BIG DATA ANALYTICS FOR CYBER SECURITY

V.S. Subrahmanian
University of Maryland, College Park
vs@cs.umd.edu, @vssubrah

Joint work with Chanyun Kang, Noseong Park, B. Aditya Prakash, Edoardo Serra
## Malware Pandemic

<table>
<thead>
<tr>
<th>Year</th>
<th>Malware Variants</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>317 Million (+26%)</td>
<td>Symantec Report 2015</td>
</tr>
<tr>
<td>2013</td>
<td>252 Million</td>
<td></td>
</tr>
</tbody>
</table>

**New Malware Variants** (Added in Each Year)

Source: Symantec
Malware is hard to detect!

Diagram:
- Host A
- Host B
- Host C
- Host D
- Host E

Infection interval:
- Infection time
- Detection time

Detection:
- 1st Discovery
- Inspect
- Release signature
- Release patch

# of infected hosts:
- t1: 1
- t2: 2
- t3: 4
- t4: 5
- t5: 4
- t6: 3
- t7: 4
- t8: 4
- t9: 3

# of detections:
- t1: 0
- t2: 0
- t3: 0
- t4: 1
- t5: 1
- t6: 0
- t7: 0
- t8: 2
- t9: 1
Key Challenge

• Statistics from Symantec WINE Dataset
  – # of Detections <<< # of Infections
Problem Statement

Using these detections

Host A
Host B
Host C
Host D
Host E

Infection interval
Infection time
Detection time

# of infected hosts
1 2 4 5 4 3 4 4 3

# of detections
0 0 0 1 1 0 0 2 1

Predict this value
Our Approaches

• Feature based prediction method
  — Proposed a set of novel features

• Epidemic model inspired by SIR model

• Ensemble method that merges the previous two methods with other state-of-the-art techniques.
1st Method

FEATURE BASED PREDICTION MODEL
Feature Based Method

Each record = (Host, Malware, File name, Infection time, Detection time)

2. Compute host-level features

‘Detection and Patch incompetence’ of each host
‘Detection and Patch ability’ of each host
‘Detection and Patch hardness’ of each malware

3. Compute country-level features

Aggregate host level features, e.g. Average

4. Train a prediction model with the features

Prediction model

Features / Infection Ratio

The expected number of infections in future

<table>
<thead>
<tr>
<th>day</th>
<th>Feature #1</th>
<th>Feature #2</th>
<th>……</th>
<th>Infected Host Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td></td>
<td></td>
<td></td>
<td>Ground Truth</td>
</tr>
<tr>
<td>d+1</td>
<td></td>
<td></td>
<td></td>
<td>Ground Truth</td>
</tr>
<tr>
<td>……</td>
<td></td>
<td></td>
<td></td>
<td>Ground Truth</td>
</tr>
<tr>
<td>d+n</td>
<td></td>
<td></td>
<td></td>
<td>Ground Truth vs. Predictions</td>
</tr>
</tbody>
</table>

~2 years data

Symantec Telemetry data
Detection/Patch Incompetence

• Each record = (Host $h$, Malware $m$, File name $f$, Infection time $i$, Detection time $t$)

• Detection time – Infection time (Detection Incompetence$^1$)
  – How good/bad is a user $h$ at detecting malware $m$?
  – How easy/hard is it to detect malware $m$?

• Patch time – Infection time (Patch Incompetence$^1$)
  – How good/bad is a user at patching a vulnerability/malware?
  – How easy/hard is it to patch a vulnerability/malware?

• Average these values for each host $\rightarrow$ host-level detection/patch incompetence

• Some other similar features, e.g., Detection time – Malware signature release time

1: These two are the most simplest features.
Detection Ability/Hardness

• Each record = (Host $h$, Malware $m$, File name $f$, Infection time $i$, Detection time $t$)

• Detection Ability (ADA) of host $h$ is the weighted sum of Detection Hardness (ADH) of malware detected by $h$.

\[
AD_{DA}(h) = \sum_{(f,m,t) \in dH(h)} w_{12}(h, f, m, t) \cdot ADH(m)
\]

A subset of WINE records, where Host = $h$

• Detection Hardness of malware $m$ is the weighted sum of Detection Ability of hosts that detected $m$.

