Security Big Data Analytics
— Big Data R&D @NICTER Project

Tao BAN
Cybersecurity Laboratory
Cybersecurity Research Institute
National Institute of Information and Communications Technology (NICT)

National Institute of Information and Communications Technology

NICTER
Network Incident analysis Center for Tactical Emergency Response

Outline

- **Big data in Cybersecurity**
- **Big data practice @NICTER**
- **Case studies**
  - **Botnet detection**
  - **Early detection of new IoT related threats**
- **Conclusions**

The Rising Cost of Cyber Crime

Provided by Ponemon Institute

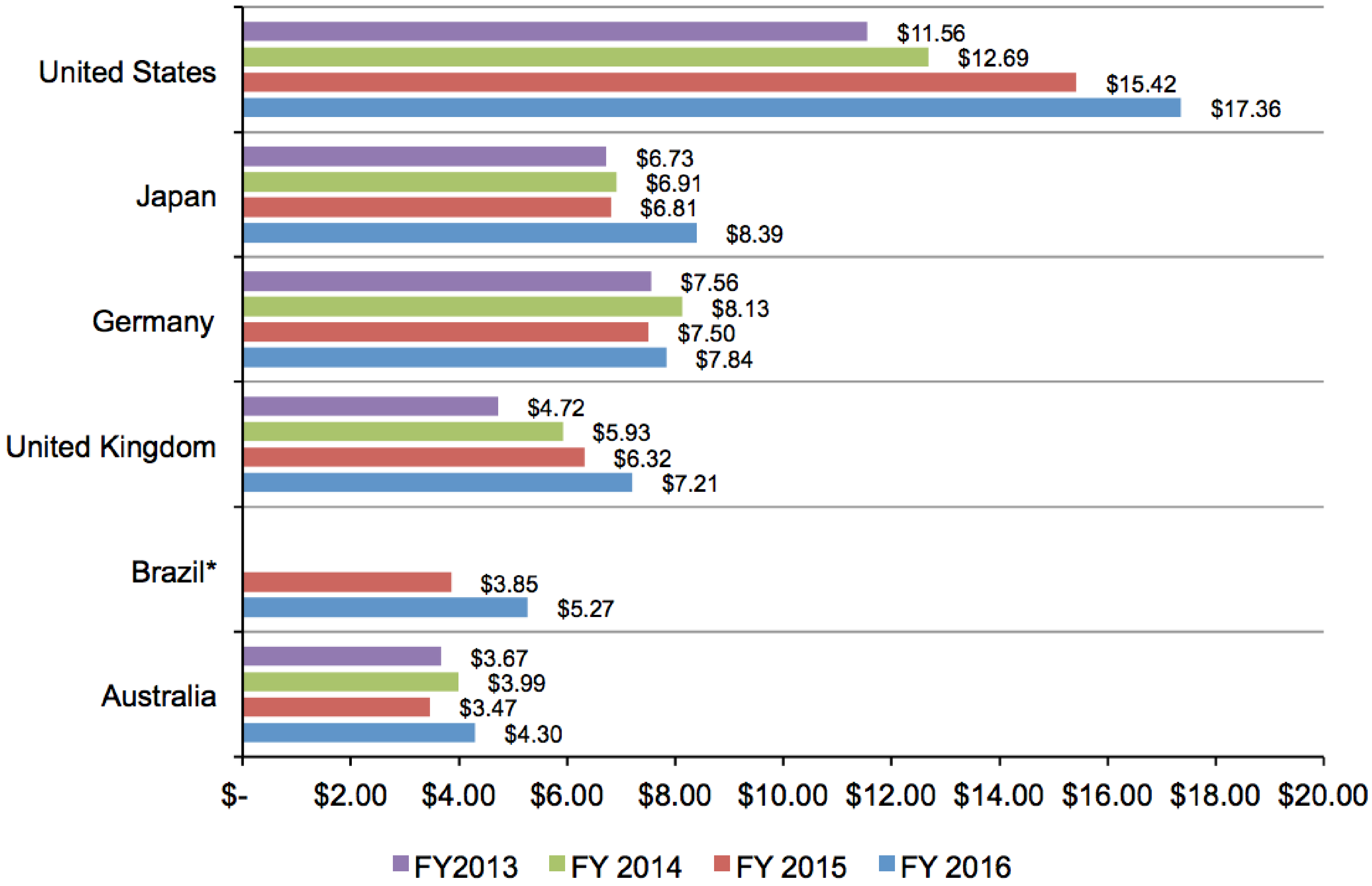


Figure 1. Total cost of cyber crime in six countries over four years
US$ millions, n = 237 separate companies

NICT National Institute of Information and Communications Technology

NICTER Network Incident analysis Center for Tactical Emergency Response

New Challenges for Cybersecurity

- New trends of new cyber attacks
  - Organized and better motivated cyber crimes
  - Drastically increasing malware programs
  - Sophisticated attacking techniques
    - APT, DRDoS, Ransomware
- Mobile security & cloud security
- IoT Security
  - Automobiles and home appliances are connected to the Internet
  - Not only digital assets but life is in danger from cyberattacks
- Big Data Problem
  - Big data is expensive
  - Analysis from a global view is unaffordable

The Importance of Security Big Data

When data can be successfully transformed to intelligence – bigger data for better intelligence – we can get smarter about security, taking a proactive rather than a reactive stance.

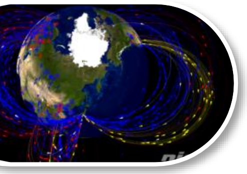
Expectations for security big data

- Better reliability and quicker response times by exploring the data correlation for a global view
- Better situation awareness by visualization tools
- More comprehensive forensic investigations and heightened defensive measures

Research Map

- NICTER and Spin-offs -

Passive



Network Incident analysis Center for Tactical Emergency Response

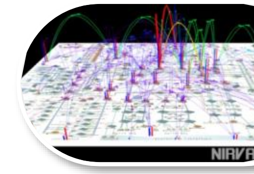
NICTER

Direct Alert Environment for Darknet And Livenet Unified Security

DAEDALUS

NICTER Real-network Visual ANAlyzer

NIRVANA



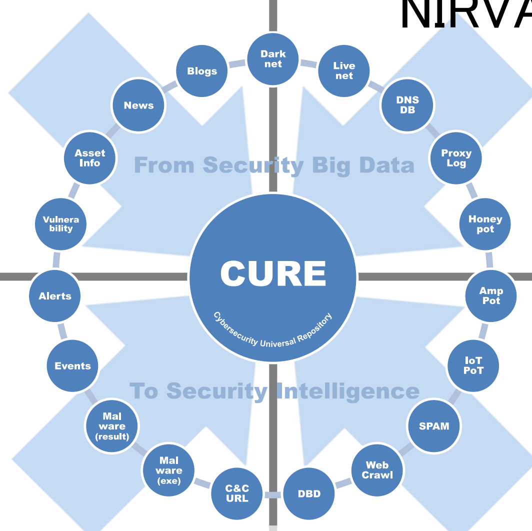
NICTER Real-network Visual ANAlyzer KAI

NIRVANA改

(KAI)

