# Machine Learning for Cyber-Security & Artificial Intelligence

## Part 1 – The Kyoto dataset

Hermes Senger

# Statistical Analysis of Honeypot Data and Building of Kyoto 2006+ Dataset for NIDS Evaluation

Jungsuk Song
National Institute of Information and Communications Technology (NICT)
song@nict.go.jp

Hiroki Takakura
Information Technology Center, Naogya University
takakura@itc.nagoya-u.ac.jp

Yasuo Okabe
Academic Center for Computing and Media Studies, Kyoto University
okabe@i.kyoto-u.ac.jp

Masashi Eto
National Institute of Information and Communications Technology (NICT)
eto@nict.go.jp

Daisuke Inoue
National Institute of Information and Communications Technology (NICT)
dai@nict.go.jp

Koji Nakao
National Institute of Information and Communications Technology (NICT)
ko-nakao@nict.go.jp

**Abstract**

With the rapid evolution and proliferation of botnets, large-scale cyber attacks such as DDoS, spam emails are also becoming more and more dangerous and serious cyber threats. Because of this, network based security technologies such as Network based Intrusion Detection Systems (NIDSs), Intrusion Prevention Systems (IPSs), firewalls have received

***Keywords*** NIDS, Honeypot Data, Kyoto 2006+ Dataset

## 1. Introduction

In general, a botnet is referred as a collection of infected hosts, *i.e.*, zombie PCs or bots, and the botnet herders use their botnets for launching large-scale cyber attacks such as

- Song, Jungsuk, et al. "Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation." *Proceedings of the first workshop on building analysis datasets and gathering experience returns for security*. 2011.

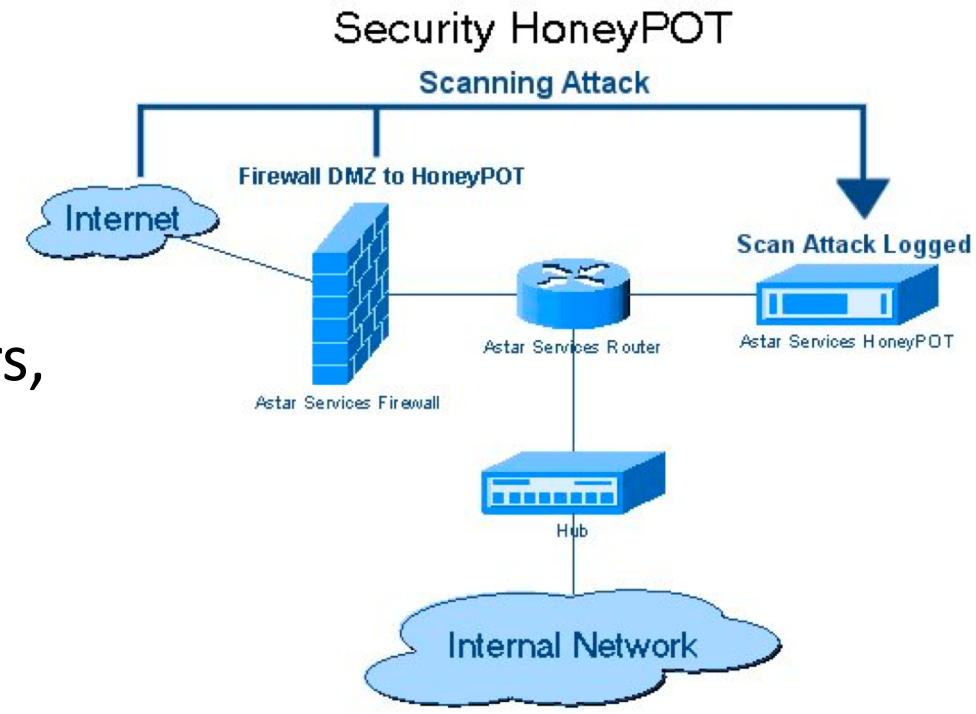
http://www.takakura.com/Kyoto_data/

# Datasets

- Traffic Data from Kyoto University's Honeypots
  - Data collected from 2006 to December 2015
  - collected from 348 Honeypots of 8 different types
  - 24 attributes + four label atribute
    - three of which were generated by commercial IDS with malware and shell script detection capabilities

# Honeypots

- Objective: to detect, deflect, or, in some manner, counteract attempts at unauthorized use of information systems.
- Consists of data (for example, in a network appears to be a legitimate part of the site that seems to contain information or a resource of value to attackers, but actually, is isolated and monitored and, enables blocking or analyzing the attackers.



