Capture and study of attackers in Darknet

Manobala Namasivayam Nirmala

Supervisors: Prof. Isabelle Chrisment,
Dr. Jerome Francois

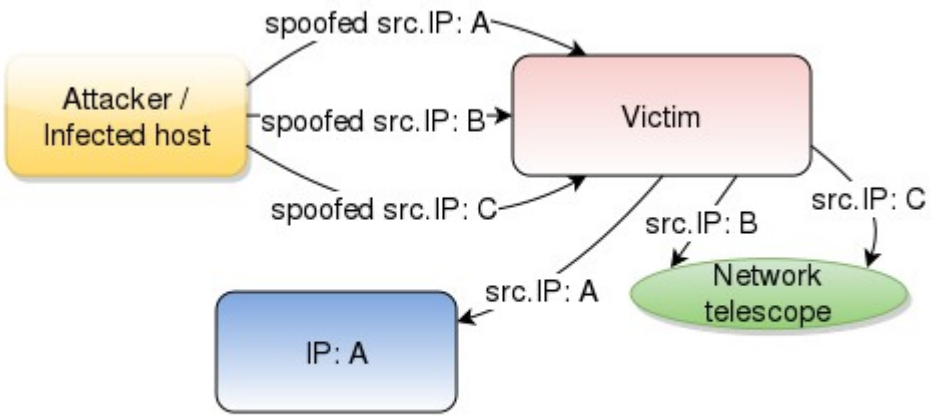
Index

Context

- The awareness in cyber-security is on the rise due to the many exploits and breaches of confidential data
- There are several tools to monitor, detect and mitigate attacks that take place
- Darknet is a monitoring method used to collect attack data and predict attacker behavior

Definition

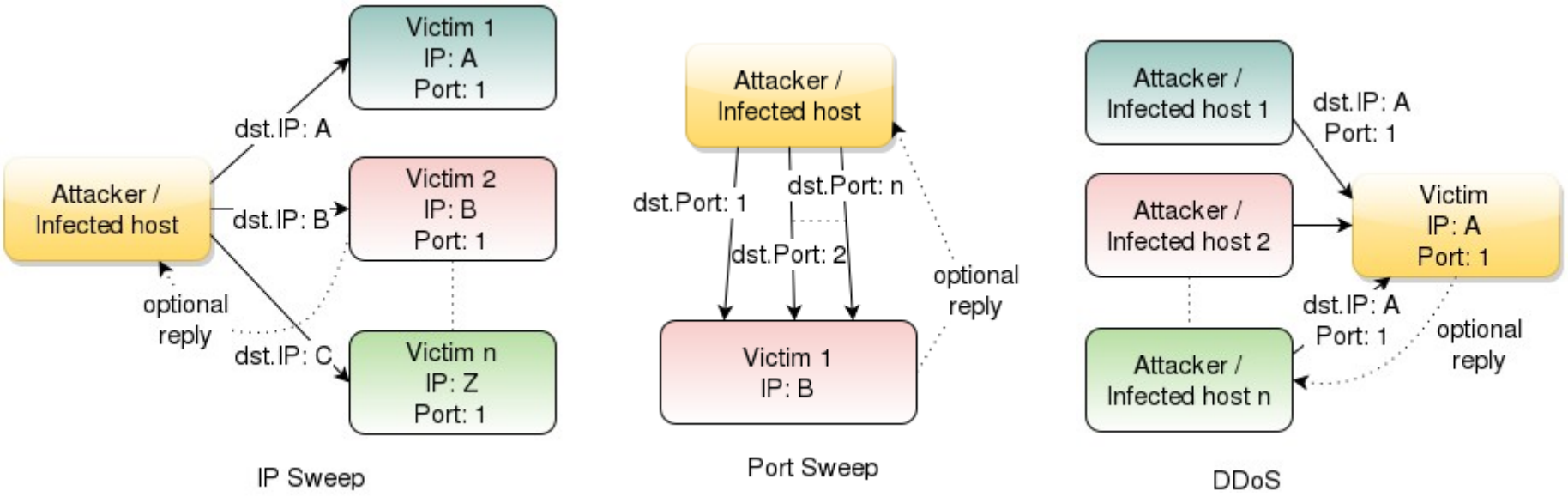
- A darknet comprises of IP addresses for which there are no associated valid services or hosts.
- Because of this fact, all encountered traffic is mostly illegitimate.