\[
AD_{DH}(m) = \sum_{(f,h,t) \in dM(m)} w_{21}(m, f, h, t) \cdot ADA(h)
\]
BiFixpoint Algorithm

Algorithm 1: BiFixpoint

Input : $\mathcal{H}, \mathcal{M}, T$ (*$T$ is a training set *)
Output: $ADA, ADH$

1. forall $h \in \mathcal{H}$, $ADA(h) \leftarrow \frac{1}{|\mathcal{H}|}$ (* initialize *)
2. forall $m \in \mathcal{M}$, $ADH(m) \leftarrow \frac{1}{|\mathcal{M}|}$
3. change $\leftarrow$ true;
4. while change do
5. \[
    ADA'(h) \leftarrow \sum_{(f,m,t) \in dH(h)} w_{12}(h, f, m, t) \ast ADH(m)
    \]
6. \[
    ADH(m) \leftarrow \sum_{(f,h,t) \in dM(m)} w_{21}(m, f, h, t) \ast ADA(h)
    \]
7. if $ADA' \sim ADA$ and $ADH' \equiv ADH$ then
8. change $\leftarrow$ false
9. else
10. \[
        ADA' \leftarrow ADA' \text{ and } ADH' \leftarrow ADH'
    \]
11. end
12. end
13. return $Ada, ADH$

We prove that convergence is always guaranteed!
Collaborative Features

• Given two similar\(^1\) hosts \(h_1\) and \(h_2\)
  – Suppose \(h_1\) was infected by \(m\).
  – \(h_2\) is likely to be infected soon with prob \(\sim sim(h_1, h_2)\).
• \(cf(h,m)\) is the estimated prob. of host \(h\) being infected by \(m\) (considering similarity).
• \(cf(C,m)\) is the sum of \(cf(h,m)\), where \(h\) is a host in country \(C\).

1: We defined various similarity measures based on calculated features.
**Time Lag Features**

- Today’s infection ratio depends on not only today’s features but also past features.
- Very high dimensional feature space

<table>
<thead>
<tr>
<th>day</th>
<th>Feature #1</th>
<th>Feature #1 (-1 day)</th>
<th>Feature #1 (-7 day)</th>
<th>.....</th>
<th>Infected Host Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>80% Training</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ground Truth</td>
</tr>
<tr>
<td>d+1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20% Test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d+n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ground Truth vs. Predictions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Recap of Features

• Features from raw values
  – Detection time – Infection time (Detection Incompetence)
  – Patch time – Infection time (Patch Incompetence)
  – Some features calculated from raw data

• Features from BiFixpoint Algorithm
  – Detection ability, Patch ability for hosts
  – Detection hardness, Patch hardness for malware

• Collaborative Features
  – Infection numbers based on host similarity

• Country Human Development Index, ...

• Time lag features

• Country level aggregation ➔ Regression Problem
2nd Method

EPIDEMIC PREDICTION MODEL
Epidemic Model

• SIR Model models the dynamics of infectious disease.
• Sometimes used for social rumor diffusion.
• Does not fit the spread of malware.
  – Recovered doesn’t precisely capture the dynamics of malware spread.
  – Transition rate is not designed for malware.
  – Network data may not always be available.

\[ b = \text{the rate at which susceptible people become infectious} \]
\[ r = \text{the rate at which infectious people recover/develop immunity} \]
DIPS Epidemic Model

• “Recovered” → “Detected” and “Patched”
• Carefully designed transition rates
• $S(t)$, $I(t)$, $D(t)$ and $P(t)$ are the number of susceptible, infected, detected and patched hosts at time $t$
• $S(t)$, $I(t)$, $D(t)$ and $P(t)$ are recursively defined.

\[
\begin{align*}
\theta(t): & \text{ patching rate} \\
\theta(t) = & \begin{cases} 0 & (t < t_p) \\
\theta_0 & (t \geq t_p) \end{cases} \\
\delta(t): & \text{ patching rate} \\
\delta(t) = & \begin{cases} 0 & (t < t_p, t_d) \\
\delta_0 & (t \geq t_p, t_d) \end{cases} \\
\beta(t): & \text{ infection rates} \\
\beta(t) = & \beta_0 \cdot \left(1 + P_a \cdot \cos \left(\frac{2\pi}{P_p} \cdot (t + P_s)\right)\right) \\
\gamma(t): & \text{ # detections in model} \\
\gamma(t) = & \gamma_0 \cdot \text{DET}(t)
\end{align*}
\]

