Analysis of Web logs
Challenges and Findings

Maria Carla Calzarossa – Luisa Massari

Università di Pavia, Italy

mcc,massari@unipv.it

http://peg.unipv.it

Outline

Web logs

Data sets

Exploratory analysis

Navigation profiles

Conclusions

Web logs

Information about Web servers traffic, usage patterns and behavior of the visitors

Challenging analysis:

- very busy servers
- *human* users vs *crawlers*
- *ethical* crawlers vs *malicious* crawlers

Useful input for many engineering and marketing activities

What do Web logs tell us???

Data sets

Logs collected during more than one year on two Web servers

- academic
- SPEC mirror

Very different in terms of potential users and traffic

SPEC log:

~ 970 MBytes & 5.1 million transactions

Academic log:

~ 50 MBytes & 240,000 transactions

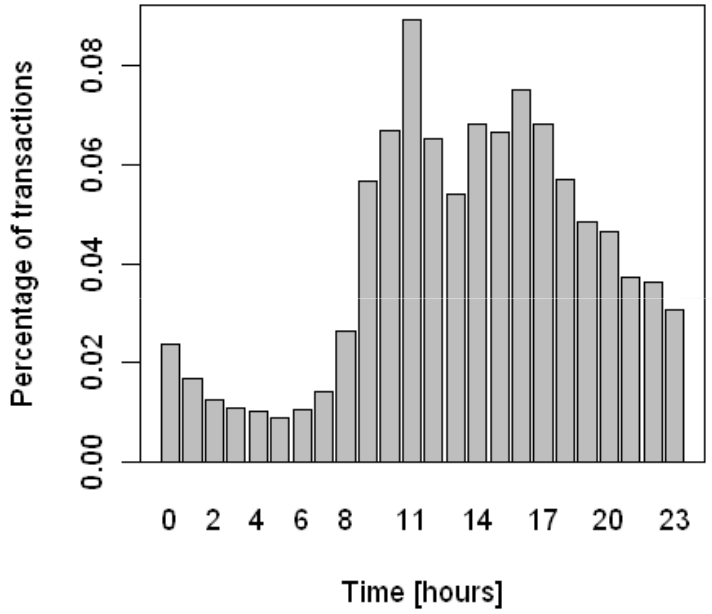
Web Logs

Extended Log File Format

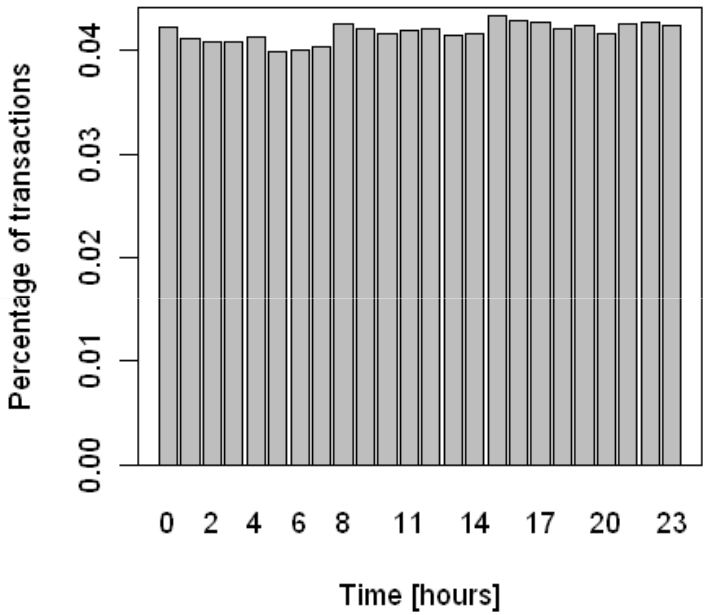
```
67.195.111.181 - - [30/May/2010:05:21:03 +0200] "GET /publications.html
HTTP/1.1" 200 3178 "http://peg.unipv.it" "Mozilla/5.0 (compatible; Yahoo!
Slurp/3.0; http://help.yahoo.com/help/us/ysearch/slurp)"

207.46.204.177 - - [30/May/2010:15:42:56 +0200] "GET / HTTP/1.1" 200 3195
"-" "msnbot/2.0b (+http://search.msn.com/msnbot.htm)"

208.65.73.109 - - [02/Jun/2010:15:54:37 +0200] "GET /IEEE93.pdf HTTP/1.1"
200 249805 "-" "Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US)
AppleWebKit/533.4 (KHTML, like Gecko) Chrome/5.0.375.55 Safari/533.4"

82.165.130.74 - - [03/Jun/2010:01:23:29 +0200] "GET /README HTTP/1.1" 404
335 "-" "Morfeus strikes again."
```

