Countering Malicious Processes with Process-DNS Association

Suphannee Sivakorn*, **Kangkook Jee**§, Yixin Sun△, Lauri Korts-Pärn§, Zhichun Li§, Cristian Lumezanu§, Zhenyu Wu§, Lu-An Tang§, Ding Li§

*Columbia University, △Princeton University,

§NEC Laboratory America Inc.

NDSS 2019

Cyber attacks and DNS services

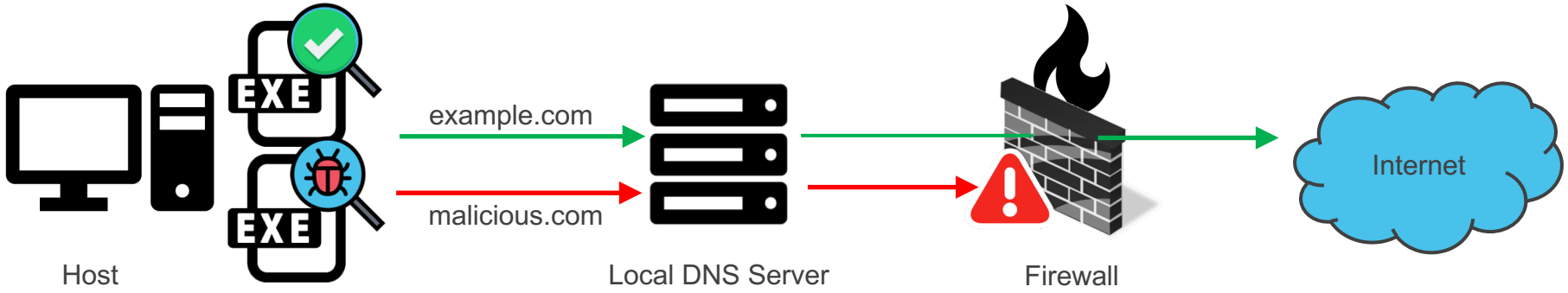
- Cyber attacks rely on the Internet
  - Command and control (C&C) e.g., delivering payload
  - Drop sites – uploading stolen information

- DNS service is critical attack vector
  - Domain names are more reliable and difficult to track
    e.g., short-lived domain name, fast-flux,
    Domain Generation Algorithm (DGA)

- Monitoring DNS traffic
  - Malicious Domain → Malicious Activity Detections

DNS-based detection systems

Malicious Domain Detections → Malicious Activity Detections



1. **Static**: Domain or IP address blacklisting

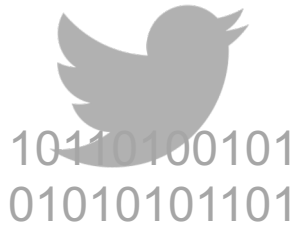
2. **Dynamic**: Anomaly detections
   E.g., diversity of resolved IPs, geographical information, name string structure (DGA), and DNS TTLs.
   [Antonakakis et al., USENIX 2010, Bilge et al., NDSS 2011, Antonakakis et al., USENIX 2011]

Stealthy DNS operations

Attempt to bypass network-level defenses

Webservice as C&C channels

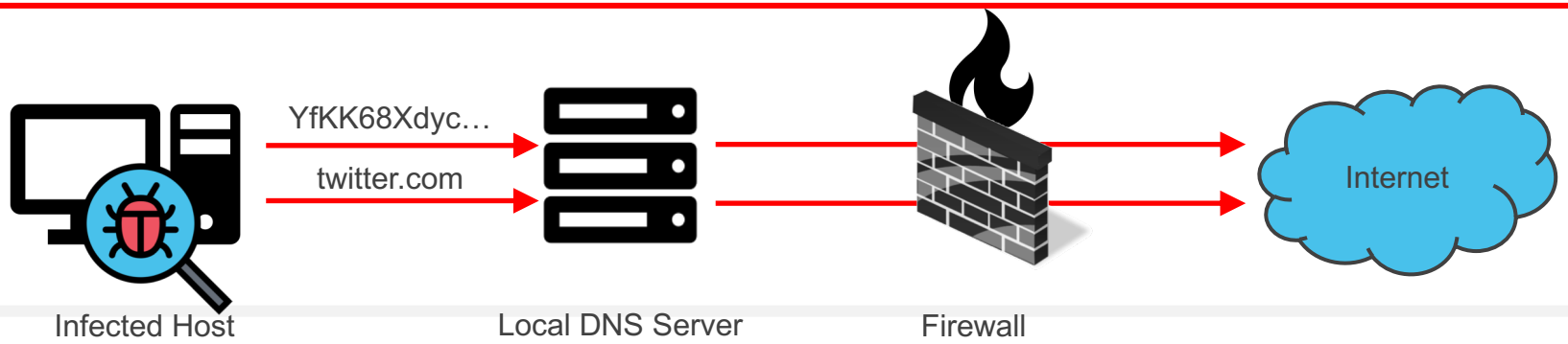
- Communications: twitter message, meme!
- Payload: githubcontent.com, wordpress.com