Backscatter

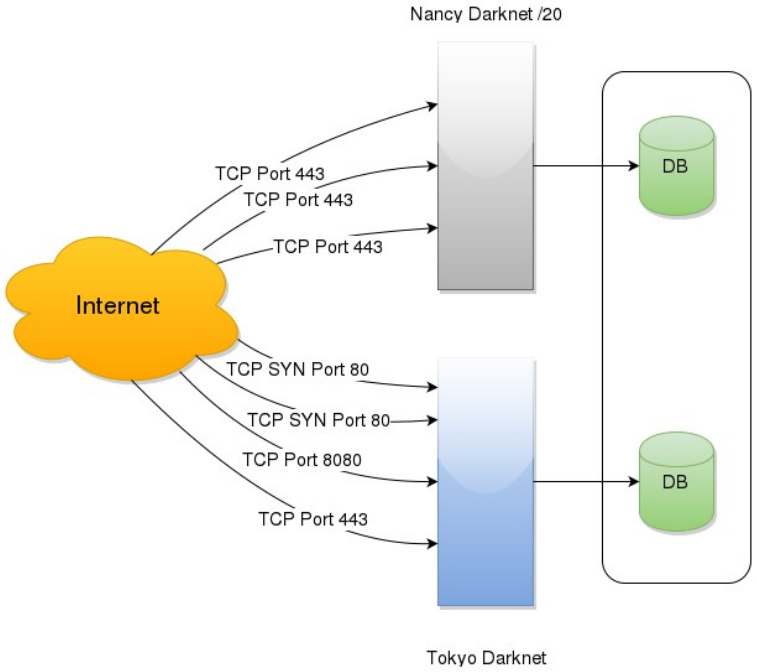
Characteristics

- IP Sweep – find reachable hosts
- Port sweep – find services running on a host
- DDoS – multiple attackers/ infected hosts attacking a particular IP address



Research Interest

- Monitoring attacker behavior
- Determine targeted services and ports
- Geographical variance in attack pattern
- Evolution of attacker behavior over time

