BIG DATA ANALYTICS FOR CYBER SECURITY

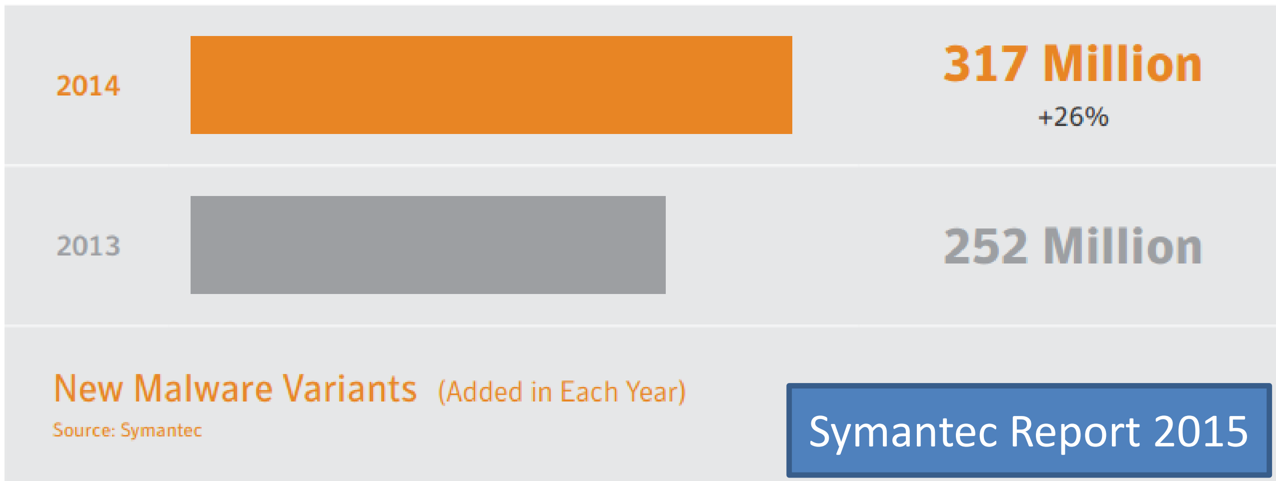
V.S. Subrahmanian

University of Maryland, College Park

vs@cs.umd.edu, @vssubrah

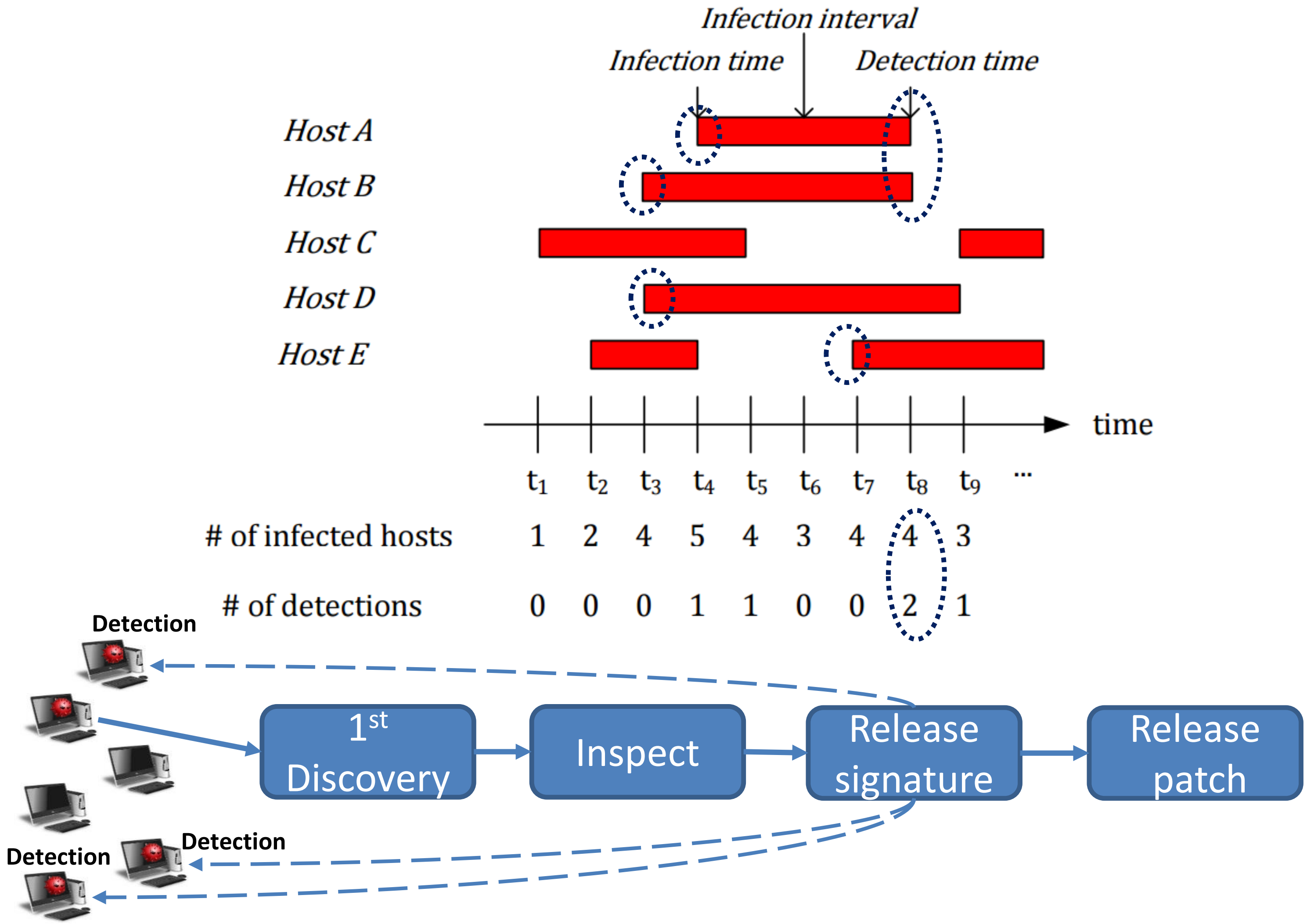
Joint work with Chanhyun Kang, Noseong Park, B. Aditya Prakash, Edoardo Serra

