# Unsupervised Clustering of Web Sessions to Detect Malicious and Non-malicious Website Users

**ANT 2011**

Dusan Stevanovic

York University, Toronto, Canada

September 19th, 2011

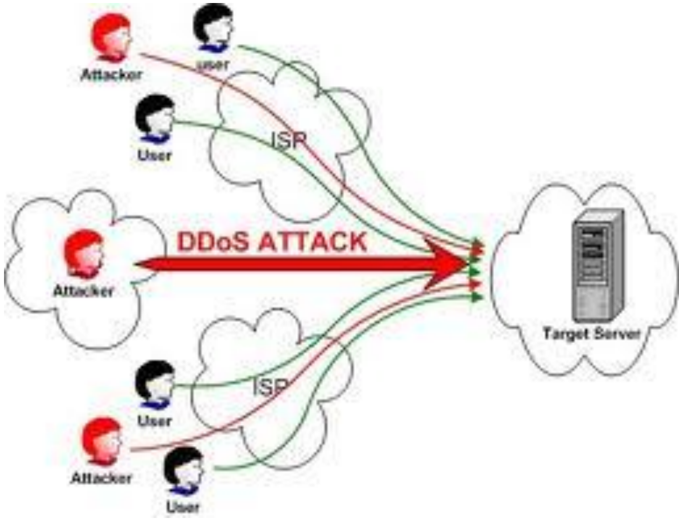
# Outline

- Denial-of-Service and Web Crawler Detection
- Related Work
- Study’s Objective
- Self Organized Maps and Adaptive Resonance Theory
- Experimental Design
- Experimental Results
- Conclusions and Final Remarks

# Denial of Service

- Security is built on top of three operational aspects of computer systems: confidentiality, integrity and availability

- (Distributed) Denial of Service (DoS) is an attack on the availability of data

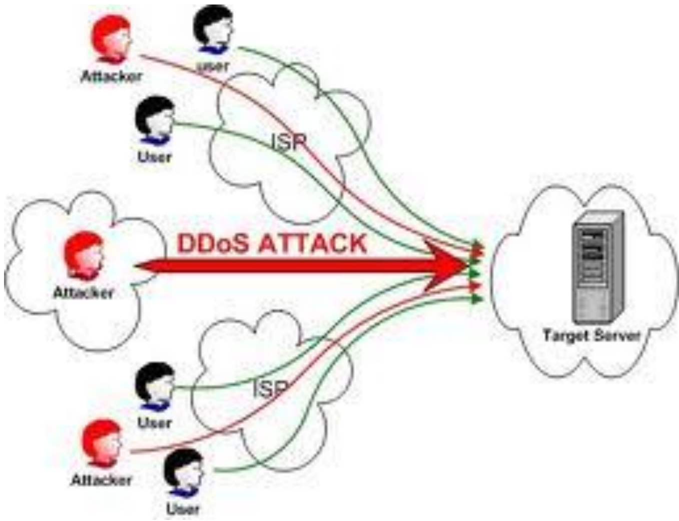


- The denial-of-service effect is achieved by sending messages to the target that interfere with its operation, and make it crash, reboot, freeze or do useless work

- Motivation can be both political and economical

# Application Layer DDoS and SPAM Mail

- So-called Application Layer DDoS attacks are caused by sending a flood of legitimate HTTP messages to the victim
- Are very hard/impossible to detect/differentiate between malicious packets and legitimate packets
- Web bots can be responsible for generating Application Layer DDoS traffic
- Or can be used to scrap email addresses from websites or click fraud