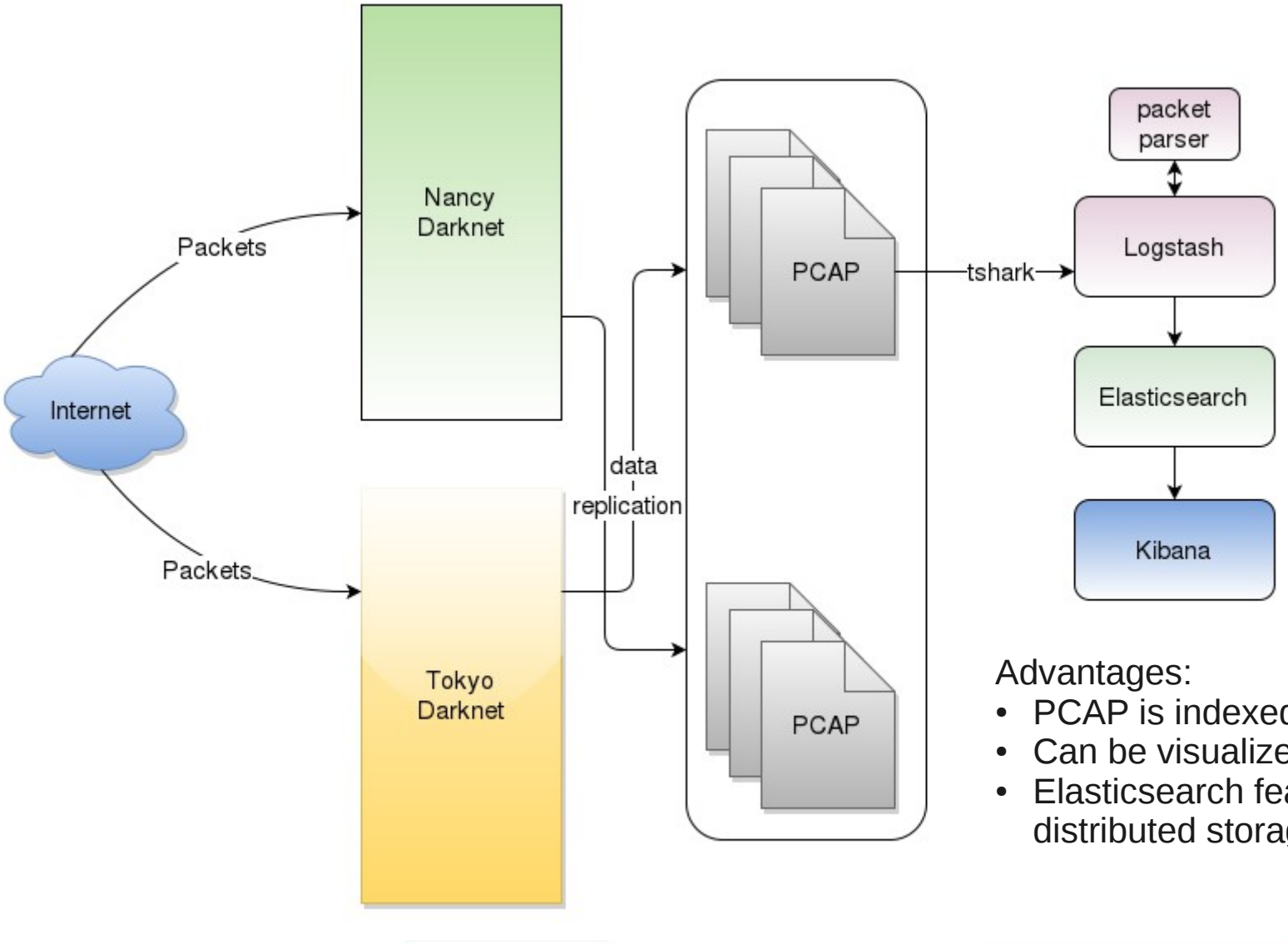


Objectives

- Index huge volume of darknet network traffic
- Extract and visualize statistics
- Extract and characterize attacks
- Compare darknet data obtained in Nancy with that of Tokyo

- Introduction
  - Context
  - Definition
  - Characteristics
  - Research Interest
  - Objective
- Contributions
  - Analyzer architecture
  - Graph representation
  - Community detection
  - Clustering
- Results
  - Dataset
  - Targeted ports
  - Extracted attacks
- Conclusion

Analyzer Architecture



Advantages:

- PCAP is indexed
- Can be visualized in kibana
- Elasticsearch features distributed storage

