Neural reputation models learned from passive DNS data

Pierre Lison
Norsk Regnesentral

Vasileios Mavroeidis
University of Oslo

IEEE Big Data Workshop on Big Data Analytics for Cyber Crime Investigation & Prevention

Boston, December 2017

Introduction

- *Blacklists* and *whitelists* (=reputation lists) are often employed to filter network traffic

- Shortcomings:
  - Complex, time-consuming (manual) process
  - Limited coverage
  - Static (can be circumvented through techniques such domain flux and fast flux networks)

➡ Can we use machine learning to automatically predict the reputation of end-point hosts?

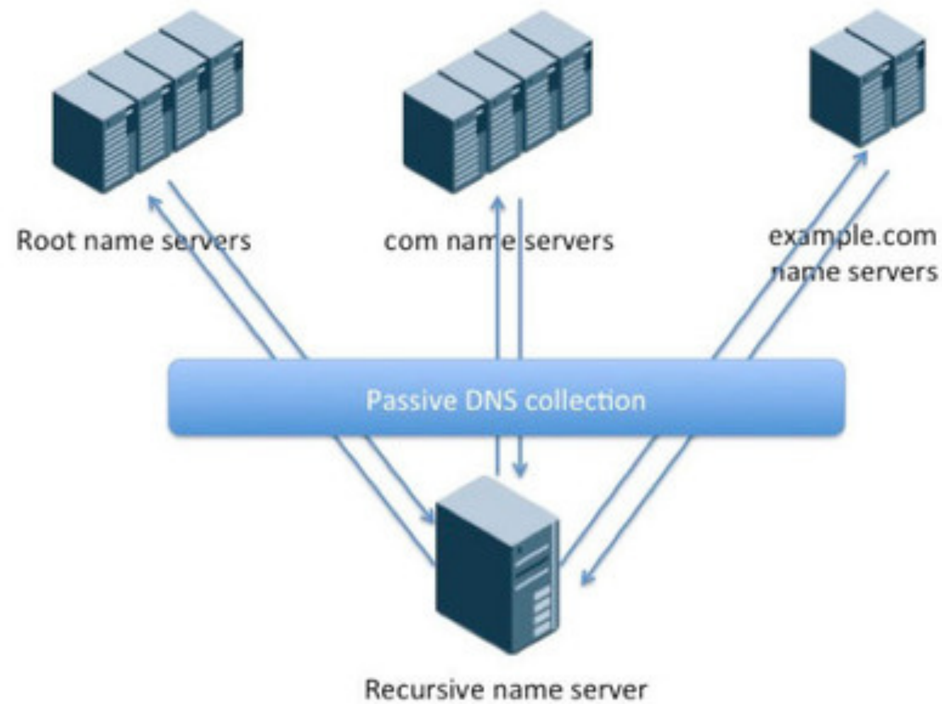
Introduction

Can we use machine learning to automatically predict the reputation of end-point hosts?

Benefits:

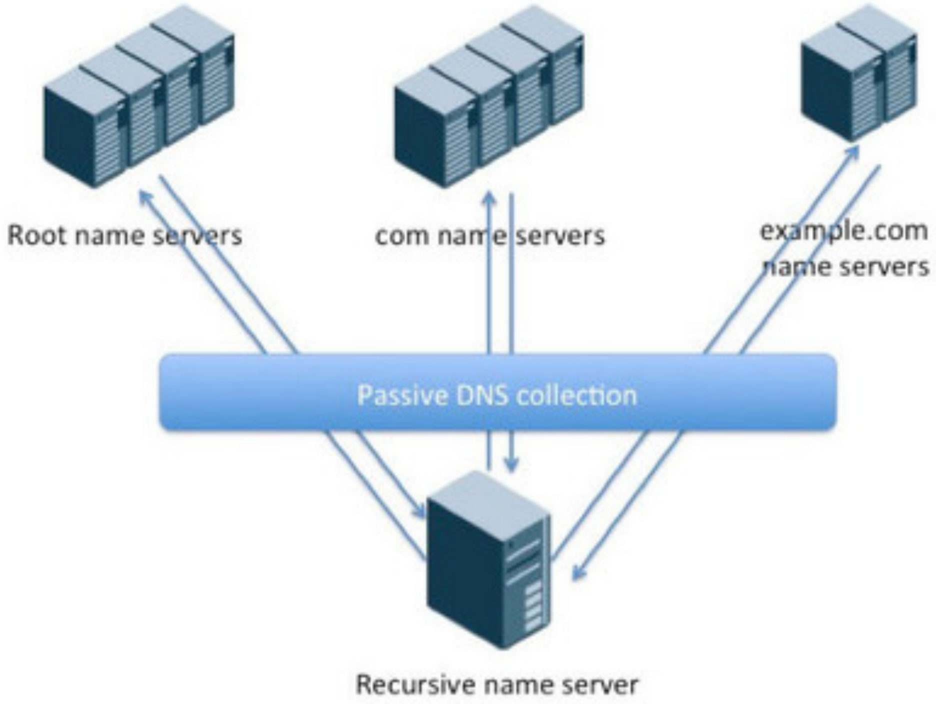
1. Ability to provide predictions in *real-time*, without human intervention
2. Less vulnerable to human errors and omissions than traditional reputation lists
3. Can provide reputation labels for any known end-point host (full coverage)

