Machine Learning Fingerprinting Methods in Cyber Security Domain: Which one to Use?

DACS Workshop
26 June 2018

Martin Laštovička
lastovicka@ics.muni.cz

Motivation

- Obsolete figerprint databases
  - ⇨ Machine learning
- Results from small static networks
  - ⇨ Dynamic wireless network
- Focus on accuracy
  - ⇨ Classification time & memory in a large network

MASARYK UNIVERSITY
Czech Republic

OS Fingerprinting Methodology

TCP/IP Parameters

✓

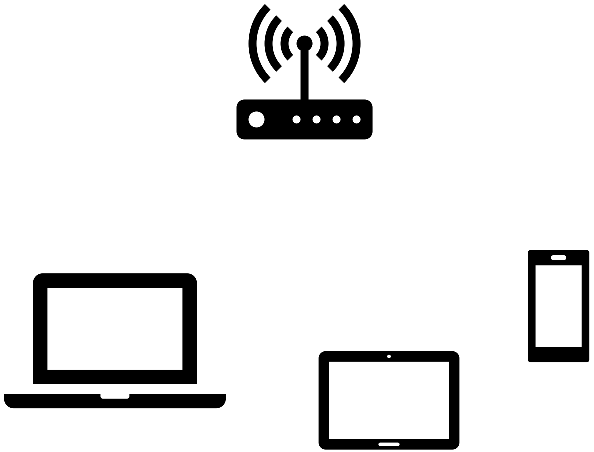
- Time To Live
- TCP Window Size
- TCP SYN Size

✘

- Checksum
- Destination Port
- Maximum Segment Size

synSize	winSize	TTL	OS
52	8192	128	Windows 10.0
52	8192	128	Windows 6.1
52	65535	128	Windows 10.0
60	65535	64	Android 6.0
60	14600	64	Android 4.4
60	29200	64	Ubuntu
64	65535	64	Mac OS X 10.12
64	65535	64	iOS 10.3

Dataset



- One week of flows + logs
  - 79 087 345 flows
  - 21 746 users
  - 25 642 unique MAC (1 692 vendor prefixes)
  - 253 374 Wi-Fi sessions
  - 6 104 unique IP addresses

Ground Truth

May 5 06:30:54 krakonos dhcpd: DHCPREQUEST	for 147.251.x.x from	98:0c:a5:x:x:x	(android-22d1bxxx)	via 147.251.x.x
May 5 06:30:54 krakonos dhcpd: DHCPACK	on 147.251.x.x to	98:0c:a5:x:x:x	(android-22d1bxxx)	via 147.251.x.x
May 5 06:31:17 krakonos dhcpd: DHCPREQUEST	for 147.251.x.x from	38:a4:ed:x:x:x	(Redmi3S-Redmi)	via 147.251.x.x
May 5 06:31:17 krakonos dhcpd: DHCPACK	on 147.251.x.x to	38:a4:ed:x:x:x	(Redmi3S-Redmi)	via 147.251.x.x
May 5 06:31:20 krakonos dhcpd: DHCPREQUEST	for 147.251.x.x from	9c:6c:15:x:x:x	(Windows-Phone)	via 147.251.x.x
May 5 06:31:20 krakonos dhcpd: DHCPACK	on 147.251.x.x to	9c:6c:15:x:x:x	(Windows-Phone)	via 147.251.x.x
May 5 06:36:24 krakonos dhcpd: DHCPREQUEST	for 147.251.x.x from	c0:f2:fb:x:x:x	(Barboras-iPhone)	via 147.251.x.x
May 5 06:36:24 krakonos dhcpd: DHCPACK	on 147.251.x.x to	c0:f2:fb:x:x:x	(Barboras-iPhone)	via 147.251.x.x