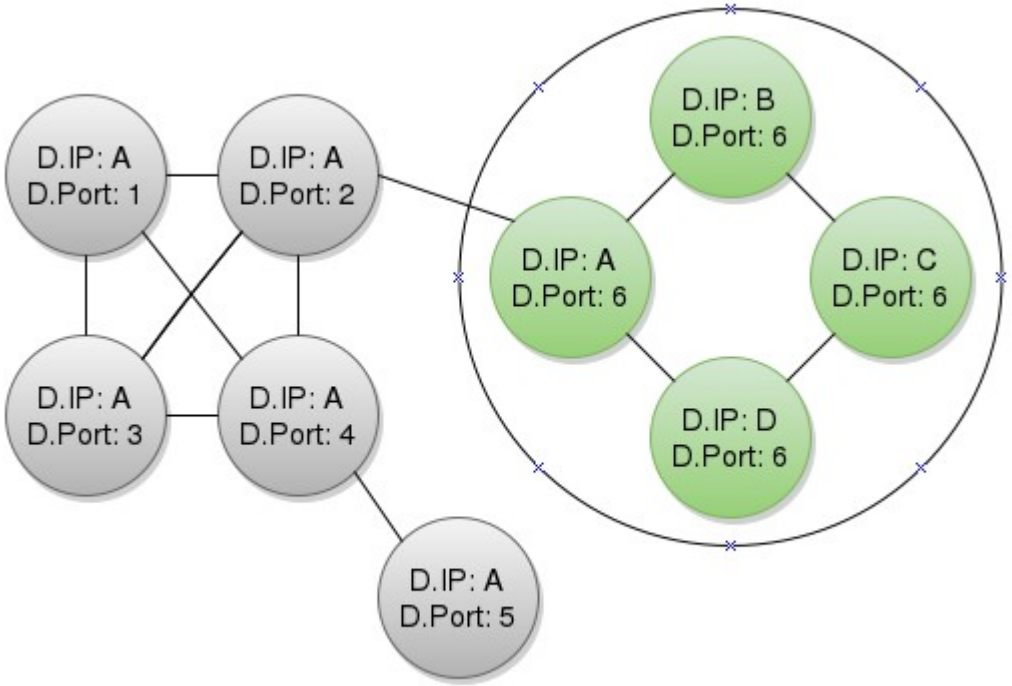
Graph representation

- N packets: $\{p_0,...,p_N\}$
- N nodes: $\{n_0,...,n_N\}$
- S edges: $\{e_0,...,e_S\}$

$$\exists e_i,\ e_i = \langle n_i, n_j \rangle \Leftrightarrow (n_i.\text{proto} = n_j.\text{proto}) \wedge (n_i.\text{time} - n_j.\text{time} = \mu) \wedge$$
$$((n_i.\text{dst\_ip} = n_j.\text{dst\_ip}) \vee \quad \text{//Port Sweep}$$
$$(n_i.\text{dst\_port} = n_j.\text{dst\_port}) \vee \quad \text{//IP Sweep}$$
$$(n_i.\text{src\_ip} = n_j.\text{src\_ip})) \quad \text{//DoS}$$

μ – Constant time

Community detection

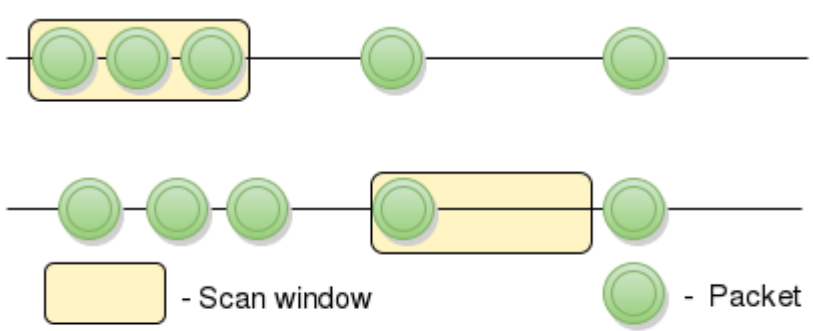


Problems:

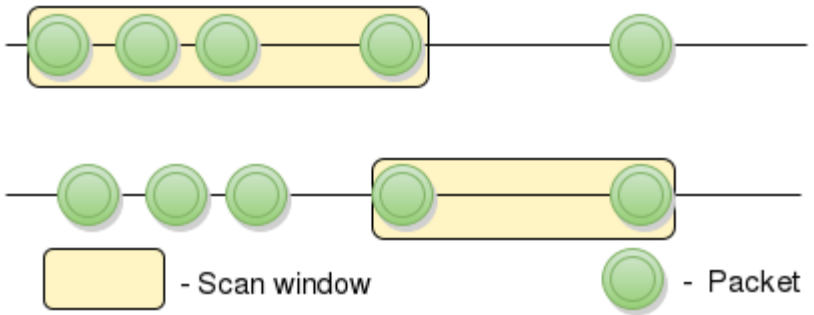
- Representing an entire day's traffic is difficult as graph size becomes unmanageable.
- Owing to the edge criteria, lot of edges are formed (approx. > 900000)

Streaming Clustering

- A dynamic stream of data is processed by a clustering algorithm.
- The drawback in using this algorithm is the inability to detect clusters spread over variable time range.