# Datasets

- Traffic Data from Kyoto University's Honeypots
  - Data collected from 2006 to December 2015
  - collected from 348 Honeypots of 8 different types
    - Windows (several releases)
    - Linux / Unix (Solaris 8, MacOS X)
    - network printer
    - home appliances (eg, TV set, HDD Recorder)
  - 24 attributes + four label atribute
    - three of which were generated by commercial IDS with malware and shell script detection capabilities

# Honeypots

**Table 1.** Overview of honeypots

| Type | Number of machines |
|---|---|
| Solaris 8 (Symantec based) | 4 |
| Windows XP (full patch) | 1 |
| Windows XP (no patch) | 5 |
| Windows XP SP2 | 2 |
| Windows Vista | 1 |
| Windows 2000 Server | 1 |
| MacOS X | 2 (one is mail server) |
| Printer | 2 |
| TV set | 1 |
| HDD recorder | 1 |
| dedicated honeypots[4] | 5 |
| SGNET honeypots[15] | 4 |
| Web Crawler | 1 |
| Balck hole sensor /24 | 1 |
| Balck hole sensor /26 | 1 |

# Data types

**Table 2.** Overall property of honeypot data

| | Number of sessions | Average number of sessions per day |
|---|---|---|
| Total | 93,076,270 | 93,638 |
| Normal | 50,033,015 | 50,335 |
| Known attack | 42,617,536 | 42,874 |
| Unknown attack | 425,719 | 428 |