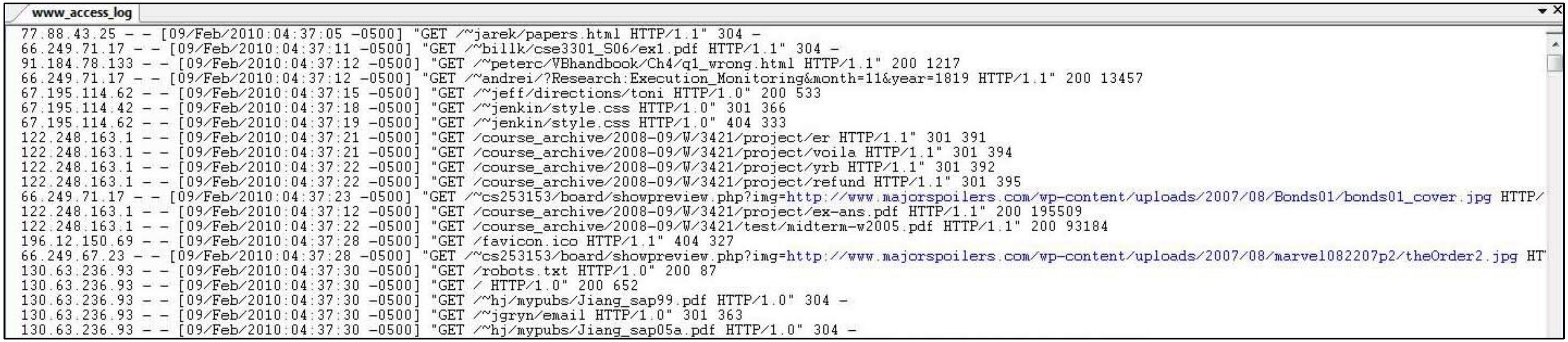
# Related Work – Session Identification

- User Navigation Session - A sequence of page requests such that no two consecutive requests are separated by more than 30 minutes

- A log file can be seen as a time-ordered set of web pages requests

  - **Fields**: IP address, Timestamp of the request, document requested, size of the file requested, HTTP code returned by the server, User Agent (browser or crawler ID), referrer page, HTTP method (GET, HEAD or POST)

  - E.g.: “*122.248.163.1 - - [09/Feb/2010:04:37:38 -0500] "GET /course_archive/2008 09/W/3421/test/testTwoPrep.html HTTP/1.1" 200 5645 Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)*”