Global

Local



Honeypot for Amplification Attack

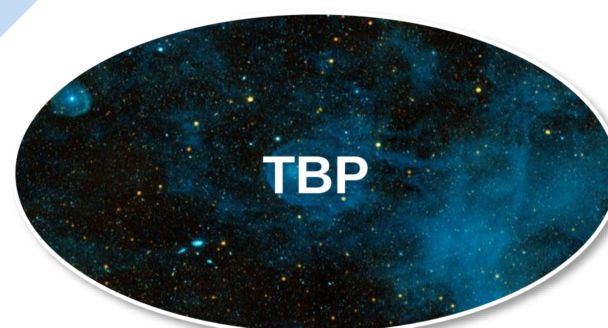
AmpPOT#

Honeypot for IoT Malware

IoTPOT#

Cybersecurity Universal REpository

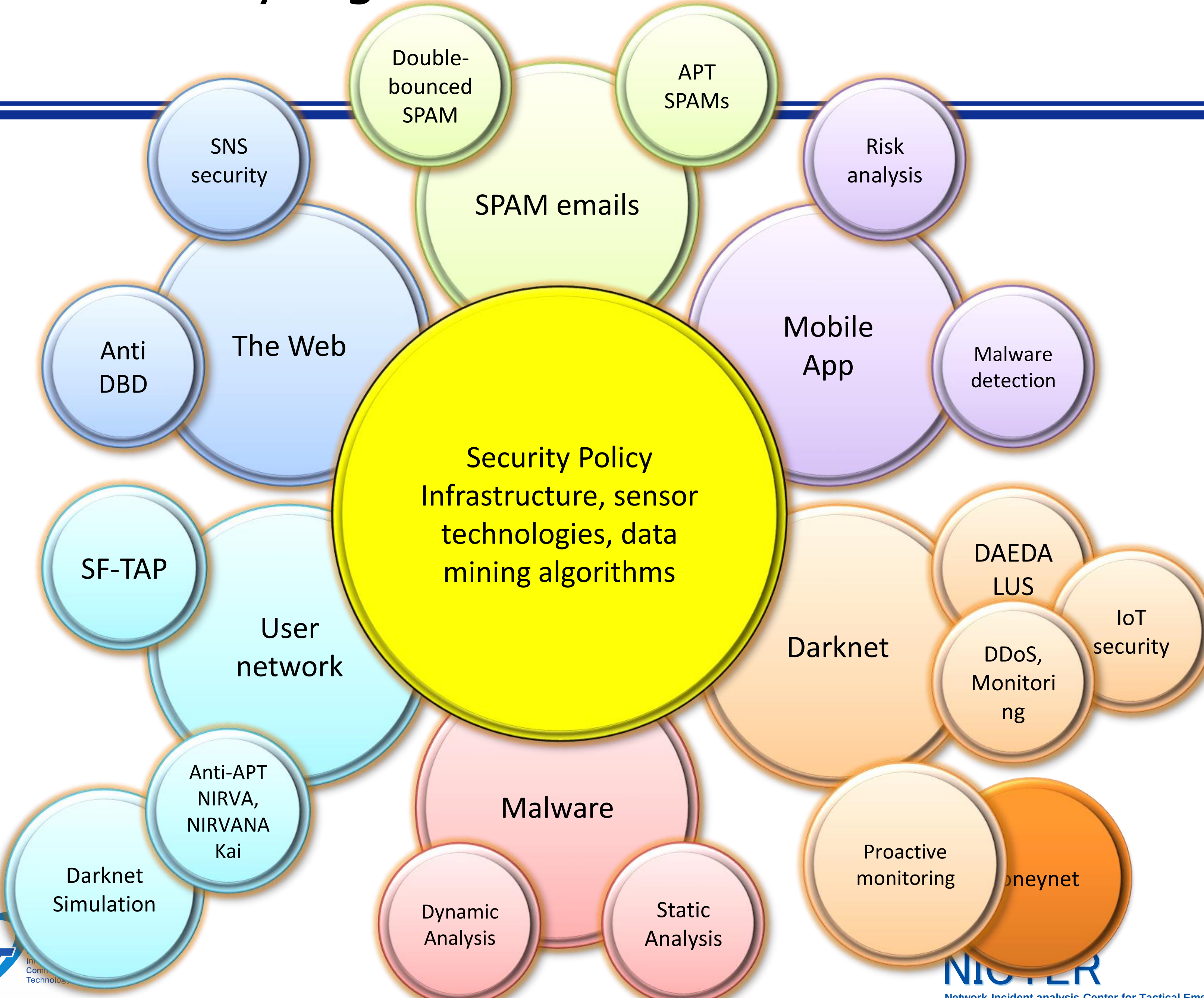
CURE



Active

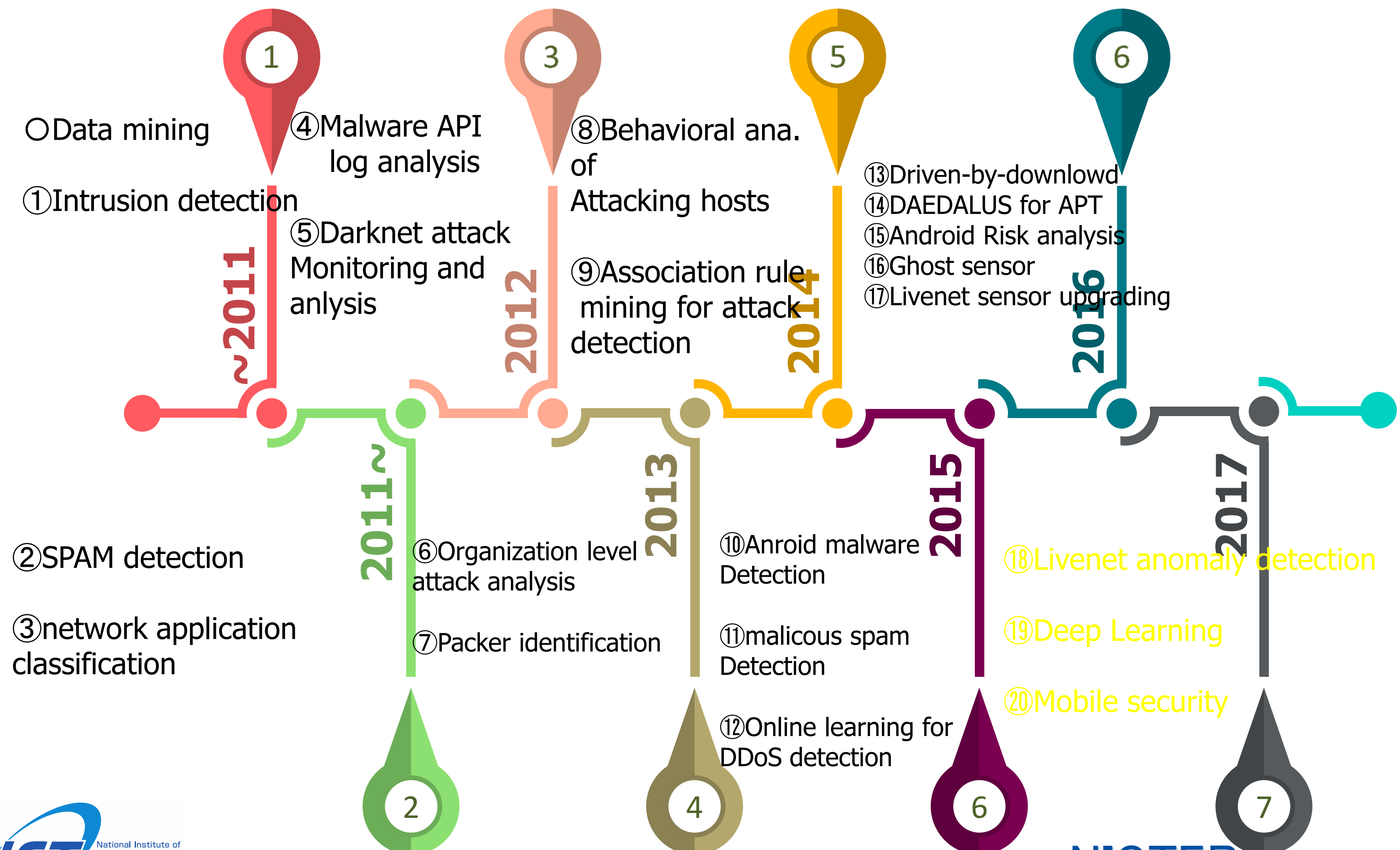
Collaborative research with Yokohama National Univ. and Saarland Univ.

Security Big Data Collected at NICT



Network Incident analysis Center for Tactical Emergency Response

Road Map for AI-based Research @CSL



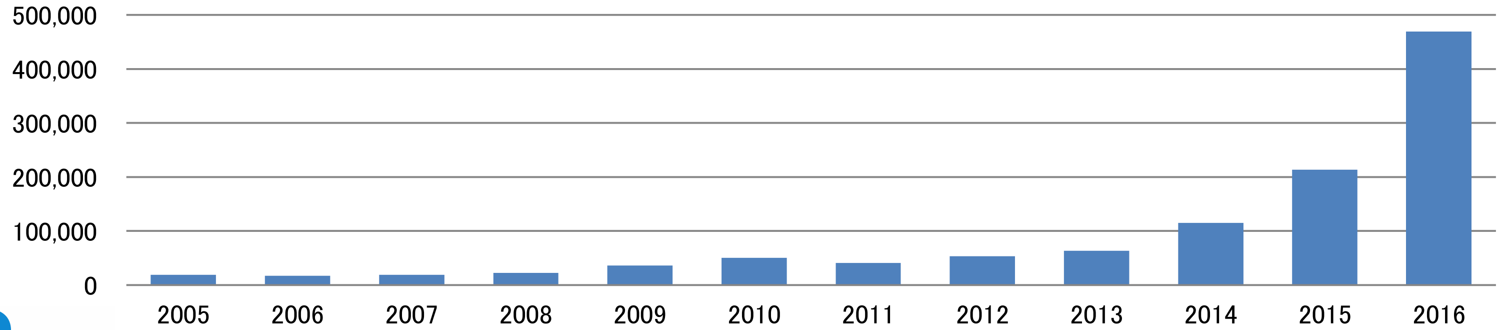
NICT National Institute of Information and Communications Technology

NICTER
Network Incident analysis Center for Tactical Emergency Response

Case Study of Darknet Traffic Analysis (1) Botnet Detection & Characterization

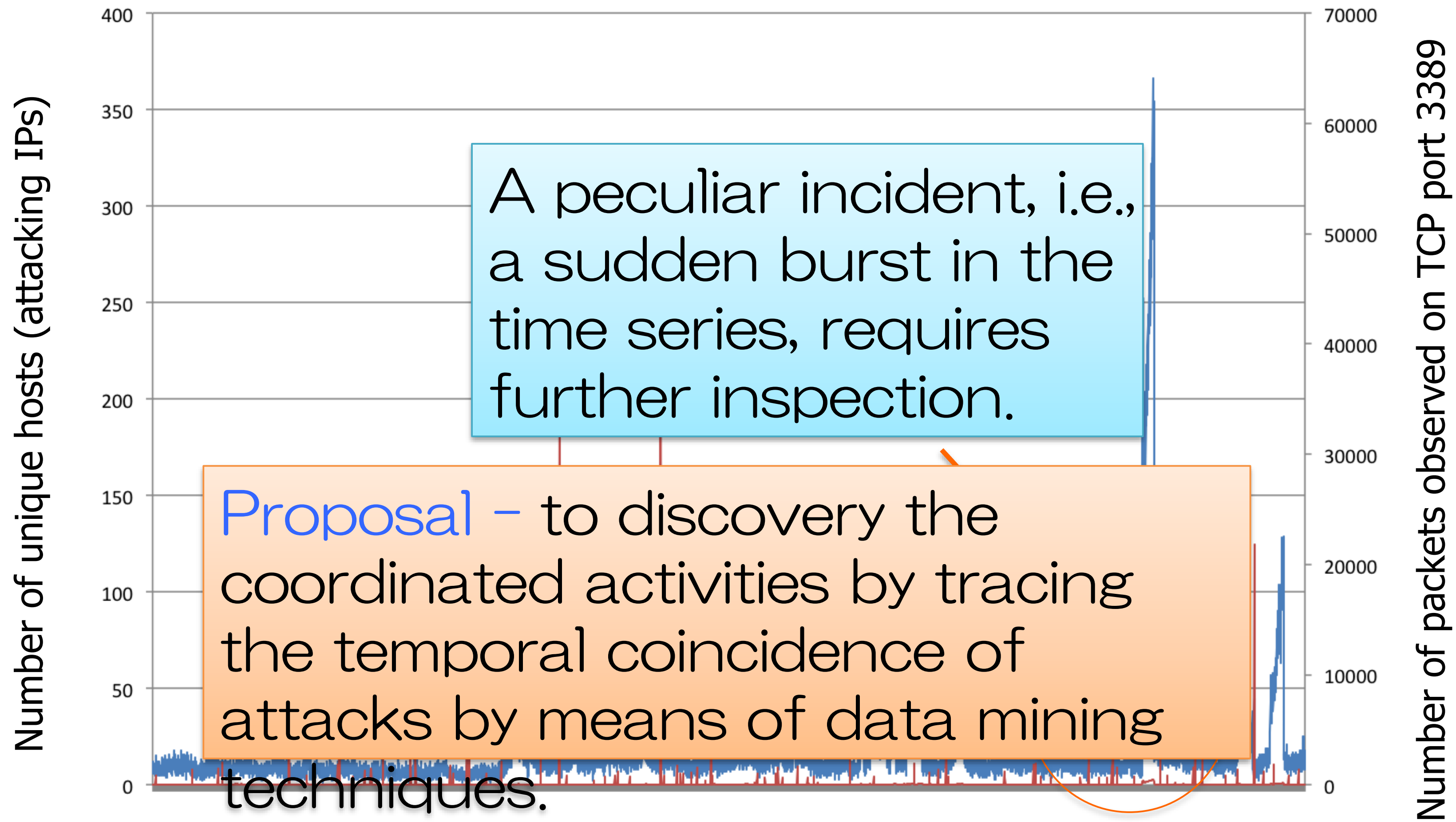
Yearly Stats of Darknet Traffic

Year	Number of packets par year	Number of IP address For darknet	Number of packets par 1 IP address per year
2005	0.31 billion	16 thousands	19,066
2006	0.81 billion	100 thousands	17,231
2007	1.99 billion	100 thousands	19,118
2008	2.29 billion	120 thousands	22,710
2009	3.57 billion	120 thousands	36,190
2010	5.65 billion	120 thousands	50,128
2011	4.54 billion	120 thousands	40,654
2012	7.79 billion	190 thousands	53,085
2013	12.9 billion	210 thousands	63,655
2014	25.7 billion	240 thousands	115,323
2015	54.5 billion	280 thousands	213,523
2016	128.1 billion	300 thousands	469,104



Number of packets par 1 IP address per year

Botnet Detection based on Darknet Monitoring

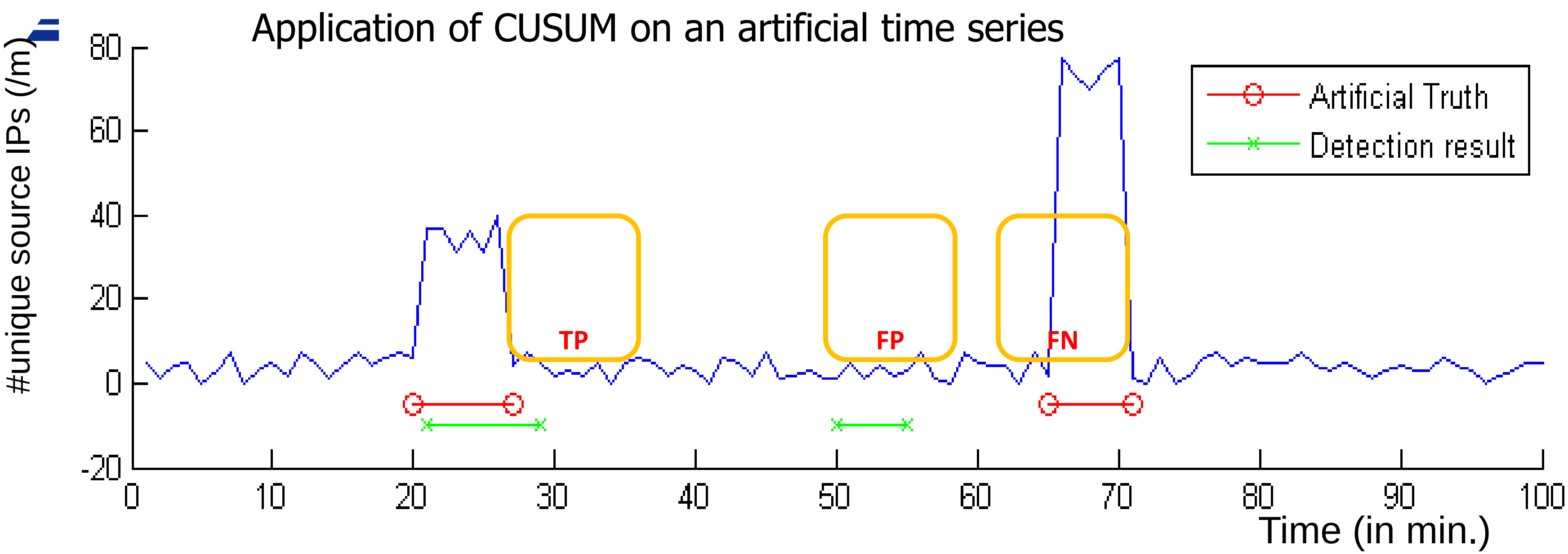


A peculiar incident, i.e., a sudden burst in the time series, requires further inspection.

Proposal – to discovery the coordinated activities by tracing the temporal coincidence of attacks by means of data mining techniques.