# Normal, Known Attacks, Unknown Attacks

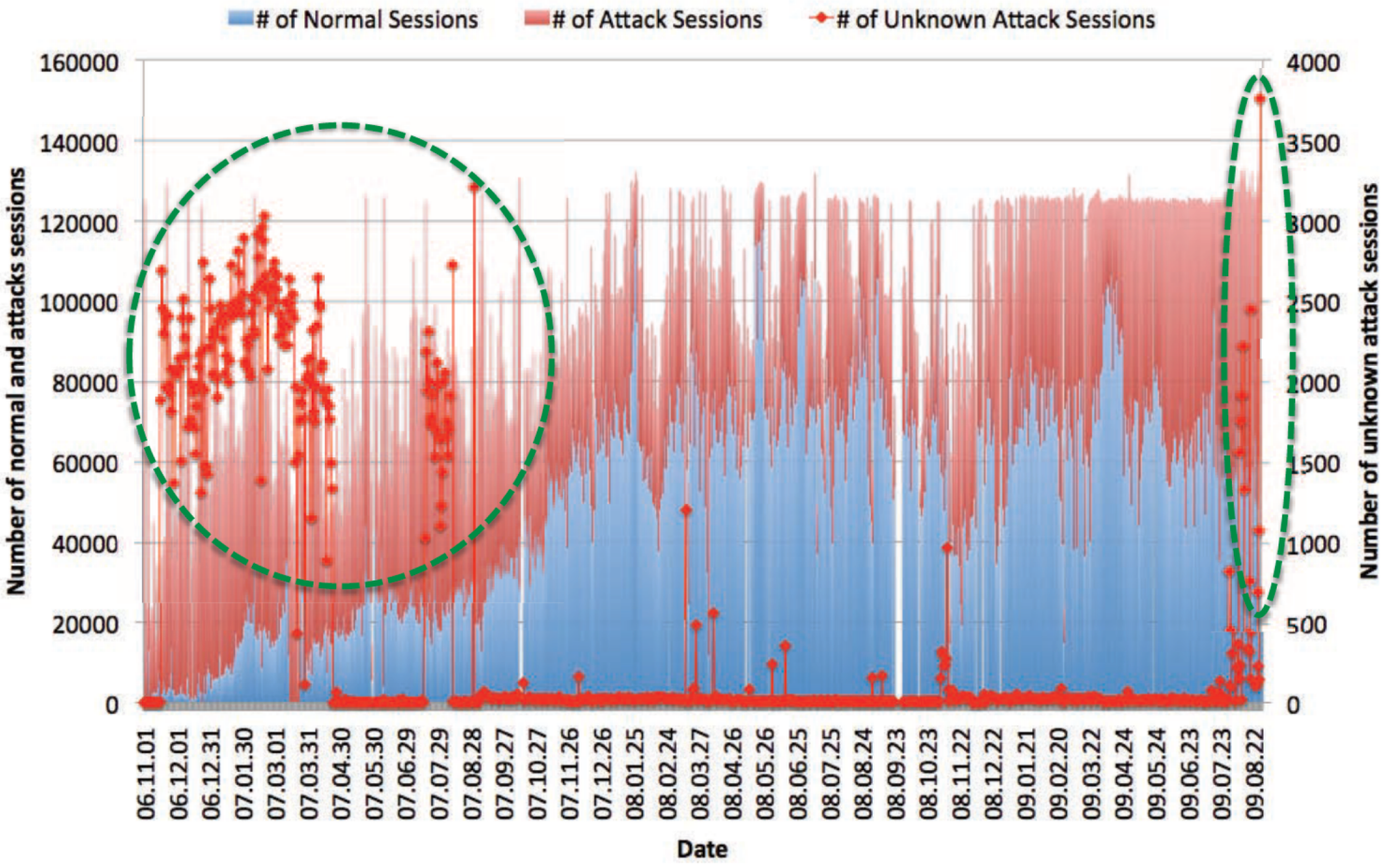



**Figure 1.