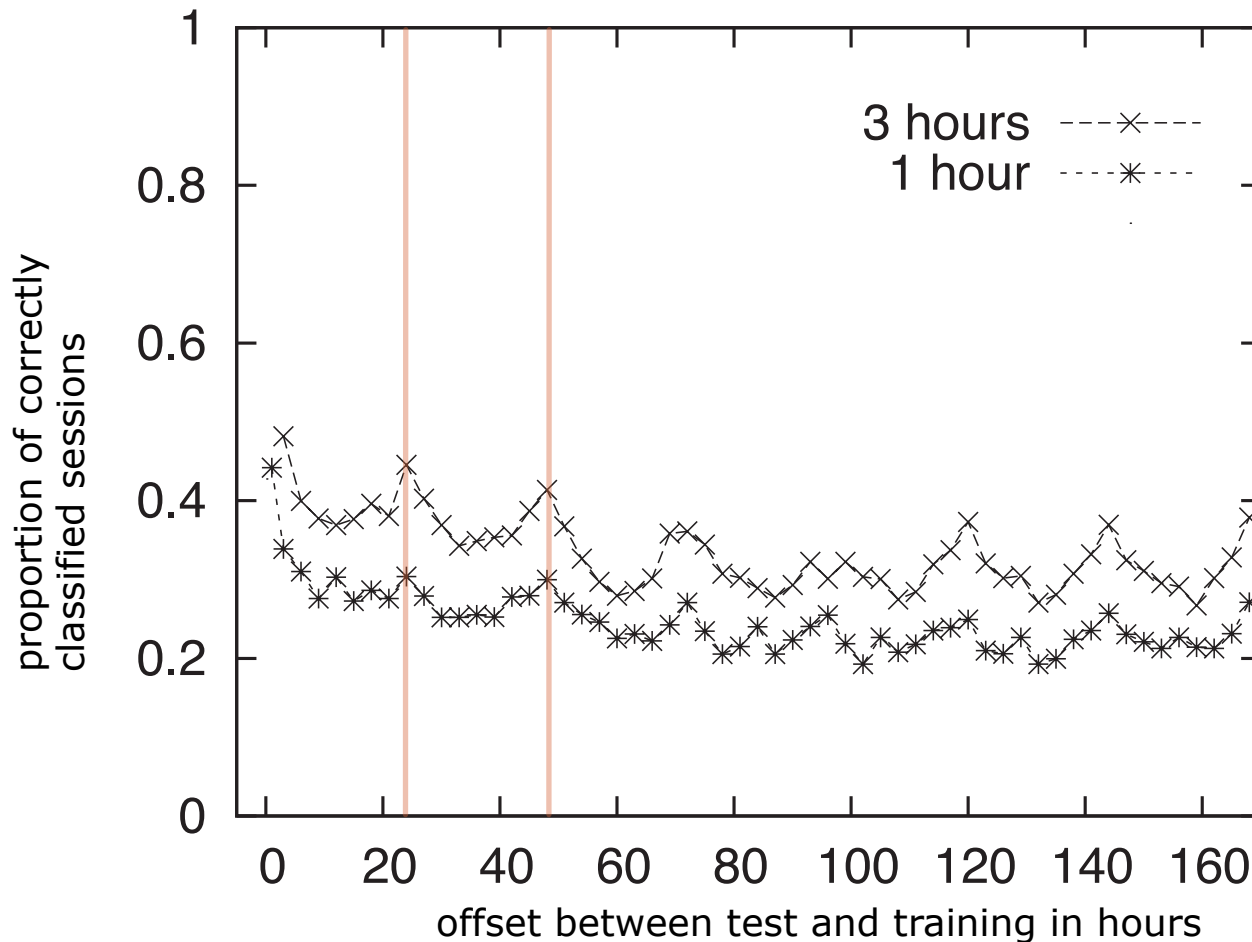
number of simultaneous users

- the fewer simultaneous users the