A case study: TCP_SYN packets statistics observed on port 33
(Data collected from 2011.7.1 to 2011.8.4 on a /16 darknet

Abrupt Change Detection : CUSUM

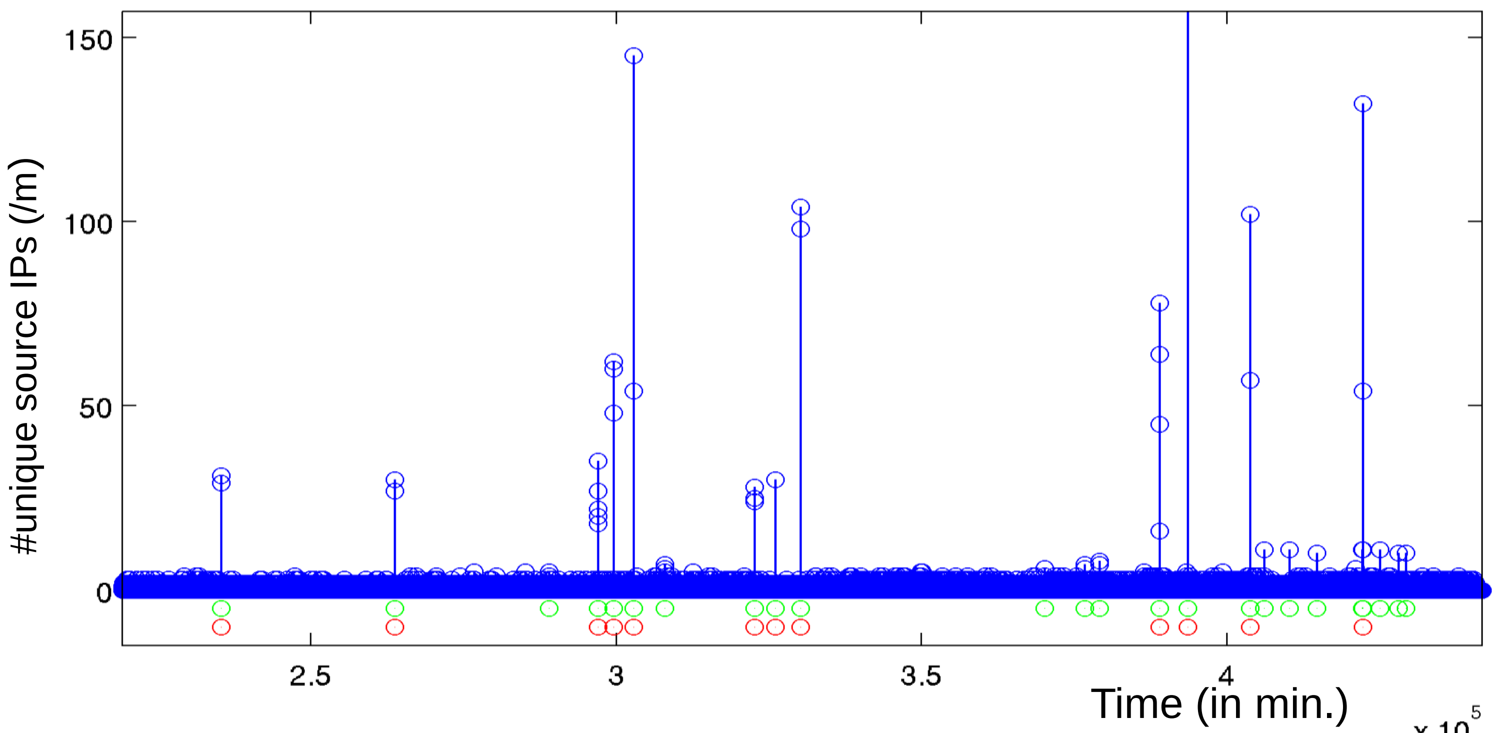


- **Step 1:** Application of a modified Cumulated Sum (CUSUM) algorithm [1] to the number of unique source IP time series for detecting the abrupt changes associated with coordinated attack events, i.e., **active epochs**, of botnets.
- **Step 2:** Filtering and justification of the epoch detection results by removing insignificant events caused by noises and justify the starting and ending times.

[1] T. L. Lai. Sequential Changepoint Detection in Quality-Control and Dynamical Systems. Journal of Royal Statistical Society - Series B. vol. 57, no. 4, pages 613–658, 1995.

Case Study : TCP Port 139

Input: TCP_SYN packets observed on destination port 139.
(Data collected in 2011 on a /16 darknet sensor.)



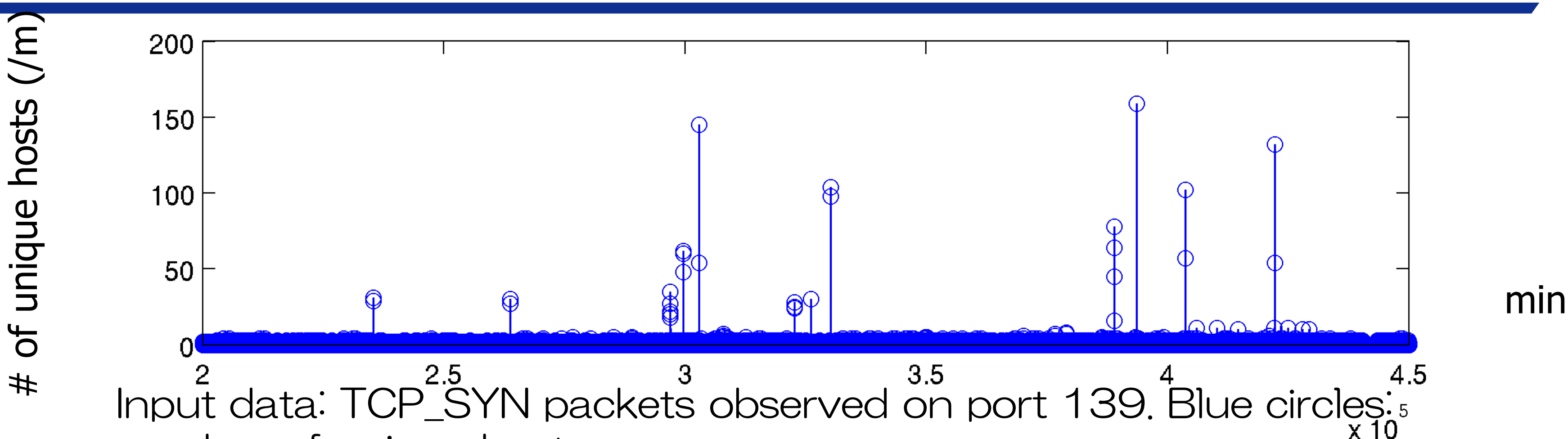
Output of step 1: Candidates of starting and ending points detected by the CUSUM algorithm, denoted by green circles under the number of unique source time series.

Output of step 2: Starting and ending points of botnet active epochs given by the filtering process applied on the output of step 1, denoted by red circles under the time series.

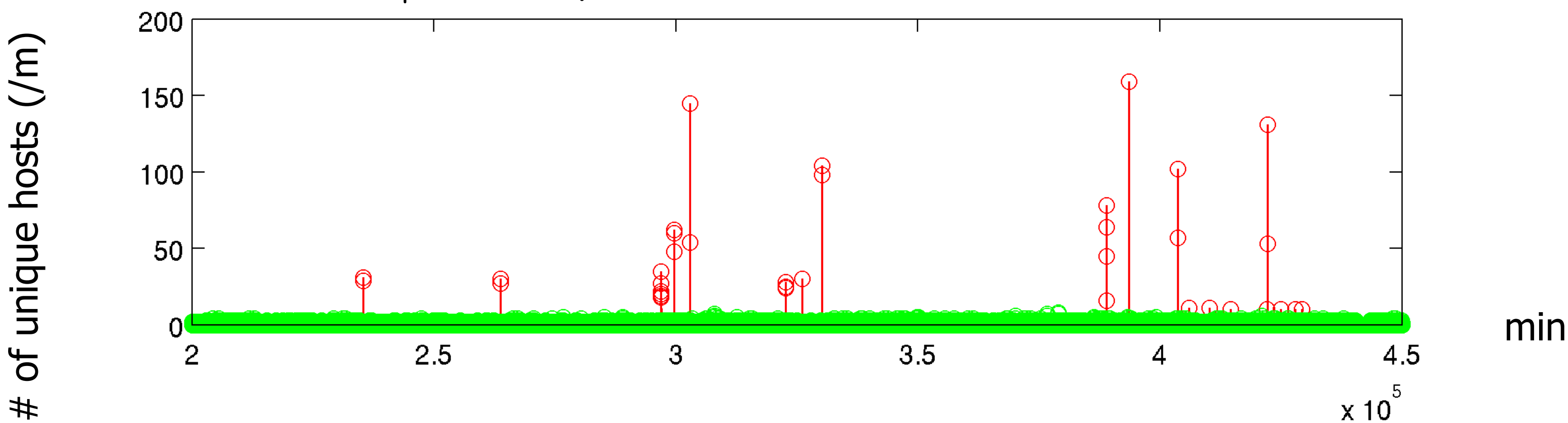
NICT National Institute of Information and Communications Technology

NICTER Network Incident analysis Center for Tactical Emergency Response

Attack Epoch Extraction @TCP Port 139

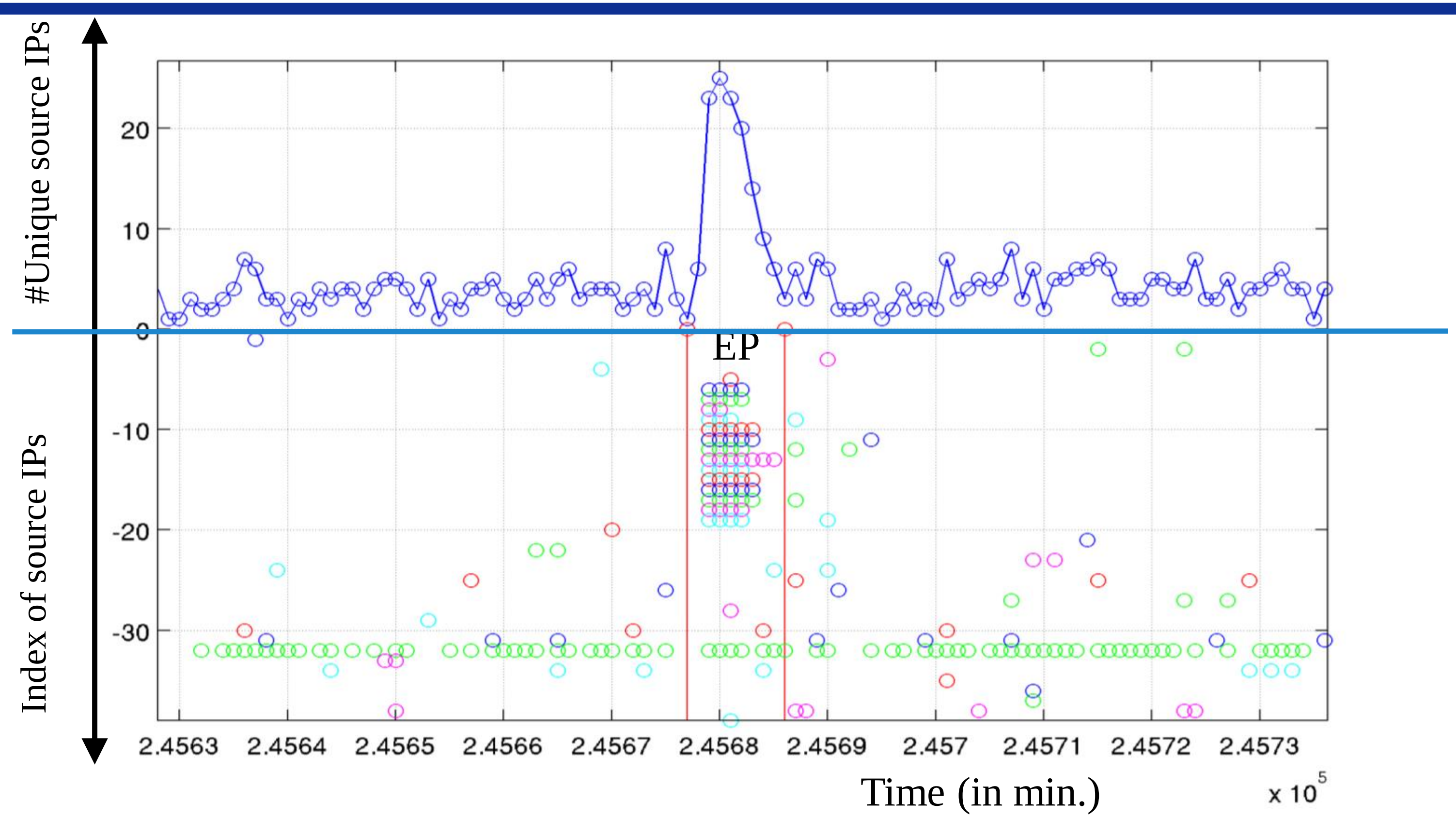


Input data: TCP_SYN packets observed on port 139. Blue circles: number of unique hosts.



Output of activity epoch detection. The input is divided into two components: red circles indicate the starting and ending time of the active epochs, and green circles indicates observations without botnet activities detected.