MASARYK UNIVERSITY
Czech Republic

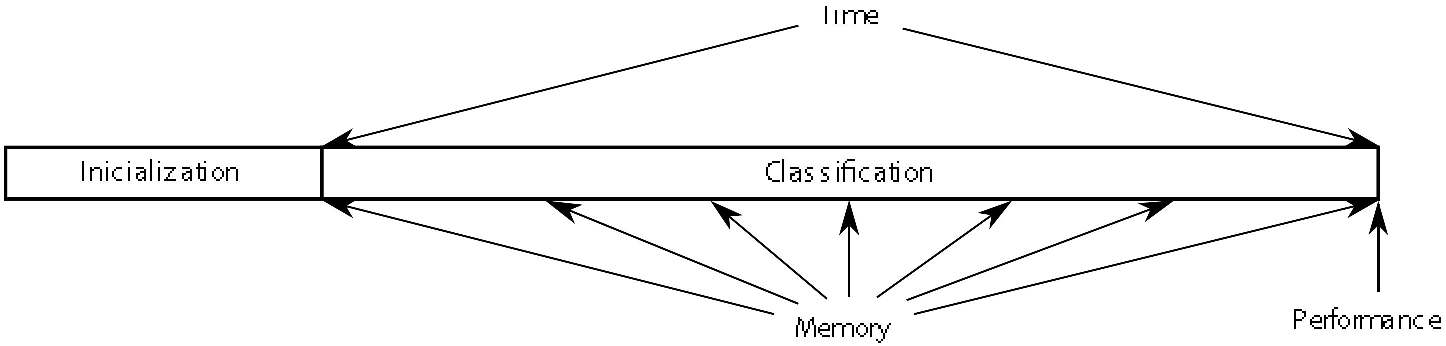
Experiment Settings

Scikit-learn Methods

- Decision tree – CART algorithm
- Naïve Bayes
- k-NN
  - K = 3
  - Euclidean distance metric
- SVM
  - Penalty C = 1
  - Gaussian kernel

Measurement

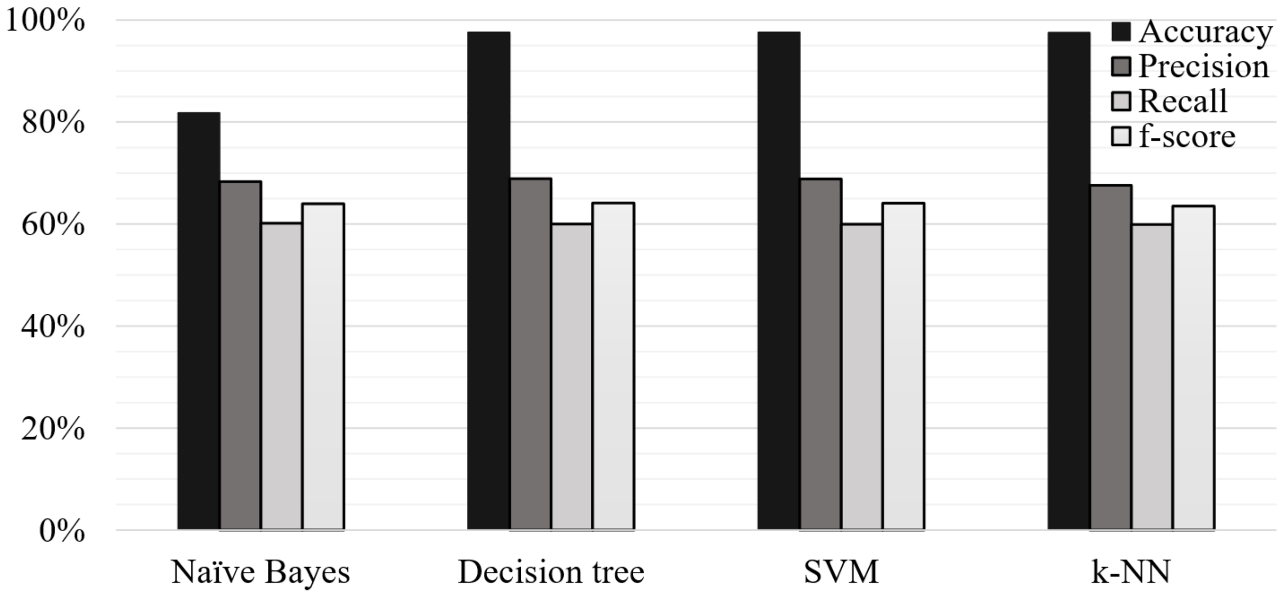
- 20 x (4 x 10) repetitions for each method
  - 4x training set size (1k – 1M)
  - 10x testing set size (1k – 10M)



