Analytics and Security Monitoring

Jérôme François, Abdelkader Lahmadi,
Frédéric Beck, Sofiane Lagraa, Loïc Rouch

jerome.francois@inria.fr RESIST Team

Outline

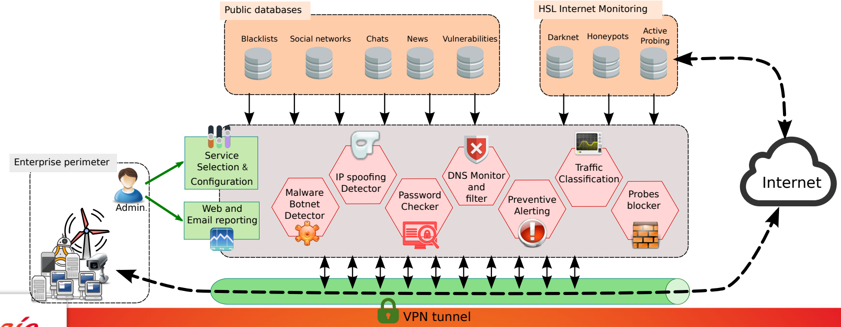
1. Introduction
2. High Security Lab (HSL)
3. Methods overview
4. Network Analytics Status

Challenges

- Why monitoring the security at an Internet-wide scale
  - Operating network security often means firewall, intrusion detection, VPN,...
  - Security risks of your own organization is not independent of the security of your neighbors
  - Knowing the risks and attacks that occur in Internet is important
  - Not only major outbreaks and vulnerability catalogs but also small events, increasing trends....
- Challenges
  - Internet traffic as a global scale is similar to noise → identify interesting/useful/valuable events
  - Correlation of Internet and internal events/logs
  - Encryption is everywhere

The Inria AMICS platform

- Make research results in security analytics available to all
  - Combine live data from monitored network, large-scale security sensors and public databases
  - VPN + customizable advanced services (botnet detection, identity spoofing, password leaks...)



Outline

1. Introduction
2. High Security Lab (HSL)
3. Methods overview
4. Network Analytics Status

A dedicated platform for security sensors and experiments

- Isolated entity within Inria
- Hosts AMICS
- A telescope with several sensors:
  - Honeypots
  - Darknet
- Tons of data, mainly network data but also system logs, malware binaries...
- Major questions:
  - Is there something valuable in all the data we collect?
  - How to extract it?

A dedicated platform for security sensors and experiments

- Isolated entity within Inria
- Hosts AMICS
- A telescope with several sensors:
  - Honeypots
  - Darknet
- Tons of data, mainly network data but also system logs, malware binaries...
- Major questions:
  - Is there something valuable in all the data we collect?
  - How to extract it? at the right time

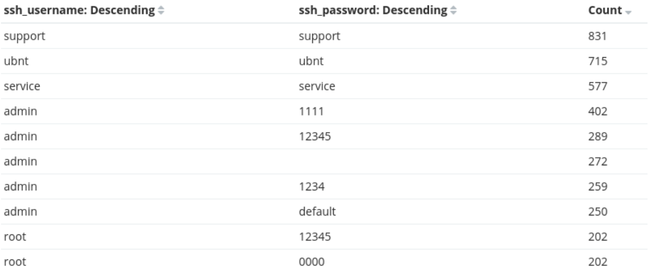
A dedicated platform for security sensors and experiments

- Isolated entity within Inria
- Hosts AMICS
- A telescope with several sensors:
  - Honeypots
  - Darknet
- Tons of data, mainly network data but also system logs, malware binaries...
- Major questions:
  - Is there something valuable in all the data we collect?
  - How to extract it?
- Once we know where to look at, it becomes evident!