Host Activities @TCP Port 139



Feature 1: rate of packets from the host observed in the epoch period (EP),

R = (N_e in EP) / N, where N is the number of packets observed in the time window (size =11EP.) embracing EP.

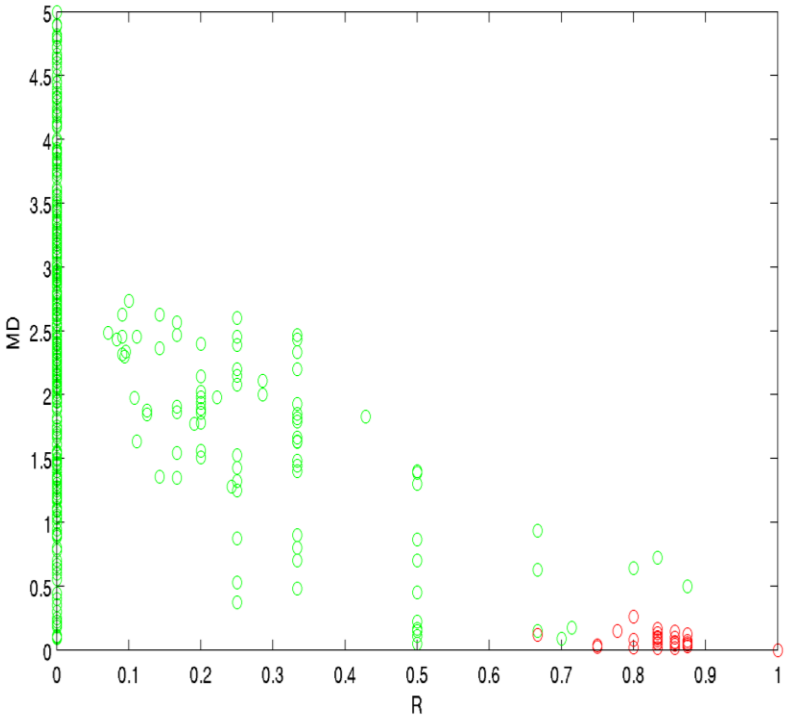
Feature 2: average deviation of all packets from the epoch normalized by EP length,

MD = mean(d_i) / length(EP) , where d_i = min(abs(t_i-Ep_s), abs(t_i-EP_e)), EP_s and EP_e are the starting and ending times of the active epoch.

NICT National Institute of Information and Communications Technology

NICTER Network Incident analysis Center for Tactical Emergency Response

Bot Classification Result



Scatter plot in the 2D space

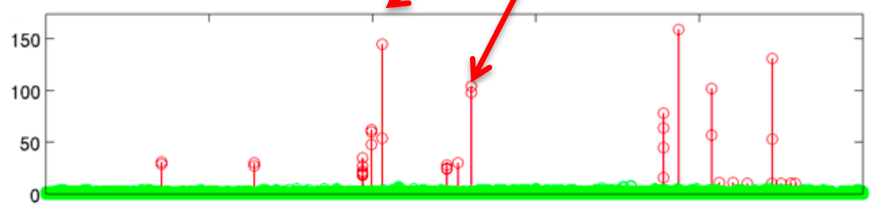
Port	Sensor	Coordination Features	Flow Features [1]
5900	1	99.59%	42.36%
5900	2	99.78%	47.21%
1433	1	100%	96.61%
1433	2	100%	92.84%
25	1	99.58%	81.91%
25	2	99.61%	89.78%
139	1	100%	79.12%
8506	1	99.44%	0
3389	1	99.90%	57.86%

G-mean values obtained by Support Vector Machine. Results of 5-fold cross validation with optimal parameters are reported.

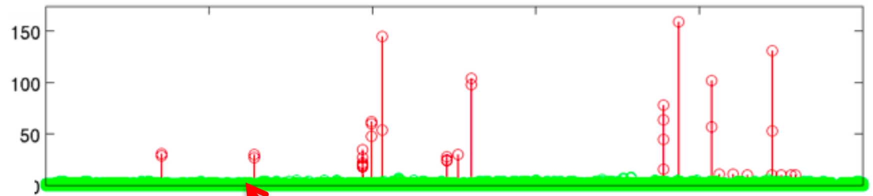
[1] T. Ban, et at., Behavior analysis of long-term cyber attacks in the darknet, ICONIP'12 Proceedings of ICONIP'12, Volume 7667, Part V, Pages 620-628.

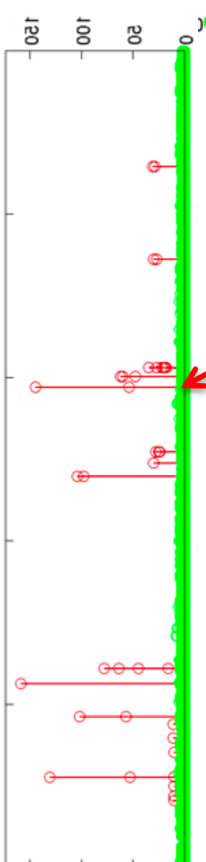
Correlation Analysis of Botnet Attacks 1

How do attacks performed at different time relate to each other? Are they from the same botnet (group of attacking hosts)?



Correlation Analysis of Botnet Attacks2

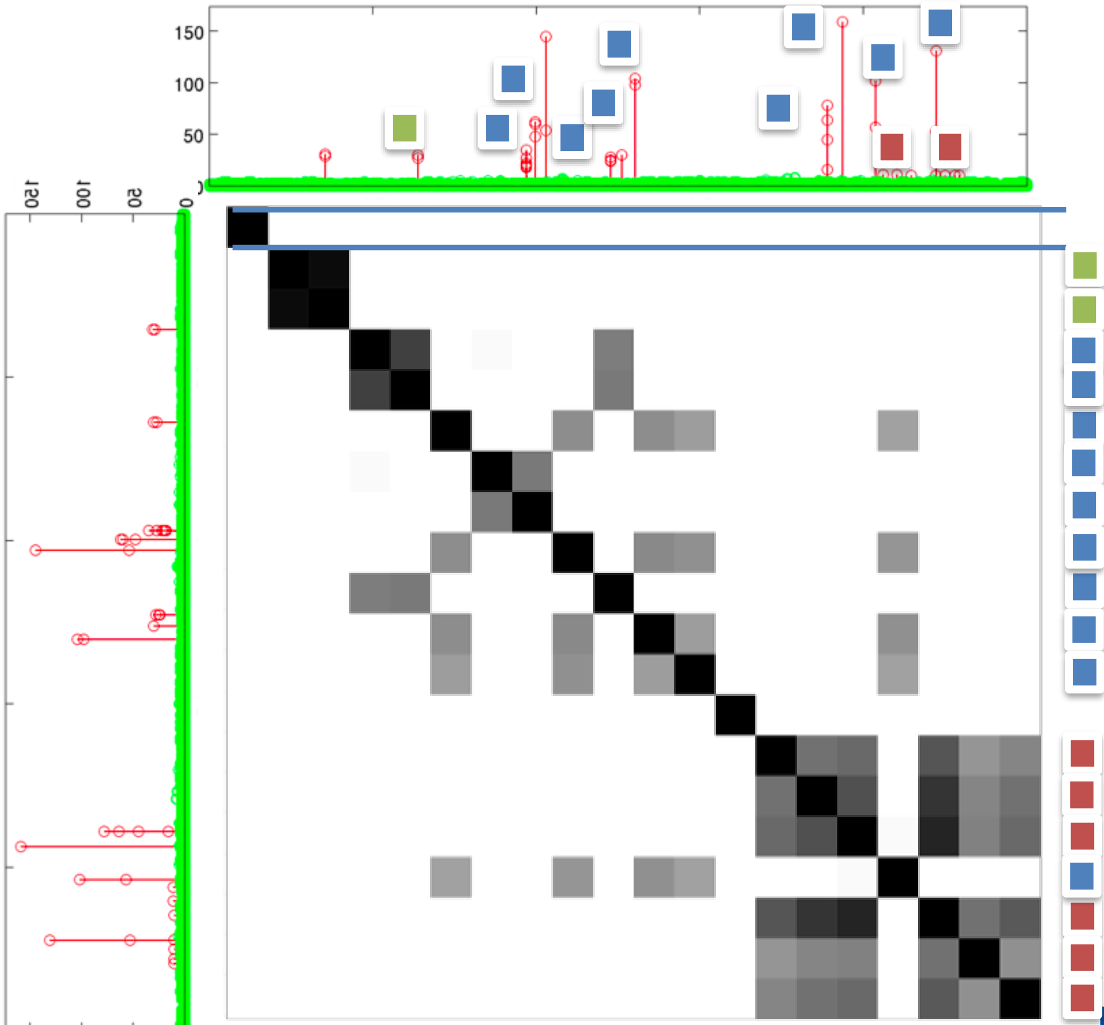




Evaluate the similarity between two events by the Jaccard distance between the set of attacking hosts, i.e.

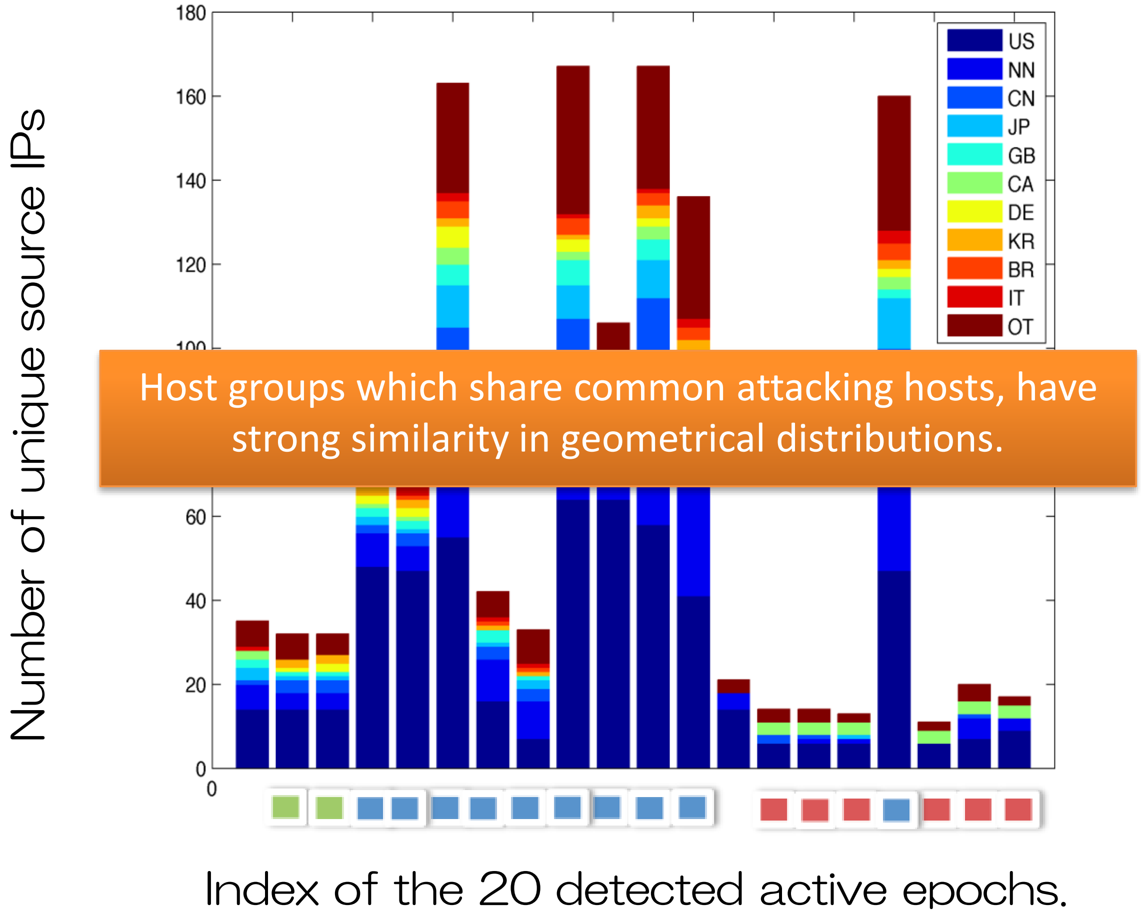
$$J(S_a, S_b) = 1 - \frac{|S_a \bigcap S_b|}{|S_a \bigcup S_b|}$$