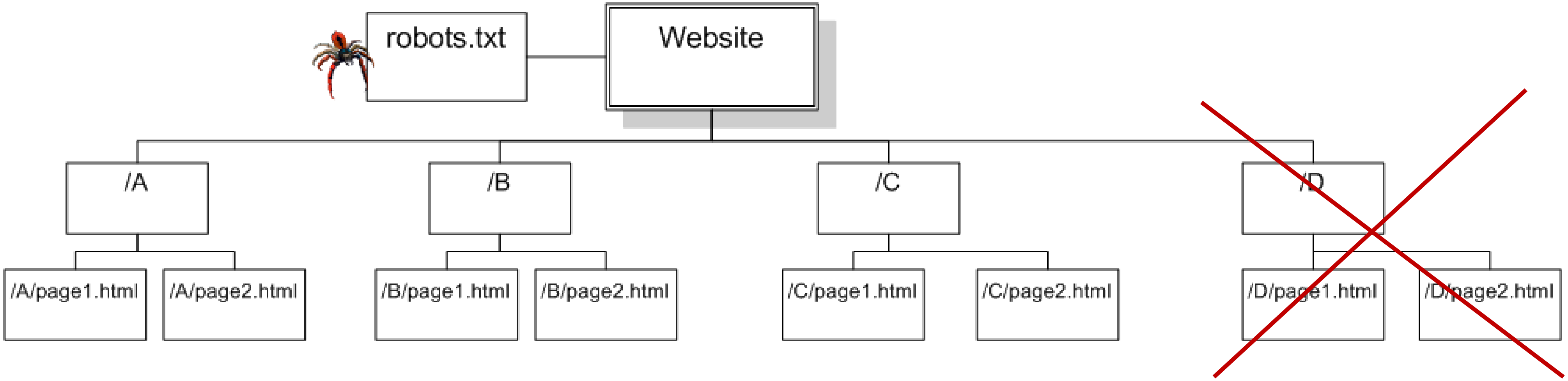


```
77.88.43.25 - - [09/Feb/2010:04:37:05 -0500] "GET /~jarek/papers.html HTTP/1.1" 304 -
66.249.71.17 - - [09/Feb/2010:04:37:11 -0500] "GET /~billk/cse3301_S06/ex1.pdf HTTP/1.1" 304 -
91.184.78.133 - - [09/Feb/2010:04:37:12 -0500] "GET /~peterc/VBhandbook/Ch4/q1_wrong.html HTTP/1.1" 200 1217
66.249.71.17 - - [09/Feb/2010:04:37:12 -0500] "GET /~andrei/?Research:Execution_Monitoring&month=11&year=1819 HTTP/1.1" 200 13457
67.195.114.62 - - [09/Feb/2010:04:37:15 -0500] "GET /~jeff/directions/toni HTTP/1.0" 200 533
67.195.114.42 - - [09/Feb/2010:04:37:18 -0500] "GET /~jenkin/style.css HTTP/1.0" 301 366
67.195.114.62 - - [09/Feb/2010:04:37:19 -0500] "GET /~jenkin/style.css HTTP/1.0" 404 333
122.248.163.1 - - [09/Feb/2010:04:37:21 -0500] "GET /course_archive/2008-09/W/3421/project/er HTTP/1.1" 301 391
122.248.163.1 - - [09/Feb/2010:04:37:21 -0500] "GET /course_archive/2008-09/W/3421/project/voila HTTP/1.1" 301 394
122.248.163.1 - - [09/Feb/2010:04:37:22 -0500] "GET /course_archive/2008-09/W/3421/project/yrb HTTP/1.1" 301 392
122.248.163.1 - - [09/Feb/2010:04:37:22 -0500] "GET /course_archive/2008-09/W/3421/project/refund HTTP/1.1" 301 395
66.249.71.17 - - [09/Feb/2010:04:37:23 -0500] "GET /~cs253153/board/showpreview.php?img=http://www.majorspoilers.com/wp-content/uploads/2007/08/Bonds01/bonds01_cover.jpg HTTP/
122.248.163.1 - - [09/Feb/2010:04:37:12 -0500] "GET /course_archive/2008-09/W/3421/project/ex-ans.pdf HTTP/1.1" 200 195509
122.248.163.1 - - [09/Feb/2010:04:37:22 -0500] "GET /course_archive/2008-09/W/3421/test/midterm-w2005.pdf HTTP/1.1" 200 93184
196.12.150.69 - - [09/Feb/2010:04:37:28 -0500] "GET /favicon.ico HTTP/1.1" 404 327
66.249.67.23 - - [09/Feb/2010:04:37:28 -0500] "GET /~cs253153/board/showpreview.php?img=http://www.majorspoilers.com/wp-content/uploads/2007/08/marvel082207p2/theOrder2.jpg HT
130.63.236.93 - - [09/Feb/2010:04:37:30 -0500] "GET /robots.txt HTTP/1.0" 200 87
130.63.236.93 - - [09/Feb/2010:04:37:30 -0500] "GET / HTTP/1.0" 200 652
130.63.236.93 - - [09/Feb/2010:04:37:30 -0500] "GET /~hj/mypubs/Jiang_sap99.pdf HTTP/1.0" 304 -
130.63.236.93 - - [09/Feb/2010:04:37:30 -0500] "GET /~jgryn/email HTTP/1.0" 301 363
130.63.236.93 - - [09/Feb/2010:04:37:30 -0500] "GET /~hj/mypubs/Jiang_sap05a.pdf HTTP/1.0" 304 -
```

# Related Work – Web Crawler Detection

- Classifying automated web clients such as robots and crawlers from web server access logs

- Differentiate between automated script and human user by analyzing web log files
  - Was robots.txt file requested?
  - Percentage of 4xx error responses / invalid requests
  - Click rate – HTML page requests over time



# Related Work – Features

- Features identified for each session include:
  - HTML to Image ratio
  - % of HEAD requests
  - % of Unassigned Referrers
  - Popularity Index
- Are there any other features that can improve the detection of crawlers?