Malware Pandemic



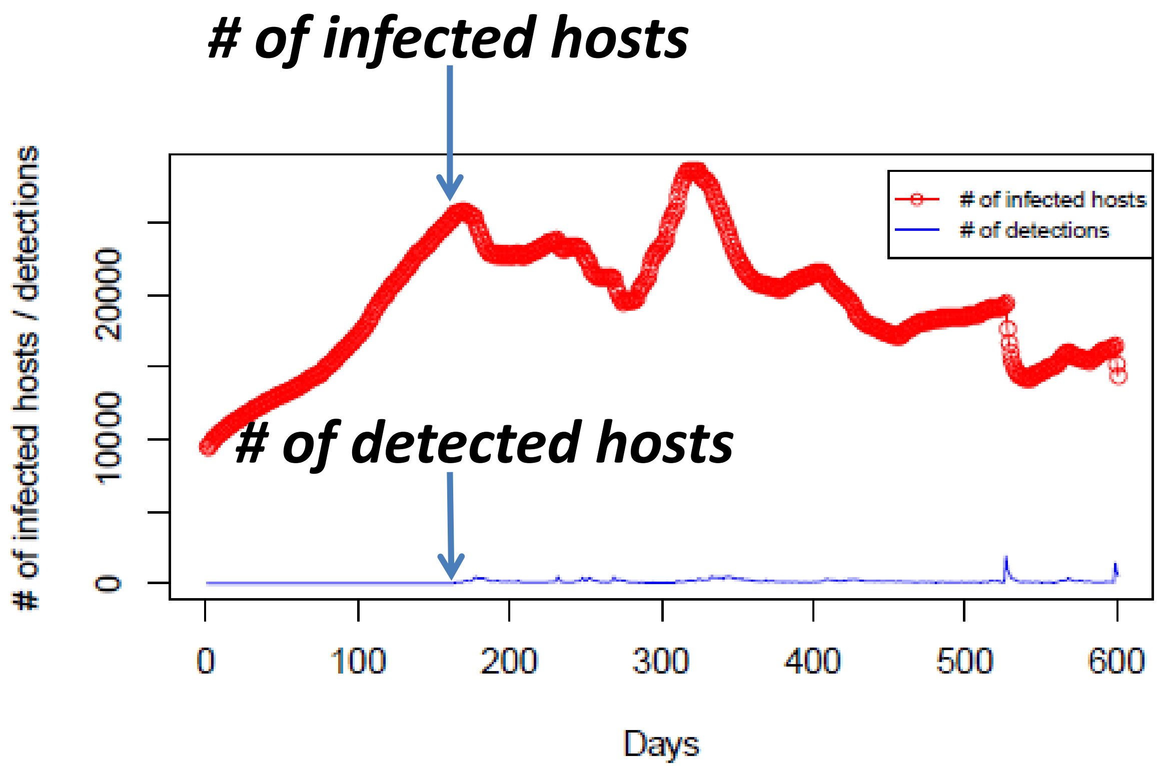
Symantec Report 2015

Malware is hard to detect!