Drop sites

- Cloud services e.g., Dropbox, Photobucket

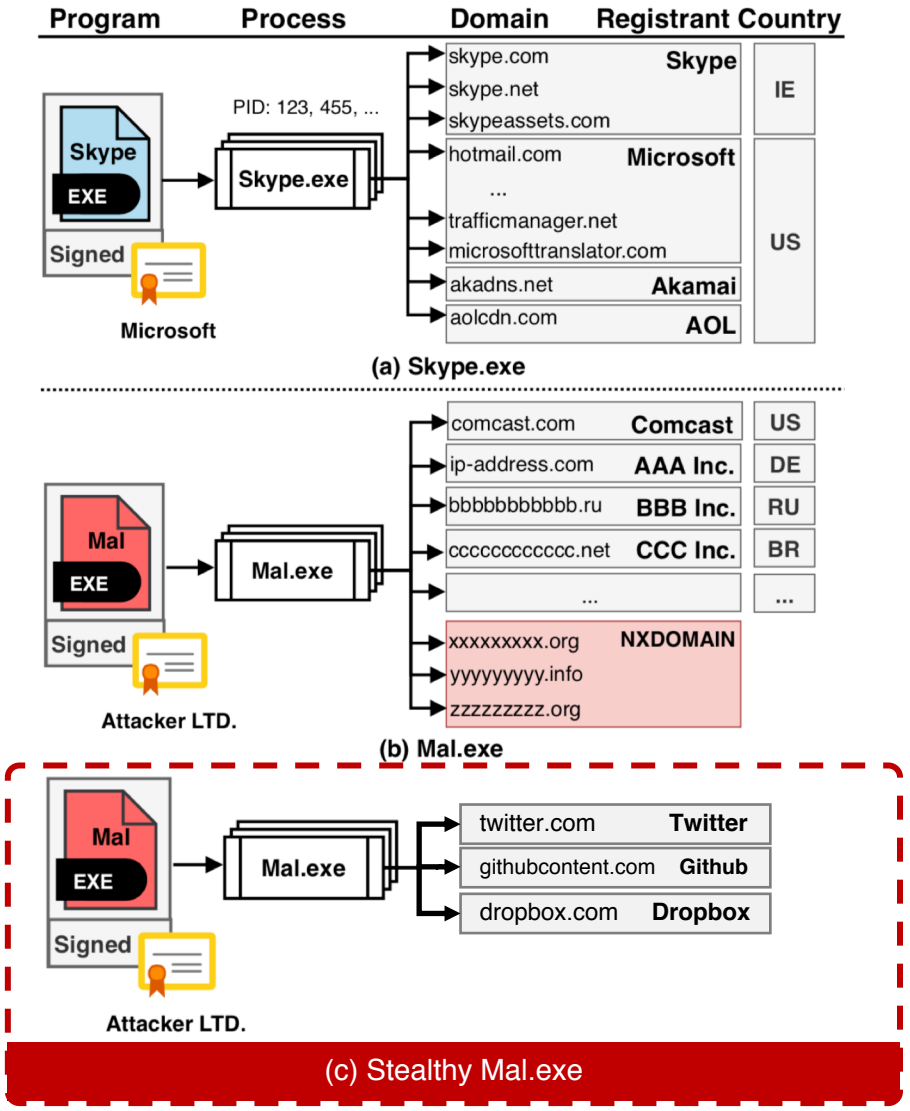
DNS-over-HTTPS: Support for Encrypted DNS query



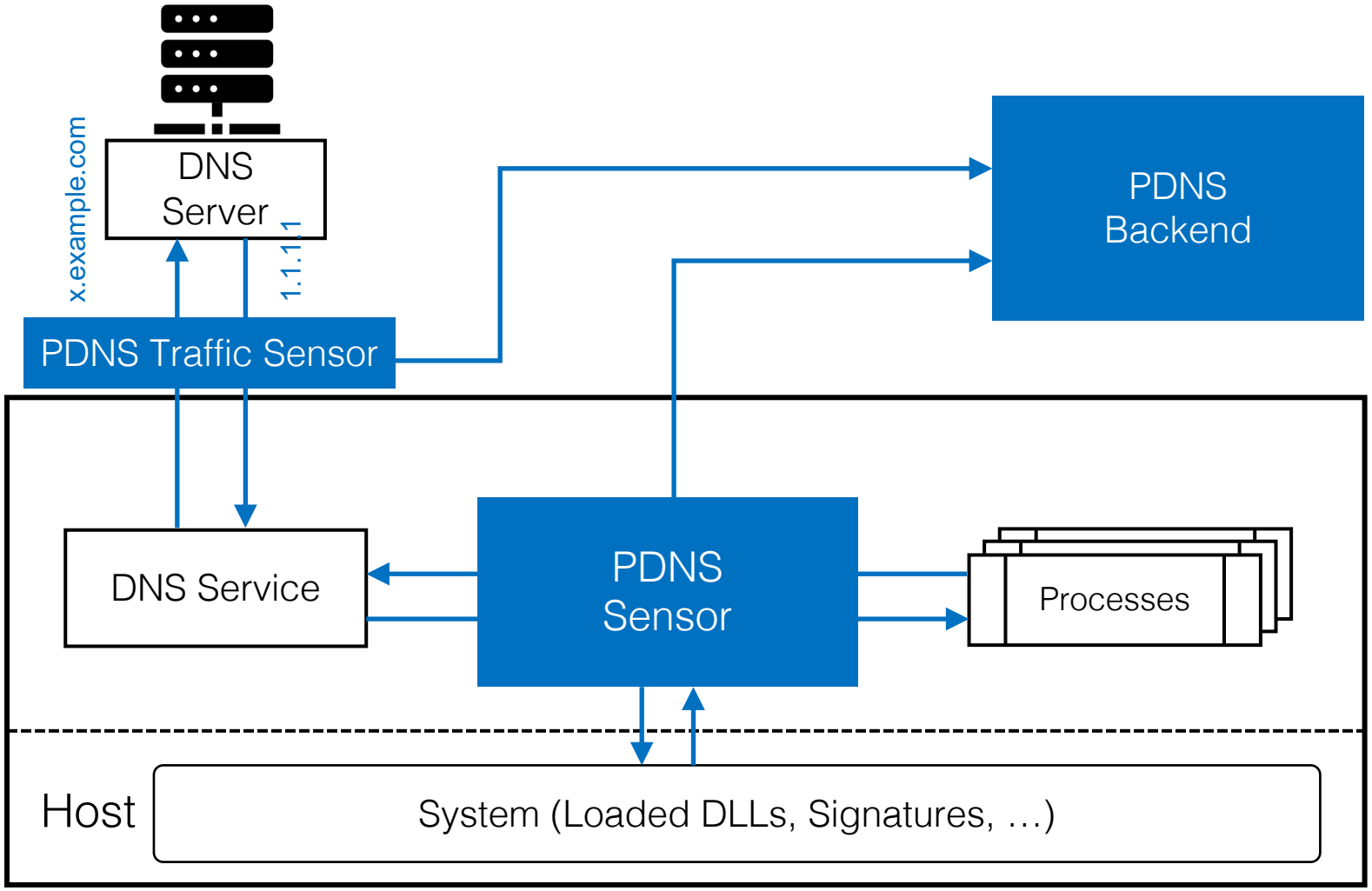
Why host-based detection system?