# Study's Objective

- Apply Unsupervised Neural Network (NN) Learning
- NN Learning can be employed to cluster web sessions
- Gain better insight into the types and distribution of visitors to a public website
- Investigate the relative differences and/or similarities between malicious web crawlers and other non-malicious visitor groups
- Employ SOM and Modified ART2 – two unsupervised NN algorithms

# SOM and Modified ART2

- Self-Organized Map (SOM) advantages:
  - Ability to produce natural clustering, i.e. clustering that is robust to statistical anomalies
  - Unlike other clustering methods, it achieve superior visualization of high-dimensional input data in 2D-representation space

- Modified Adaptive Resonance Theory 2 (Modified ART2) advantages:
  - Exposure to new training data does not destroy previously learned information
  - Ability to identify underrepresented but significant clusters

# Classification Features

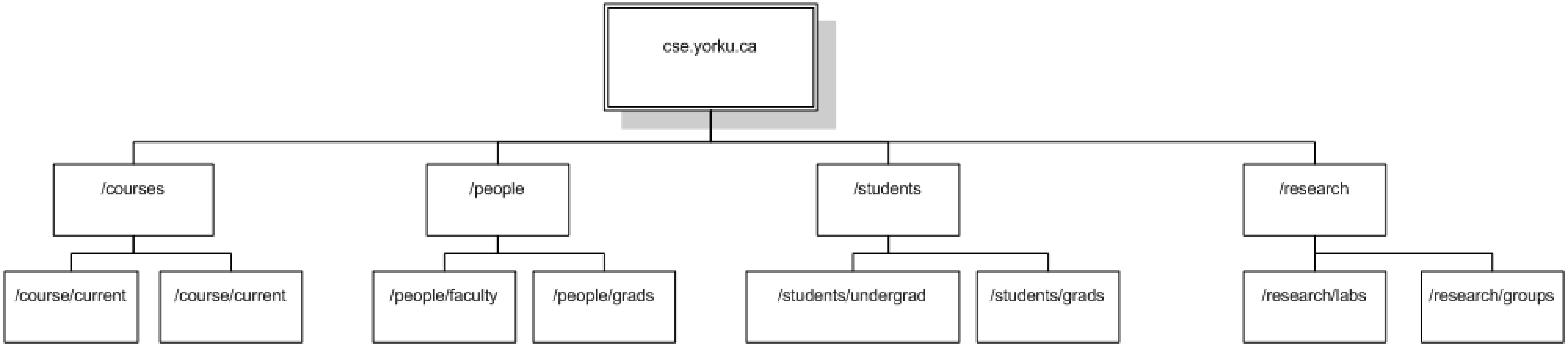
- Traditional features used in the Experiments:

1. Click Number (# of request in a session)
2. HTML to Image ratio
3. % of PDF
4. % of HEAD requests
5. % of ERROR requests (with HTTP code between 400 and 500)
6. % of Unassigned Referrers
7. Popularity Index

# Classification Features – Novel Features

8. Consecutive Sequential HTTP Request Ratio – human users using browser would typically request an HTML file and relevant image files and scripts while a web crawler would request only the HTML page
   - A series of requests for web pages matching pattern '/cshome/course/*.*
   - '/cshome/index.html' and then 'cshome/courses/index.html' would not be considered consecutive

9. Standard Deviation of Page Request Depth - should be low for web robot sessions since a web robot should scan over a narrower directory structure of a web site than a human users
   - '/cshome/courses/index.html' – depth = 3
   - '/cshome/calendar.html' – depth = 2



# Experimental Design – Log Pre-processor/Analyzer

- Builds training datasets by pre-processing York University’s CSE department server log file
  - 4 weeks of server activity