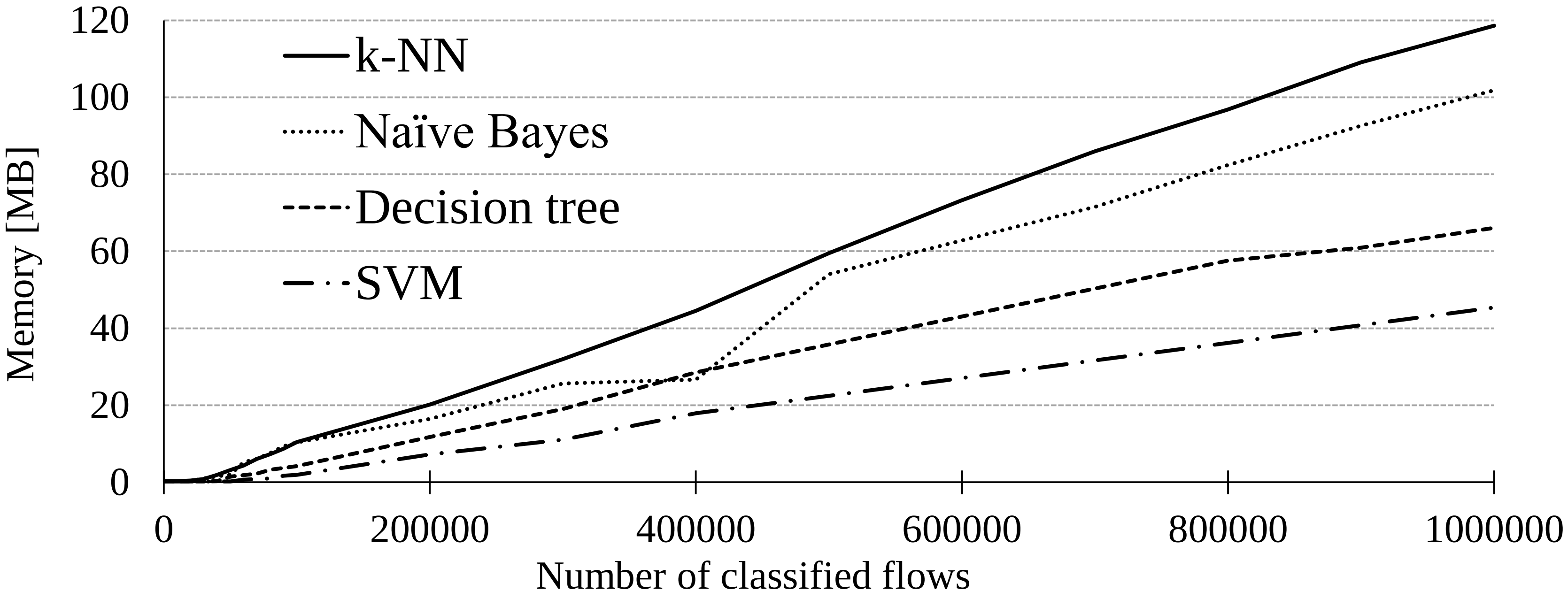
MASARYK UNIVERSITY
Czech Republic

Results

Performance Measures



Memory Consumption



Time Complexity in Seconds

Training size	10k			100k			1M		
Samples	10k	1M	10M	10k	1M	10M	10k	1M	10M
Decision tree	0.0004	0.075	0.810	0.0005	0.083	0.901	0.0005	0.087	0.938
Naïve Bayes	0.001	0.229	4.237	0.001	0.231	4.366	0.001	0.232	4.385
k-NN	0.298	29.82	294.9	2.519	250.1	2494	42.71	4359	42925
SVM	0.374	37.44	368.6	3.407	341.8	3378	33.50	3314	34374

Conclusion

- Memory not a problem
- Performance measures are similar
- Time extremely dependent on model complexity
- Decision trees are best suited for OS fingerprinting in large networks
  - MU network up to 10k flow/s (6.1k avg)

MASARYK UNIVERSITY
Czech Republic

Discussion

Martin Laštovička
lastovicka@ics.muni.cz

Brno Ph.D. Talent Scholarship Holder
Funded by the Brno City Municipality

Brno Ph.D. Talent