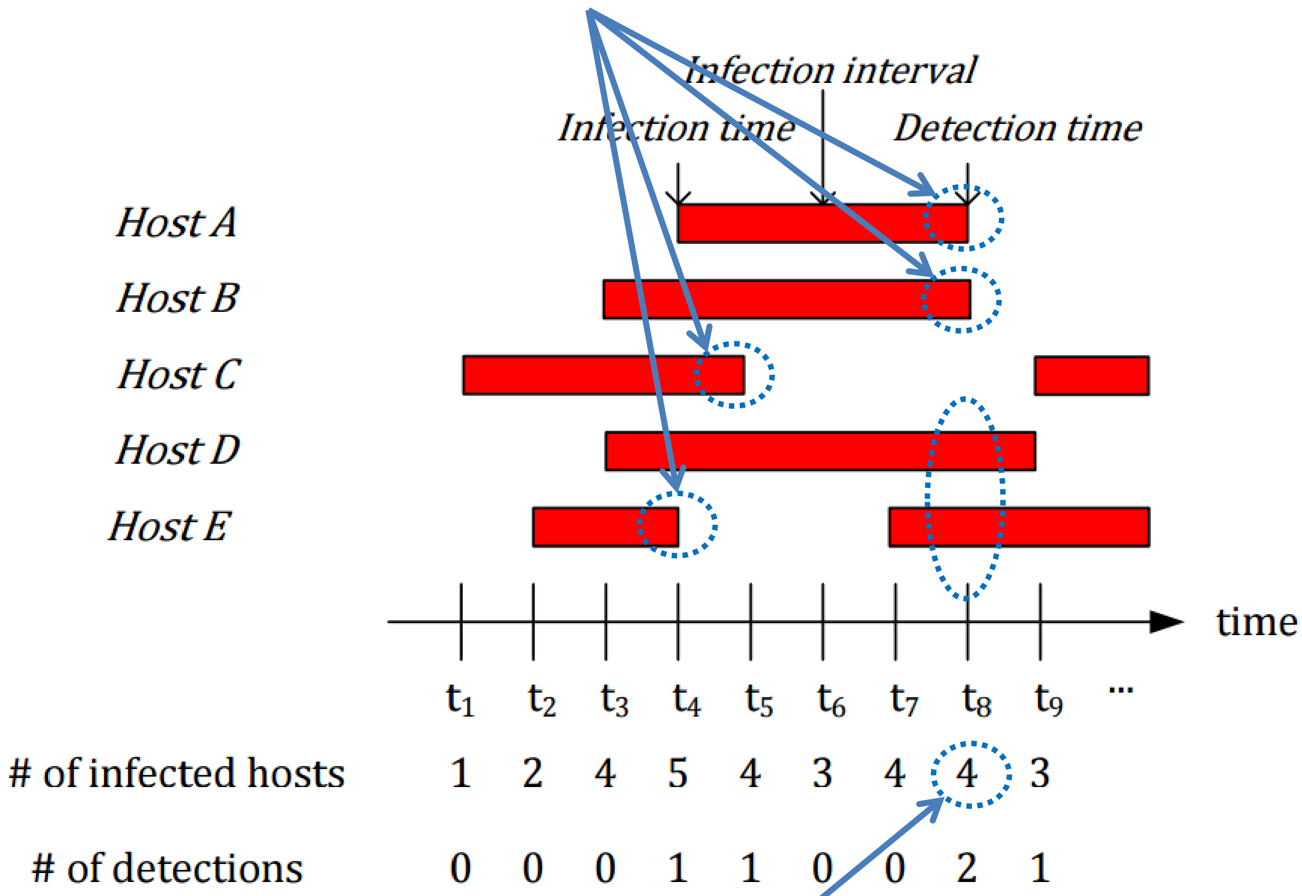
Key Challenge

- Statistics from Symantec WINE Dataset
  - # of Detections <<< # of Infections



Problem Statement

Using these detections



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

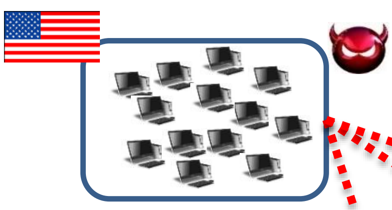
Each record= (Host, Malware, File name, Infection time, Detection time)

Symantec Telemetry data

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

~2 years data

	day	Feature #1	Feature #2	……	Infected Host Ratio
80% Training	d				Ground Truth
	d+1				
	⋮				
20% Test	d+n				Ground Truth vs. Predictions
	⋮				

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 → 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.

$$ADA(h) \quad = \sum_{(f,m,t)\in \boxed{dH(h)}} w_{12}(h,f,m,t)\cdot ADH(m)$$