Fixed scan window
misses packets



Adaptive scan window
captures packets effectively

Adaptive streaming clustering

Cluster list – cl, dead cluster list – dcl

```
foreach read_packet p
    foreach c in cl
        if c.fwd + c.end < p.time
            move c from cl to dcl
        else
            find = false
            if p is_related to c
                c.add(p)
                find = true
                c.fwd = (p.time – c.end + c.fwd)/fwd.factor
                c.end = p.time
            if !find
                cl.create_new_cluster(p)
                start = end = p.time
                fwd = 1800000
```

$$\exists e_i,\ e_i = \langle n_i, n_j \rangle \Leftrightarrow (n_i.\text{proto} = n_j.\text{proto}) \wedge (n_i.\text{time} - n_j.\text{time} = \mu) \wedge$$

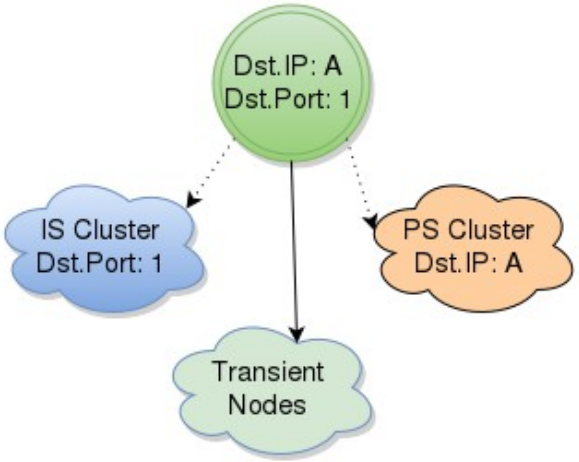
$$((n_i.\text{dst\_ip} = n_j.\text{dst\_ip}) \vee \quad \text{//Port Sweep}$$

$$(n_i.\text{dst\_port} = n_j.\text{dst\_port}) \vee \quad \text{//IP Sweep}$$

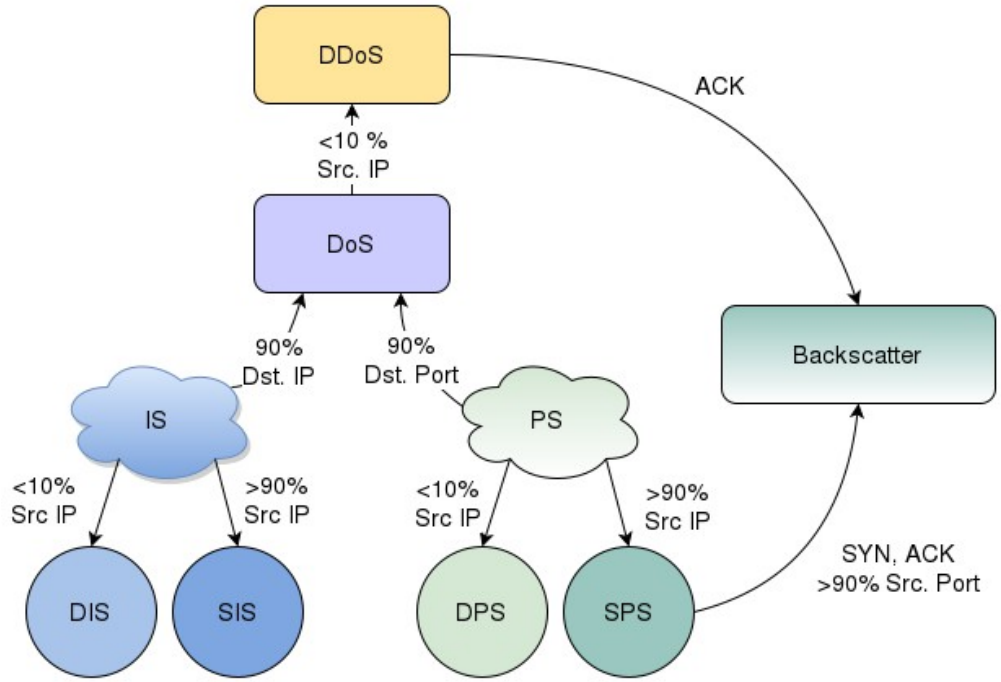
$$\sout{(n_i.\text{src\_ip} = n_j.\text{src\_ip}))} \quad \sout{\text{//DoS}}$$