Not the network-based

- Process and domain query relationship
  - Monitoring at finer granularity
  - Process-level detection
- Expands host-level features
  - Code Signing, software publisher, loaded DLLs, command line arguments
- Individualized network-level features
  - Domain name, domain registration duration, domain TTL seen from process-level



PDNS: Process and DNS association



- Windows Model
- PDNS (Host) Sensor
  - DNS traffics
  - Process information
- PDNS Traffic Sensor
  - Network DNS traffics
  - Validated with DNS traffics from host agent
- PDNS Backend
  - Centralizing data
  - Feather extraction e.g., WHOIS, IP Location
  - Train, Classify and Report

Datasets

- Benign dataset
  - Deployed PDNS sensors on 126 Windows workstations on our enterprise (6 months)
  - ~130M DNS requests
  - 455K processes (643 unique programs and 1543 unique hashes)
  - Cross-checked all binary hashes with VirusTotal service

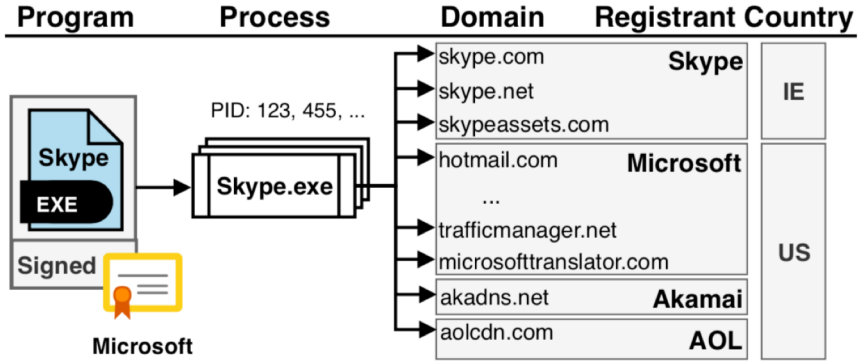
- Malicious dataset

Dataset	Source	# Samples	Reported year	# DNS queries	# Processes
Train	VirusSign	12k	2015	1.1M	5k
	VirusShare, VXVault	4k	2017	600k	2k
Train-total		16k		1.7M	7k
Test	VirusShare	2k	2018	10k	1.5k

  - Behavioral-based: executed malware in our sandbox environment
  - Selected only malware with activities

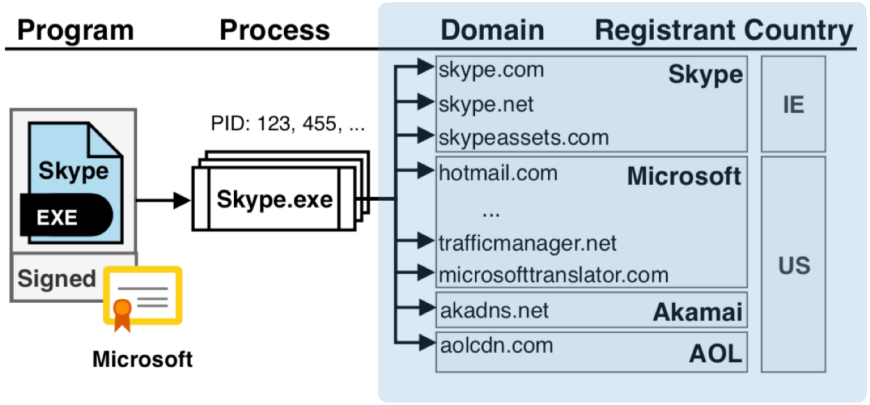
Feature engineering

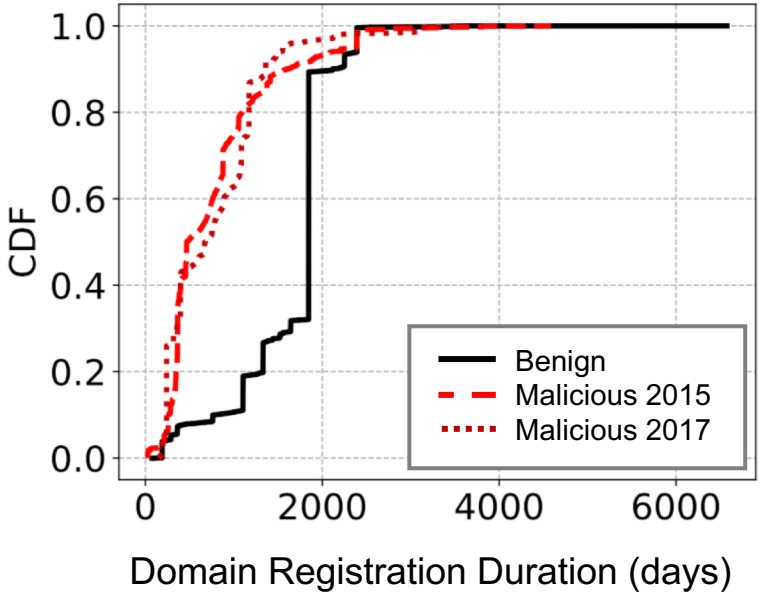
	Data Source	Feature Category	# Features
External Sources	Domain	Domain Duration	12
	WHOIS	Domain Registrant	5
		Location	2
	WHOIS IP	AS Number	4
	IP	Location	2
	Geolocation		
Internal Sources	DNS activity	Domain and Hostname	22
		Resolve Failure Rate	1
	System	Code Signing	1
		Loaded DLLs	3
Integrated		Location	2
		Publisher and Registrant	4
Total			61



Feature engineering

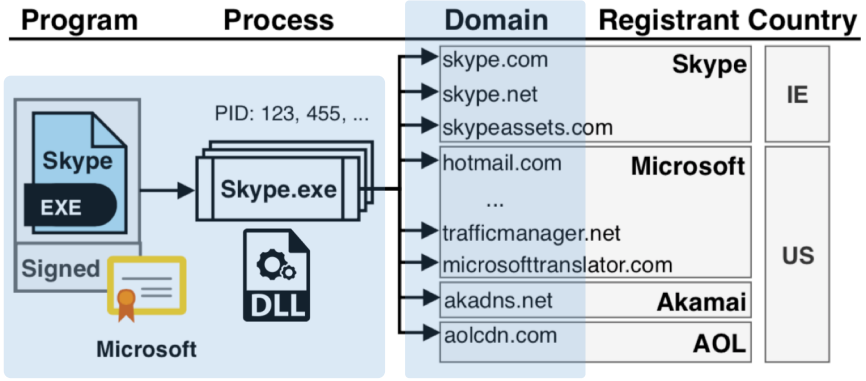
	Data Source	Feature Category	# Features
External Sources	Domain	Domain Duration	12
	WHOIS	Domain Registrant	5
		Location	2
	WHOIS IP	AS Number	4
	IP	Location	2
	Geolocation		
Internal Sources	DNS activity	Domain and Hostname	22
		Resolve Failure Rate	1
	System	Code Signing	1
		Loaded DLLs	3
Integrated		Location	2
		Publisher and Registrant	4
Total			61