Exploratory analysis: hourly traffic

academic server



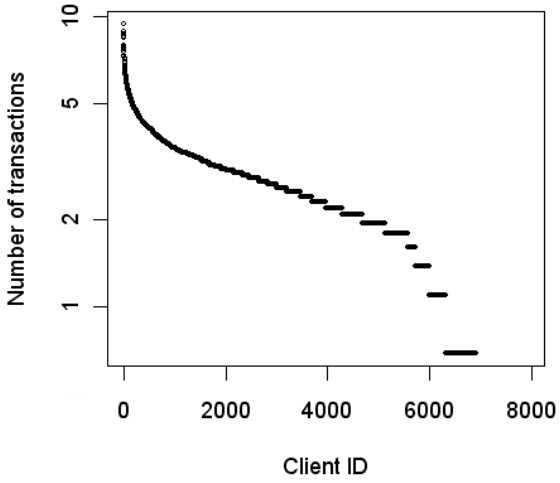
SPEC server



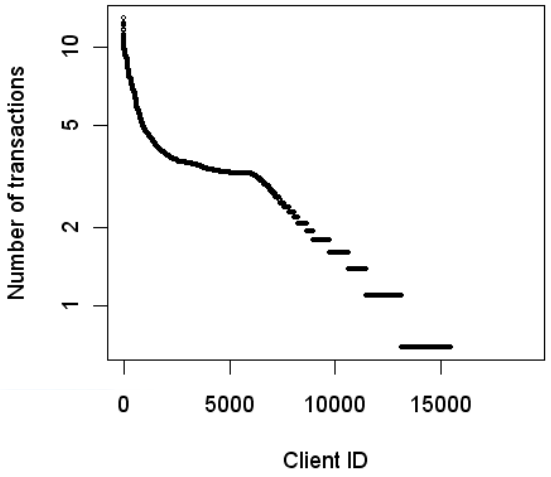
What is the behavior of individual clients???

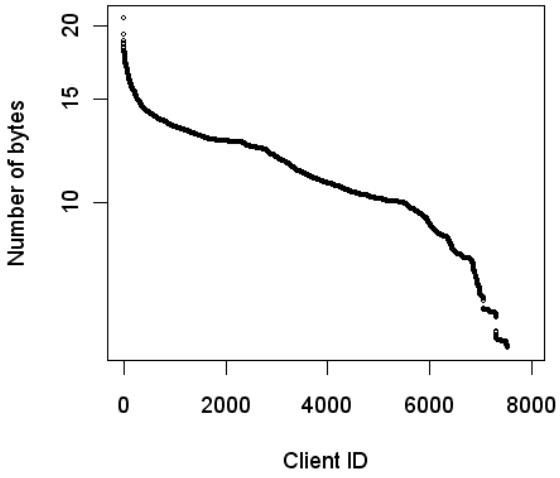
Client behavior

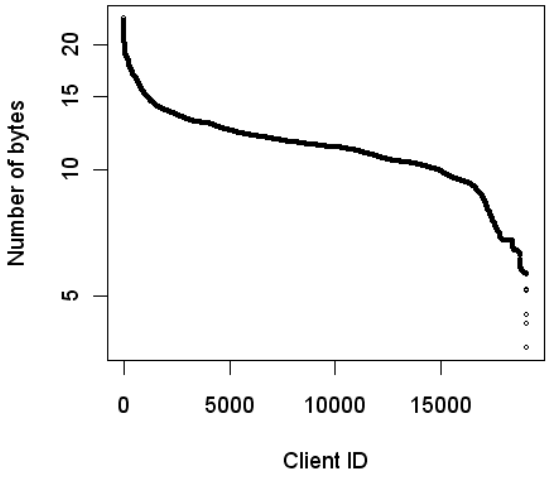
academic server



SPEC server







Who are the crawlers?