Correlation Analysis of Botnet Attacks3



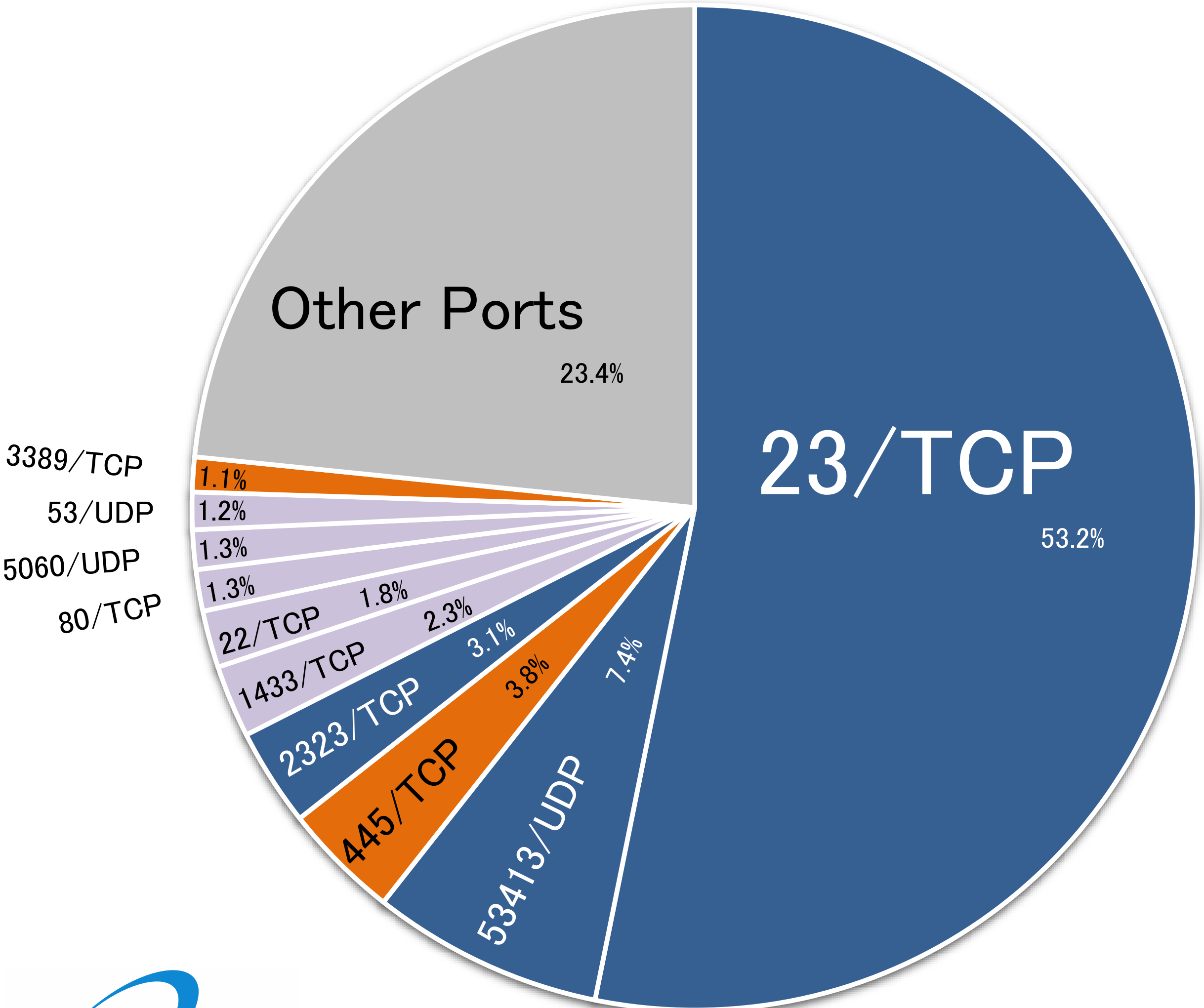
Variations on Geo-distribution on Port 139

Stacked plot of geo-locations of source IPs in the active epochs detected on destination port 139, 2011.



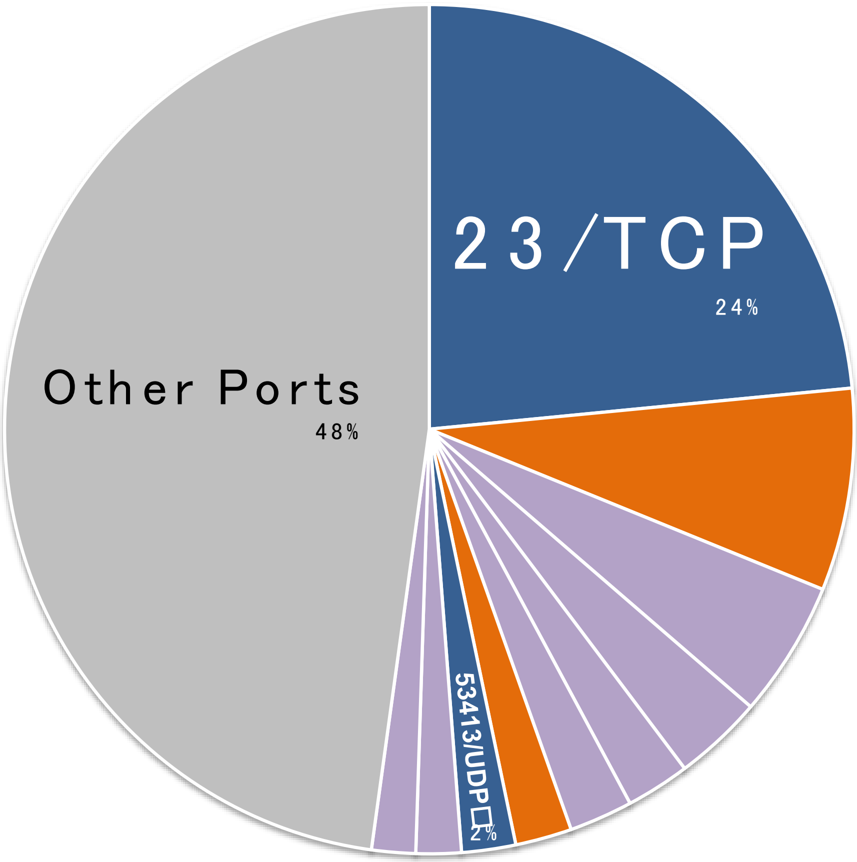
Case Study of Darknet Traffic Analysis (2)
Early Detection of New IoT Threats

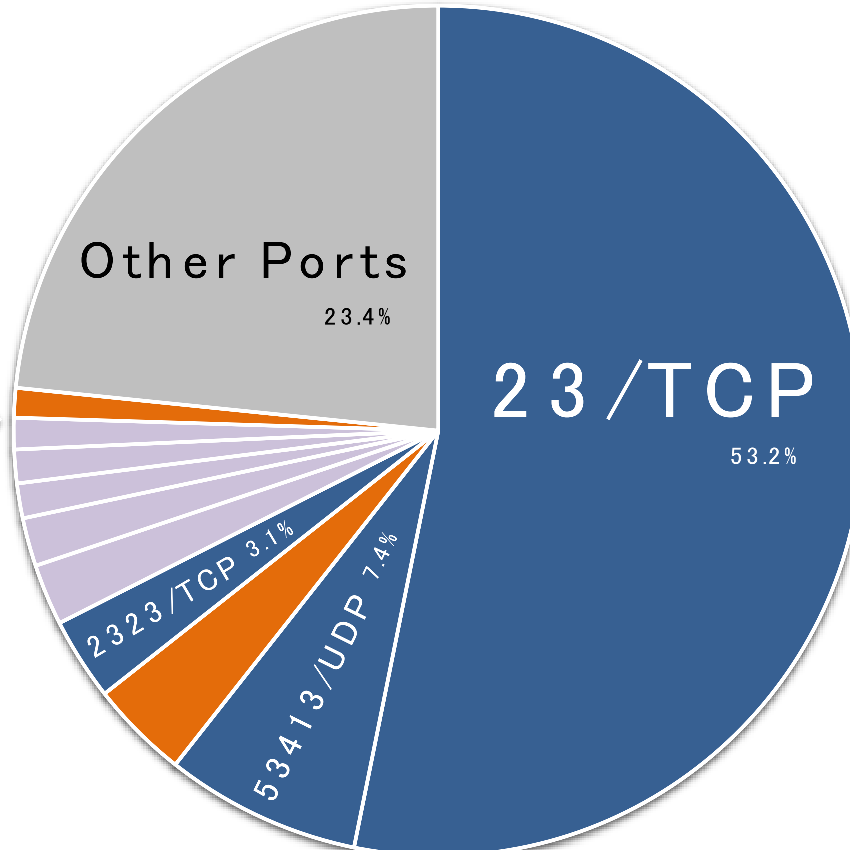
Distribution of Port Numbers (2016)



Port	Target Service
23/TCP	IoT (Web Camera, etc.)
53413/UDP	IoT (Netis Router)
445/TCP	Windows (Server Service)
2323/TCP	IoT (Web Camera, etc.)
1433/TCP	SQL
22/TCP	SSH
80/TCP	HTTP
5060/UDP	SIP
53/UDP	DNS
3389/TCP	Windows (RDP)

Transition from 2015 to 2016





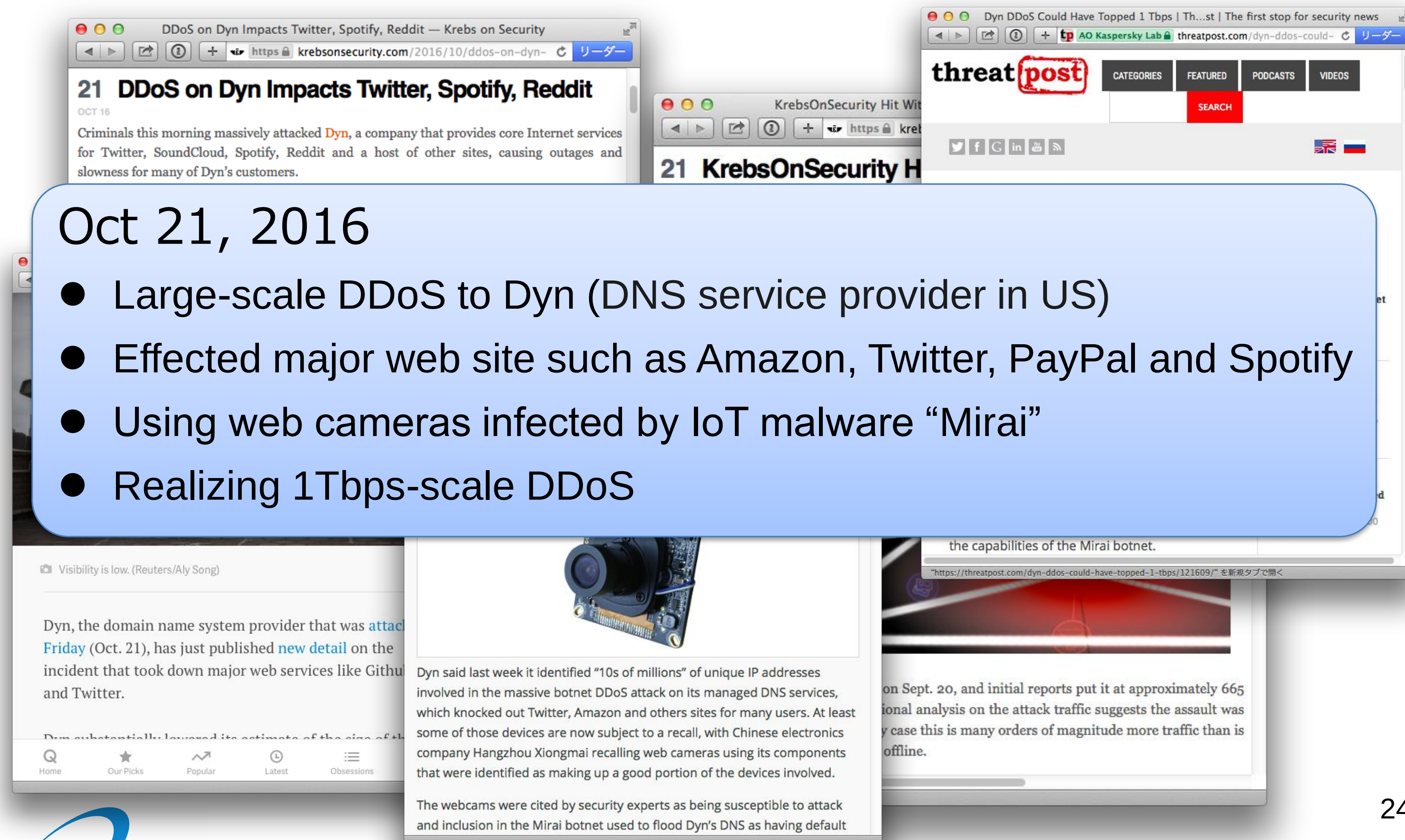
2015: IoT > 26%

(23/TCP + 53413/UDP)

2016: IoT > 64%

(23/TCP + 2323/TCP + 53413/UDP)