** Distributions of normal, known attack and unknown attack sessions in honeypot data.

# Source distribution

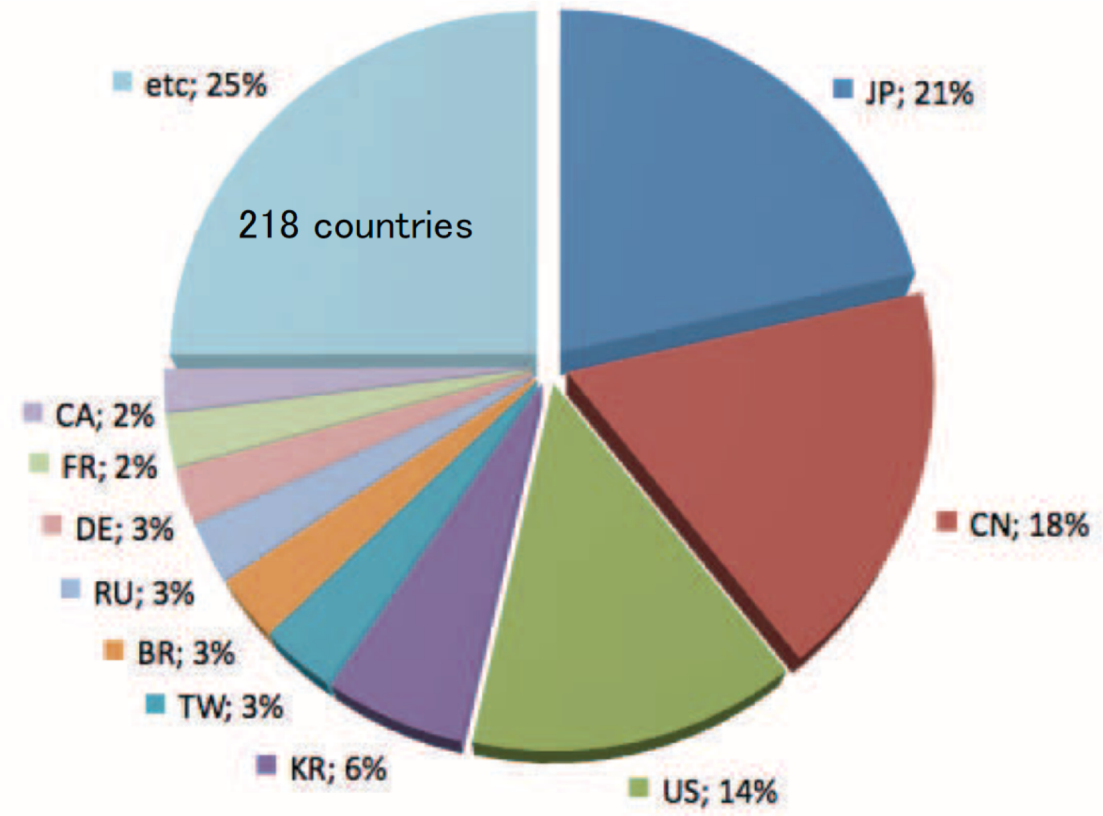
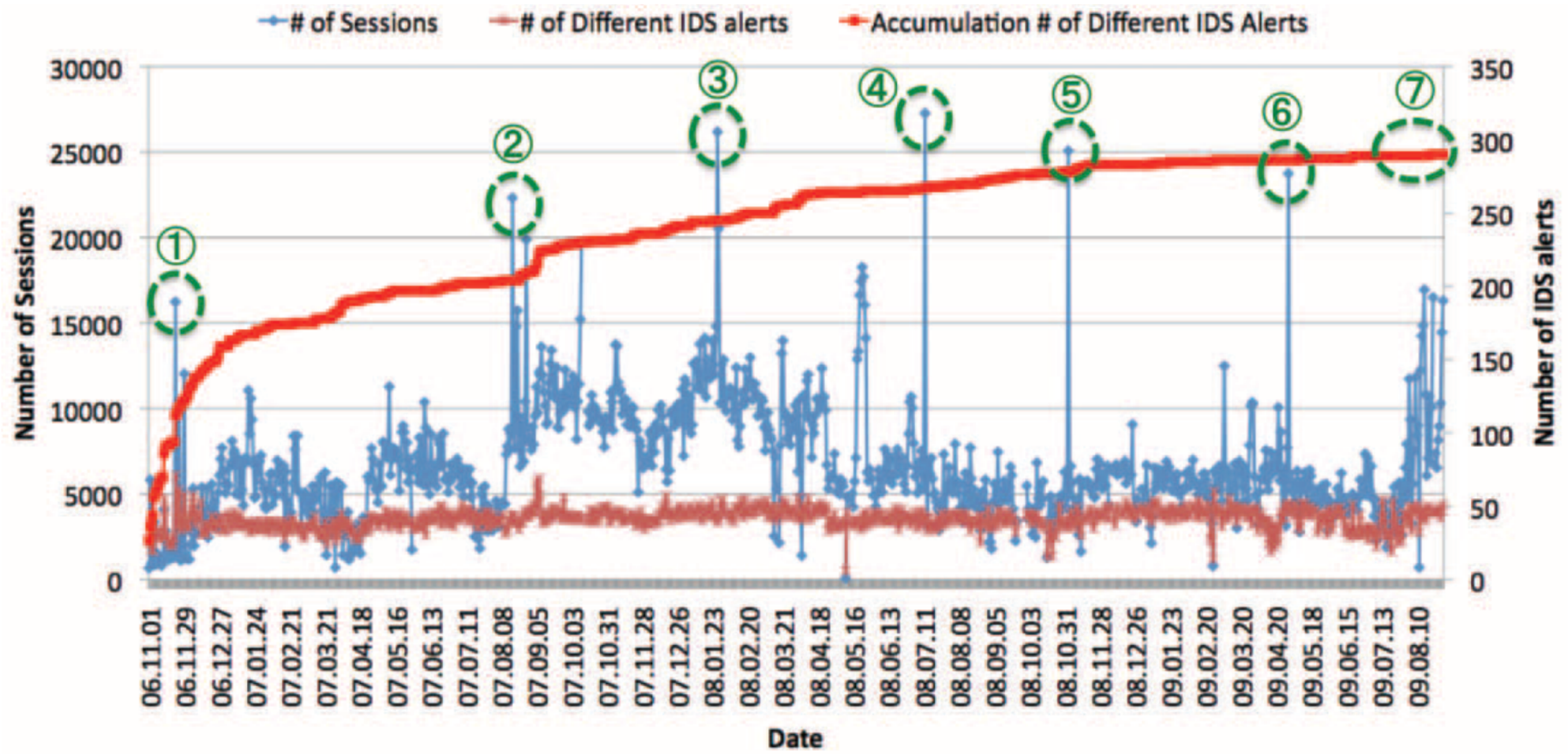
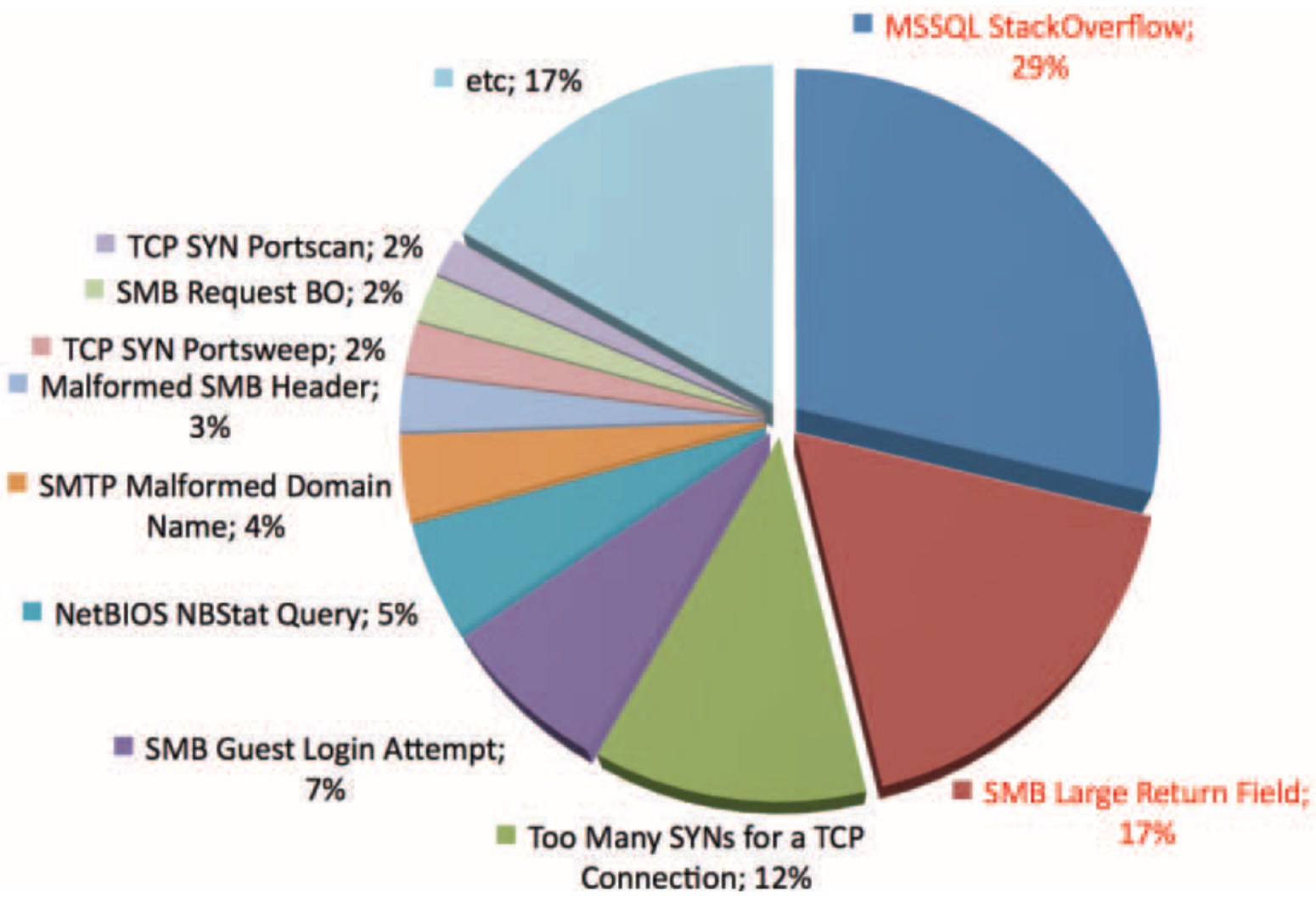