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.

$$ADH(m) \quad = \sum_{(f,h,t)\in dM(m)} w_{21}(m,f,h,t)\cdot ADA(h)$$

BiFixpoint Algorithm

Algorithm 1: BiFixpoint

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

forall $h \in \mathcal{H}$, $ADA(h) \leftarrow \frac{1}{|\mathcal{H}|}$ (* initialize *)
forall $m \in \mathcal{M}$, $ADH(m) \leftarrow \frac{1}{|\mathcal{M}|}$

Uniform initialization

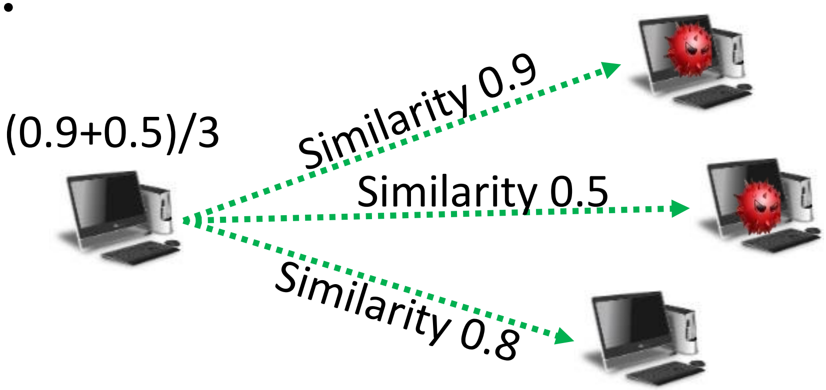
change $\leftarrow$ true;
while *change* **do**
    $ADA'(h) \leftarrow \sum_{(f,m,t) \in dH(h)} w_{12}(h,f,m,t) * ADH(m)$
    $ADH(m) \leftarrow \sum_{(f,h,t) \in dM(m)} w_{21}(m,f,h,t) * ADA(h)$
    if $ADA' \sim ADA$ *and* $ADH' \equiv ADH$ **then**
        change $\leftarrow$ false
    else
        $ADA \leftarrow ADA'$ and $ADH \leftarrow ADH'$
    end
end
return ADA, ADH