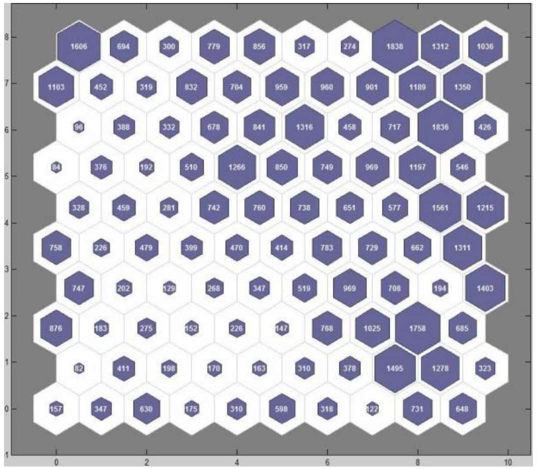
- Pre-label each session depending on the contents of User Agent String
  - User Agent string info can be found on web sites www.user-agents.org and botsvsbrowsers.com

| | |
|---|---|
| **# of Human Sessions** | **53640** |
| **# of Well-behaved Crawler Sessions** | **7607** |
| **# of Malicious Crawler Sessions** | **287** |
| **# of Unknown Visitor Sessions** | **4042** |
| **Total Number of Sessions** | **65576** |

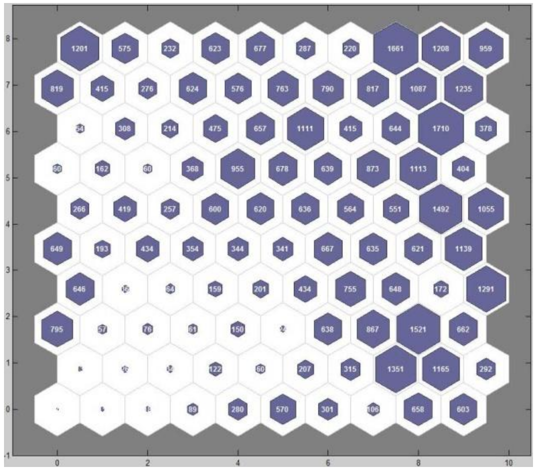
# Experimental Design – Clustering Parameters

- SOM implementation provided within MATLAB as a part of Neural Network Toolbox software package
- SOM comprising 100 neurons in 10-by-10 hexagonal arrangement
- Modifed ART2 algorithm parameters: $\rho_{max} = 1.5$, $\Delta\rho = 0.1$ and $n_{max} = 5$
- All input vectors were normalized prior to being fed to SOM and Modified ART2

