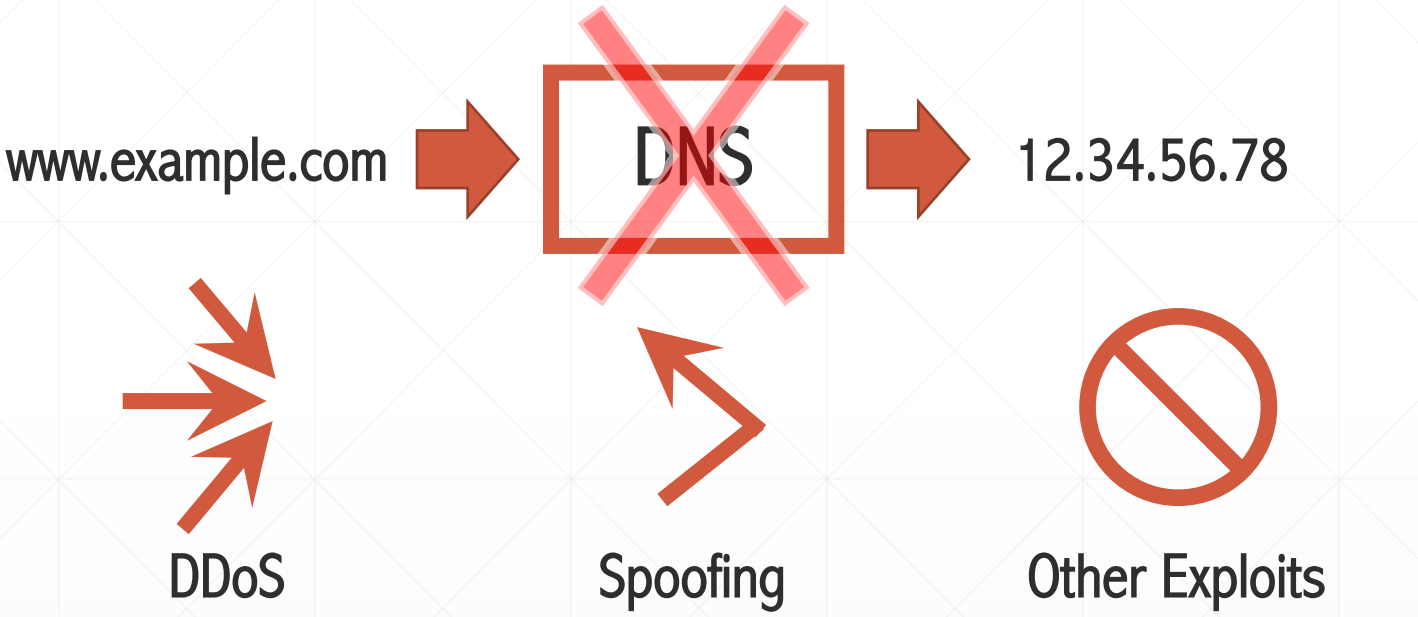
A Practical Machine Learning-Based Framework to Detect DNS Covert Communication in Enterprises

Ruming Tang[1], Cheng Huang[2], Yanti Zhou[3], Haoxian Wu[2], Xianglin Lu[1], Yongqian Sun[4], Qi Li[1], Jinjin Li[3], Weiyao Huang[3], Siyuan Sun[3], and Dan Pei[1]

[1]Tsinghua University [2]Bizseer Technologies [3]Bank of Communications, China [4]Nankai University, China

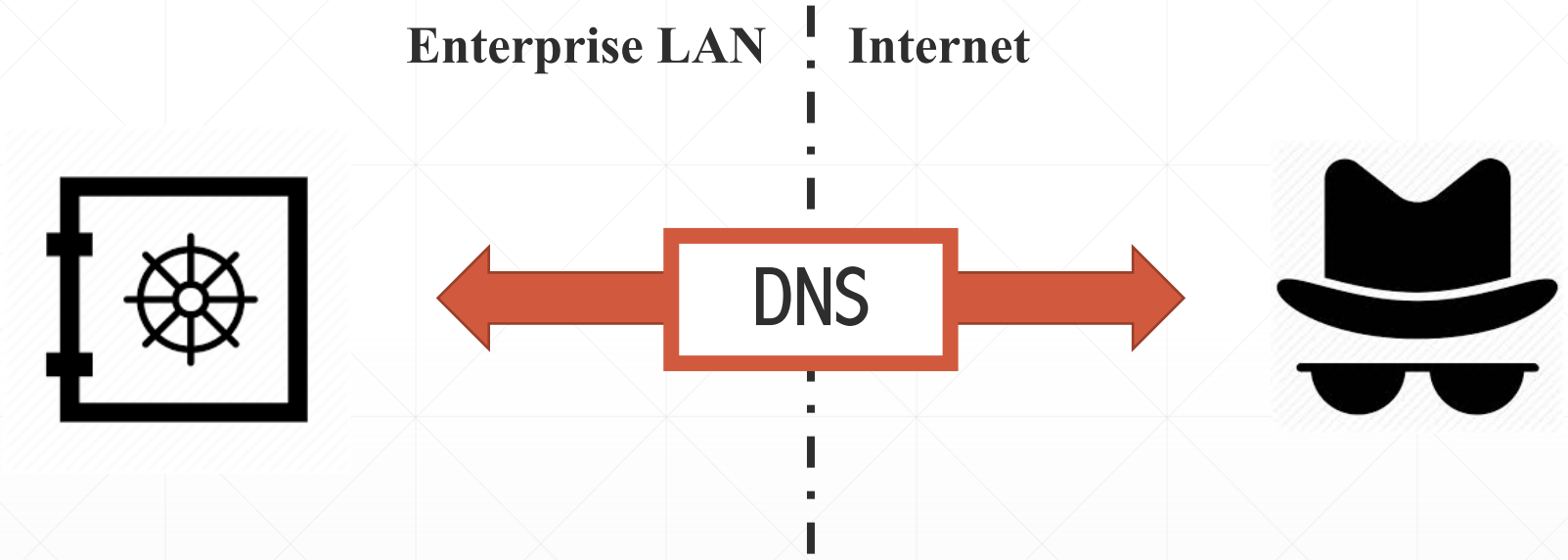
EAI SecureComm 2020 -16th EAI International Conference on Security and Privacy in Communication Networks, October 21-23, 2020, Cyberspace