Recursive calculation

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 ~ $sim(h_1, h_2)$.
- $cf(h,m)$ 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

	day	Feature #1	Feature #1 (-1 day)	Feature #1 (-7 day)	……	Infected Host Ratio
80% Training	d					Ground Truth
	d+1					
	⋮					
20% Test	d+n					Ground Truth vs. Predictions
	⋮					

Recap of Features

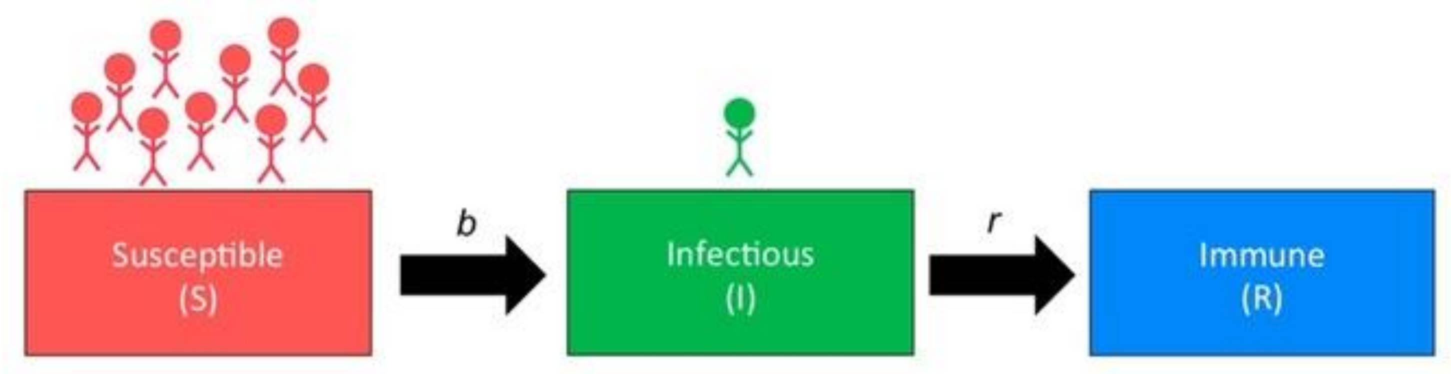
- Features from raw values
  - Detection time – Infection time (Detection Incompetence)
  - Patch time – Infection time (Patch Incompetence)
  - Some features calculated from raw data
- Features from BiFixpoint Algorithm
  - Detection ability, Patch ability for hosts
  - Detection hardness, Patch hardness 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 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 = the rate at which susceptible people become infectious
r = the rate at which infectious people recover/develop immunity

DIPS Epidemic Model

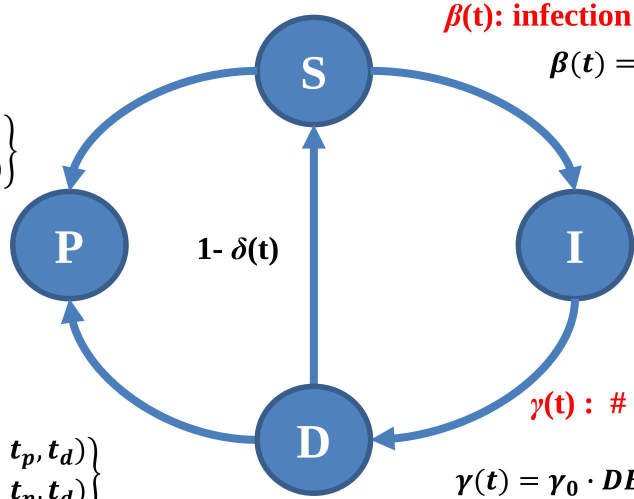
- "Recovered" → "Detected" and "Patched"
- Carefully designed transition rates
- S(t), I(t), D(t) and P(t) are the number of susceptible, infected, detected and patched hosts at time t
- S(t), I(t), D(t) and P(t) are recursively defined.