- $S(t+1) = S(t) - \beta(t) \cdot S(t) \cdot I(t) + (1 - \delta(t)) \cdot D(t) - \theta(t) \cdot S(t)$
- $I(t+1) = I(t) + \beta(t) \cdot S(t) \cdot I(t) - \gamma_0 \cdot \text{DET}(t)$
- $D(t+1) = \gamma_0 \cdot \text{DET}(t)$
- $P(t+1) = P(t) + \delta(t) \cdot D(t) + \theta(t) \cdot S(t)$
How to predict with DIPS

• Find the optimal set of parameters with Least Square Method to minimize the sum of \((\text{true-prediction})^2\)

• Train with the target country-malware pair.
  – Initialization \(\rightarrow\) local optimal \(\rightarrow\) not stable learning

• Learning algorithm (two phases)
  – First, train the parameters with all countries and malware
  – Second, train again only for the target country-malware

Train with the first 80% infection/detection history

Detections, \(DET(t)\), for the last 20%

Infections, \(I(t)\)
DIPS - Susceptible

\[ \beta(t) = \beta_0 \cdot (1 + P_a \cdot \cos \left( \frac{2\pi}{P_p} \cdot (t + P_S) \right)) \]

- \( S \rightarrow I \) in between \( t \) and \( t+1 \): \( \beta(t) \cdot S(t) \cdot I(t) \)
- \( D \rightarrow S \): \( (1 - \delta(t)) \cdot D(t) \)
- \( S \rightarrow P \): \( \theta(t) \cdot S(t) \)
- \( S(t+1) = S(t) - \beta(t) \cdot S(t) \cdot I(t) - \theta(t) \cdot S(t) + (1 - \delta(t)) \cdot D(t) \)

1: This is from SIR model.
DIPS - Detected

\[ \delta(t) = \begin{cases} 0 & (t < t_p, t_d) \\ \delta_0 & (t \geq t_p, t_d) \end{cases} \]

\[ \gamma(t) = \gamma_0 \cdot DET(t) \]

- I \rightarrow D: \gamma_0 \cdot DET(t), where DET(t) is the true detection numbers at time t
- D \rightarrow S: (1 - \delta(t)) \cdot D(t)
- D \rightarrow P: \delta(t) \cdot D(t)
- D(t) = \gamma_0 \cdot DET(t)
DIPS-exp Epidemic Model

- Modeling of “Birth” of the SIR model
- $\sigma(t)$ is added.

$\sigma(t)$: inflow of additional hosts

$\delta(t)$: patching rate
$\theta(t) = \begin{cases} 0 & (t < t_p) \\ \theta_0 & (t \geq t_p) \end{cases}$

$\beta(t) = \beta_0 \cdot (1 + P_a \cdot \cos \left( \frac{2\pi}{P_p} \cdot (t + P_S) \right))$

$\delta(t)$: patching rate
$\delta(t) = \begin{cases} 0 & (t < t_p, t_d) \\ \delta_0 & (t \geq t_p, t_d) \end{cases}$

$\gamma(t) : \#$ detections in model
$\gamma(t) = \gamma_0 \cdot DET(t)$

- $S(t+1) = S(t) - \beta(t) \cdot S(t) \cdot I(t) + (1 - \delta(t)) \cdot D(t) - \theta(t) \cdot S(t) + \sigma(t)$
- $I(t+1) = I(t) + \beta(t) \cdot S(t) \cdot I(t) - \gamma_0 \cdot DET(t)$
- $D(t+1) = \gamma_0 \cdot DET(t)$
- $P(t+1) = P(t) + \delta(t) \cdot D(t) + \theta(t) \cdot S(t)$
3rd Method

ENSEMBLE PREDICTION MODEL
Combine Prediction Models

- Combine Feature Method and DIPS.
- Use DIPS prediction results as additional features.
Not Enough Training Data

• To predict number of hosts infected by malware $m$, train jointly with similar malware

• Discover similar malware with Dynamic Time Warping to calculate time-series similarity

• Lots of noise
Robust Regression

• Need a robust regression

• Gaussian Process Regression
  – Very strong Bayesian regression method
  – Less parametric (Parameters are calculated from data with maximum likelihood.)