Outline

1. Introduction
2. High Security Lab (HSL)
   - Honeypot
   - Darknet
3. Methods overview
4. Network Analytics Status

Top SSH password attemps



ssh_username: Descending	ssh_password: Descending	Count
support	support	831
ubnt	ubnt	715
service	service	577
admin	1111	402
admin	12345	289
admin		272
admin	1234	259
admin	default	250
root	12345	202
root	0000	202

- Very usual and meaningful passwords
- But some were not well known at the time we discovered them

December 2016: Mirai botnet

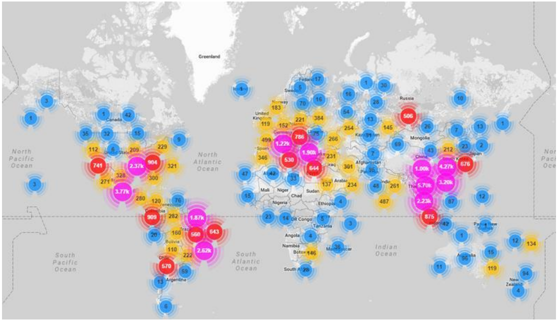


Figure: src: https://www.incapsula.com

- Few passwords tested with some of them observed in our SSH honeypot before the large attack occurs
- Prediction of next targets → derive automatically the semantic of tested passwords

Outline

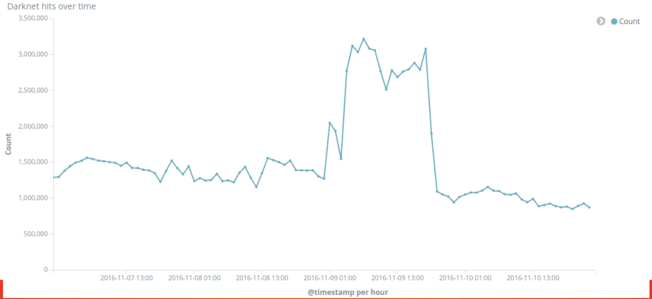
1. Introduction
2. High Security Lab (HSL)
   - Honeypot
   - Darknet
3. Methods overview
4. Network Analytics Status

Darknet

- An entire subnetwork to monitor unsolicited traffic
  - theory : no packets should arrive
  - reality +6 million packets per day since nov. 2014
- Internet background noise (Internet Background Radiation)
- What are the observed IP packets?
  - Scans by malware or attackers trying to identify a target
  - Backscatter (reflection of DDoS attacks)
  - DNS reflection attacks attempts, misconfigurations...

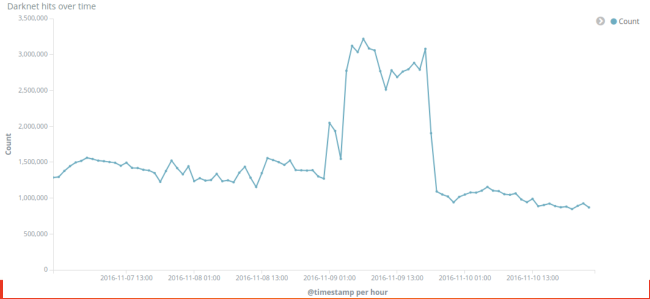
Example

- The anomaly is evident here
- How can it be explained?
    - look at the traffic which counts the most in the abnormal period
    - → a very particular port/service



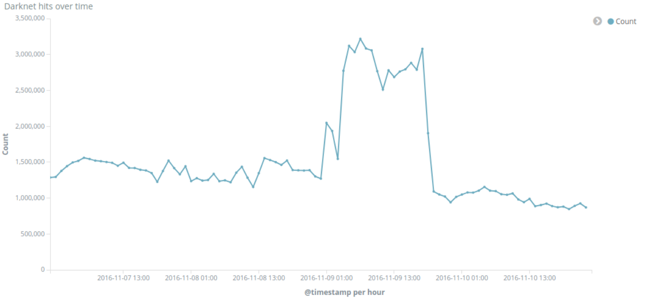
Example

- The anomaly is evident here
- How can it be explained?
    - look at the traffic which counts the most in the abnormal period
    - → a very particular port/service
- so a major attack against this service occurs?