θ(t): patching rate

$$\theta(t) = \begin{cases} 0 & (t < t_p) \\ \theta_0 & (t \ge t_p) \end{cases}$$

β(t): infection rates

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



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

δ(t): patching rate

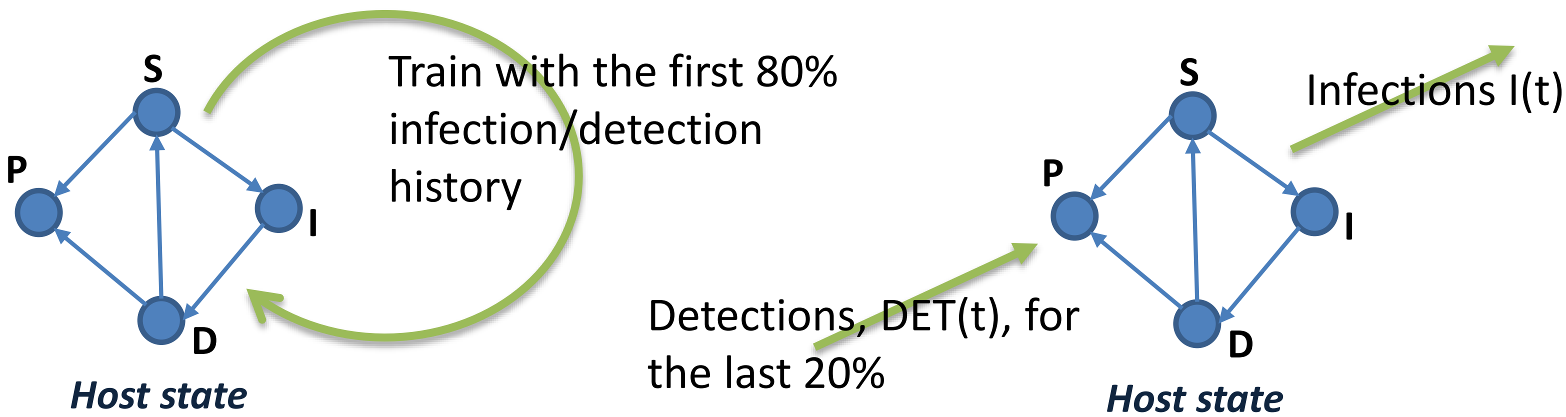
$$\delta(t) = \begin{cases} 0 & (t < t_p, t_d) \\ \delta_0 & (t \ge t_p, t_d) \end{cases}$$

γ(t) : # detections in model

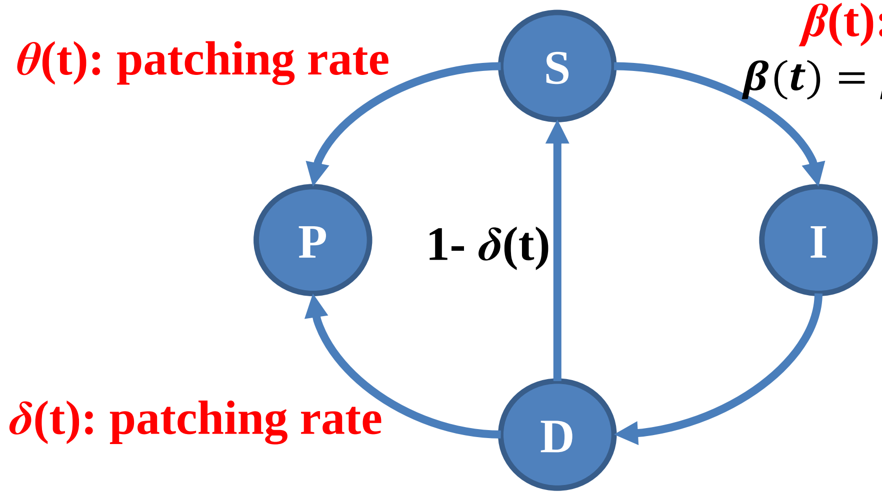
$$\gamma(t) = \gamma_0 \cdot DET(t)$$

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 → local optimal → 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