better it works



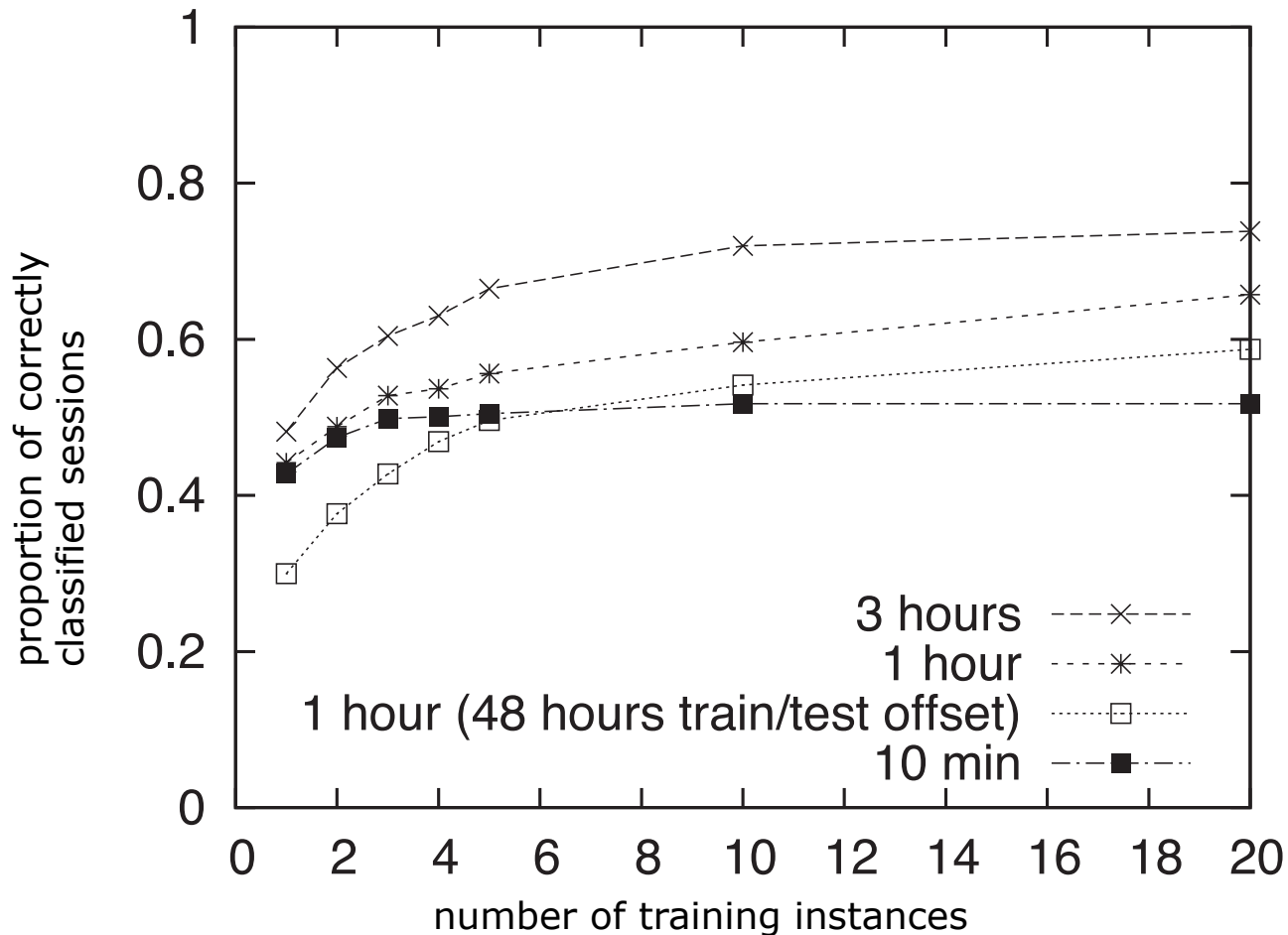
offset between test and training sessions