- Used a large passive DNS from Mnemonic:
  - *567 million* aggregated DNS queries collected over four years
  - *Server-to-server* communication (less privacy concerns)



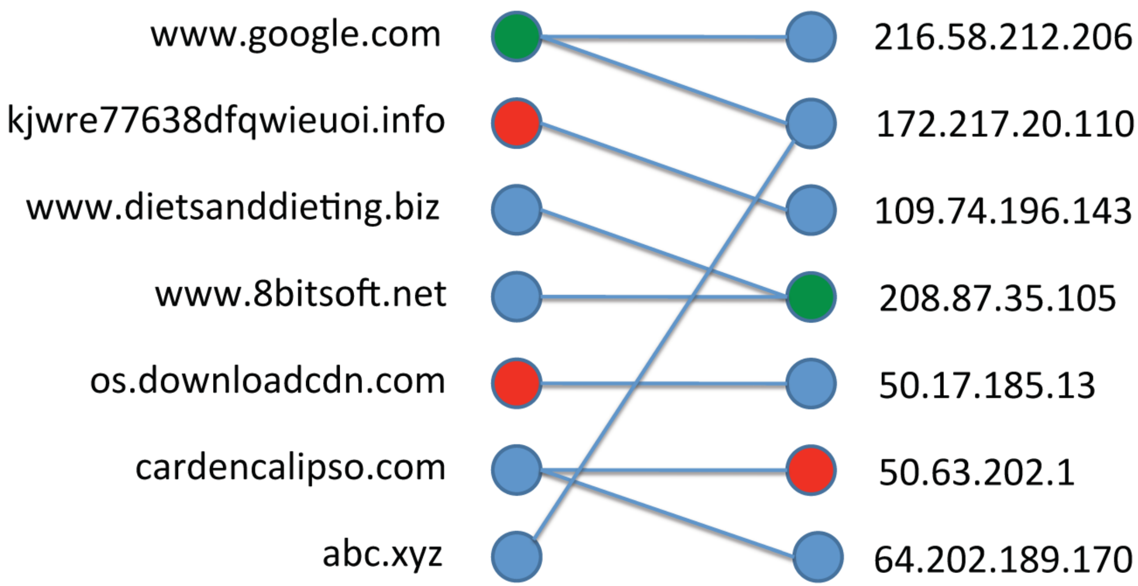
Reputation models

- Can we automatically predict the reputation of domain names and IP addresses from DNS data?
- Used a large passive DNS from Mnemonic:
  - *567 million* aggregated DNS queries collected over four years
  - *Server-to-server* communication (less privacy concerns)



Data

Labelled dataset of **378 million** records (**122 M** records labelled as benign, 9 M records as malicious and 201 K records as sinkhole)