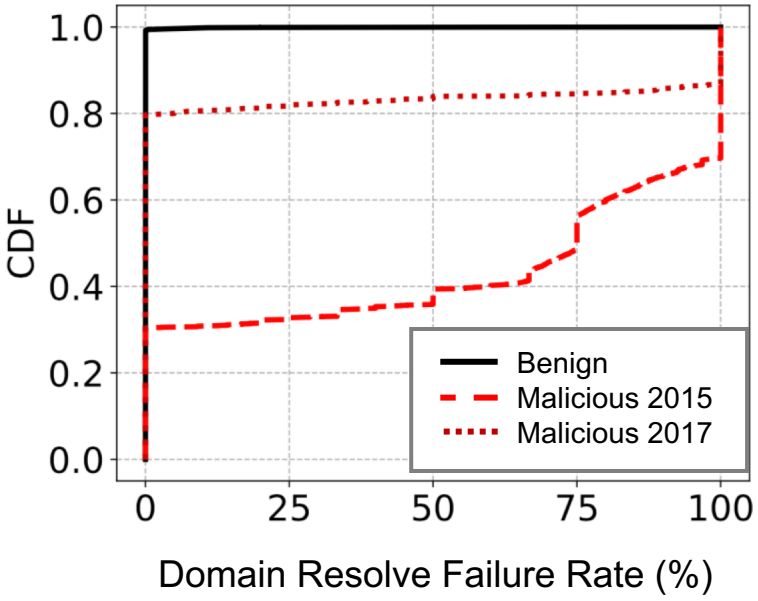




Feature engineering

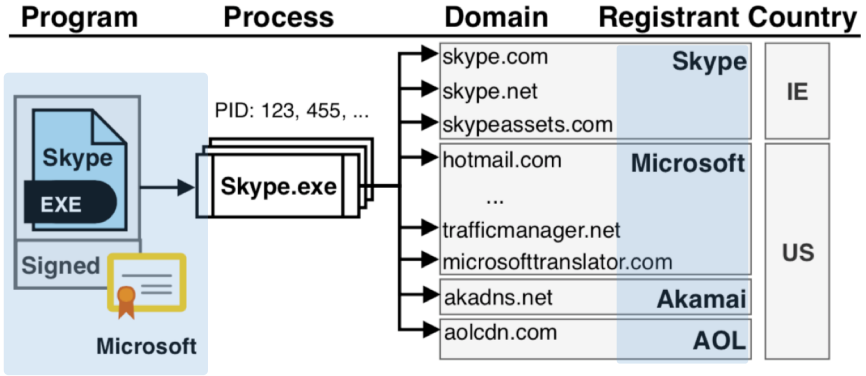
	Data Source	Feature Category	# Features
External Sources	Domain	Domain Duration	12
	WHOIS	Domain Registrant	5
		Location	2
	WHOIS IP	AS Number	4
	IP	Location	2
	Geolocation		
Internal Sources	DNS activity	Domain and Hostname	22
		Resolve Failure Rate	1
	System	Code Signing	1
		Loaded DLLs	3
Integrated		Location	2
		Publisher and Registrant	4
Total			61

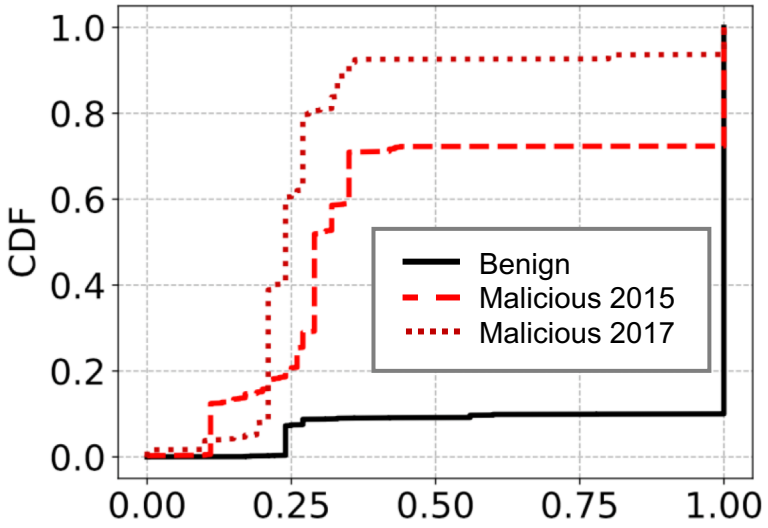




Feature engineering

	Data Source	Feature Category	# Features
External Sources	Domain WHOIS	Domain Duration Domain Registrant Location	12 5 2
	WHOIS IP IP Geolocation	AS Number Location	4 2
Internal Sources	DNS activity System	Domain and Hostname Resolve Failure Rate Code Signing Loaded DLLs	22 1 1 3
Integrated		Location **Publisher and Registrant**	2 4
Total			61





Software Publisher and Domain Registrant Similarity Score (Levenshtein distance) [0.0 – 1.0]

Training the model

Data normalization

Unbalance data:
SMOTE technique
(over-sampling +
under-sampling)

Classifiers

- Logistic Regression (LR)
- K-Nearest Neighbor (KNN)
- Random Forest (RF)
- Linear Support Vector (LinearSVC)
- Deep Neural Network (DNN) with NN Multi-layer Perceptron classifier