Domain Name System

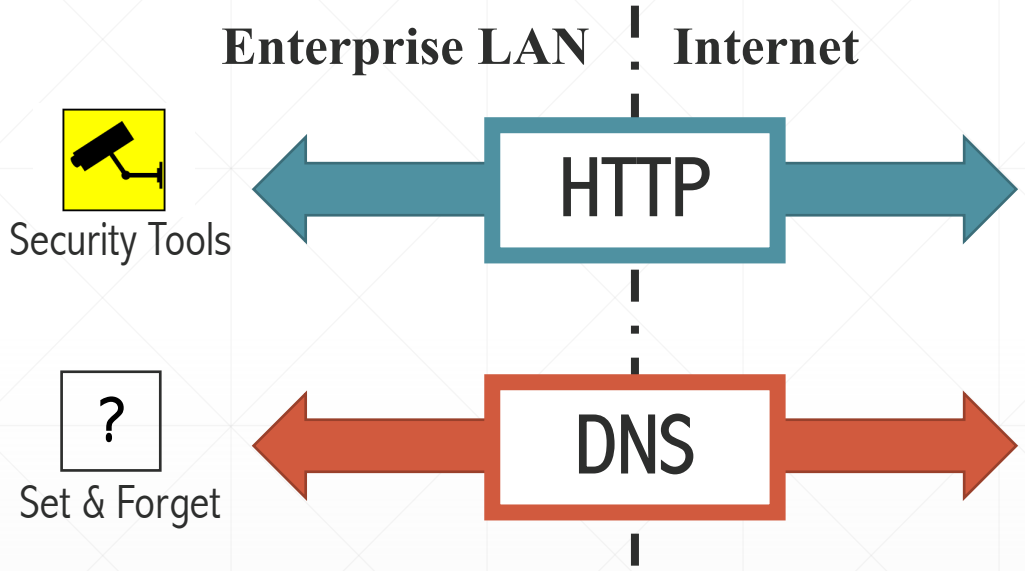


Attacks against DNS infrastructure itself are much easier to be noticed

Domain Name System

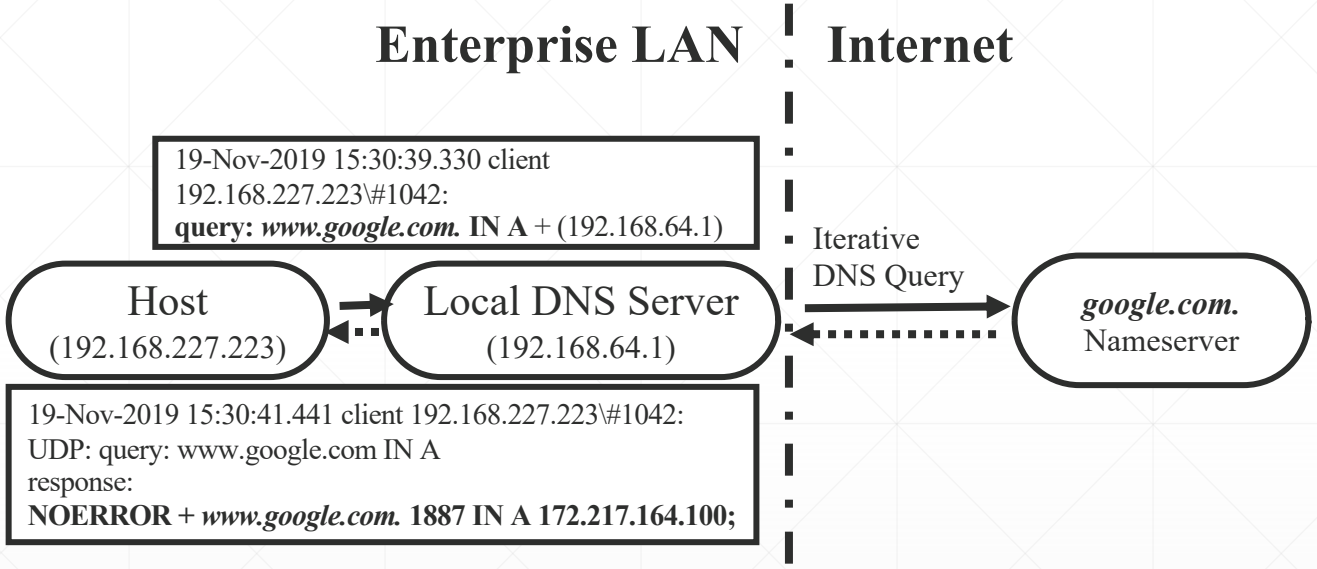


Domain Name System

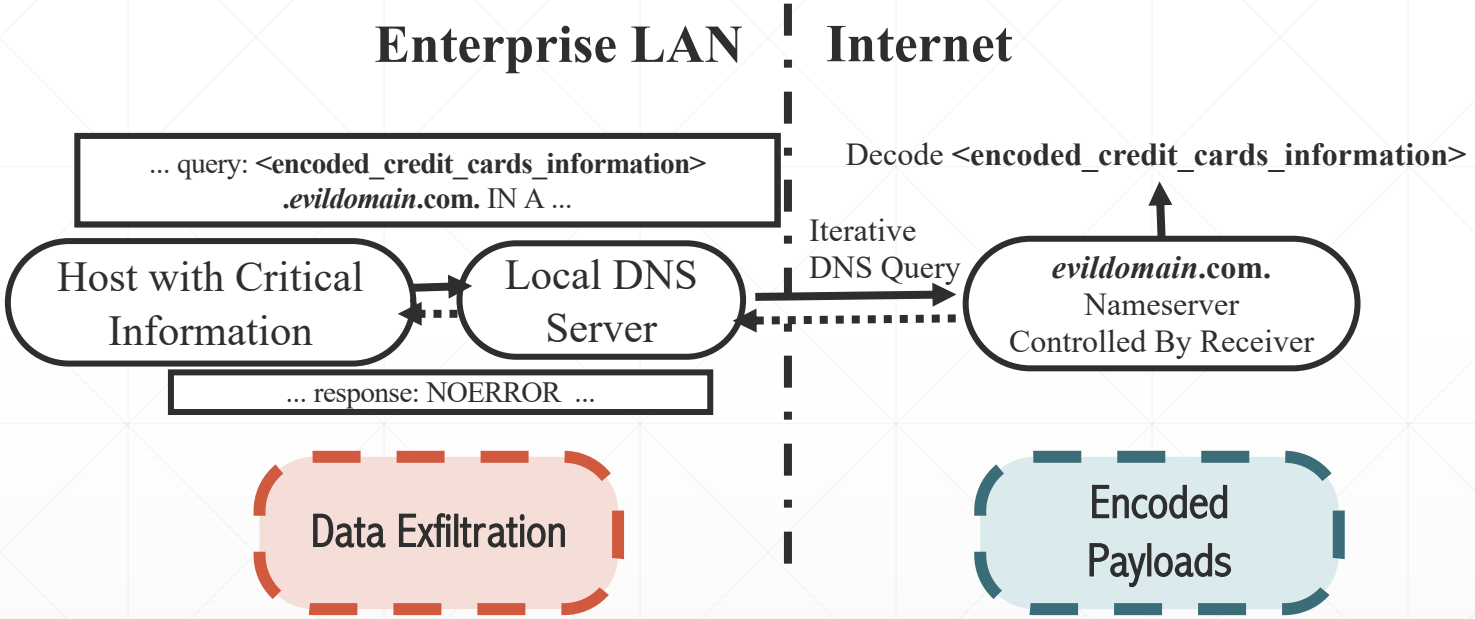


Attacks to transmit messages through DNS channel are difficult to be detected

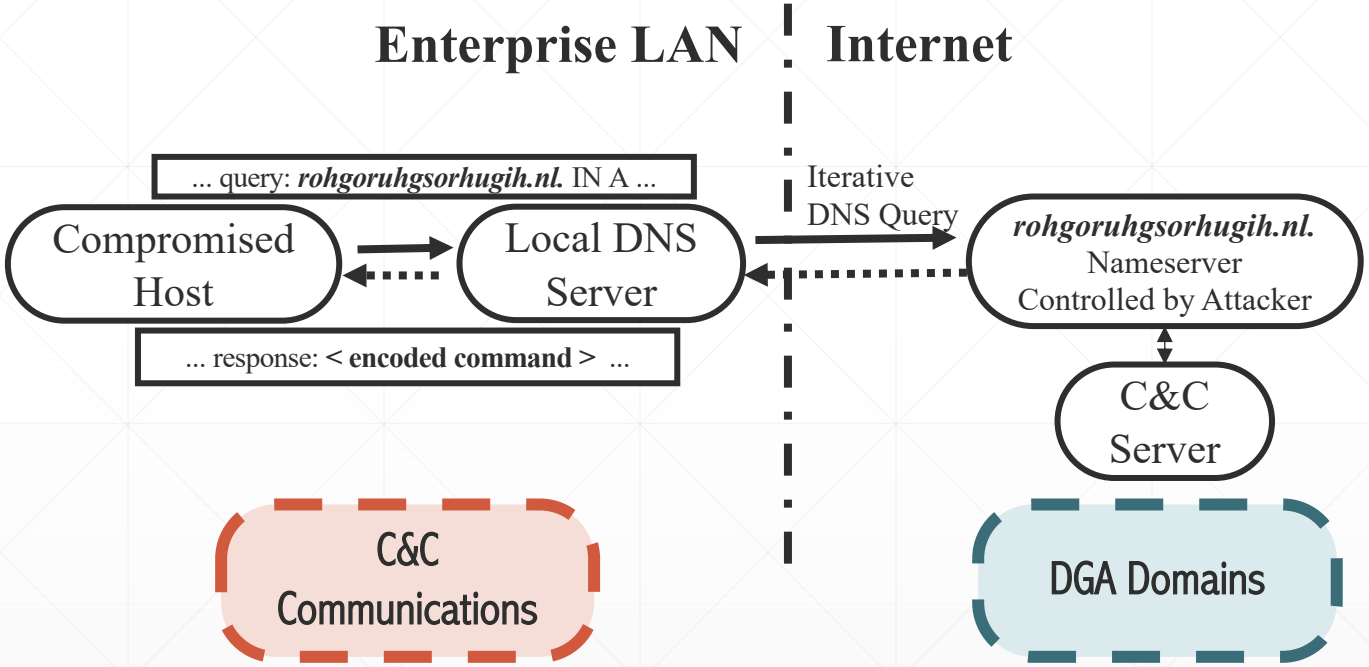
DNS Lookups


Enterprise LAN
Internet
19-Nov-2019 15:30:39.330 client 192.168.227.223\#1042:
query: www.google.com. IN A + (192.168.64.1)
Iterative DNS Query
Host (192.168.227.223)
Local DNS Server (192.168.64.1)
google.com. Nameserver
19-Nov-2019 15:30:41.441 client 192.168.227.223\#1042:
UDP: query: www.google.com IN A
response:
NOERROR + www.google.com. 1887 IN A 172.217.164.100;