**Figure 2.** National distribution of attack source IP addresses.

# Statistics of IDS Sessions



# Distribution of IDS Alerts



# Location of Top 10 Source IPs

**Table 3.** Locations of top 10 source IP addresses in Japan

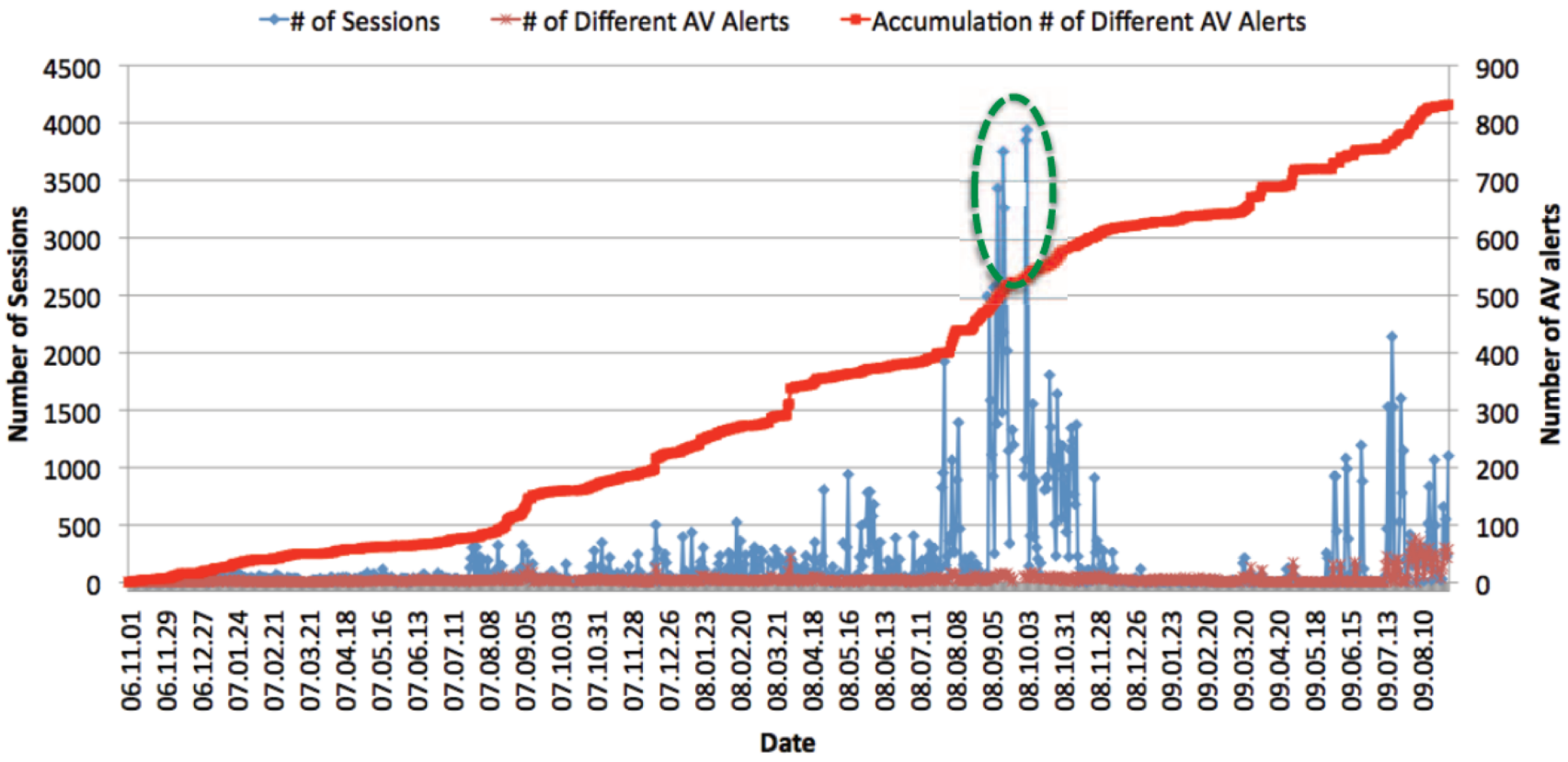
| IP address | Count | Location |
|---|---|---|
| x.x.x.1 | 1,932,303 | Darknet |
| x.x.x.2 | 530,039 | L2 Switch (Unicast Flooding) |
| x.x.x.3 | 377,599 | Darknet |
| x.x.x.4 | 355,607 | L2 Switch (Unicast Flooding) |
| x.x.x.5 | 170,182 | Honeypot (Windows 2k) |
| x.x.x.6 | 131,115 | Darknet |
| x.x.x.7 | 118,006 | No honeypot (MacOS X) |
| x.x.x.8 | 105,832 | Honeypot (Fedora Core) |
| x.x.x.9 | 100,824 | No honeypot |
| x.x.x.10 | 92,509 | Honeypot (Original WinXP) |

