# Traffic capture Kyoto University's Honeypots

**Dataset files**

* <http://www.takakura.com/Kyoto_data/data_with_IP/>

We used following file for checking: <http://www.takakura.com/Kyoto_data/data_with_IP/2009/200901.tar.gz>

**Dataset entry description**

**Popular features taken from KDD dataset**

Attribute 1 (Duration) : the length (number of seconds) of the connection

Attribute 2 (Service) : the connection’s service type, e.g., http, telnet, etc

Attribute 3 (Source bytes) : the number of data bytes sent by the source IP address

Attribute 4 (Destination Bytes): the number of data bytes sent by the destination IP

Address

Attribute 5 (Count) : the number of connections whose source IP address and

destination IP address are the same to those of the current

connection in the past two seconds

Attribute 6 (Same srv rate) : % of connections to the same service in Count feature

Attribute 7 (Serror rate) : % of connections that have “SYN” errors in Count feature

Attribute 8 (Srv serror rate) : % of connections that have “SYN” errors in Srv count(the

number of connections whose service type is the same to

that of the current connection in the past two seconds) feature

Attribute 9 (Dst host count) : among the past 100 connections whose destination IP

address is the same to that of the current connection, the number of connections whose source IP address is also the same to that of the current connection

Attribute 10 (Dst host srv count): among the past 100 connections whose destination IP

address is the same to that of the current connection, the

number of connections whose service type is also the

same to that of the current connection

Attribute 11 (Dst host same src port rate):% of connections whose source port is the same to

that of the current connection in Dst host count feature

Attribute 12 (Dst host serror rate): % of connections that have “SYN” errors in Dst host count

feature

Attribute 13 (Dst host srv serror rate):% of connections that “SYN” errors in Dst host srv count

feature

Attribute 14 (Flag) : the state of the connection at the time the summary was

written (which is usually when the connection terminated).

The different states are summarized in the below section.

**Features derived from other AVs, sanitised IP/ports and labelling**

Attribute 15 (IDS detection) : Reflects whether IDS(Intrusion Detection System)

triggered an alert for the connection;

* ‘0’ means any alerts were not triggered,
* an arabic numeral(except ‘0’) means the different kinds of the alerts. Parenthesis indicates the number of the same alert observed during the connection.

We used Symantec IDS[3] to extract this feature.

Attribute 16 (Malware detection): indicates whether malware, also known as malicious

software, was observed in the connection;

* ‘0’ means no malware was observed,
* a string indicates the corresponding malware observed at the connection.

We used ‘clamav’ software to detect malwares.

Parenthesis indicates the number of the same malware

observed during the connection.

Attribute 17 (Ashula detection): means whether shellcodes and exploit codes were used in

the connection by using the dedicated software[4];

* ‘0’ means no shellcodes and exploit codes were observed,
* an arabic numeral(except ‘0’) means the different kinds of the shellcodes or exploit codes. Parenthesis indicates the number of the same shellcode or exploit code observed during the connection.

Attribute 18 (Label) : indicates whether the session was attack or not;

* ‘1’ means the session was normal,
* ‘-1’ means known attack was observed in the session,
* ‘-2’ means unknown attack was observed in the session.

Attribute 19 (Source IP Address): indicates the source IP address used in the session. Due

to the security concerns, the original IP address on IPv4

was properly sanitized to one of the Unique Local IPv6

Unicast Addresses (private IP addresses)[5]. Also, the

same private IP addresses are only valid in the same

month: if two private IP addresses are the same within the

same month, it means their IP addresses on IPv4 were

also the same, but if two private IP addresses are the

same within the different month, their IP addresses on

IPv4 are also different.

Attribute 20 (Source Port Number): indicates the source port number used in the session.

Attribute 21 (Destination IP Address) : indicates the source IP address used in the session. Due

to the security concerns, the original IP address on IPv4

was properly sanitized to one of the Unique Local IPv6

Unicast Addresses (private IP address)[5]. Also, the same

private IP addresses are only valid in the same month: if

two private IP addresses are the same within the same month, it means their IP addresses on IPv4 were also the same, but if two private IP addresses are the same within the different month, their IP addresses on IPv4 are also different.

Attribute 22 (Destination Port): indicates the destination port number used in the

session.

Attribute 23 (Start Time) : indicates when the session was started.

Attribute 24 (Duration) : indicates how long the session was being established.

**Example entry**

0.00 13 0 0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 RSTRH 0 0 0 -1 fd59:1382:86f1:a0a3:7a40:30d3:3039:7c29 4918 fd59:1382:86f1:a7a3:34cb:45b1:2b1e:197b 80 00:00:02 0.000000

**Labelling field**

Attribute 18 (Label)

Possible values

‘1’ means the session was normal,

‘-1’ means known attack was observed in the session,

‘-2’ means unknown attack was observed in the session.

**Questions to answer using ML**

1. Known attacks happened on the network?
2. What were the IPs and ports in case of unknown attacks?
3. What is the percentage of session in which known attacks took place in total sessions?

**Reference**

* <http://www.takakura.com/Kyoto_data/>