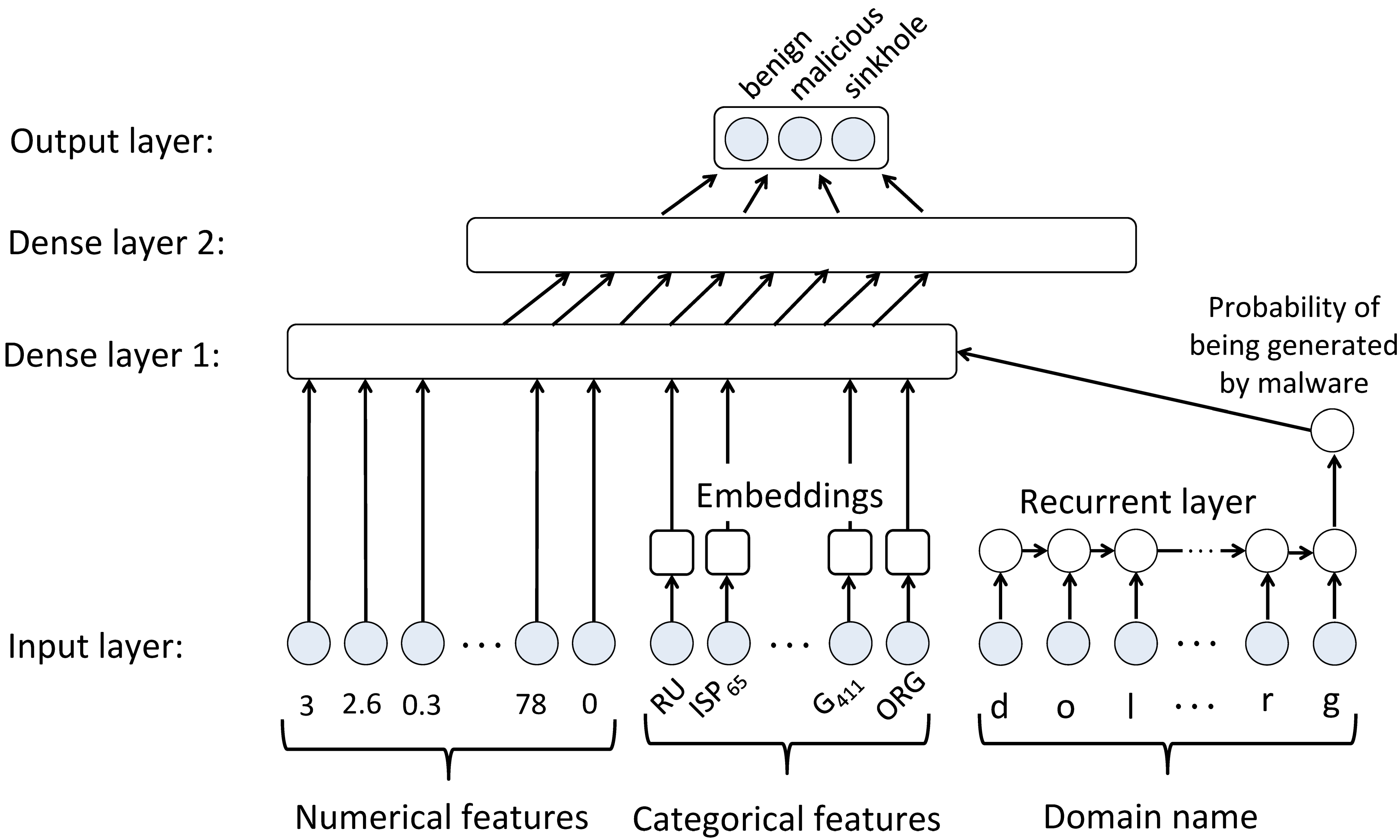
We enriched the passive DNS data with:

- ► Reputation labels from existing blacklists and whitelists
- ► IP location(geoname identifiers) and ISP data

Features

- Numerical features derived from the records:
  - Lifespan, number of queries (for record, domain or IP), number of distinct countries or ISP, TTL values, etc.
- Categorical features:
  - ISP, geolocation, top-level domain, etc.
- Ranking features from Alexa
- Features extracted from neighbouring records
  - Number of records at distance 1 and of reputation X
- Sequence of characters from the domain

Neural model



Results

Model	Benign			Malicious			Sinkhole			Accuracy
	P	R	F_1	P	R	F_1	P	R	F_1	
nb_domain_queries < 10	0.98	0.44	0.61	0.10	0.87	0.19	0.0	0.0	0.0	0.54
Logistic regression	0.97	0.97	0.97	0.60	0.65	0.62	0.51	0.26	0.35	0.944
Neural net (with 1 hidden layer)	0.99	0.99	0.99	0.93	0.93	0.93	0.99	1.00	0.99	0.990
Neural net (with 2 hidden layers)	1.00	0.99	0.99	0.92	0.95	0.93	0.98	1.00	0.99	0.990
Neural net (with 3 hidden layers and two passes)	1.00	**1.00**	1.00	0.97	0.96	**0.96**	0.99	0.96	0.98	**0.995**

In this setting, the neural net is first trained on the labelled dataset and applied to predict the reputation of unlabelled records, which are then used to get better estimates of the "neighbour" features. The model is then trained again on these new feature values.

Conclusion

- Neural networks can be successfully used to predict the **reputation** of end-point hosts
  - Detection of DGA from the domain names
  - Detection of malicious records from passive DNS
- Can be integrated in software tools for cyber-threat intelligence
- Current work:
  - Consolidate experimental results
  - Submission of journal article