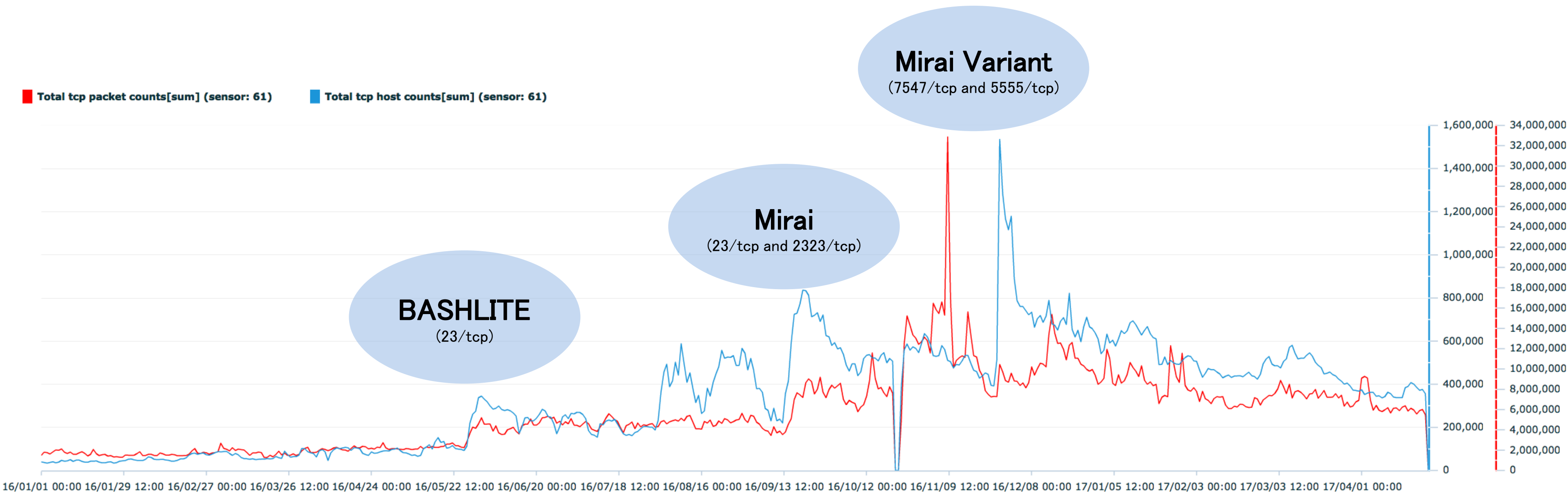
Large-scale DDoS by Mirai



Oct 21, 2016

- Large-scale DDoS to Dyn (DNS service provider in US)
- Effected major web site such as Amazon, Twitter, PayPal and Spotify
- Using web cameras infected by IoT malware "Mirai"
- Realizing 1Tbps-scale DDoS

Darknet Traffic TO FR Sensor



TCP Packets and Unique Hosts per Day (January 2016 – April 2017)

Association Rule Learning

- **Association Rule Learning** is a method to discover interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness --- Wikipedia
- **An association rule**: X➔Y
- Early application: market basket analysis

Transaction No.	Item 1	Item 2	Item3	…
101(Alice)	Bread	Milk	Jam	
102(Bob)	Rice ball	Tea	Lunchbox	
…				

- Bread ➔ Milk &Jam
- Rice ball &Tea ➔ Lunchbox

Rule Evaluation – Support

Support: **the frequency in which the items in LHS and RHS co-occur.**

$$\text{Support rate} = \frac{\text{No. of transactions containing items in LHS and RHS}}{\text{Total No. of transactions in the dataset}}$$

Transaction No.	Item 1	Item 2	Item 3	…	Count
100	Bread	Milk	Jam	Beer	1
101	Bread	Milk			1
102	Bread	Jam	Beer		1
103	Bread	Jam			1

Support(Bread) = 4
Support(Milk) = 2
Support(Bread, Milk) = 2

Is buy(bread) leading to buy(milk)
or buy(milk) leading to buy(bread)?

Rule Evaluation – Confidence

Confidence can be interpreted as an estimate of the conditional probability P(Y|X), the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

$$\text{Confidence} = \frac{\text{No. of transactions containing both LHS and RHS}}{\text{No. of transactions containing LHS}}$$

Transaction No.	Item 1	Item 2	Item 3	…	Count
100	Bread	Milk	Jam	Beer	1
101	Bread	Milk			1
102	Bread	Jam	Beer		1
103	Bread	Jam			1

- confidence for buy(Bread) → buy(Milk) = 2/4 = 50%
- confidence for buy(Milk) → buy(Bread) = 2/2 = 100%
- So buy(Milk) → buy(Bread) is a more important rule in terms of confidence.

Association Rule Learning Algorithms

- **Apriori:** the best-known algorithm
  - Find all itemsets that have minimum support (frequent itemsets, also called large itemsets).
    - Extensively used the Apriori principle: if an item set is frequent, then all of its subsets must also be frequent.
  - Use frequent itemsets to generate rules.
    - E.g., a frequent itemset
      {Bread, Milk, Butter} [sup = 3/7]
      and a rule from the frequent itemset
      Bread$\rightarrow$ Milk, Butter [sup = 3/7, conf = 3/3]
- **FP-growth algorithm:** an improved algorithm proposed to overcome the bottlenecks of Apriori.
  - Does not create candidate of frequent itemsets;
  - The FP-tree is stored in the main memory.

Darknet Sensor Statistics

Number of packets: > 100M
Number of hosts: > 5M

Sensor ID	A	B	C	D	E	F
Type	I	II	I	II	II	II
Size	/16	/16	/18	/18	/18	/17
#Pkt/IP	161.51	190.18	193.86	281.77	414.97	406.66
#Host/IP	85.55	118.48	118.97	161.07	175.40	230.87
#Ports	65536	63227	65224	30728	29651	46678
Port 1	23	8	8	445	445	445
Port 2	8	23	23	23	23	23
Port 3	29735	3389	29735	8	8	8
Port 4	29991	80	3389	3389	3389	3389
Port 5	30247	29735	29991	21060	30759	30759
Port 6	30503	8080	80	60557	80	80

Experiment Setting

- One day (1st. Sept. 2012) packet data collected from darknet sensor A (/16). Each transaction is a set of destination ports attacked by a single IP, regardless of the DHCP problem.

Attack No.	DPort 1	DPort 2	DPort 3	…	Occurrence
100	23	210	1526		441
101	23	210	1526	12345	32
102	23	210	1522	2040	7
103	23	210	1522	3351	23
104	23	1522	8		3

- Other features are also explored, e.g., destination sensor ID, used protocol, tcp flags, sequence IDs, etc.
- FP-growth is used to extract the rules.
- Parameter setting: support = 200, confidence = 80%.

Results on Destination Ports (1)

Frequent itemsets related to Port 80 (8/560)

ID	DPort 1	DPort 2	DPort 3	DPort 4	Occur.
①	80				2932
②	80	8			747
③	80	443			786
④	80	13			715
⑤	80	8	443		741
⑥	80	8	13		713
⑦	80	13	443		712
⑧	80	8	13	443	711

P8: unassigned

P13: Daytime protocol

P80: Hypertext Transfer Protocol (HTTP)

P443: Hypertext Transfer Protocol over TLS/SSL (HTTPS)

Association rules

No.	Rule	Sup.	Conf.
①	80→8	747	27.5%
②	8→80	747	4.7%
③	80→13	715	24.3%
④	13→80	715	94.7%
⑤	80,443→8	741	94.3%
⑥	8,443→80	741	95.45%
⑦	8,80→443	741	99.2%
⑧	13,443→80	712	95.3%
⑨	80,443→13	712	90.6%
⑩	13,80→443	712	99.6%
⑪	8,13→80	713	95.2%
⑫	8,80→13	713	95.4%
⑬	13,80→8	713	99.7%
⑭	13,8,443→80	711	95.4%
⑮	8,80,443→13	711	96.0%
⑯	13,80,443→8	711	99.9%
⑰	8,13,80→443	711	99.7%

Results on Destination Ports (2)

No.	Rule	Sup.	Conf.
①	210➔23	20047	98.66%
②	23➔210	20141	98.20%
③	23,1526➔210	1150	99.57%
④	210,1526➔23	1422	99.44%
⑤	210,8010➔23	1150	99.57%
⑥	23,8010➔210	1156	99.05%
⑦	210,3351➔23	1343	99.33%
⑧	23,3351➔210	1341	99.48%

- **Service on P23: Telnet protocol-unencrypted text communications.**
- **Service on P210: ANSI Z39.50, an international standard client-server, application layer communications protocol for searching and retrieving information from a database over a TCP/IP computer network.**

Results on Other Features

No.	Rule	Sup.	Conf.
①	TCP_ACK ➔TCP_SYN	868	94.58%
②	TCP_ACK, ICMP➔TCP_SYN	809	98.64%
③	TCP_ACK, TCP_SYN ➔ICMP	821	97.20%
④	TCP_ACK ➔TCP_RST	868	93.20%
⑤	TCP_RST, UDP➔TCP_SYN	284	99.30%
⑥	TCP_RST ➔TCP_SYN	817	82.86%

- **As the causal packet type, TCP_ACK packets seems to carry much information of the attacking tools.**
- **Together with port information, packet type may be applied as signatures for some malware programs.**

Signatures Confirmed

- **The reported sets of simultaneously attacked ports**
  - **80, 8, 13, 443**
  - **23, 210**

  are discovered to be associated with the Carna botnet [2]
  - **The botnet is composed of more than 400,000 compromised devices which scan the IPV4 space continuously using an advanced network scanning tool.**
  - **The scan logs are released by the master of the botnet.**

[2] C. Stocker and J. Horchert, "Mapping the internet: A hacker's secret internet census," ***Spiegel Online*****, 22/3 2013.**

Correlation between the Sensors

- **High correlation is discovered on the sensors, which are distributed in separated networking environments: companies and universities.**

ID	A	B	C	D	E	F
A	506805					
B	36798	90512				
C	44870	26205	159907			
D	13385	9905	10810	63693		
E	14099	10649	11690	27832	62003	
F	20149	14138	15461	16257	16563	57703

Attacking hosts observed across the sensors

Preliminary Results

ID	Sensor 1	Sensor 2	Sensor 3	Sensor 4	Occur.
1	E	F			16563
2	D	F			16257
3	D	E			27832
4	D	E	F		11486
5	B	F			14138
6	B	E			10649
7	C	B			26205
8	C	B	F		12353
9	A	B	F		13775
10	A	B	E		10408
11	A	C	F		14833
12	A	C	E		11242
13	A	C	D		10366
14	A	C	B		24826
15	A	C	B	F	12258

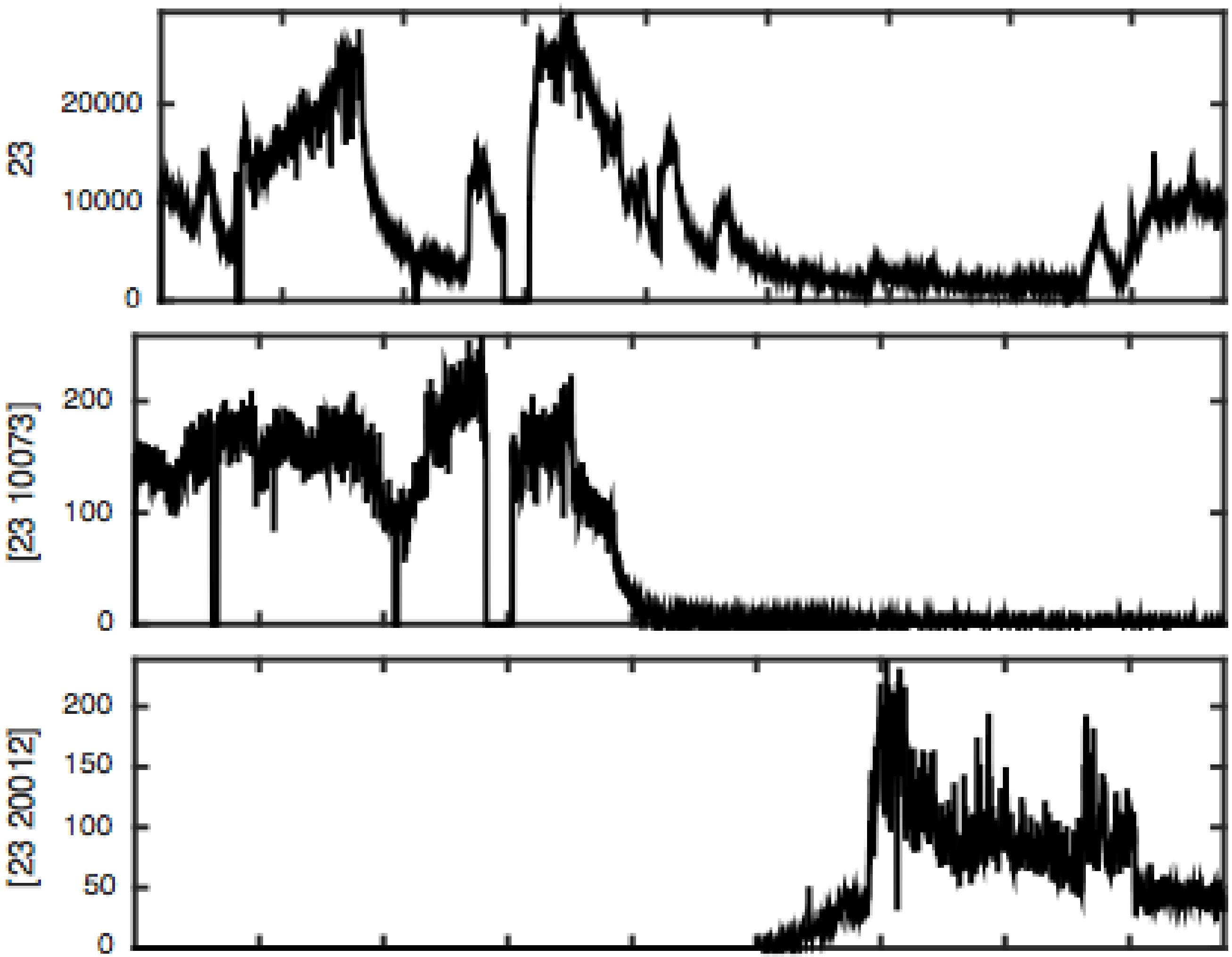
Frequent itemsets discovered among the six sensors.