About 12% of the clients identified as (ethical) crawlers

- ~ 15% of transactions of academic server
- > 90% of transactions of SPEC server

Top crawlers: Google, Microsoft and Yahoo (80% of crawlers traffic)

Crawlers tend to reduce their impact on server resources

- inter-reference time between consecutive requests of a given client rather large
- no bursts
- re-visit patterns

Navigation profiles

Traffic of each client and its temporal distribution

Parameters describing the distribution of the requests across months, days and hours

Similarities among clients and periodic patterns

Multivariate analysis techniques applied in combination:

- *Principal Component Analysis* to reduce the number of dimensions
- *Hierarchical clustering* to classify clients according to their patterns
- *Correspondence Analysis* to display clients and parameters

Principal Components Analysis

Traffic of individual clients (crawlers vs other clients) described by 26 parameters

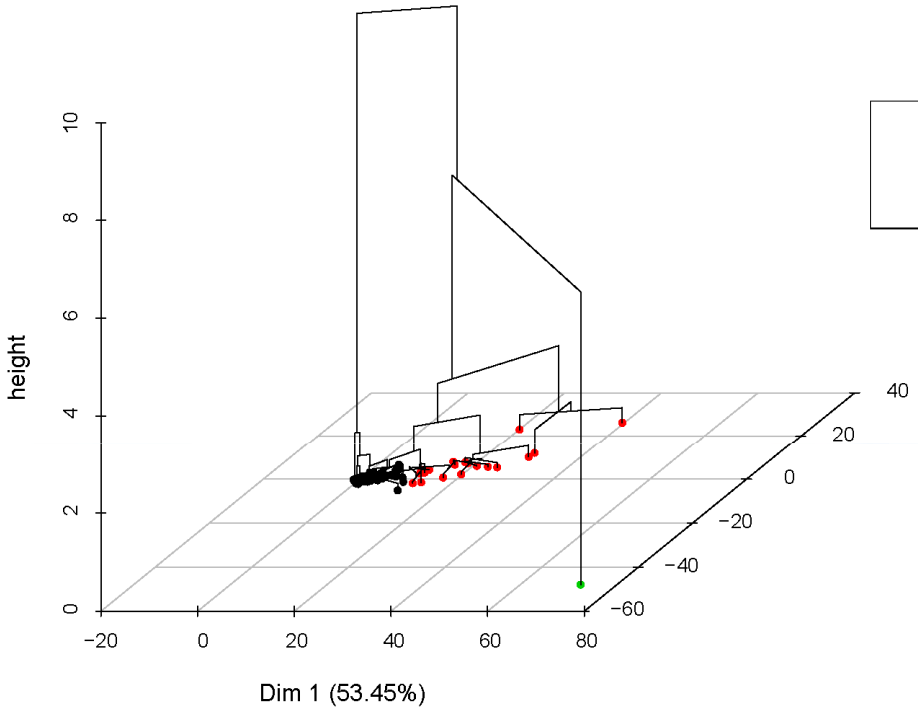
For crawlers, two PCs summarize 70% of the variance:

- first PC: traffic evenly distributed during the 24 hours
- second PC: contrast between day and night traffic

For other clients, two PCs account for 55% of the variance:

- first PC: business hours traffic
- second PC: “sporadic” traffic

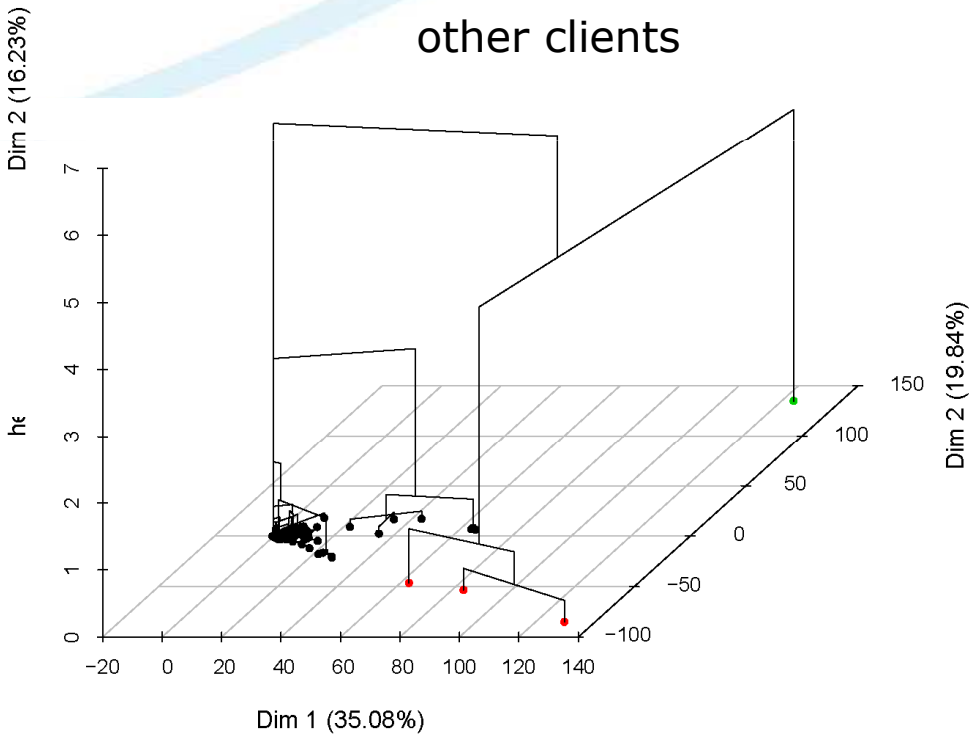
Hierarchical clustering



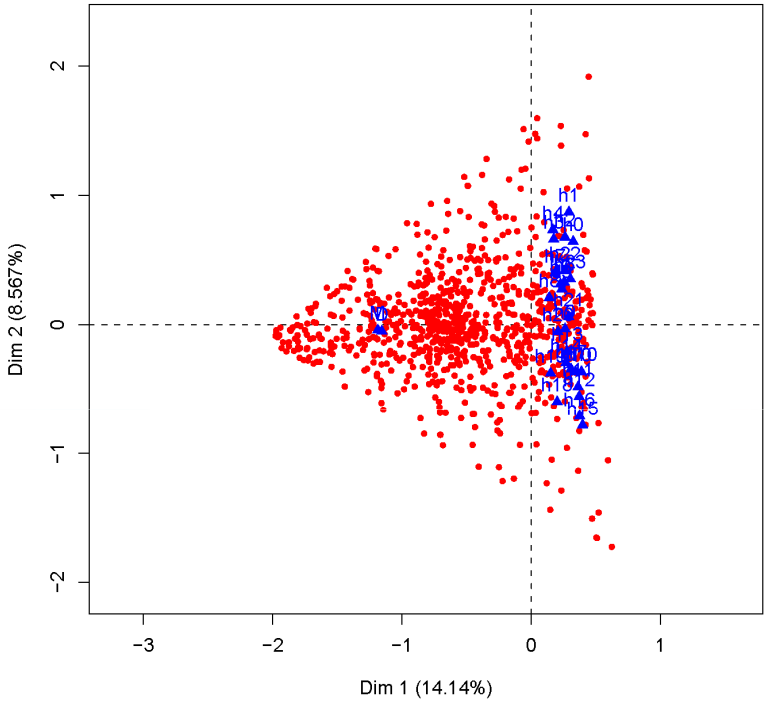
crawlers

cluster 1
cluster 2
cluster 3

other clients

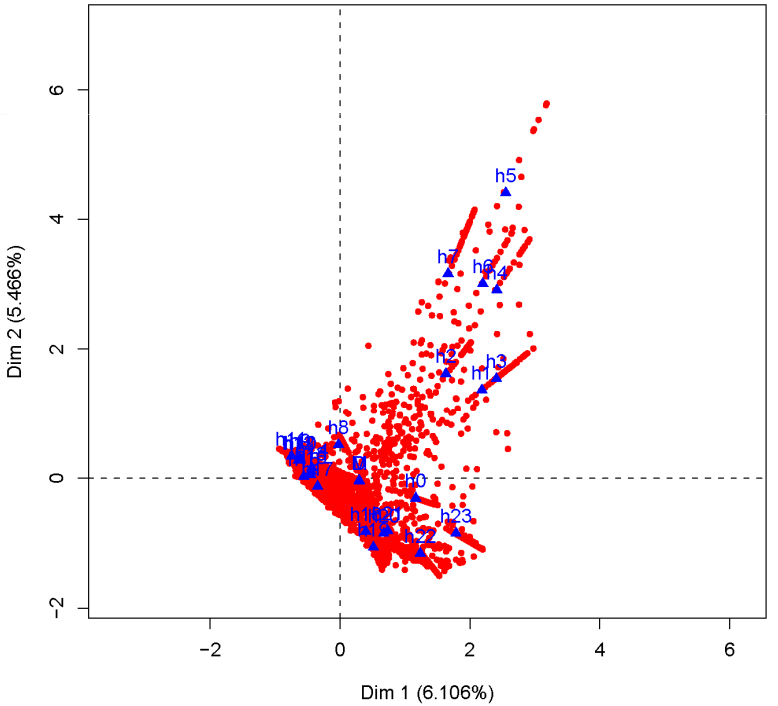


Correspondence Analysis



crawlers

other clients



Conclusions

Web logs store many important and useful information

Access patterns and navigation profiles of crawlers are rather homogeneous

Human users and unidentified crawlers are characterized by different behavior

Some clients visit a site to exploit vulnerabilities

Future work:

- better identification of the clients
- classification of malicious crawlers
- development of proactive policies