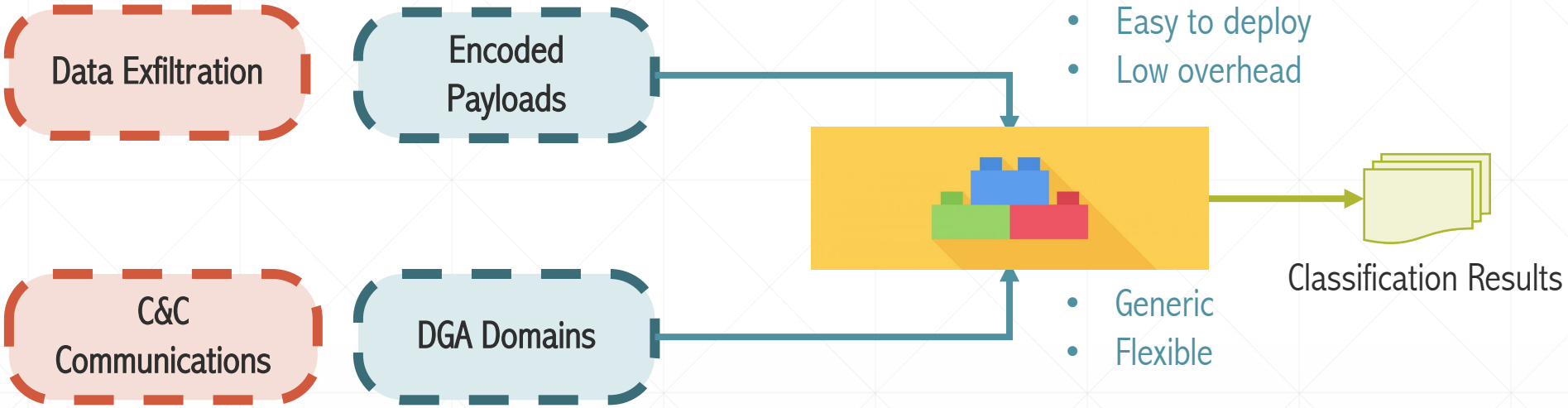
DNS Covert Communications



DNS Covert Communications



Idea



- Be able to detect multiple types of threats
- Easy to deploy
- Low overhead

- Generic
- Flexible

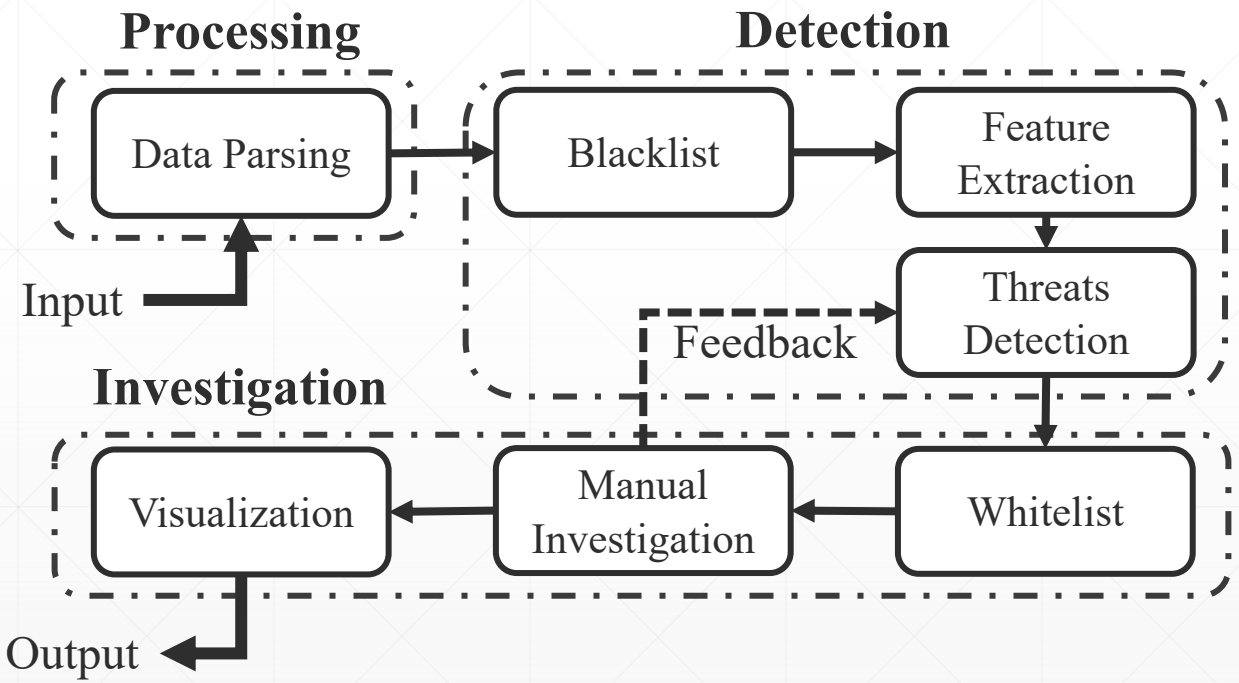
Modular machine learning detection models

- Including supervised & unsupervised models
- Most suitable model will be applied
- Each model can be adjusted individually
- Low overhead on model tuning or re-training