# Allert counts during 6 days

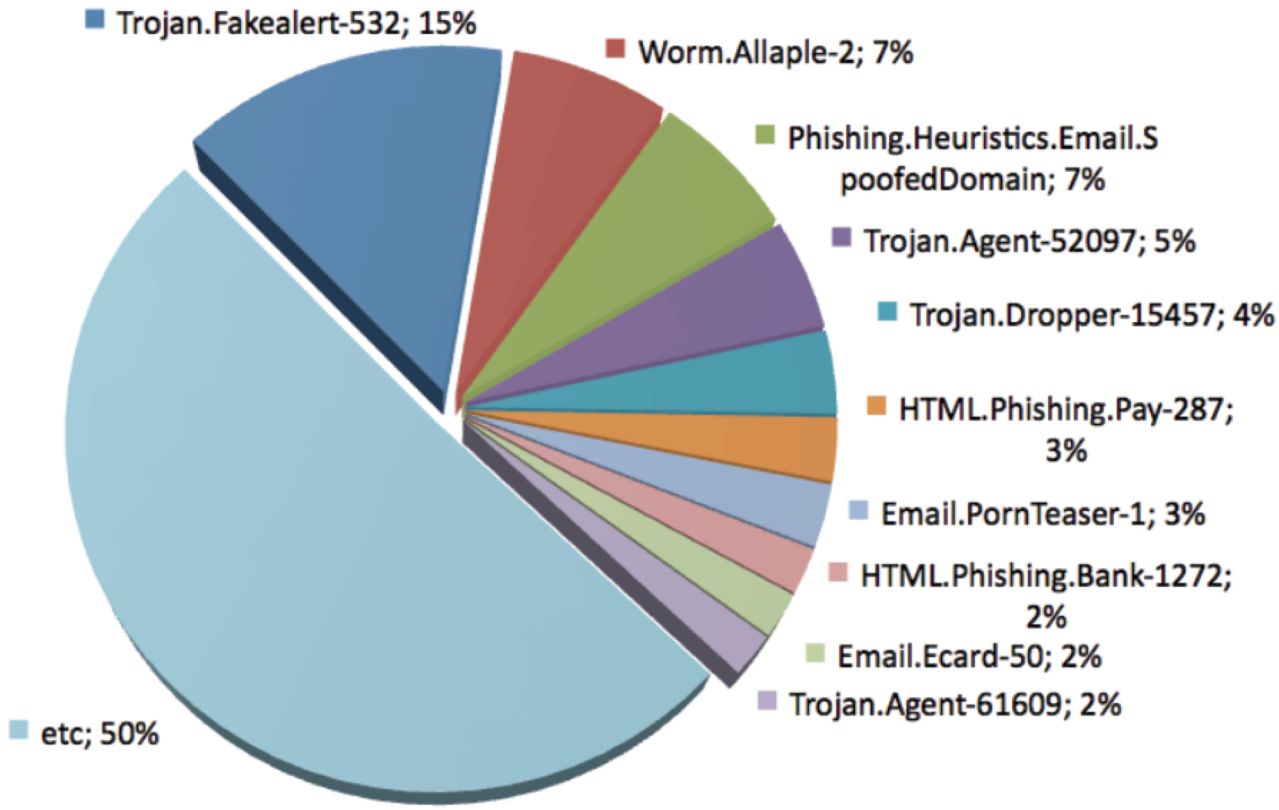
**Table 4.** IDS alerts observed during 6 days

| Date | Signature name | Count |
|---|---|---|
| ① | P2P BitTorrent Activity | 1,802 |
| | P2P Edonkey Start Upload Request | 2,867 |
| | Too Many SYNs for a TCP Connection | 3,011 |
| | Emule File Traffic Detected | 5,586 |
| | P2P eMule Hello | 5,369 |
| | P2P Emule Kademlia Request | 8,100 |
| ② | Too Many SYNs for a TCP Connection | 1,341 |
| | Out-of-Sequence TCP RST Packet | 4,779 |
| | Out-of-Sequence TCP SYN Packet | 13,859 |
| ③ | Too Many SYNs for a TCP Connection | 13,364 |
| | MS SQL Stack BO | 4,685 |
| ④ | Too Many SYNs for a TCP Connection | 22,223 |
| | MS SQL Stack BO | 2,508 |
| ⑤ | Too Many SYNs for a TCP Connection | 21,285 |
| | Unauthenticated OSPF | 5,893 |
| ⑥ | Repeated TCP SYN with Diff ISN and TTL | 6,820 |
| | MS SQL Stack BO | 10,264 |