- each user tends to act similar at the same time of the day



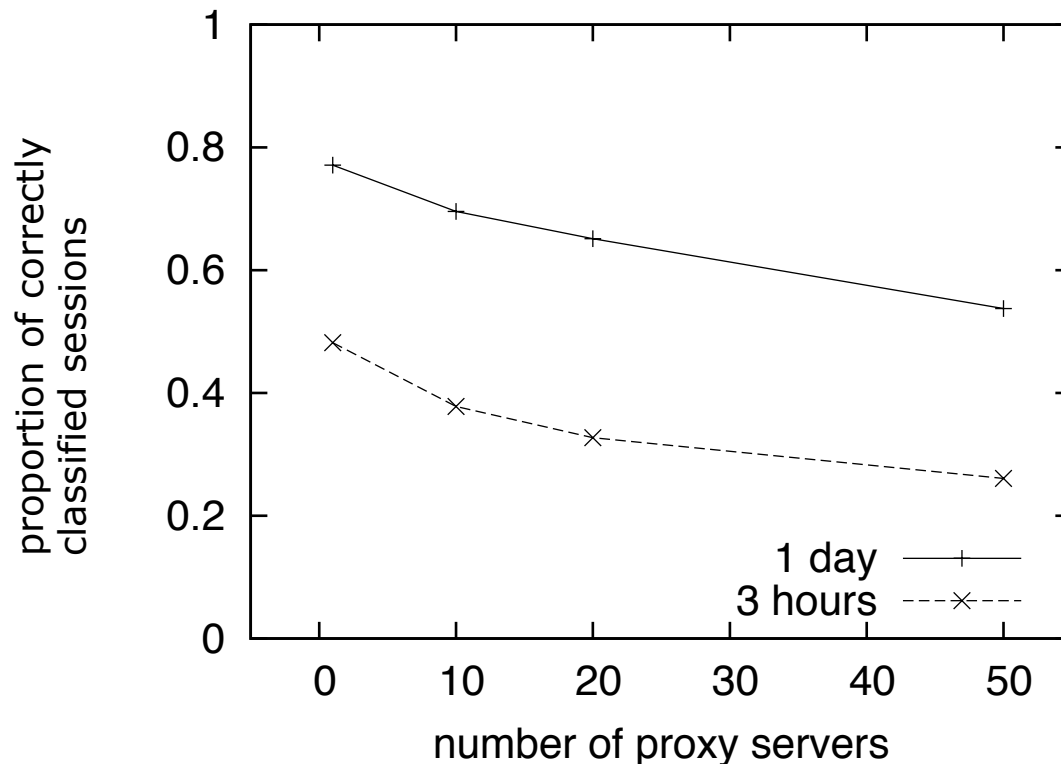
number of training instances

- more training instances are better, but only few are needed



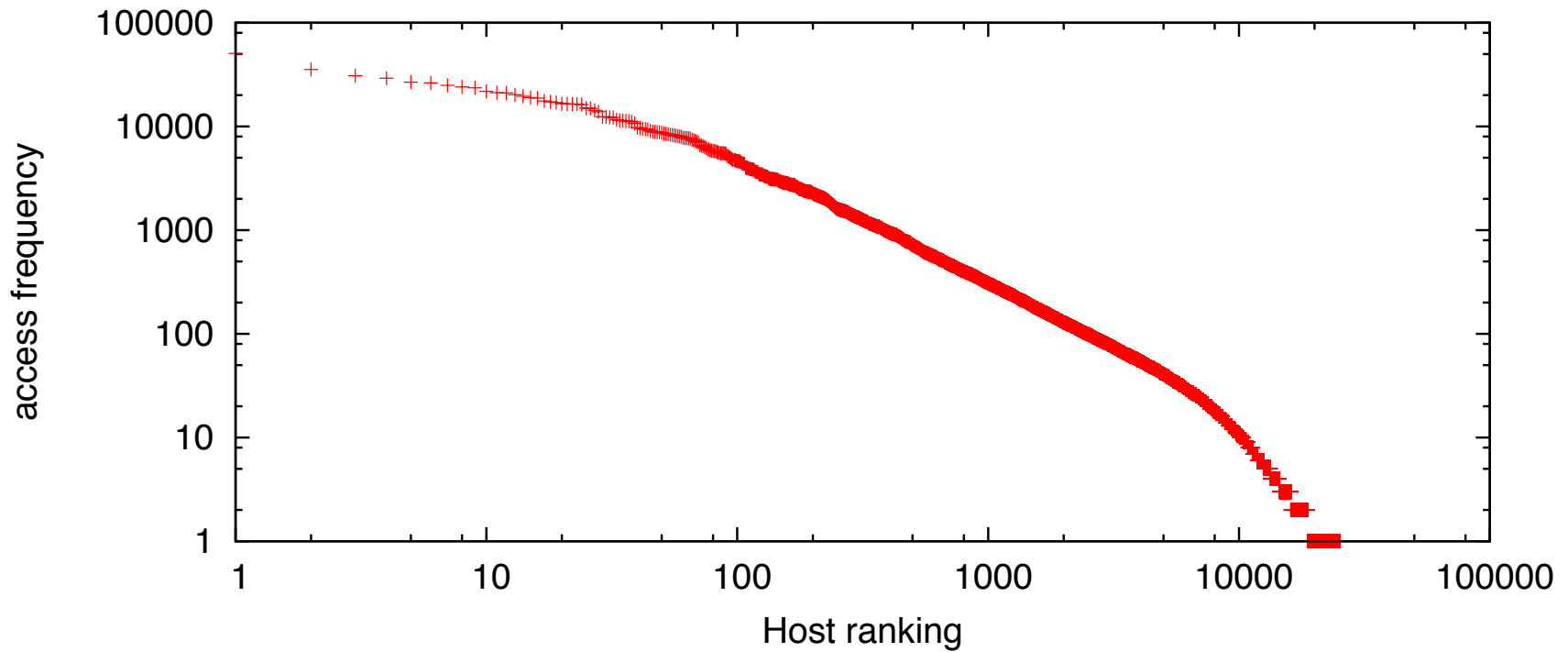
countermeasures

- using multiple, non-colluding proxy servers works
 - but is not practicable (at this early stage)
- more distribution schemes conceivable