Parameter tuning

Hyperparameter tuning
GridSearchCV

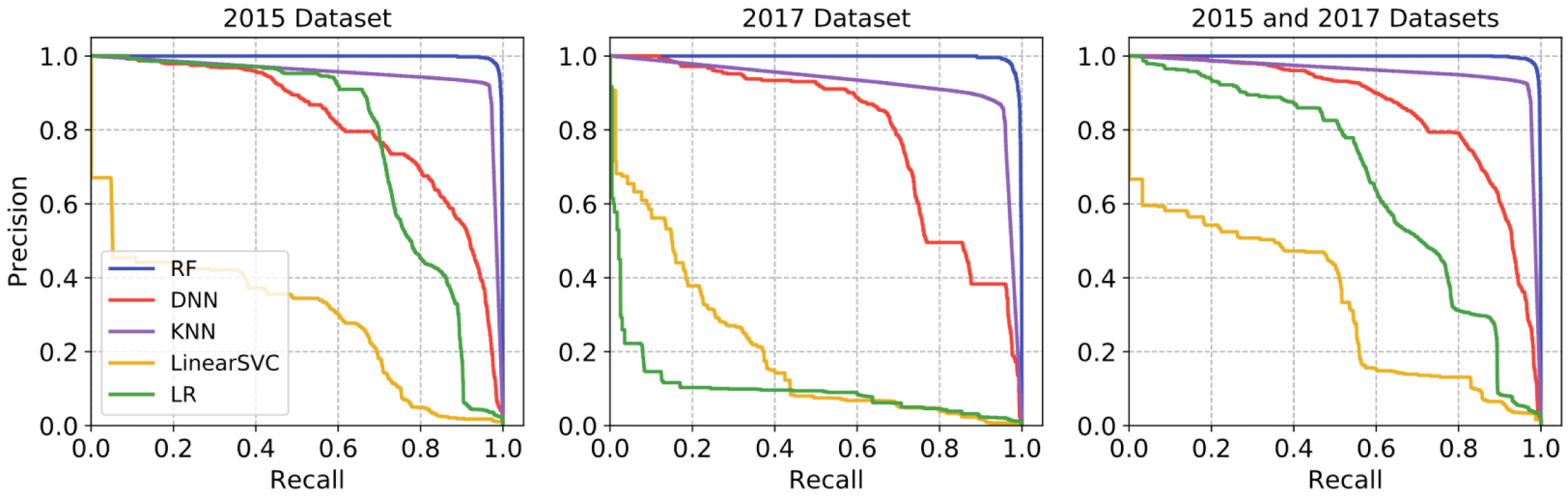
Evaluating metrics

- 10-fold cross validation
- TP_rate, FP_rate
- ROC, precision-recall

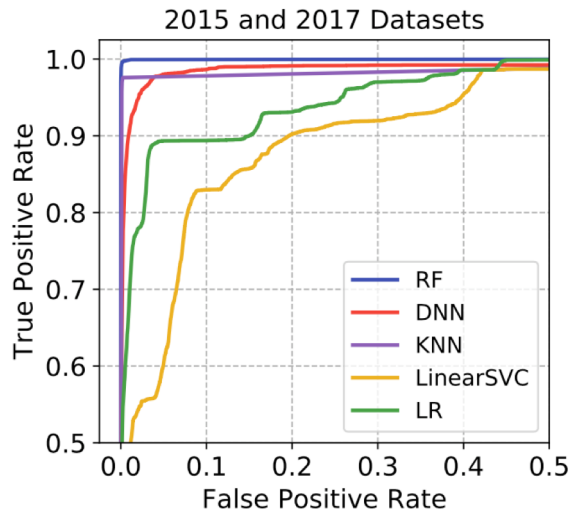
- Feature Importance:
  MDI
  (mean decrease in impurity)

Detection accuracy

Precision-recall



ROC

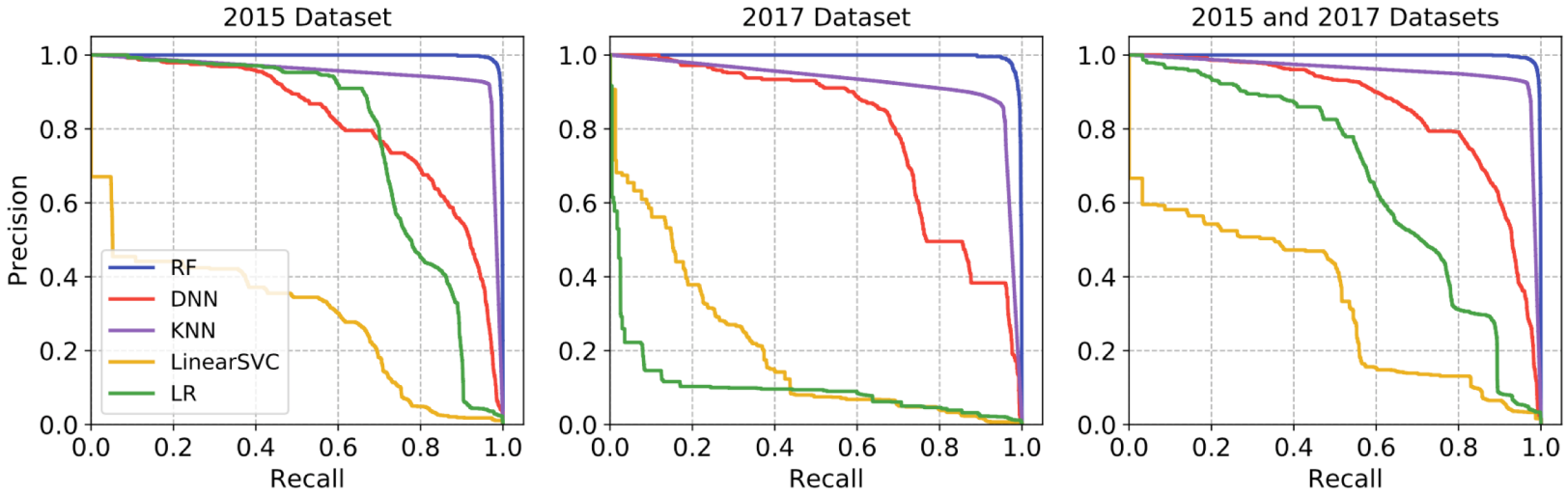


RF performs the best:

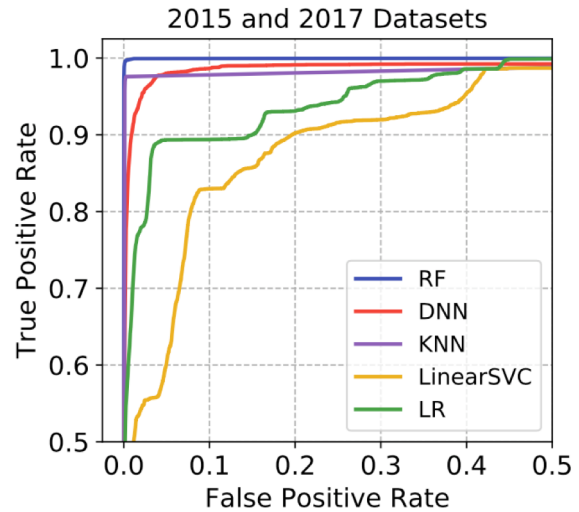
- On average, TP_rate = 98.55%, FP_rate = 0.03%
- Accuracy ≥ 0.99, precision ≥ 0.98, recall ≥ 0.97, F1 score ≥ 0.98

