Big Data Analytics for Host Misbehavior Detection

Miguel Pupo Correia

joint work with Daniel Gonçalves, João Bota (Vodafone PT)

2016 European Security Conference

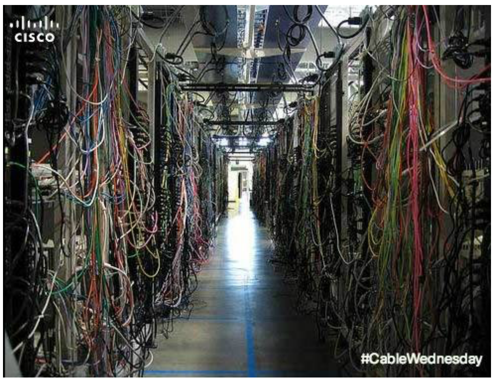
June 2016

TÉCNICO LISBOA

inesc id lisboa

Motivation

- Networks are complex, many attacks happen, how to know if there are compromised hosts?



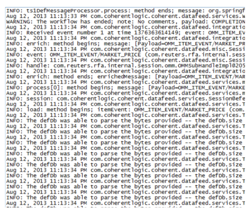
2

Motivation

- What do compromised hosts do?
  – Distributed denial of service attacks
  – Exfiltrating confidential data
  – Sending spam
  – Mapping the network
  – Contact bot command&control centers
  – etc.

The Problem

- **How to discover malicious hosts?**
- Information can be extracted from **logs**
  – files with data about events
- but in complex networks:
  – Logs are huge: <u>big data</u>
  – Logs are heterogeneous: DHCP servers, authentication servers, firewalls, etc., etc.
- so <u>data mining</u> is needed



The Problem

- Problem may be considered **intrusion detection**
  1. *Misuse-based detection*
     - looking for bad patterns (signatures)
  2. *Anomaly-based detection*
     - looking for deviations from good patterns (models)
  3. *Policy-based detection*
     - looking for violations of good patterns
- but
  - 1. and 3. require defining what is bad/good behavior
  - 2. requires large dataset with good behavior
  - **Where to get them with evolving threats?**

5

Solution in a Nutshell (I)

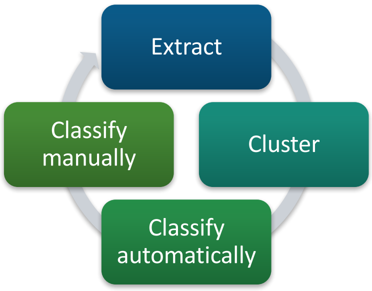
1. Extract features from the logs using MapReduce
   - Features: characteristics, attributes, e.g., num. bytes sent
   - MapReduce allows parallelism, using several servers/cores
2. Obtain *automatically* groups of hosts with similar behavior using **clustering**
   - **unsupervised machine learning**
   - reduces the size of what needs to be classified: clusters
   - condenses the relevant data

6

Solution in a Nutshell (II)

3. Detect misbehavior *automatically* using **classifiers**
   - **supervised machine learning**
4. Classify *manually* the missing clusters
   - **Humans must be kept in the loop due to the evolving nature of threats**

- Repeat, e.g., daily



7

THE APPROACH: PREPARATION

8

Two Phases of the Approach

- **Preparation**: definition and configuration of the detection mechanism
- **Runtime**: Execution of the detection mechanism in runtime

Data Normalization

- How to identify hosts? Name, IP, MAC?
  - We used: MAC and name
  - Dynamic IPs translated to MACs using DHCP log
- Repeated entries in logs?
  - Remove copies
- Dates in different time zones?
  - Translate to a single one

Feature Selection

- Feature engineering is critical in DM / ML; we need:
  - features that allow distinguishing good from bad behavior
  - without knowing which => use a superset, no assumptions
- Types of features and examples (for Tf = 1 day)
  - Session-based, e.g., Number of long sessions
  - Authentication-based, e.g., Number of authentication tries
  - Connection-based, e.g., Num. of TCP packets sent blocked
  - Endpoint-based, e.g., Number of IP addresses with bad reputation contacted

Features

	Session-based
1	Number of sessions
2	Number of long sessions
3	Fraction of sessions of long duration
4	Burst bytes sent
5	Burst bytes received
	Authentication-based
6	Number of admin authentications tries
7	Number of failed admin authentications tries
8	Fraction of admin authentications tries
9	Burst of admin authentications tries
10	Number of authentication tries
11	Number of failed authentication tries
12	Fraction of failed authentication tries
13	Burst of authentication tries
	Connection-based
14-15	Number of packets sent blocked/allowed
16-17	Number of packets received blocked/allowed
18	Burst of packets sent
19	Burst of packets received
20	Fraction of packets sent blocked
21	Fraction of packets received blocked
22-24	Number of TCP/UDP/ICMP packets sent blocked
25-27	Fraction of TCP/UDP/ICMP packets sent blocked
	Endpoint-based
28	Number of IP addresses in the top of malicious subnets
29	Number of IP addresses with bad reputation
30	Number of external IP addresses not contacted last T_f period
31	Number of internal IP addresses not contacted last T_f period
32	Number of external IP address locations not found last T_f period
33	Number of external IP addresses in the malicious AS list
34	Number of external IP addresses in the spam AS list