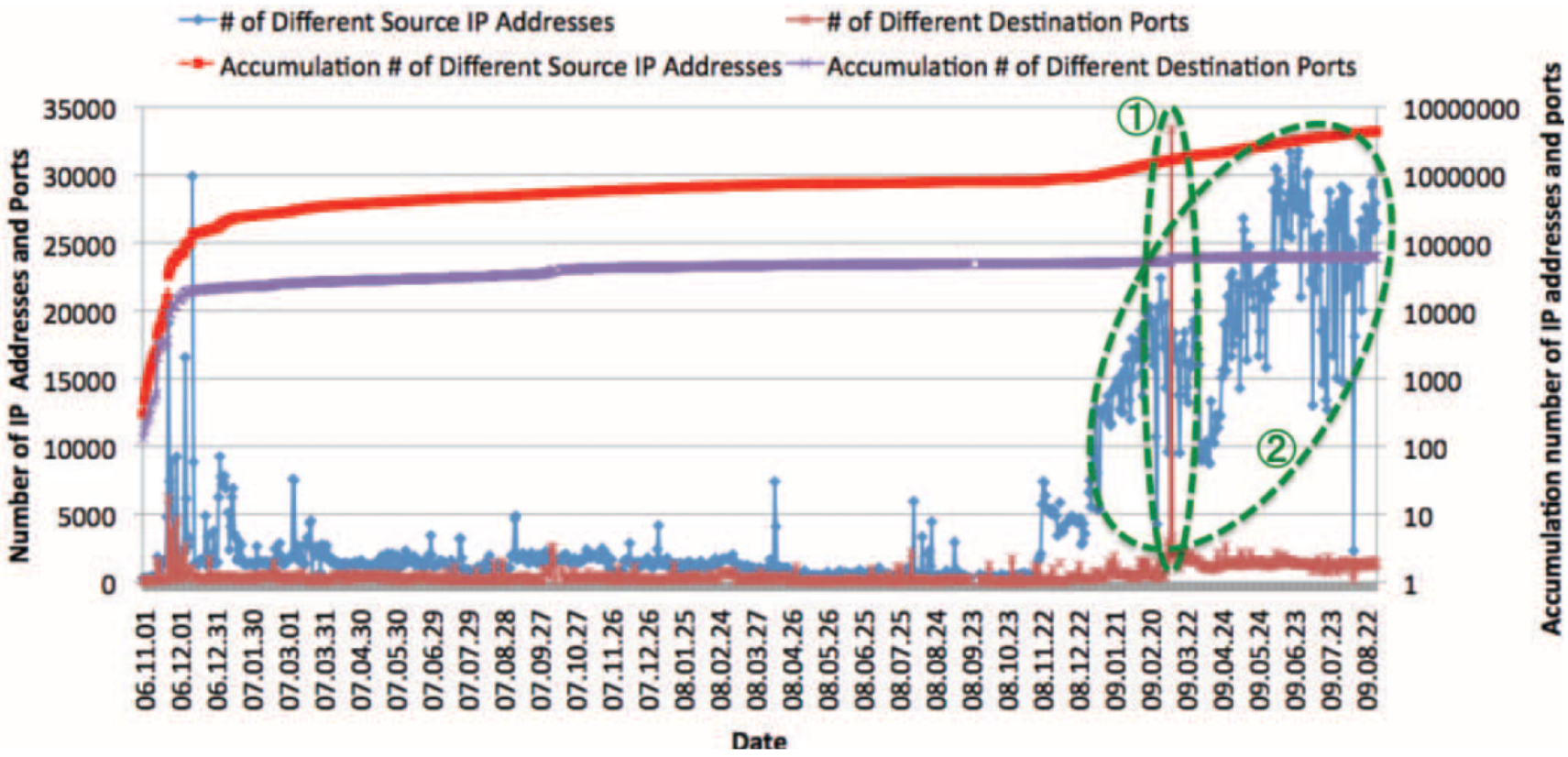
# Statistics of AV detected sessions



# Distribution of AV allert types



# Statistics of Source IP addresses and Destination Ports



# Distribution of Destination Ports