Detection accuracy – unseen dataset

Precision-recall



ROC



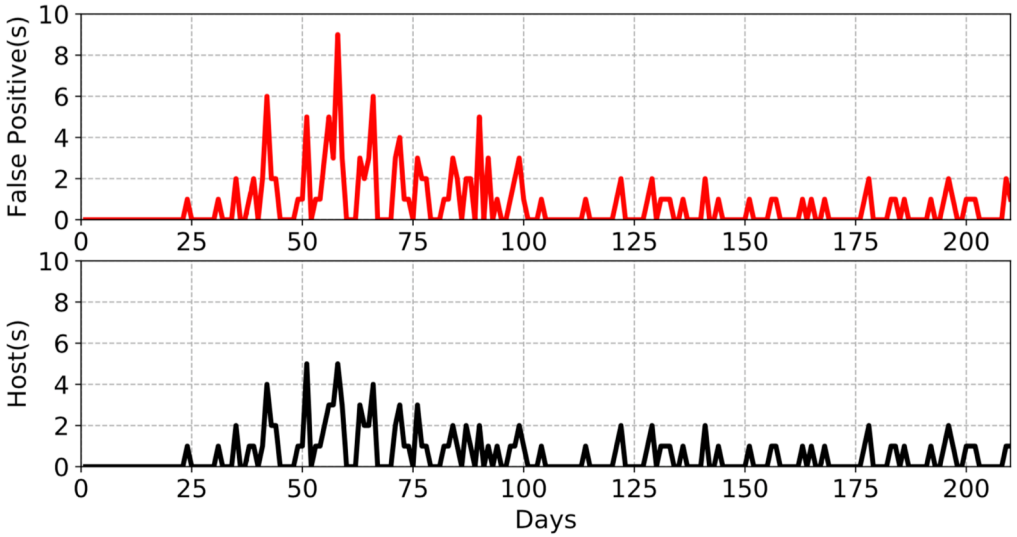
Dataset	Source	# Samples	Reported year	# DNS queries	# Processes
Train	VirusSign	12k	2015	1.1M	5k
	VirusShare, VXVault	4k	2017	600k	2k
Train-total		16k		1.7M	7k
Test	VirusShare	2k	2018	10k	1.5k

Unseen malware:

TP_rate = 98.03%

False positives

- Over 6 months: 146 unique processes
  - 45 distinct hosts
    (out of 126 deployed hosts)
  - On average 0.7 false positives per day
- False positives:
  1. Command-line processes
     e.g., pythonw.exe, java.exe, javaw.exe
  2. Browsers



- Security incidents e.g.,

videodl.exe

↑ domain requests
↑ geo-locations
✕ signed

googlevideo.com, youtube.com, openload.com,
spadesplus.com, pages.ebay.com, flyreport.com, chinawomendating.asia

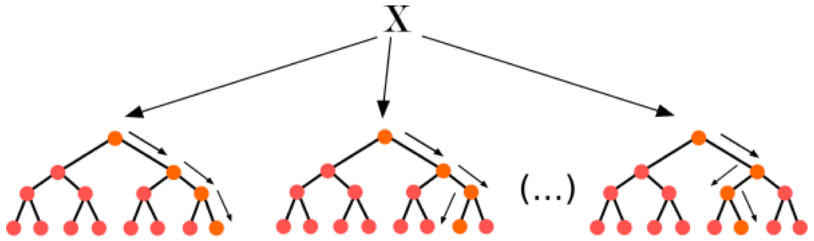
Feature importance

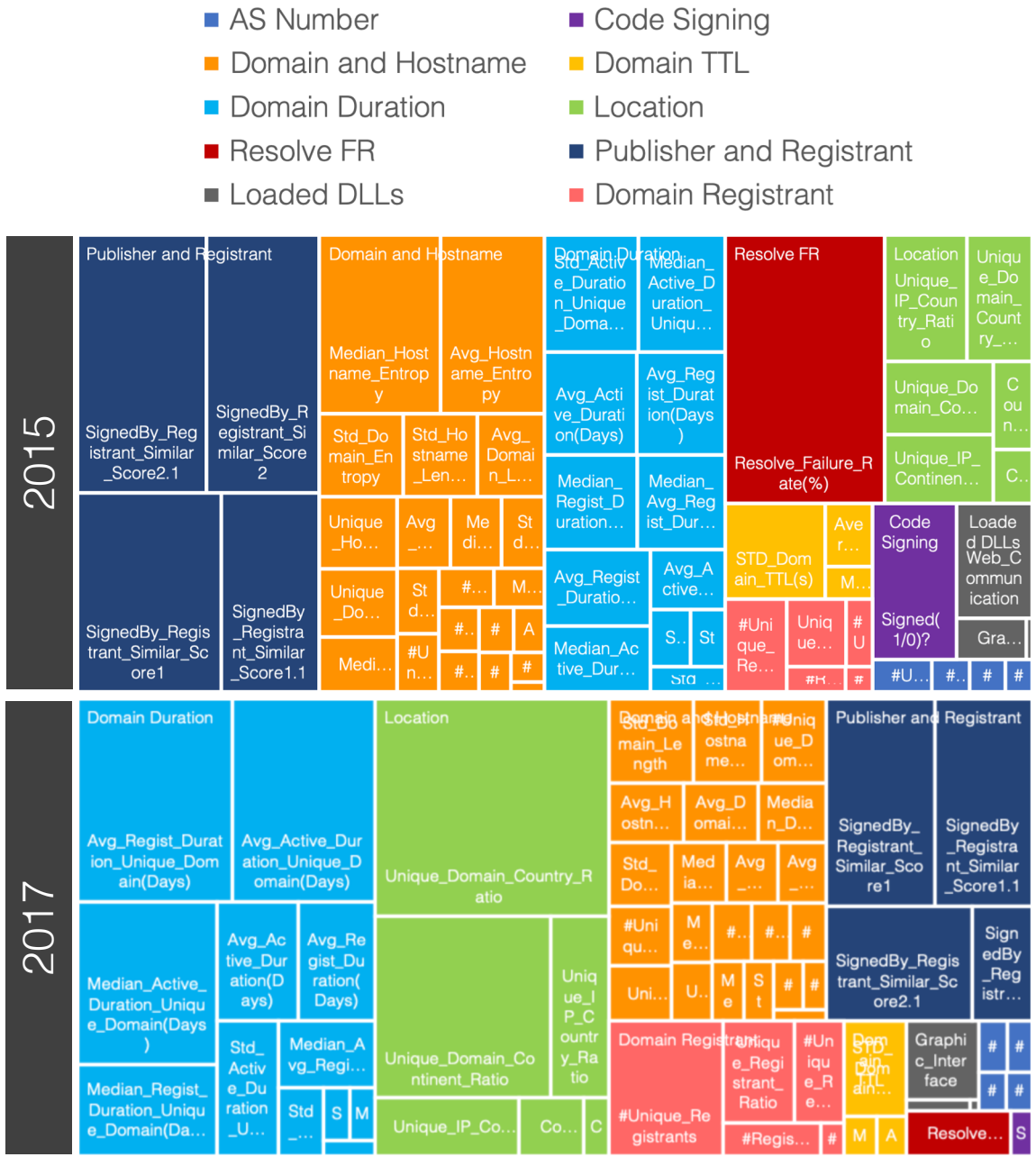
	Data Source	Feature Category	# Features
External Sources	WHOIS	Domain Duration Domain Registrant Location	12 5 2
	WHOIS IP IP Geolocation	AS Number Location	4 2
Internal Sources	DNS activity System	Domain and Hostname Resolve Failure Rate Code Signing Loaded DLLs	22 1 1 3
Integrated		Location Publisher and Registrant	2 4
Total			61