THE APPROACH: RUNTIME

Feature Extraction

- MapReduce framework (Hadoop)
  - allows parallelizing log processing:
    - one *mapper* per file extracts features
    - *reducer* provides a single output
  - allows taking computation to the nodes that keep the logs
    - if they allow it
- Caches for external data
  - Autonomous System Numbers
  - Suspicious IP, subnets

Clustering

- Means creating groups of entities (hosts) that are similar in terms of features
  - features are normalized to the interval [0,1]
- We use a probabilistic clustering algorithm: Expectation-Maximization (EM)
  - doesn't need prior knowledge of the feature distribution
  - appropriate to cluster large data sets
  - num. of clusters is an input: small percentage of hosts per cluster, except clusters that represent common behaviors

15

Cluster Classification

- Manual – first time and for unclassifiable clusters
  - small number of clusters, so feasible (not thousands of hosts)
  - features marked as primary, secondary, low-relevance
  - feature values classified as VH, H, M, L, VL
  - clusters are assigned a class
- Automatic
  - based on a Naive Bayes algorithm
  - assigns clusters to classes automatically
  - typ. several classes: normal server, normal PC,...

16

EXPERIMENTAL EVALUATION

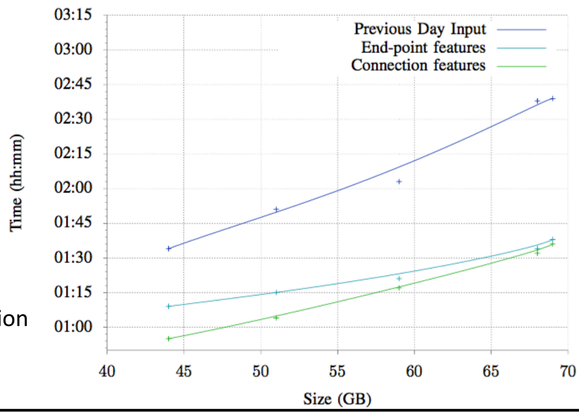
Overview

- Time period = 1 day
- ~300 GB logs for 5 consecutive days
- logs of firewalls, DHCP and authentication servers
- Code in Java
- Hadoop for data processing
- WEKA for machine learning algorithms
- Data processed in a 32-core server
- Number of clusters fixed to 23

Data Processing

Log size per day per log source

Log Source \ Day	1	2	3	4	5
Firewall type 1	51 GB	44 GB	69 GB	68 GB	59 GB
Firewall type 2	18 MB	18 MB	18 MB	18 MB	18 MB
DHCP	14 MB	14 MB	14 MB	14 MB	14 MB
Authentication serv.	222 MB	202 MB	201 MB	197 MB	210 MB



Time for feature extraction versus size of the logs

Classifying the Clusters Manually

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
#	1	55	566	9	207	72	78	61	697	25	86	27	53	1091	54	109	83	459	169	35	34	205	89
%	0.02	1.29	13.27	0.21	4.85	1.69	1.83	1.43	16.34	0.59	2.02	0.63	1.24	25.58	1.27	2.56	1.95	10.76	3.96	0.82	0.80	4.80	2.0
Feature																							
1	VL	VL	VL	VL	VL	VL		VL	VL	VL		VL	VL	VL	VL	VL		L	VL		VL	VL	
2	VL	VL	VL	VL	VL	VL	L	VL	VL	VL		VL	VL	VL	VL	L		L	VL		VL	VL	L
4	VL	VL	VL	VL	VL	VL		VL	VL	VL		VL		VL	VL			VL	VL		VL	VL	
5	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL		VL	VL		VL	VL	VL		VL	VL	VL
6																							
7																							
9																							
10	L		VL	VL	VL		VH		VL	VL	VH			VL	VH			VH	VL		VL	VL	
11	VL	VL	VL	VL	VL	VL		VL		VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL
13	VL	VL	VL	VL	VL	VL		VL	VL	VL		VL	VL	VL	VL	VL	VL	VL	VL		VL	VL	
14	VL	VL		VL	L		VL	VL	VL	VL	VL	VL		VL	VH	VL	L	VL	VL	VH	VL	VL	
15	VL	VL	VL	VL	VH		VL	VL	VL	VL	VL	VL	VH	VL		VL		VL				VL	
16	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL		VL	VL	VL		VL	VL		VL	VL	VL
17	VL	VL	VL	VL		VL	VL	VL	VL	VL	VL	VL		VL	VL	VL		L			VL	L	VL
18	VL		VL	VL	VH			L	H	VL	VL		VH	VL	VH			VL	M		VH		
19	VL	VL	VL	VL	VH	VL	VL	VL	VL		VL	VL		VL	VL	VL		VL			VL	M	VL
22	VL	VL		VL	VL		VL	VL	VL	VL	VL	VL		VL		VL		VL	VL	VH	VL	VL	
23	VL	VL	VL	VL	L		VL	VL	VL	VL	VL	VL		VL	VH	VL	VL	VL	VL		VL	VL	
24	VL	L		VL	VL	VL	VL	VL	VL	VL	VL	L	VL	VL		VL	VL	VL	VL			VL	
28																							
29	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL		VL	VL	VL	VL	VL	VL	VL		VL	VL
30	VL	VL	VL	VL	VL		VL	VL	VL	VL	VL	VL		VL	VL	VL	VL	VL	VL			VL	VL
31	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL		VL	VL	VL
32	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL		VL	VL	VL	VL	VL	VL	VL	VL	VL	VL
33	VL	VL	VL	VL	VL		VL	VL	VL	VL	VL	VL	VH	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL
34	VL	VL	VL	VL	VL		VL	VL	VL	VL	VL	VL	VH	VL		VL		VL	VL			VL	VL