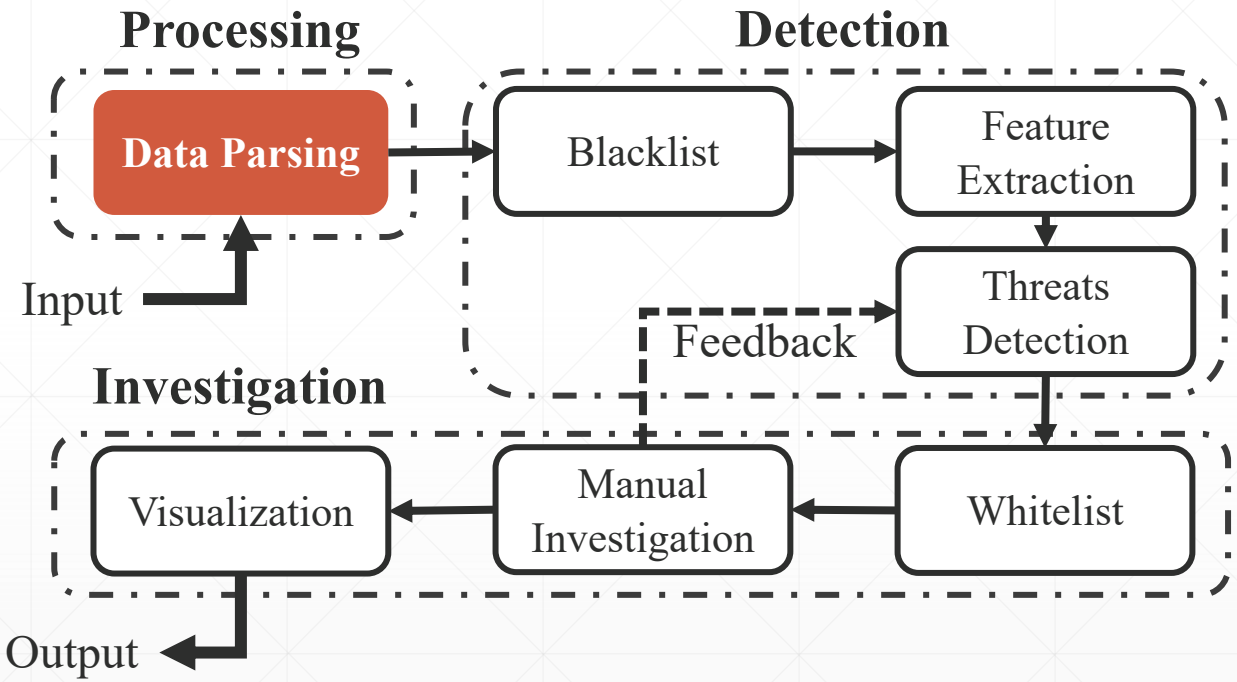
System Overview

D²C² (D2C2): Detecting DNS Covert Communication

A practical framework with modular detection models
to detect covert communication in DNS traffic in enterprise environments