Strong Association Rules

ID	Rule	Support	Confidence
1	$B, F \rightarrow C$	14138	87.37%
2	$B, F \rightarrow A$	14138	97.43%
3	$B, E \rightarrow A$	10649	97.73%
4	$C, F \rightarrow A$	15461	95.94%
5	$C, E \rightarrow A$	11690	96.17%
6	$C, D \rightarrow A$	10810	95.89%
7	$A, C, F \rightarrow B$	14833	82.64%
9	$A, B, F \rightarrow C$	13775	88.99%
10	$C, B, F \rightarrow A$	12353	99.23%
11	$C, B \rightarrow A$	26205	94.74%

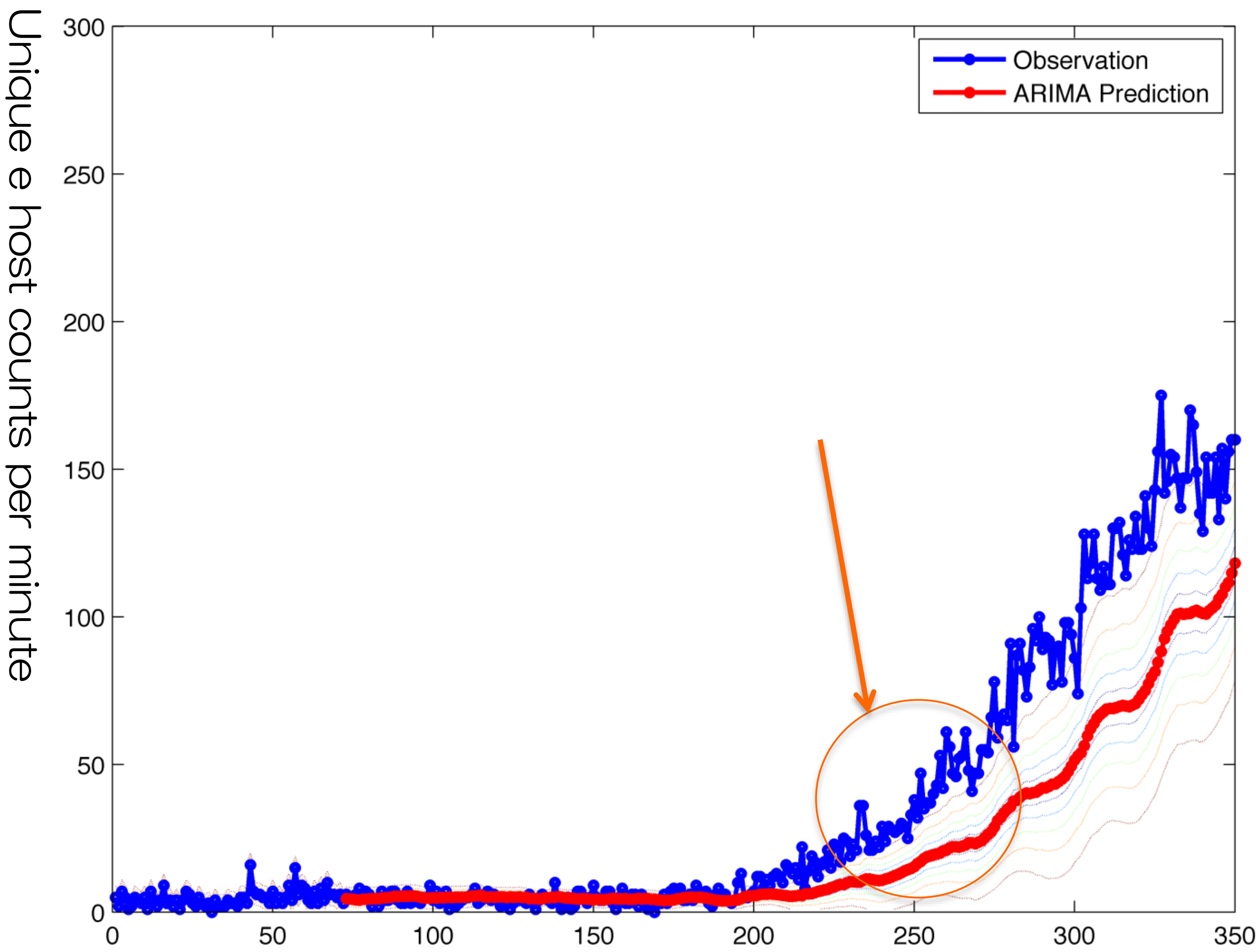
Strong association rules (support = 10000, confidence = 80%)

Long Term Observations of Attack Patterns (combination of destination ports)

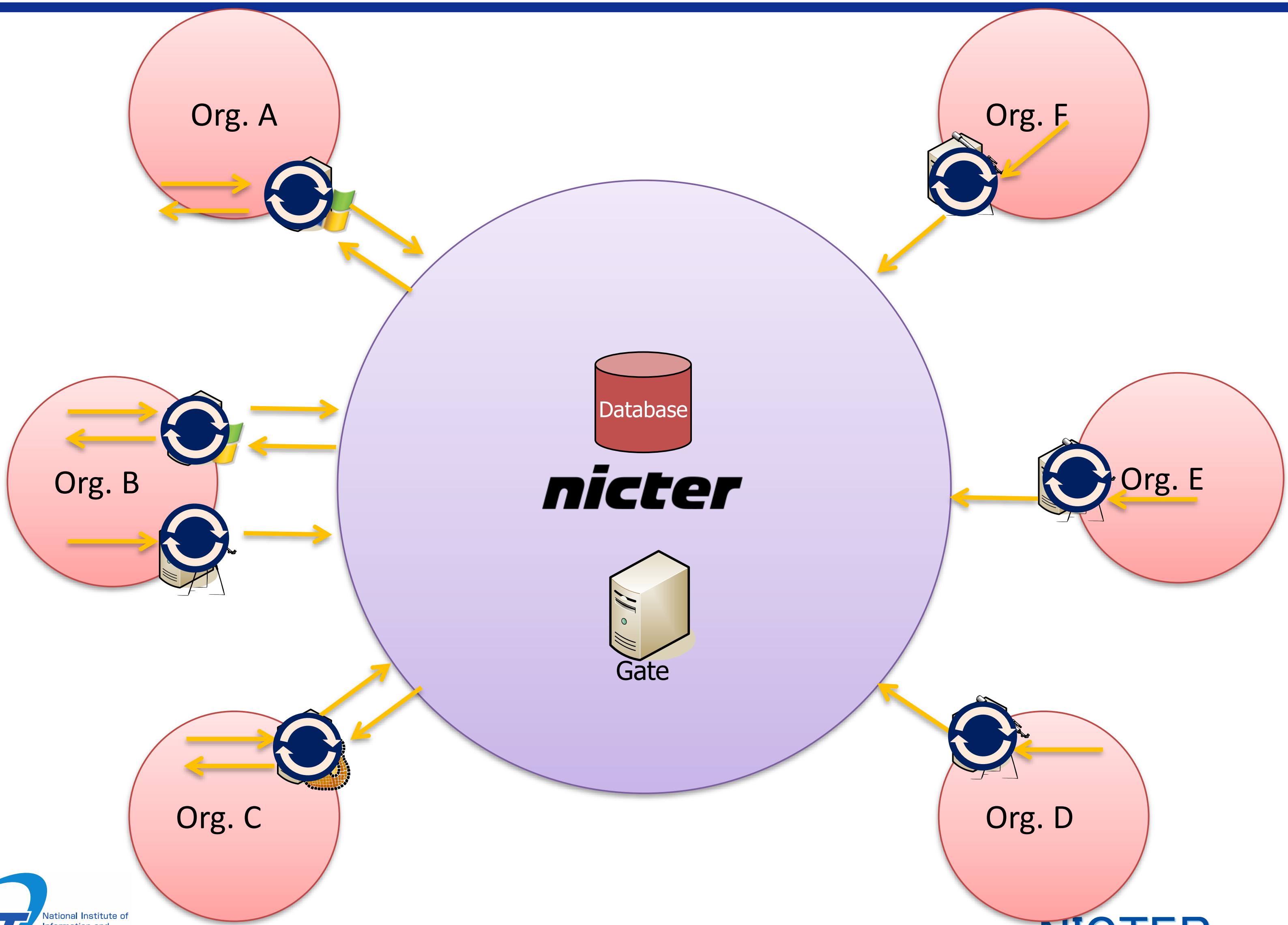


2015/1/1 ~ 2015/12/31

Abrupt Changes on the Time Series Indicates Pandemic Incidents



Use the Detection Information for Better Information Collection (Ghost Sensor)



National Institute of Information and Communications Technology

NICTER
Network Incident analysis Center for Tactical Emergency Response

Conclusions

- **Security big data are essential to fight with cyberattacks and protect the organizations and end users.**
- **Machine learning methods have been proved promising for counterattack cyber challenges.**
- **Aggregation of human intelligence and AI are the most practical practice in the current cyber age.**
- **Big data research call forth more international collaboration as the remedy of lack of data and intelligence.**