Feature importance for tree-based model

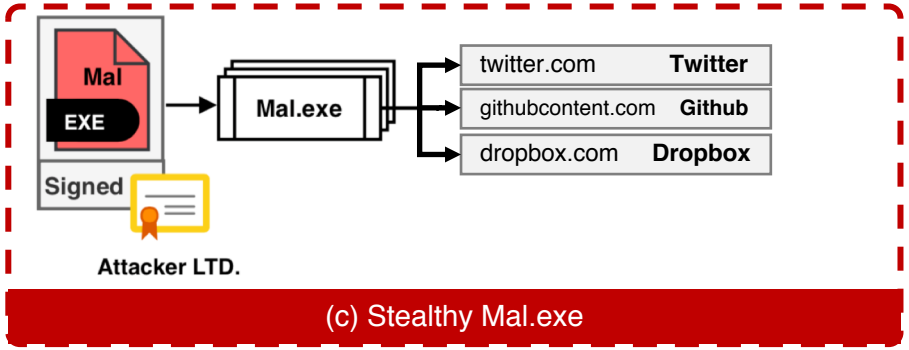
Mean Decrease in Impurity (MDI):

summing total impurity reduction at all tree nodes where the variable appears (Breiman et al., 1984)



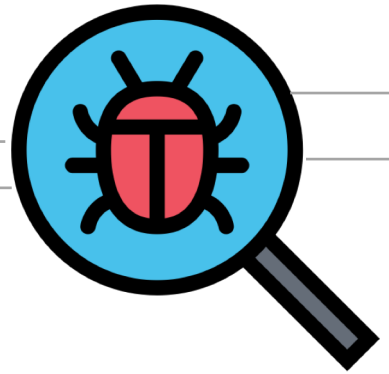


Stealthy attack detection



1. Discovered by FireEye researchers 2015
2. All domains connected is legitimate, mimicking day-to-day users

HAMMERTOSS



3. Visit twitter, retrieve hidden URLs to page in githubcontent obtaining payload

4. Delegates Internet Explorer (IE) processes to fetch contents from github, twitter

5. Extract information and upload to designated cloud services

PDNS vs Hammertoss

- Executes in our sandbox environment, collecting domain traffics
- **Zero-day fashion:** our training does not contain Hammertoss

Detects the host is infected!

Alerted ~53% of delegated IE processes

Short-lived

Publisher and Registrant relation

PDNS Take away

- PDNS: End-point Process and DNS association monitoring system
  - Captures domain queried at the process-level
  - Enhances visibility, context and relationship of queried domains and processes
  - Detects stealthy malware/processes

- Malware Evasion
  - Forged DNS activities, loaded DLLs, and avoid DNS queries
  - Enhances host-based features
  - Always improves and updates the model!

Questions?

Take away

PDNS: End-point Process and DNS association monitoring system

- Captures domain queried at the process-level
- Enhances visibility, context and relationship of queried domains and processes
- Detects stealthy malware/processes

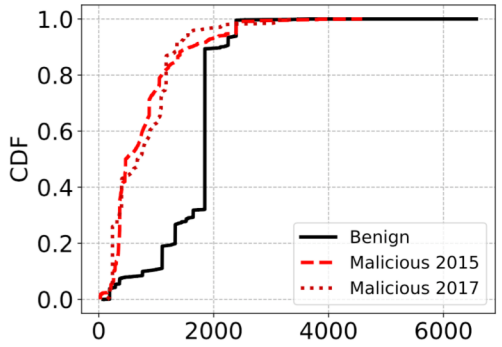
Trained and evaluated PDNS with real-world and large-scale data

- Deployed on over 126 hosts our enterprise environment
- Executed large collection of malware from multiple sources

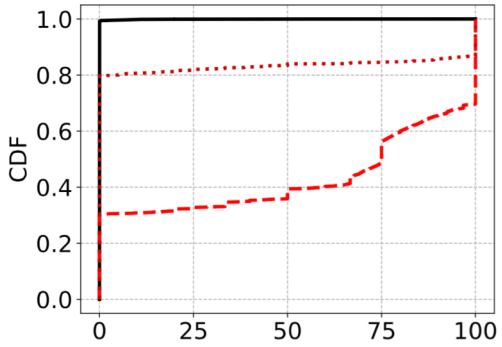
PDNS detection achieves a high accuracy

- Feature engineering: differentiate the benign and malware behaviors
- Reports the TP_rate 98% and FP_rate 0.03%