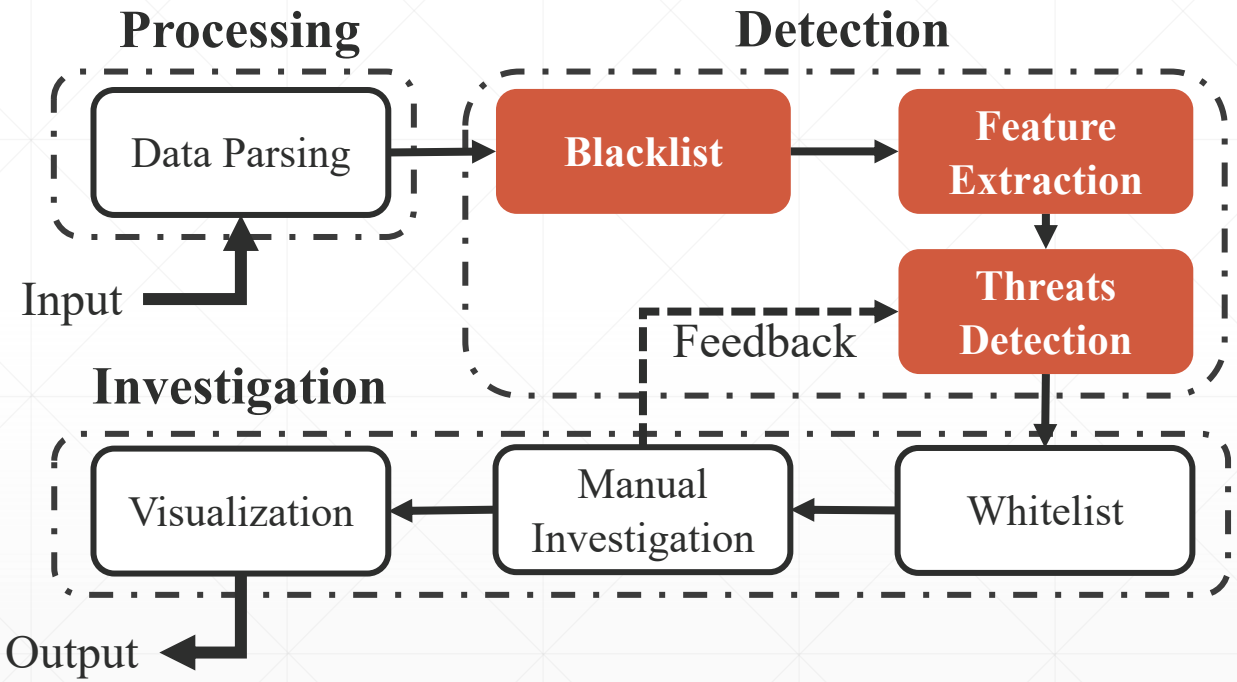
System Overview



Processing Stage

- Parsing raw data
- Extracting user demographics

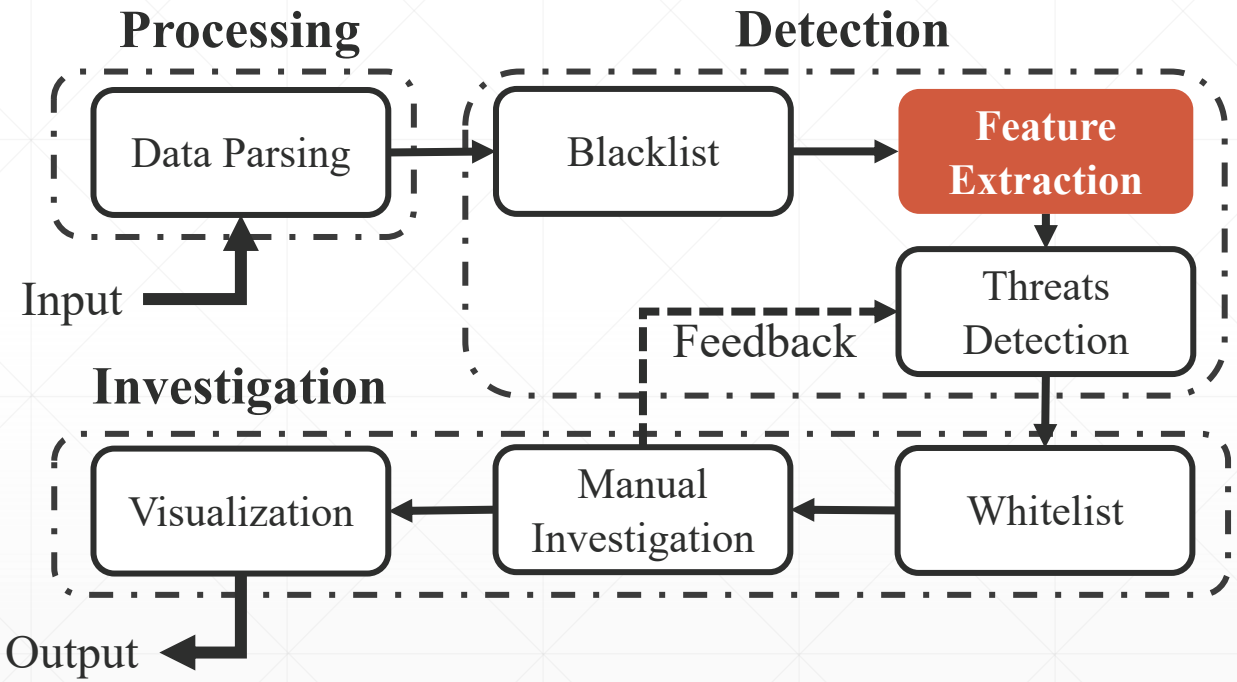
System Overview



Detection Stage

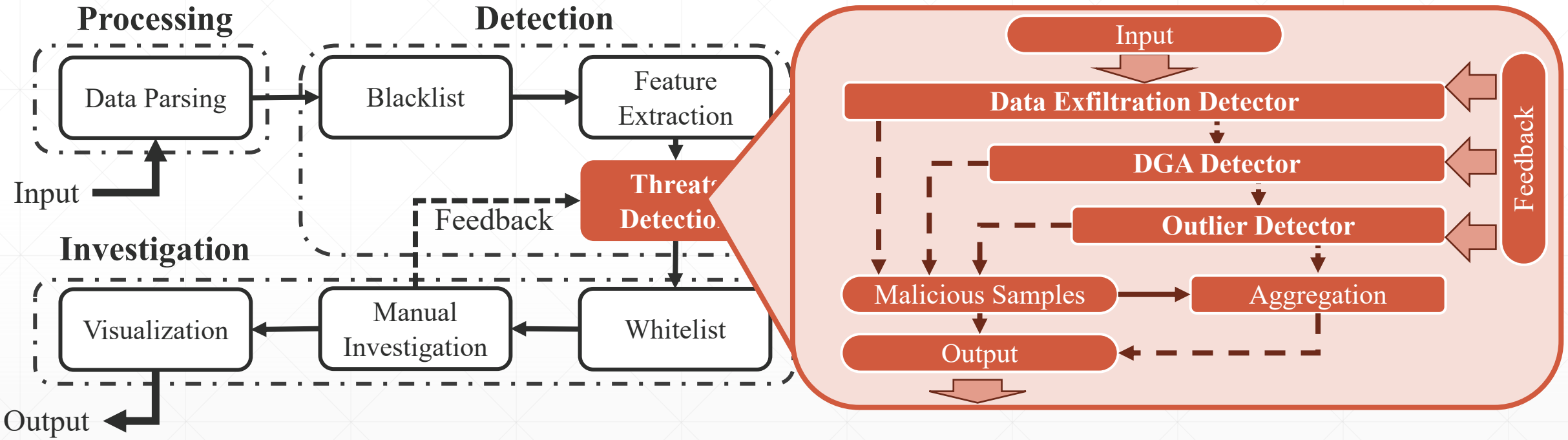
- Blacklist filters the logs
- Extracting features from logs
- Detecting anomalies by machine-learning models

System Overview



#	Feature	Type
1	Length of domain name.	integer
2	Length of subdomain.	integer
3	No. of labels.	integer
4	Longest label length.	integer
5	Contains one-character label.	boolean
6	Contains IPv4.	boolean
7	Has "WWW" prefix.	boolean
8	Alphabet size.	integer
9	No. of uppercase characters.	integer
10	The ratio of digits.	float
11	Ratio of hexadecimal parts.	float
12	Ratio of vowels.	float
13	Ratio of underscore.	float
14	Ratio of repeat characters.	float
15	Ratio of consecutive consonants.	float
16	Ratio of consecutive digits.	float
17	Shannon entropy [16].	float
18	Gibberish score [26].	float
19	Bigram of domain name.	vector