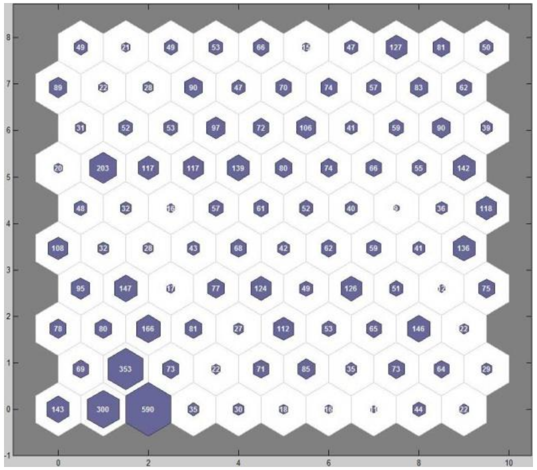
# Experimental Results – SOM Clustering



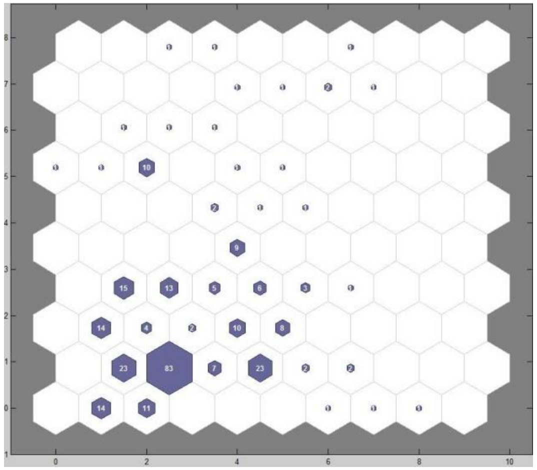
All Session Neuron Hits



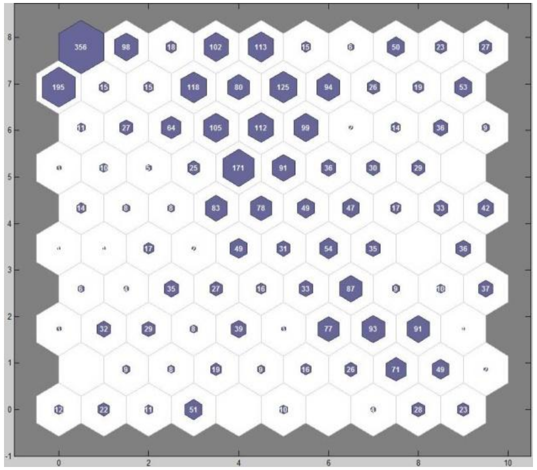
Human visitor Session Hits



Well-behaved Crawler Session Hits

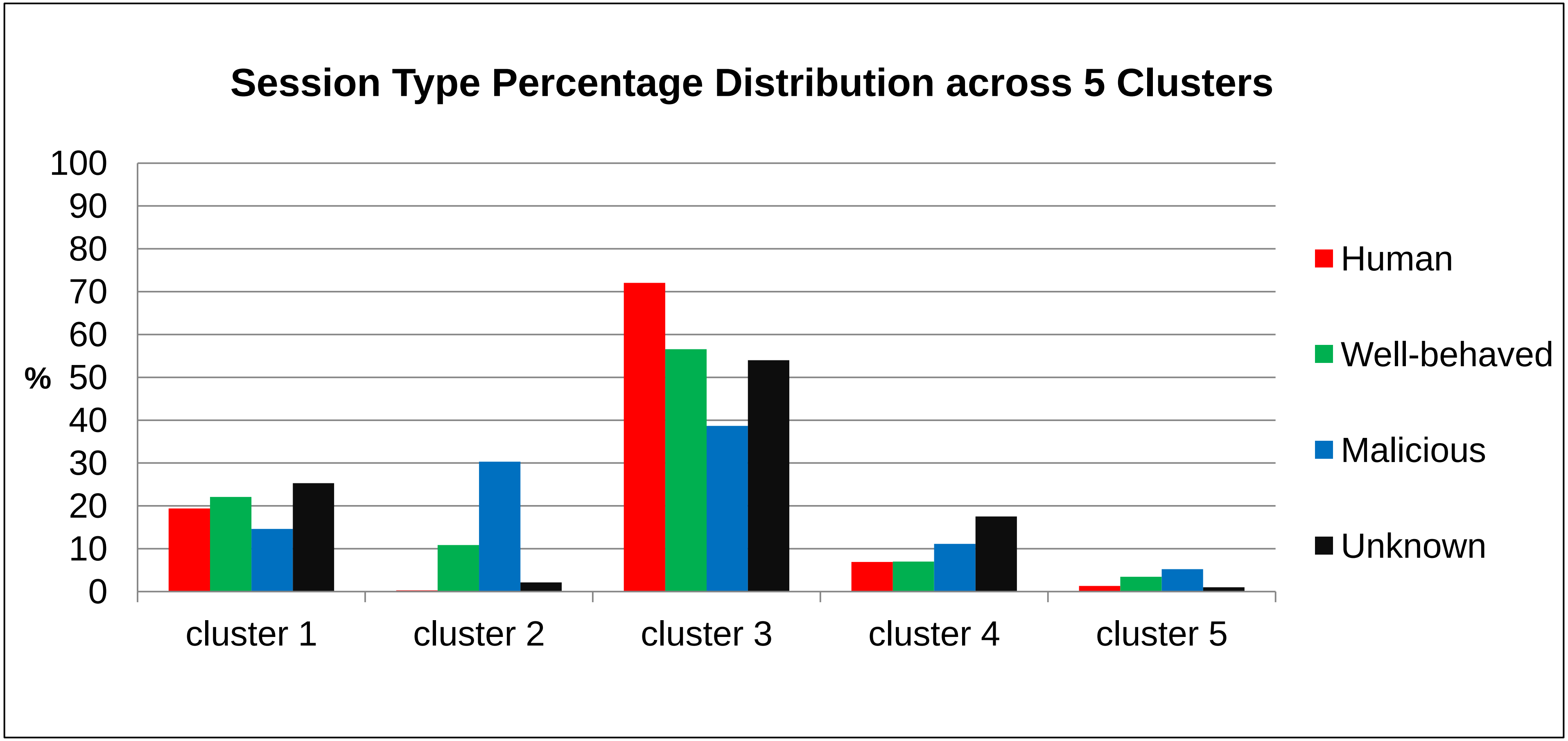


Malicious Crawler Session Hits



Unknown Visitor Session Hits

# Experimental Results – Modified ART2 Clustering



# Conclusions and Final Remarks

- There exists a pretty good separation between malicious and non-malicious web users in terms of their browsing behaviour
- While human visitors tend to follow rather similar browsing patterns malicious web crawlers exhibit a range of browsing strategies
- Moreover, nearly 52% of malicious web crawlers exhibit very much ‘human-like’ browsing behaviour
- With a higher level of sophistications, these crawlers could pose a serious challenge for future web-site security systems
- Also 10% of sessions labelled as belonging to humans exhibit malicious-like browsing behaviour

# Questions?

Thank you!

# References

1. C. Wilson, Botnets, Cybercrime, and Cyberterrorism: Vulnerabilities and Policy Issues for Congress, Foreign Affairs, Defense, and Trade Division, United States Governemnt, CRS Report for Congress, 2008.

2. Prolexic Technologies, Evolving Botnet Capabilities - and What This Means for DDoS, White Paper, 2010.