Linear Regression
\[ \hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p \]

Ridge Regression
\[ \min_w \|Xw - y\|_2^2 + \alpha \|w\|_2^2 \]

Lasso Regression
\[ \min_w \frac{1}{2n_{samples}} \|Xw - y\|_2^2 + \alpha \|w\|_1 \]

Linear combination of weighted features + regularization term
## ESM Model

<table>
<thead>
<tr>
<th>Feature #1</th>
<th>Feature #2</th>
<th>DIPS output</th>
<th>......</th>
<th>Infected Host Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>80% Training (m)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80% Training (m1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20% Test (m)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Symantec WINE dataset**

- Susceptible
- Patched
- Detected
- Infected

**Host state**

- Malware downloaded
- Signature released

**Features**

**ESM**

- GPR

**The expected number of infections in future**

**Cluster of Similar Malware**
Experiment Environment

• Top 50 Most Infectious Malware, Top 40 Country in GDP per capita → 2000 Predictions
• 1.45M unique hosts, 2.99M records
• FBP
• DIPS, DIPS-exp
• FUNNEL: state-of-the-art epidemic model
• ESM0 (FBP + DIPS + DIPS-exp + Similar Malware)
• ESM1 (ESM0 + FUNNEL)
Measurements

- $\text{MAE}^* = |\text{true infections} - \text{predicted infections}|$
- $\text{MSE} = (\text{true infection ratio} - \text{predicted infection ratio})^2$
- $\text{RMSE} = \sqrt{\text{MSE}}$
- $\text{NRMSE}$
- $\text{Pearson Correlation Coefficient}$
FUNNEL (prior art)

• State of the art epidemic model for human disease
• C.C. between truths and predictions are very bad.
Feature Based Prediction
DIPS

Not enough training data
ESM0
## Error Values

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE*</th>
<th>RMSE</th>
<th>NRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBP</td>
<td>73.74</td>
<td>0.00170</td>
<td>0.179</td>
</tr>
<tr>
<td>Funnel</td>
<td>127.83</td>
<td>0.00269</td>
<td>0.226</td>
</tr>
<tr>
<td>DIPS</td>
<td>32.36</td>
<td>0.00083</td>
<td>0.165</td>
</tr>
<tr>
<td>DIPS-Exp</td>
<td>36.56</td>
<td>0.00096</td>
<td>0.223</td>
</tr>
<tr>
<td>ESM₀</td>
<td>39.41</td>
<td>0.00115</td>
<td>0.150</td>
</tr>
<tr>
<td>ESM₁</td>
<td>41.84</td>
<td>0.00118</td>
<td>0.151</td>
</tr>
<tr>
<td>FBP⁻ Funnel</td>
<td>79.01</td>
<td>0.00189</td>
<td>0.179</td>
</tr>
</tbody>
</table>
Prediction results

Days

infected host ratio

0.000

0.055

0.045

0.040

0.035

0.030

0.025

0.020

0.015

0.010

0.005

0.000

0 50 100 150 200 250 300 350 400 450 500 550

ESM  FUNNEL  TRUE  Training/Testing

# of hosts in coun...  554969

ESM Performance

NRMSE  0.10995
RMSE  0.00295
MAE*  1283.57
Correlatio...  0.89259

FUNNEL Performance

NRMSE  0.29253
RMSE  0.00786
MAE*  3988.68
Correlatio...  0.66307

MAE is computed with # of infected hosts
Prediction results

Days

Infected host ratio

ESM, FUNNEL, TRUE, Training/Testing
References

Contact Information

V.S. Subrahmanian
Dept. of Computer Science & UMIACS
University of Maryland
College Park, MD 20742.

vs@cs.umd.edu
@vssubrah
www.cs.umd.edu/~vs/