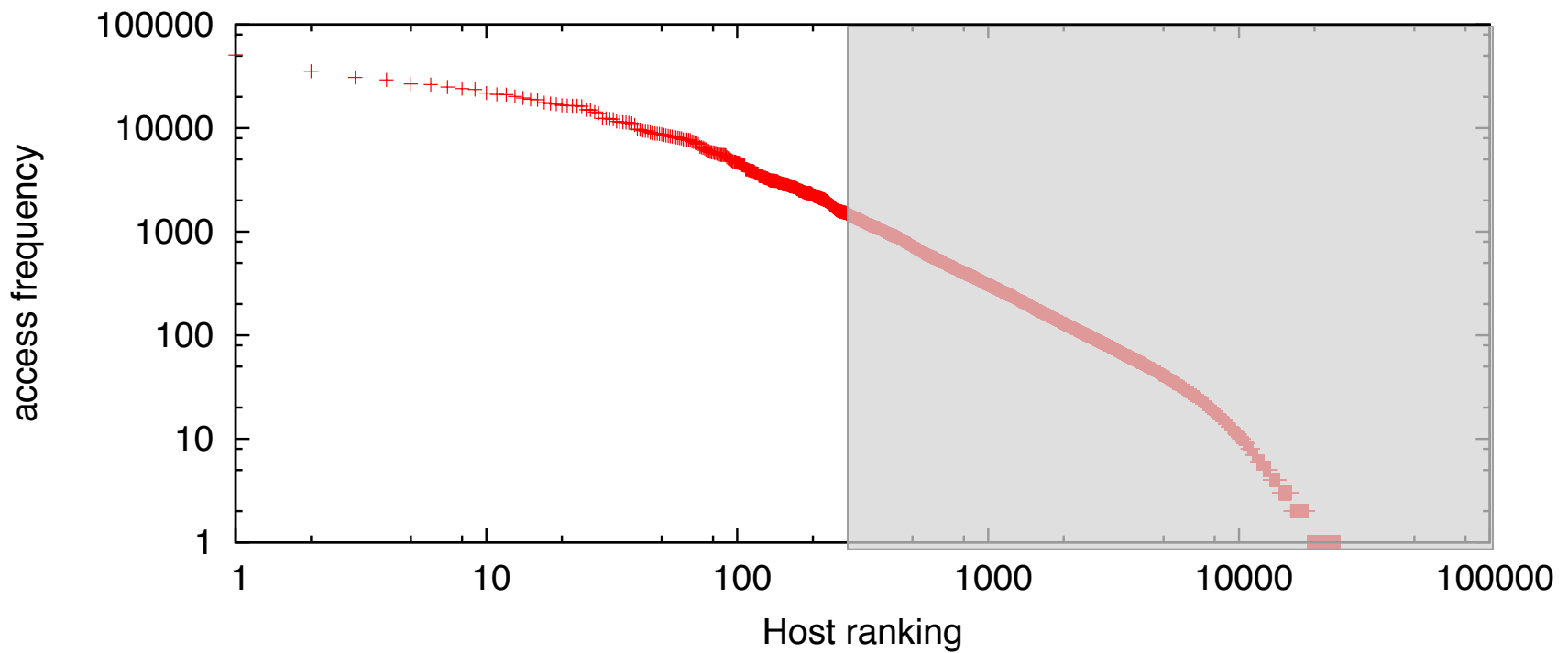
countermeasures

- analyzing a part of the host frequency distribution



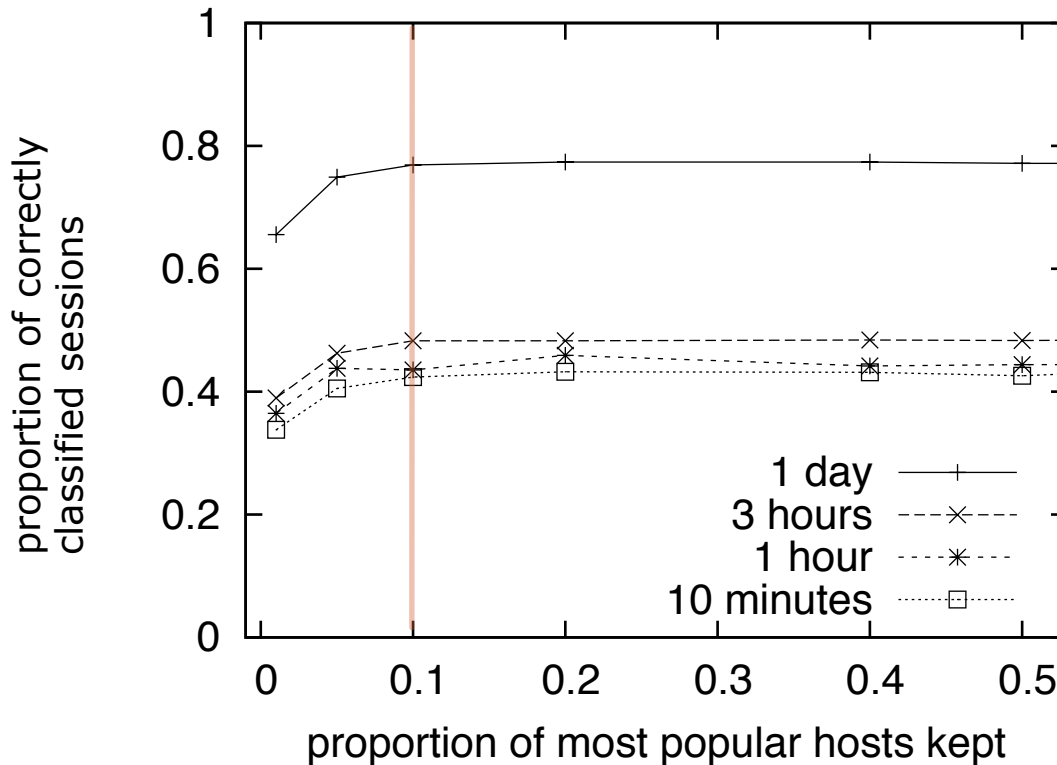
countermeasures

- analyzing a part of the host frequency distribution
 - keep the most popular hosts



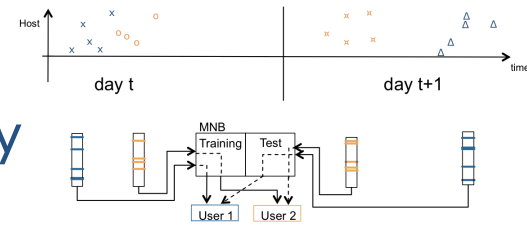
countermeasures

- analyzing a part of the host frequency distribution
 - keep the most popular hosts
 - can not prevent from user re-identification



conclusion and discussion

- re-identification as a feasible attack
- evaluated on a privacy preserving case study



- works well for small closed groups
- not only relevant for proxy-servers

- improvements in using context information

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700]
"GET http://www.ab.com/index.html HTTP/1.0"
200 2326
```



```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700]
"GET http://www.ab.com/index.html HTTP/1.0"
200 2326
```



- improvements in gathering more realistic sessions