3. Y. Xie and S.-Z. Yu, Monitoring the Application-Layer DDoS Attacks for Popular Websites, IEEE/ACM Transactions on Networking, vol. 17, no. 1, pp. 15-25, Feb. 2009.

4. 4G. Oikonomou and J. Mirkovic, Modeling Human Behavior for Defense against Flash-Crowd Attacks, in In Proceedings of IEEE International Conference on Communications, Dresden, Germany, 2009, pp. 1-6.

5. P. Hayati, V. Potdar, K. Chai, and A. Talevski, Web spambot detection based on web navigation behaviour, in International Conference on Advanced Information Networking and Applications, Perth, Australia, 2010, pp. 797-803.

6. C. Bomhardt, W. Gaul, and L. Schmidt-Thieme, Web Robot Detection - Preprocessing Web Logfiles for Robot Detection, in In Proc. SISCLADAG, Bologna, Italy, 2005.

7. K. Park, V. Pai, K. Lee, and S. Calo, Securing Web Service by Automatic Robot Detection, in Proceedings of the annual conference on USENIX '06 Annual Technical Conference, Berkeley, CA, 2006, pp. 23-29.

8. T. Kohonen, Self-Organizing Maps, 3rd ed. New York: Springer-Verlag, Berlin Heidelberg, 2001.

9. N. Vlajic and H. C. Card, Vector quantization of images using modified adaptive resonance algorithm for hierarchical clustering, IEEE Transactions on Neural Networks, vol. 12, no. 5, pp. 1147-1162, Sep. 2001.