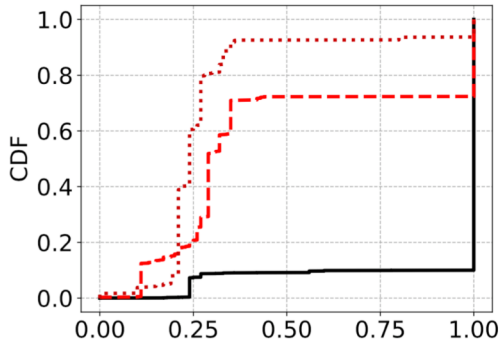
Feature Engineering



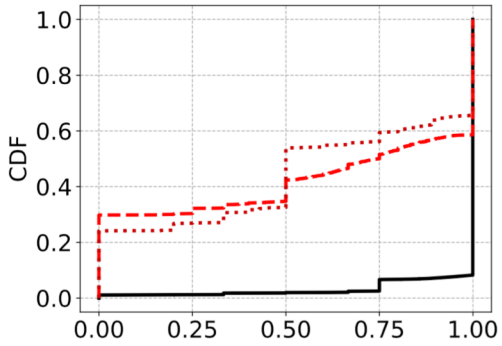
(a) Domain renewed duration (days)



(b) Domain Resolve FR (%)

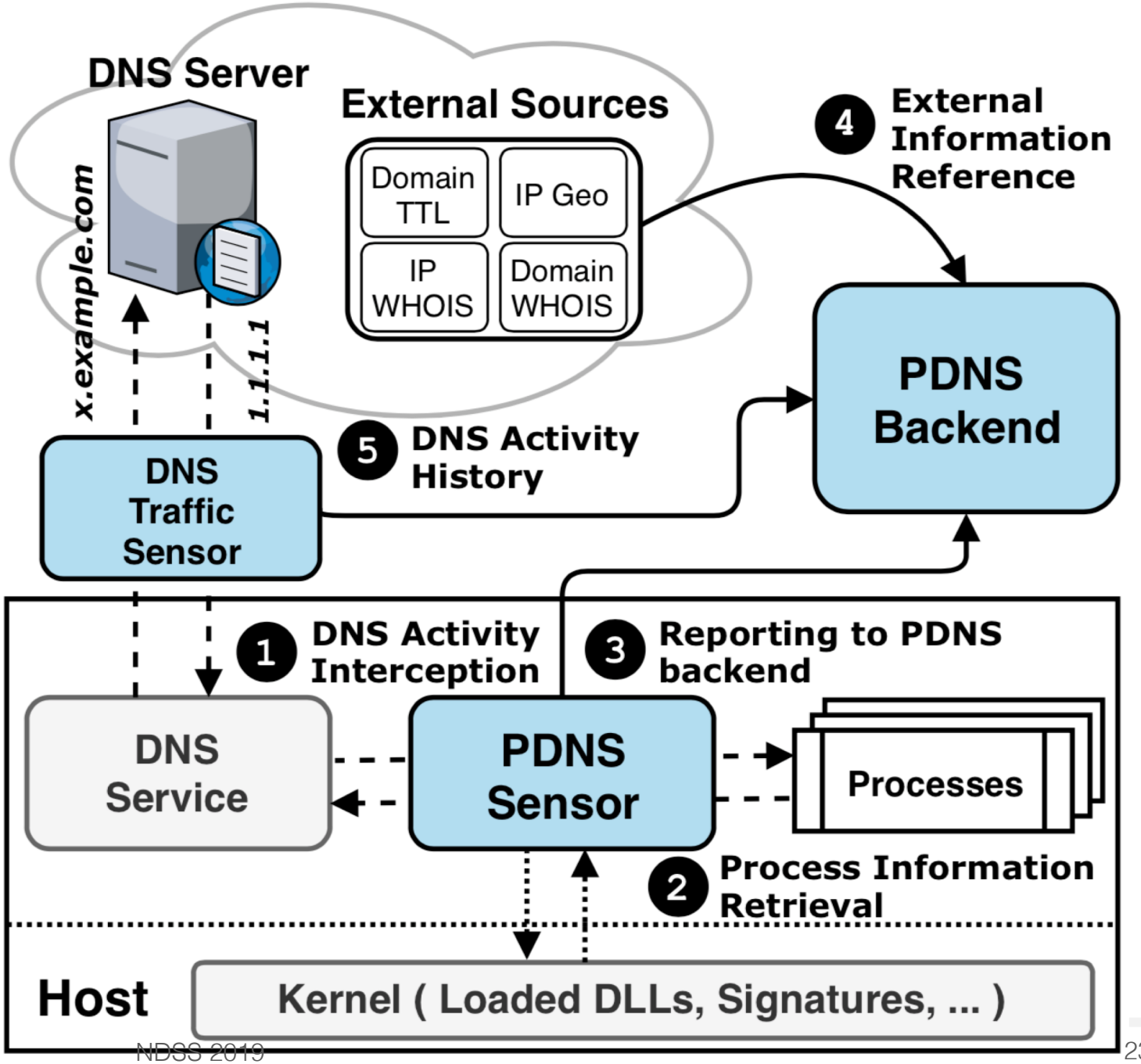


(c) Software publisher and domain registrant similarity score



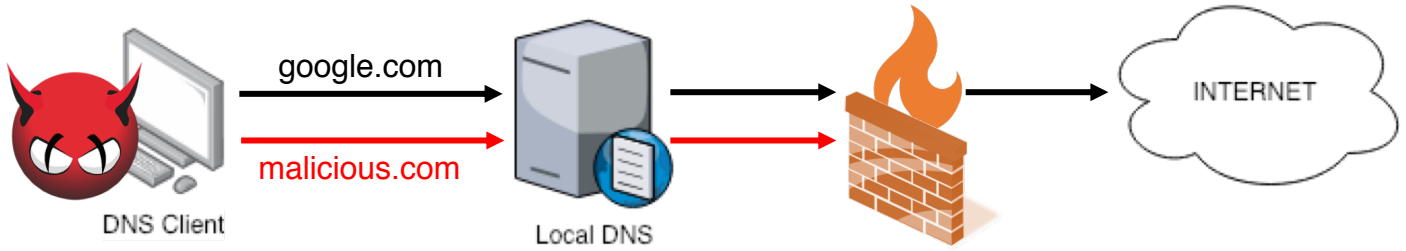
(d) IP and Domain geo-correlation score

PDNS: Process-DNS association



DNS-based Detection Systems

Network-based malicious domain detections



Malicious Domain Detections → Malicious Activity Detections

DNS-based Detection Systems

Network-based malicious domain detections

1. **Static**: Domain or IP address blacklisting

2. **Dynamic**: Anomaly detections
   e.g., diversity of resolved IPs, geographical information, name string structure (DGA), and DNS TTLs.
   [Antonakakis et al., USENIX 2010, Bilge et al., NDSS 2011, Antonakakis et al., USENIX 2011]