Cluster extraction and labeling

- Transient node
  - Node that can belong to one of several clusters
- Clusters after expiration are kept in memory until transient nodes are resolved



Probability of multiple cluster association

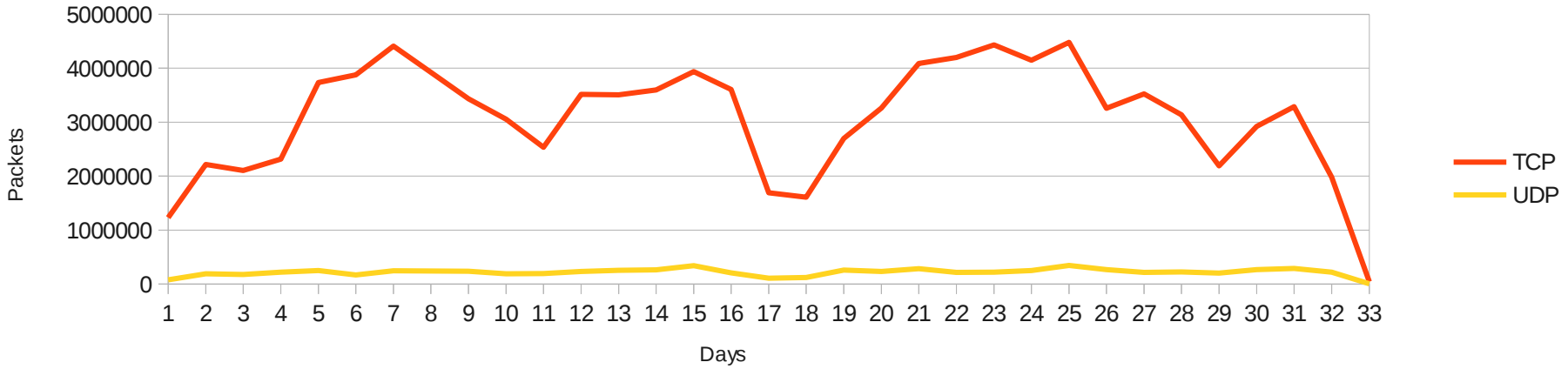


Cluster Labeling

- Introduction
  - Context
  - Definition
  - Characteristics
  - Research Interest
  - Objective
- Contributions
  - Analyzer architecture
  - Graph representation
  - Community detection
  - Clustering
- Results
  - Dataset
  - Targeted ports
  - Extracted attacks
- Conclusion

Dataset

- Network data gathered during the month of January 2015 from darknets in Nancy and Tokyo was used for analysis.
- Data analysis on Jan 16, 2015 could not be performed due to power failure.
- Intel Core i5-4590 CPU(16GB ram) was used for computation.