System Overview



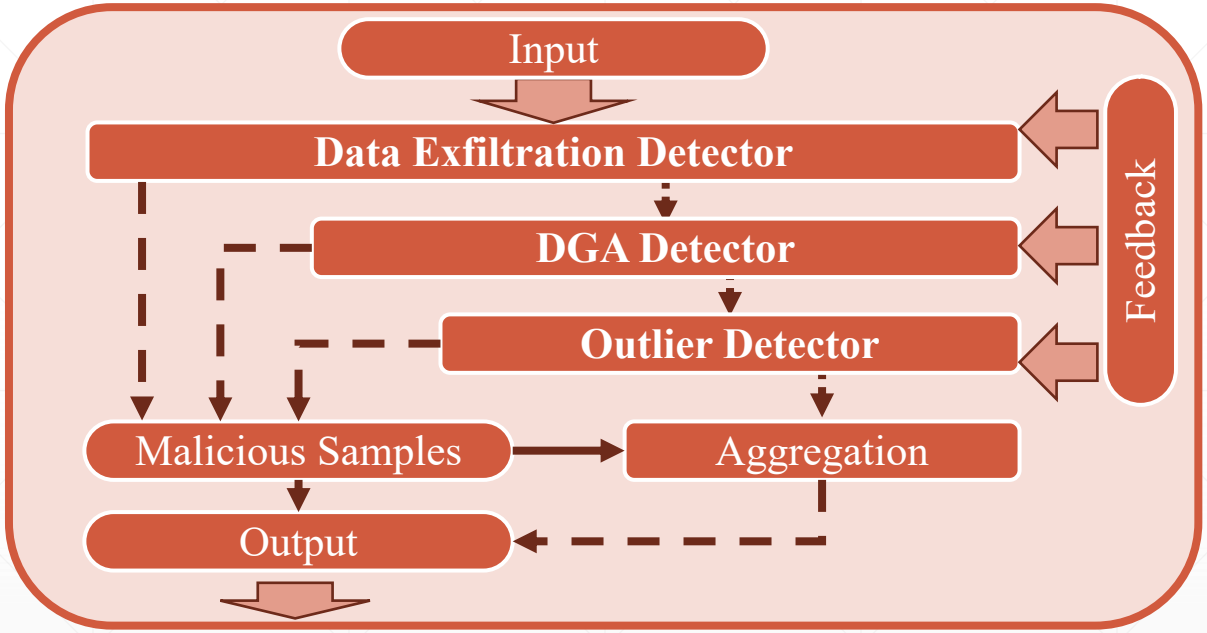
System Overview

Threat Scenarios

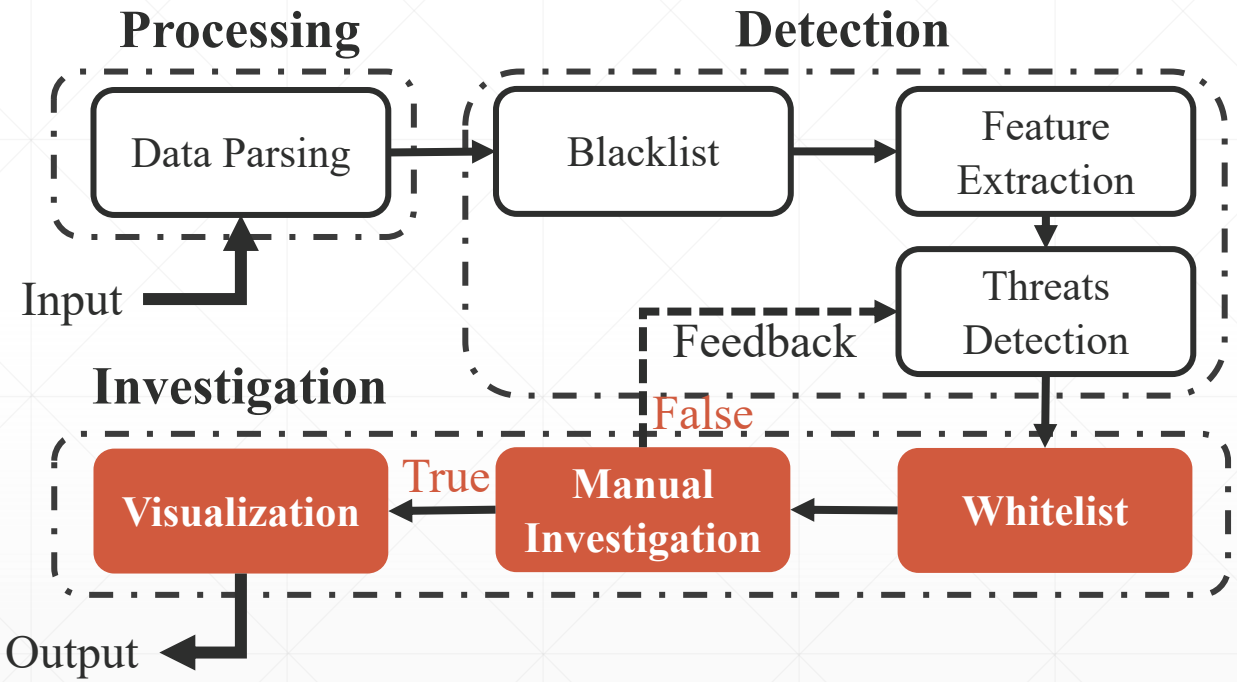
- Data Exfiltration
- C&C Communication (commonly seen as DGA)
- Other rare threats

Modular Detectors

- Supervised for known and common attacks
- Unsupervised for unknown and rare attacks
- Running in series

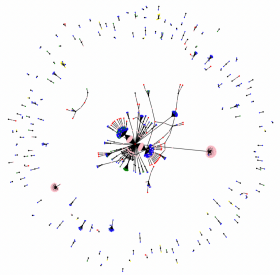


System Overview



Investigation Stage

- Whitelist filters the logs
- Manually check results & feedback
- Visualized results on detected results



Evaluation Setup

Deployment

- In a large enterprise environment
- Including servers in IDC & desktops/laptops in office networks
- More than 25,000 hosts
- 100 millions DNS logs per day on average

Data

- One-month deployment data trace of over 5 billion logs
- A labeled dataset of 764k logs contains known attack examples to test performance & overhead
- Blacklist
- Whitelist

Model

- Random forest
- Support Vector Machine
- Multi-layer perceptron

- Isolation forest
- X-means

Dataset Statistics

Deployment Data Trace

- One-month dataset of over 5 billion logs
- 3 billion queries
- 90% queries are A/AAAA queries