13 - Suspicious AS

15 - Blocked UDP and Authentication Tries

20 - Blocked TCP

Cluster description in terms of hosts it contains (total 4265) and primary features

20

Classifying the Clusters Manually

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13
#	1	55	566	9	207	72	78	61	697	25	86	27	53
%	0.02	1.29	13.27	0.21	4.85	1.69	1.83	1.43	16.34	0.59	2.02	0.63	1.24
Feature													
1	VL	VL	VL	VL	VL	VL		VL	VL	VL		VL	VL
2	VL	VL	VL	VL	VL	VL	L	VL	VL	VL		VL	VL
4	VL	VL	VL	VL	VL	VL		VL	VL	VL		VL	
5	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	
6													
7													
9													
10	L		VL	VL	VL		VH		VL	VL	VH		
11	VL	VL	VL	VL	VL	VL		VL		VL	VL	VL	VL
13	VL	VL	VL	VL	VL	VL		VL	VL	VL		VL	VL
14	VL	VL		VL	L		VL	VL	VL	VL	VL	VL	
15	VL	VL	VL	VL	VH		VL	VL	VL	VL	VL	VL	VH
16	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	
17	VL	VL	VL	VL		VL	VL	VL	VL	VL	VL	VL	
18	VL		VL	VL	VH			L	H	VL	VL		VH
19	VL	VL	VL	VL	VH	VL	VL	VL	VL		VL	VL	
22	VL	VL		VL	VL		VL	VL	VL	VL	VL	VL	
23	VL	VL	VL	VL	L		VL	VL	VL	VL	VL	VL	
24	VL	L		VL	VL	VL	VL	VL	VL	VL	VL	L	VL
28													
29	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	
30	VL	VL	VL	VL	VL		VL	VL	VL	VL	VL	VL	
31	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL
32	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	VL	
33	VL	VL	VL	VL	VL		VL	VL	VL	VL	VL	VL	VH
34	VL	VL	VL	VL	VL		VL	VL	VL	VL	VL	VL	VH

13 - Suspicious AS

Cluster 13 – 53 hosts:
- problematic features with VH:
  - num. of packets sent
  - bursts of packets sent
  - num. of external IPs contacted in malicious ASs
  - and in spam AS lists
- all the other clusters have VL in the last 2 features!

Cluster description in terms of hosts it contains (total 4265) and primary features

21

Suspicious clusters – bots

- Cluster 15 – 54 hosts
  - problematic features VH:
    - num. of authentication tries
    - num. of packets sent blocked by the firewall
    - bursts of packets sent
    - num. of UDP packets sent blocked by the firewall
- Cluster 20 – 35 hosts
  - problematic features VH:
    - num. of packets sent blocked by the firewall and
    - TCP packets sent blocked by the firewall

22

Conclusions

- Our approach allows identifying malicious entities in a semi-automatic way based on large logs...
- ...without having to say how entities misbehave
- Uses clustering (unsupervised ML) to reduce the size of the problem and
- a classifier (supervised ML) to automatize classification
- Keeps humans in the loop; mandatory due to the evolving nature of threats

Thank you

Learn more:
Big Data Analytics for Detecting Host Misbehavior in Large Logs
Daniel Gonçalves, João Bota, Miguel Correia
14th TrustCom, Aug. 2015

Miguel Pupo Correia
miguel.p.correia@tecnico.ulisboa.pt
http://www.gsd.inesc-id.pt/~mpc/

TÉCNICO LISBOA

inesc id lisboa