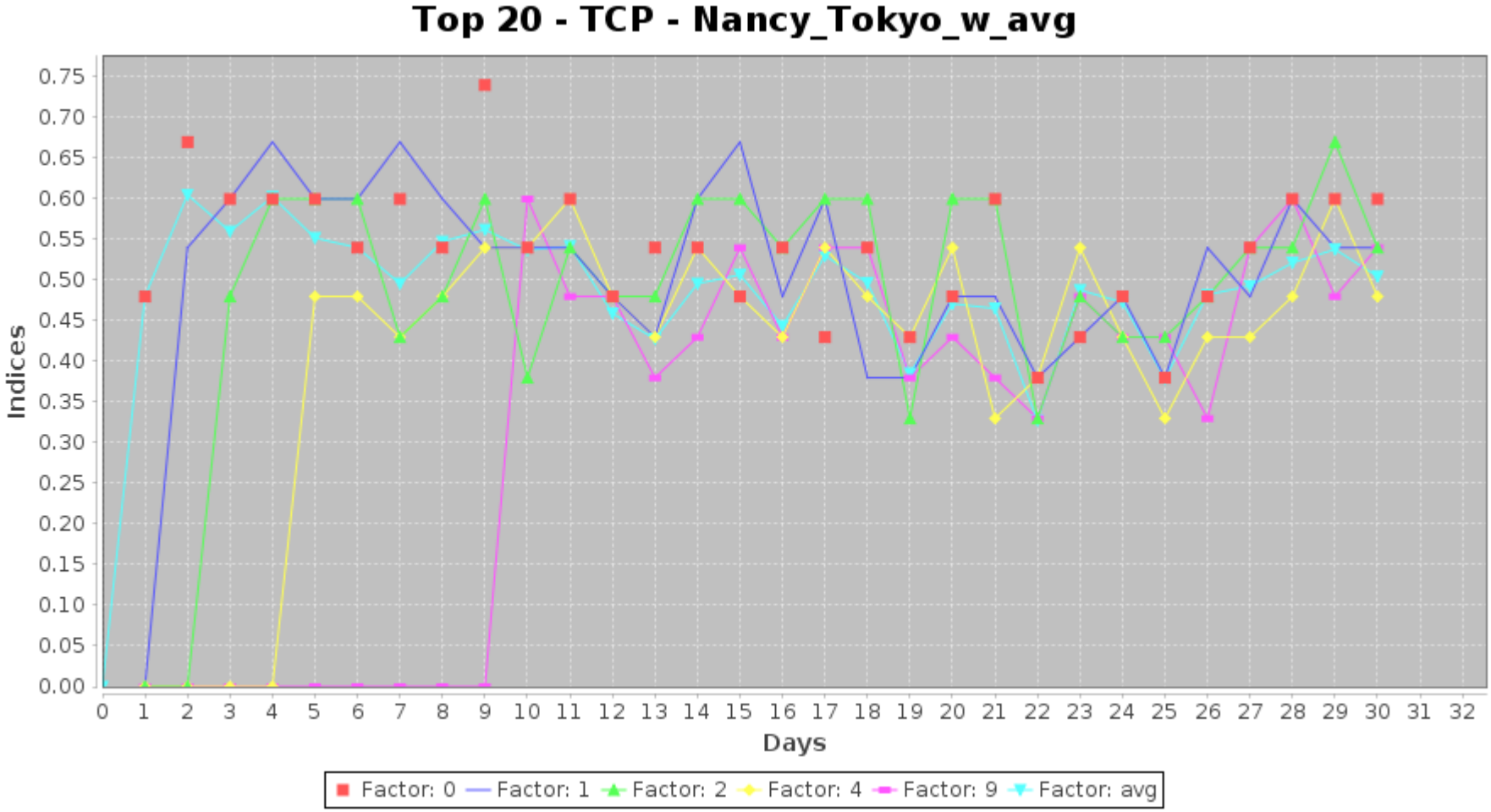


Targeted ports

- To compare the similarities between top targeted ports across the considered period Jaccard index was used.
- N days : $\{n_1,...,n_N\}$
- Top ports of a day in Nancy : $\{n_1(1),...,n_N(i)\}$
- Top ports of a day in Tokyo : $\{t_1(1),...,t_N(i)\}$

