Metadata-based Malicious Cyber Discovery

Mehadi Hassen
Ph.D. Student in Computer Sciences
Florida Institute of Technology

Dr. Phil Chan
Associate Professor,
Florida Institute of Technology

Dr. Marco Carvalho
Associate Professor,
Florida Institute of Technology

The Challenge

- To invent and prototype approaches for identifying high interest, suspicious and likely malicious behaviors from meta-data that challenge the way we traditionally think about the cyber problem. As C3E, we value innovation and paradigm shifting approaches above incremental improvements to existing anomaly techniques.

GaTech Malware Passive DNS Data

- DNS queries (and replies)
  - from suspicious/malicious programs
- Data set
  - Date, hashcode/program, domain name, IP address

Distribution of Number of DNS Queries per Program

# of queries per program	2011	2012	2013
1	0.492	0.514	0.456
2	0.159	0.183	0.148
3	0.122	0.154	0.207
4	0.076	0.071	0.121
5	0.066	0.018	0.024
6 or more	0.085	0.060	0.033

- About half of the programs have only 1 query
- 3-9% of the programs have 6 or more queries

Our Approach to Characterizing DNS Behavior of Malicious Programs

1. Identify features for program behavior
2. Find patterns from
   a) any number of DNS queries (per program)
   b) larger number of DNS queries (per program)

Features

DNS Behavior of Malicious Programs

1. Suspicious domain names

2. Suspicious IP addresses

3. Suspicious combinations of domain names and IP addresses

Suspicious Domain Names

- Known malicious domain names
  - Blacklist from maliciousdomains.com
    - Domain name, reason, date entered, date for next review…
  - (multiple blacklists on the web)

- Unresolved domain names
  - DNS did not have a reply

Suspicious IP addresses

- Fake and suspicious
  - DNS might return a fake IP
    - 1.1.1.1, 2.2.2.2, …
  - Addresses for loopback, network, broadcast, private/internal network, …

- Cannot be mapped to a country
  - Unknown country or reserved
  - Data from software77.net/geo-ip

- Mapped to a foreign country (not USA)

Suspicious Combinations of Domain Names and IP addresses

- Multiple domain names
  - are resolved to the same IP address

- One domain name
  - is resolved to IP addresses in different countries

Fraction of Programs with Feature(s)

	2011	2012	2013
Programs	2158919	3299863	3424589
atLeastOneFeature	0.965	0.950	0.936
ipMultiDomains	0.805	0.849	0.831
notUSA	0.726	0.762	0.634
domainMultiCountries	0.319	0.390	0.223
fakeIP	0.260	0.331	0.215
domainUnresolved	0.226	0.319	0.204
noCountry	0.028	0.018	0.019
malwareDomain	0.018	0.009	0.012

Observations

- 93+% -- at least one of the 7 features

- .9 to 1.8% -- on the maliciousdomains.com blacklist
  - Relying on a blacklist might not be sufficient

- 3.5 to 6.4% -- none of the features
  - Need more features

- Ranking of features is consistent over 3 years

Patterns from Any Number of Queries

Learning a Model for Malicious Behavior

- Given
  - A set of malicious programs described by features
    - A program has a feature
      - if any of its queries has the feature
- Find
  - A concise list of patterns (model) that describes the programs

Patterns

- Allow wild card (don't care) for features
  - E.g., notUSA & fakeIP
    - Wild card for the other features
    - (Different from feature combinations, all features are T or F)

- Generalized to cover different feature combinations

Correlation (Quality) of a Pattern

- Mutual Information ("Total Correlation")

$$P(A,B,\ldots)\log\left(\frac{P(A,B,\ldots)}{P(A)P(B)\ldots}\right)$$

  - *P(A,B,…)*
    - Observed joint probability
  - *P(A)P(B)…*
    - Expected joint probability (if A,B,… are independent)

- Higher mutual information => more correlation

Algorithm Outline

1. Sort patterns in descending mutual information
2. While more programs/hashcodes and patterns
   a) If programs match the best pattern
      i. Remove the programs
      ii. Add the pattern to the model
   b) Update the best pattern to the next best

Top 3 Learned Patterns (empty=wild card)

	2011			2012			2013		
ipMultiDomains	T	T		T		T	T	T	T
notUSA	T	T	T	T	T	T	T	T	T
domainMultiCountries		T	T			T		T	
fakeIP	T	T	T	T	T	T	T	T	T
domainUnresolved	T			T	T		T		
noCountry									
malwareDomain									
MutualInfo	0.613	0.330	0.276	0.708	0.633	0.298	0.640	0.195	0.176

- Two of the top 3 most correlated patterns
  - are consistent over 3 years

Evaluating the Learned Models

	2011	2012	2013
	Training set	Test set	
# of patterns in model	23		
Coverage	.964	.950	
		Training set	Test set
# of patterns in model		27	
Coverage		.950	.936
			Training set
# of patterns in model			24
Coverage			.936

- Missed programs have none of the features
- No normal programs in the test set
  - Models could be overfiting and have false coverage
    - Can be reduced by increasing features & threshold for mutual info

Patterns from
Larger Number of Queries

Query Sequence

- Query
  - Represented by feature combinations
- Query sequence (n-gram)
  - In the order issued by the program
  - Trigrams and pentagrams
- Consider programs with at least 3 or 5 queries

Top-5 Trigrams

Sym	Feature
C	domainMultiCountries
D	ipMultiDomains
N	notUSA
U	domainUnresolved

- Top trigram in all 3 years:
  - CD,CD,CD (.20, .20, .45)
    - Rank 1 in all 3 years
- Others in Top 5 in 2 years:
  - CDN,CDN,CDN (.09, .09)
  - U,U,CU (.09, .11)
  - U,U,CD (.08, .10)
  - U,U,U (.08, .11)

Top-5 Pentagrams

Sym	Feature
C	domainMultiCountries
D	ipMultiDomains
N	notUSA
U	domainUnresolved

- Top 5 in all 3 years:
  - U,U,U,U,U (.19, .41, .11)
    - Rank 1 in 2012
- Top 5 in 2 years:
  - CD,CD,CD,CD,CD (.28, .24)
    - Rank 1 in 2011 & 2013
  - U,U,U,U,CD (.17, .36)

Clustering Programs Based on Query Sequences

- Programs clustered based their top query trigram and pentagram sequence
- Distance function
  - Hamming Distance
  - Edit distance
- Centroid
  - Sequence at minimum overall distance from others in a cluster.

Cluster Results

- Select top clusters centroids
  - Size of cluster and average distance between centroid and elements.

- Correspond to top trigram and pentagram sequence.

Concluding Remarks

- Additional features
  - from more information on the context of data
  - domain names requested by more programs are less suspicious
  - finer-grain (e.g. country name, % of queries with feature)
- DNS data from normal programs
  - can help evaluate the models more effectively
- Scalability (patterns from any # of queries)
  - Sampling “good” patterns with a randomized alg (e.g. LERAD)
- Markov Models of the query sequences
  - cluster the programs based on it.

Thank You

Questions?