Example

- The anomaly is evident here
- How can it be explained?
  - look at the traffic which counts the most in the abnormal period
  - → a very particular port/service
- so a major attack against this service occurs?
  - look at the date = last US president election



Challenging problems

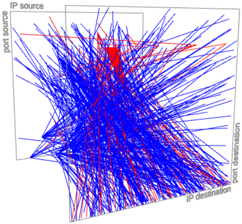
- Relevant information may not be technical (politics, sport events, etc.)
- Security data analytics is not about numerical values but also text (NLP)
- Multiple data sources have to be correlated
- Dependences within data can be complex
- Data can be encrypted [NOMS 2016]

Outline

1. Introduction
2. High Security Lab (HSL)
3. Methods overview
4. Network Analytics Status

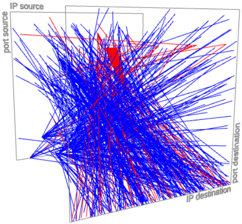
Topological Data analysis [IEEE WIFS 2016]

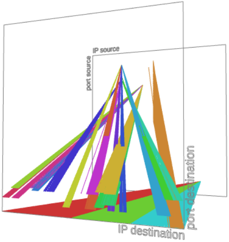
- Apply Mapper method from TDA on darknet traffic to extract attack patterns (scanning, DDoS)



Topological Data analysis [IEEE WIFS 2016]

- Apply Mapper method from TDA on darknet traffic to extract attack patterns (scanning, DDoS)

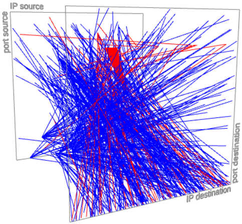


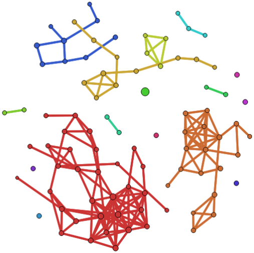


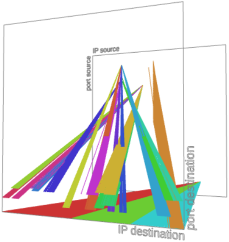
- with scans, DDoS

Topological Data analysis [IEEE WIFS 2016]

- Apply Mapper method from TDA on darknet traffic to extract attack patterns (scanning, DDoS)







- with scans, DDoS
- through an intermediate graph representation built thanks to a clustering algorithm

Mapper method details

- Input : feature vectors of darknet packets (the timestamp, the source and destination IP addresses and ports, and the protocol)
- Parameters: number of intervals (resolution), overlapping percentage (zoom)

1. Filter function f (identity): $\mathbb{R}^6 \rightarrow \mathbb{R}^6$
2. Put data into overlapping bins : $f^{-1}(a_i, b_i)$
3. Cluster each bin using DBSCAN and a distance function
4. Create a graph
    - Vertex: a cluster of a bin
    - Edge: nonempty intersection between clusters

Outline

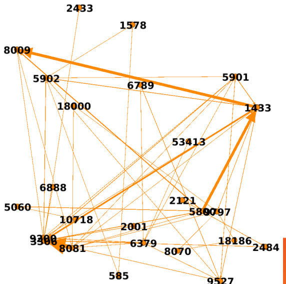
1. Introduction
2. High Security Lab (HSL)
3. Methods overview
4. Network Analytics Status

Need for network-specific ML

- Common errors
  - suppose that there is no necessity to customize the model with context-specific information (*e.g.* the structure and semantics of data)
  - use blackbox approaches (It is actually very hard to benchmark the best algorithms to use)
- Distances between network flows (Euclidian distance?)
  - Not all features are numeric
  - Numeric features are not in the same space
  - Usual distance may not catch the real semantic (e.g. port numbers)

TCP/UDP Port similarities

- Towards a distance/similarity metrics between port numbers
  - security → leverage attacker semantics from darknet monitoring
  - graph mining (community detection) over scans [IM/ANNET 2017]
    - Database service ports: **mysql**: 3306, **redis**: 6379, **ms-sql-s**: 1443 (Microsoft-SQL-Server), **radg**: 6789 (GSS-API for the Oracle), **ttc-ssl**: 2484 (Oracle TTC SSL)
    - Medical service ports: **ohsc**: 18186 (Occupational Health SC), and **biimenu**: 18000 (Beckman Instruments, Inc)



Predicting the next target

- Scanning = early step of an attack
- Defeating scan is thus primordial
- how scans are performed
  - vertically, horizontally with some randomness → stochastic modeling
  - pre-established list of services based on some context / semantic → attack behavior graph modeling
- well defined models → simple/regular ML techniques can (even) be efficient

Analytics and Security Monitoring

*Jérôme François, Abdelkader Lahmadi,
Frédéric Beck, Sofiane Lagraa, Loïc Rouch*

jerome.francois@inria.fr RESIST Team