Historical Labeled Data Trace

- Labeled & verified by operators
- 764k logs

Blacklist/ Whitelist

- By operators & website ranking

Types	# of Queries (Responses)	Total	%
A	2,310,206,811 (2,175,715,764)	4,485,922,575	75.98%
AAAA	443,000,848 (441,857,308)	884,858,156	14.98%
PTR	245,185,527 (244,886,490)	490,072,017	8.30%
SOA	5,751,338 (5,722,695)	11,474,033	0.19%
SRV	5,651,489 (5,611,368)	11,262,857	0.19%
NS	4,790,185 (4,788,276)	9,578,461	0.16%
TXT	3,392,785 (3,389,870)	6,782,655	0.11%
CNAME	630,267 (630,246)	1,260,513	0.02%
MX	327,305 (320,792)	648,097	0.01%
Other	958,983 (963,691)	1,922,674	0.03%
Total	3,019,895,538 (2,883,886,500)	5,903,782,038	–

Evaluation on labeled dataset

Dataset

- In the same enterprise
- 764k logs

Label

- Labeled & verified by operators

Model

- Random forest
- Multi-layer perceptron
- Support Vector Machine
- Isolation forest
- X-means

Results

- RF & MLP have better performance
- iForest has better performance

Evaluation Results

Detector		Precision	Recall	Accuracy	F1
D-Exfil	**RF**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	MLP	**0.9999**	**0.9995**	**0.9995**	**0.9993**
	SVM	0.9997	0.9998	0.9998	0.9997
D-DGA	**RF**	**0.9580**	**0.9787**	**0.9945**	**0.9682**
	MLP	**0.9290**	**0.9660**	**0.9910**	**0.9471**
	SVM	0.8049	0.9558	0.9765	0.8793
D-Outlier	**iForest**	**0.8495**	**0.9190**	**0.9988**	**0.8829**
	X-Means	0.6708	0.5371	0.9981	0.5965

Evaluation on labeled dataset

Dataset

- In the same enterprise
- 764k logs

Label

- Labeled & verified by operators

Results

- Average input: 1,200 logs/s
- All models have fast processing speed
- SVM's speed drops when size increases

Model

- Random forest
- Multi-layer perceptron
- Support Vector Machine
- Isolation forest
- X-means

Evaluation Results

Detector		Precision	Recall	Accuracy	F1
D-Exfil	**RF**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	MLP	**0.9999**	**0.9995**	**0.9995**	**0.9993**
	SVM	0.9997	0.9998	0.9998	0.9997
D-DGA	**RF**	**0.9580**	**0.9787**	**0.9945**	**0.9682**
	MLP	**0.9290**	**0.9660**	**0.9910**	**0.9471**
	SVM	0.8049	0.9558	0.9765	0.8793
D-Outlier	**iForest**	**0.8495**	**0.9190**	**0.9988**	**0.8829**
	X-Means	0.6708	0.5371	0.9981	0.5965

Overhead

Model		Processing Speed (logs/s)
Supervised	**RF**	49344.9
	MLP	9210.2
	SVM	24150.2*
Unsupervised	**iForest**	9149.0
	X-Means	4090.6

$O(n^2)$

Deployment Results

Dataset

- One-month dataset of over 5 billion logs
- Models chosen based on performance & overhead

Model

- Random forest
- Multi-layer perceptron
- Isolation forest

Evaluation Results

Detector		Precision	#TP/day	#FP/day
D-Exfil	**RF**	**0.9755**	**155.6**	**3.9**
	MLP	0.9934	1070.0	7.1
D-DGA	**RF**	**0.9986**	**3958.9**	**5.6**
	MLP	0.9764	3871.0	93.5
D-Outlier	**iForest**	**0.9214**	**29.3**	**2.5**
Total (RF + iForest)		**0.9971**	**4143.8**	**12.0**

Deployment Results

Dataset

- One-month dataset of over 5 billion logs
- Models chosen based on performance & overhead

Model

- Random forest
- Multi-layer perceptron
- Isolation forest

Evaluation Results

Detector		Precision	#TP/day	#FP/day
D-Exfil	**RF**	**0.9755**	**155.6**	**3.9**
	MLP	0.9934	1070.0	7.1
D-DGA	**RF**	**0.9986**	**3958.9**	**5.6**
	MLP	0.9764	3871.0	93.5
D-Outlier	**iForest**	**0.9214**	**29.3**	**2.5**
Total (RF + iForest)		**0.9971**	**4143.8**	**12.0**

Results

- High precision over 0.92
- Low False Alerts

Deployment Results

Dataset

- One-month dataset of over 5 billion logs
- Models chosen based on performance & overhead

Model

- Random forest
- Multi-layer perceptron
- Isolation forest

Evaluation Results

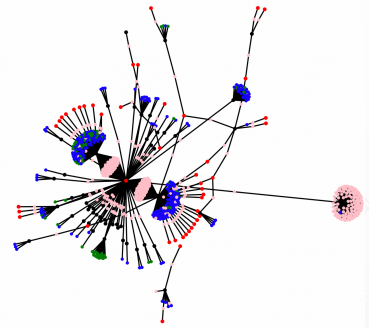
Detector		Precision	#TP/day	#FP/day
D-Exfil	**RF**	**0.9755**	**155.6**	**3.9**
	MLP	0.9934	1070.0	7.1
D-DGA	**RF**	**0.9986**	**3958.9**	**5.6**
	MLP	0.9764	3871.0	93.5
D-Outlier	**iForest**	**0.9214**	**29.3**	**2.5**
Total (RF + iForest)		**0.9971**	**4143.8**	**12.0**

Results

- High precision over 0.92
- Low False Alerts

Visualization

- Detected 4,000 anomalous logs/day
- Found 7 compromised hosts based on merging and visualization



Summary

- A practical, flexible and end-to-end ML-based framework
  - Detecting threats in enterprise environments, generic, easy to deploy
  - Modular detection models, flexible

- Deployment in a real-world enterprise
  - One-month dataset of over 5 billion logs
  - 4,000 **anomalous** logs detected per day and high precision
  - Low overhead
  - Visualized results, 7 compromised hosts found

Thanks!
And Questions.

Ruming Tang: trm14@tsinghua.org.cn