Jaccard index : $\frac{n_i \cap t_{j+f}}{n_i \cup t_{j+f}}$

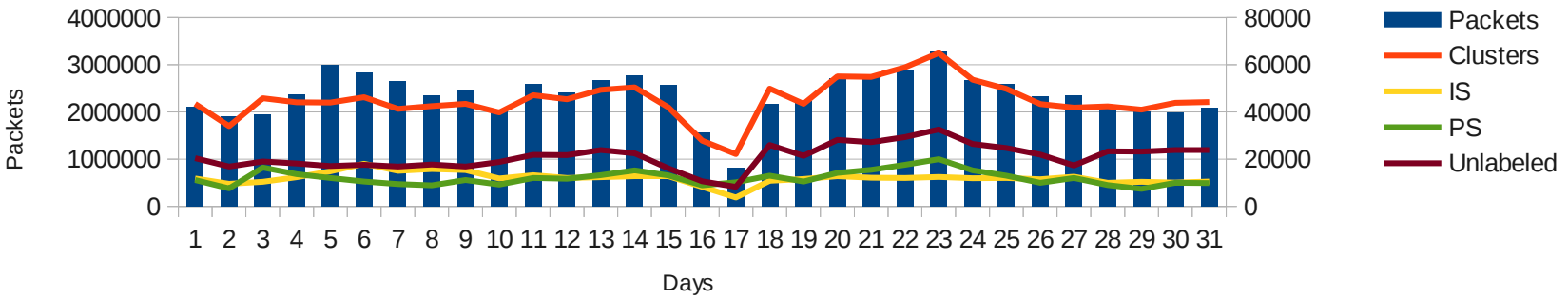
Targeted ports – TCP – Jaccard Index



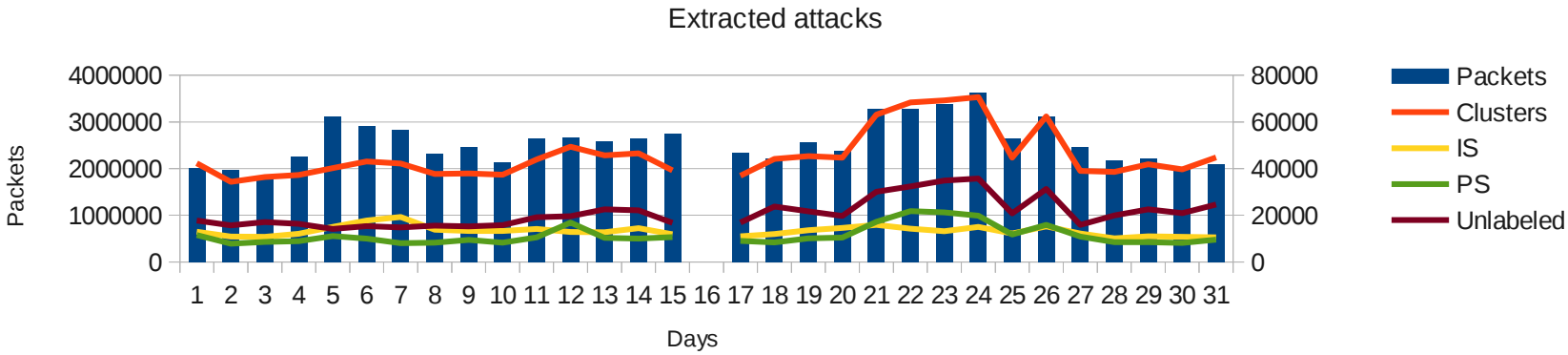
Protocol	Port	Service
TCP	2222	Rockwell-csp2 , BackDoor.Botex, SweetHeart, Rootshell, Way
UDP	27397	W32.Chaim
UDP	34555	[trojan] Trinoo

Extracted attacks

Nancy



Tokyo



The existence of a higher number of unlabeled clusters imply the need for labeling using the decision tree seen earlier.

- Introduction
  - Definition
  - Characteristics
  - Research Interest
  - Objective
- Contributions
  - Analyzer architecture
  - Graph representation
  - Community detection
  - Clustering
- Results
  - Dataset
  - Cluster labeling
  - Targeted ports
  - Extracted attacks
- Conclusion

Conclusion

- Darknet data was successfully indexed in elasticsearch.
- Statistical analysis of darknet data was done
- An adaptive clustering algorithm was implemented
- Analysis was done to determine the difference in data between Tokyo & Nancy
- From analyzing the traffic pattern, in addition to exploits, even some of the standard services are targeted by attackers
- Future work
  - Implement cluster labeling