International Darknet Traffic Sharing

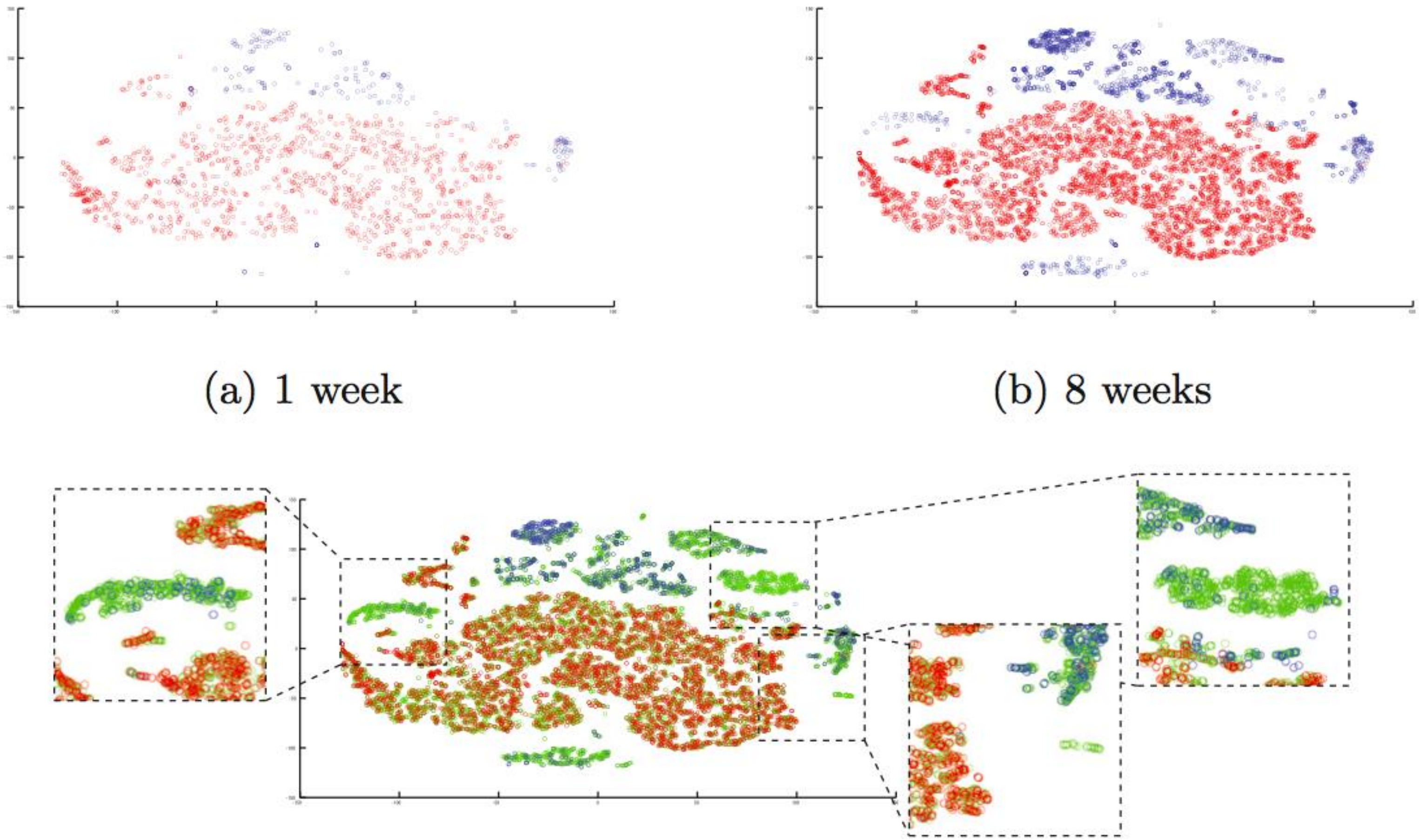
References

- Siti Hajar Aminah Ali, Seiichi Ozawa, Tao Ban, Junji Nakazato, Jumpei Shimamura: A neural network model for detecting DDoS attacks using darknet traffic features. IJCNN 2016: 2979-2985
- Tao Ban, Shaoning Pang, Masashi Eto, Daisuke Inoue, Koji Nakao, Runhe Huang: Towards Early Detection of Novel Attack Patterns through the Lens of a Large-Scale Darknet. UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld 2016: 341-349
- Hironori Nishikaze, Seiichi Ozawa, Jun Kitazono, Tao Ban, Junji Nakazato, Jumpei Shimamura: Large-Scale Monitoring for Cyber Attacks by Using Cluster Information on Darknet Traffic Features. INNS Conference on Big Data 2015: 175-182
- Tao Ban, Lei Zhu, Jumpei Shimamura, Shaoning Pang, Daisuke Inoue, Koji Nakao: Behavior Analysis of Long-term Cyber Attacks in the Darknet. ICONIP (5) 2012: 620-628

DDoS-event Detection in the Darknet

- Goals
  - Early detection and warning of DDoS attacked hosts.
  - Differentiating victim scanners from active scanners.
  - Extend the intelligence learned from conventional attacks to newly targeted services – E.g. DRDoS attacks.

New Attack Patterns Appear In the Darknet (DDoS)

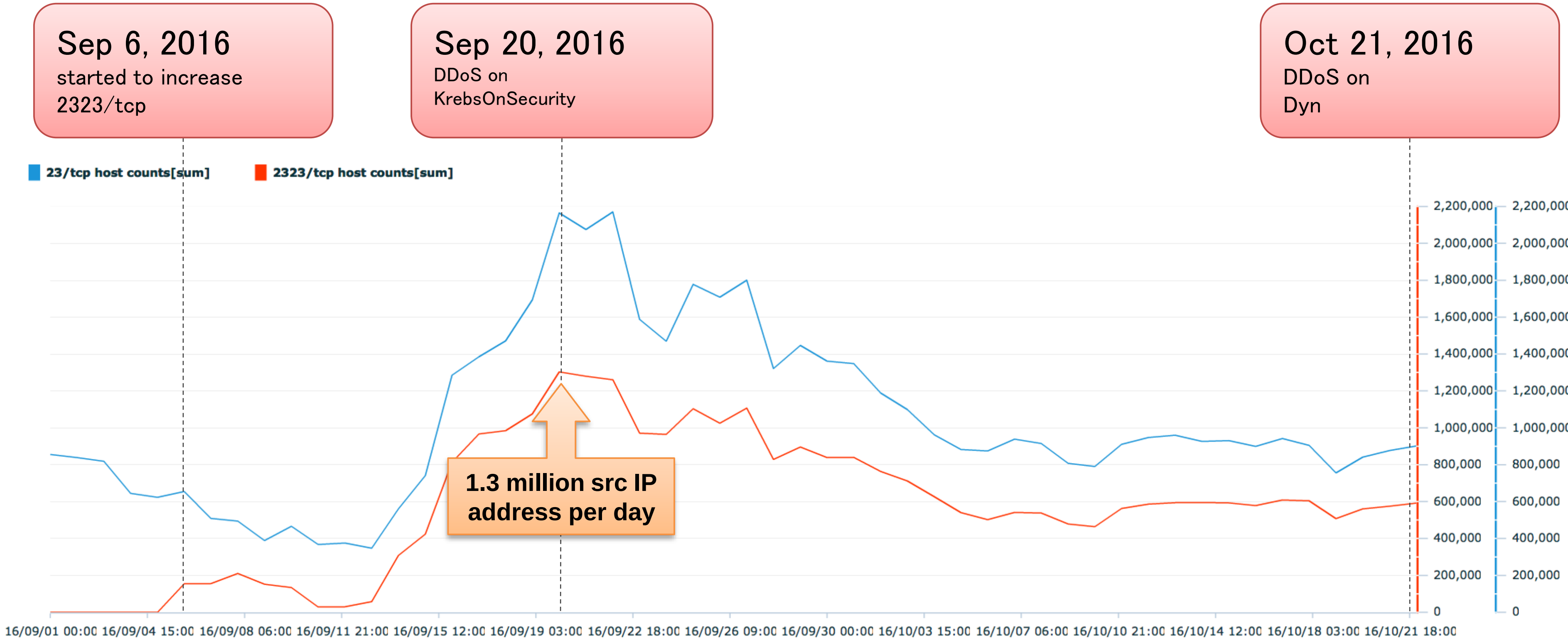


(a) 1 week (b) 8 weeks

(c) 6 months

Mirai in Darknet

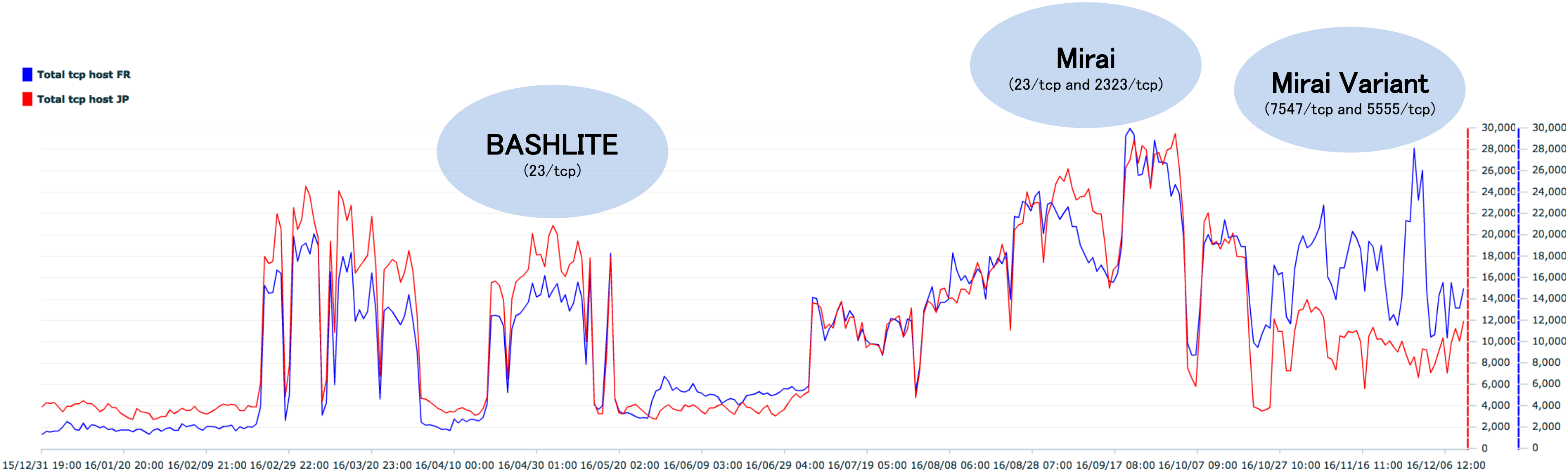
- Scanning to Telnet (23/tcp and 2323/tcp)
- Intrusion using simple IDs and Passwords
- Source codes are uploaded on GitHub



Number of Unique Hosts on 23/tcp and 2323/tcp

(Sep 1, 2016 – Oct 21, 2016)

Darknet Traffic FROM FR and JP

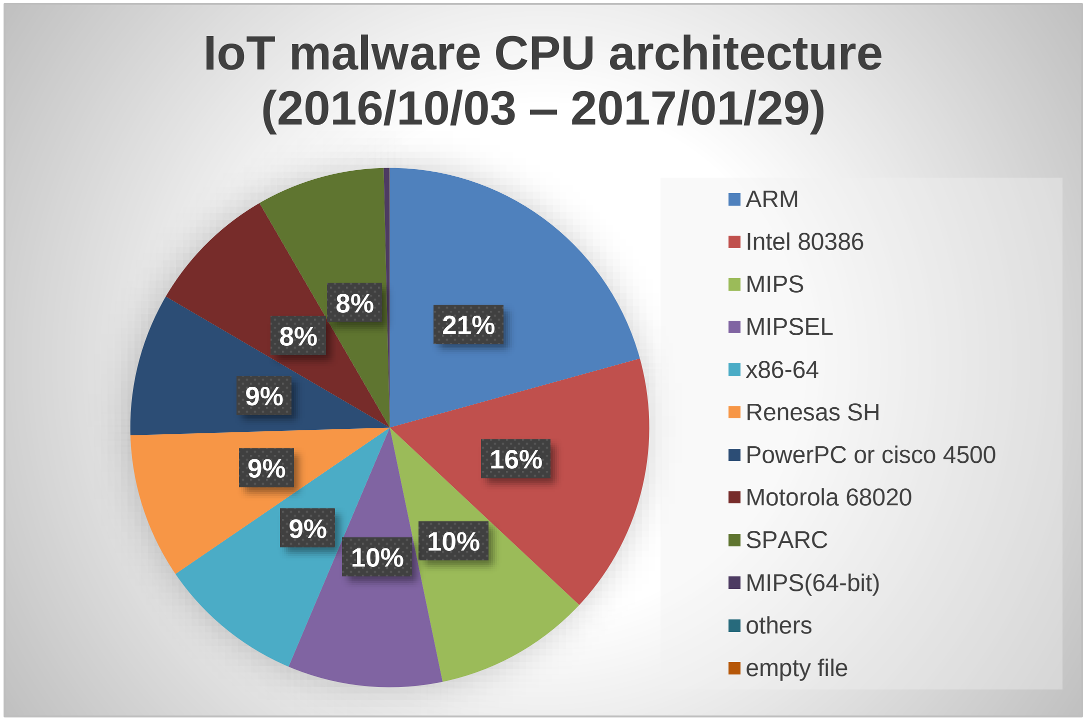


TCP Unique Hosts per Day (January 2016 – Dec 2016)

Preliminary Analysis on IoT Malware (1/2)

- Investigate the ratio of packed IoT malware using LYDA 2007*.
- Malware samples are captured by IoTPOT developped by YNU.

CPU ARCH	CNT
ARM	2714
Intel 80386	2130
MIPS	1279
MIPSEL	1263
x86-64	1191
Renesas SH	1187
PowerPC or cisco 4500	1165
Motorola 68020	1075
SPARC	1048
MIPS (64-bit)	46
others	2
empty file	1



*R. Lyda et al. "Using entropy analysis to find encrypted and packed malware," IEEE Security & Privacy 5.2 (2007).

Next Step

- **Cross analysis of IoT malware between FR and JP**

- **Deploy new honeypot systems for sharing new data**
  - ✓ IoTPOT [1]
  - ✓ AmpPot [2]

- **Joint paper**

- **Joint budget**

[1] Yin Minn Pa Pa, Shogo Suzuki, Katsunari Yoshioka, and Tsutomu Matsumoto, Takahiro Kasama, Christian Rossow, “IoTPOT: Analysing the Rise of IoT Compromises,” 9th USENIX Workshop on Offensive Technologies (USENIX WOOT 2015).

[2] Lukas Krämer; Johannes Krupp, Daisuke Makita, Tomomi Nishizoe, Takashi Koide, Katsunari Yoshioka, Christian Rossow, “AmpPot: Monitoring and Defending Against Amplification DDoS Attacks,” 18th International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2015).