DIPS - Susceptible



β(t): infection rate

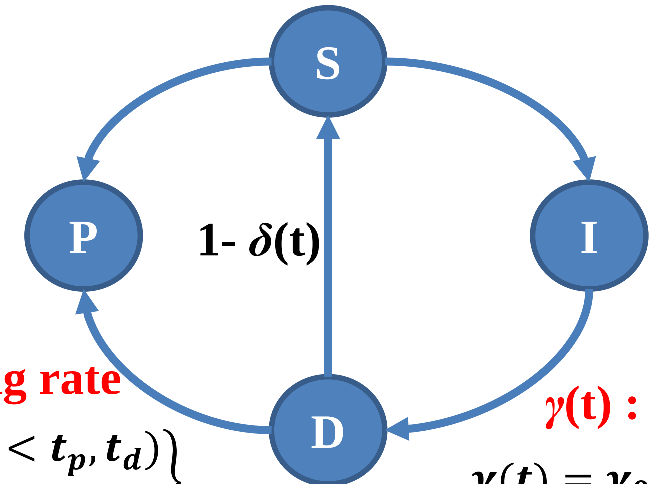
$$\beta(t) = \beta_0 \cdot \left(1 + P_a \cdot \cos\left(\frac{2\pi}{P_p} \cdot (t + P_S)\right)\right)$$

β_0: base infection rate
P_a: amplitude of cycle
P_p: period of cycle
P_S: phase shift of cycle

- S→I in between t and t+1: β(t)·S(t)·I(t)[1]
- D→S: (1- δ(t))·D(t)
- S→P: θ(t)·S(t)
- S(t+1) = S(t) – β(t)·S(t)·I(t) – θ(t)·S(t) + (1- δ(t))·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→D: $\gamma_0 \cdot DET(t)$, where DET(t) is the true detection numbers at time *t*
- D→S: $(1-\delta(t)) \cdot D(t)$
- D→P: $\delta(t) \cdot D(t)$

- $D(t) = \gamma_0 \cdot DET(t)$

DIPS-exp Epidemic Model

- Modeling of "Birth" of the SIR model
- $\boldsymbol{\sigma(t)}$ is added.

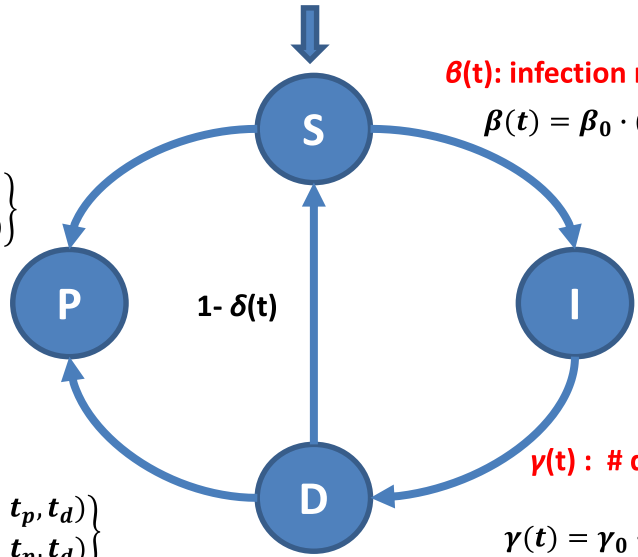
$\boldsymbol{\sigma(t)}$: inflow of additional hosts

ϑ(t): patching rate

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

β(t): infection rates

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



- $S(t+1) = S(t) - \beta(t) \cdot S(t) \cdot I(t) + (1 - \delta(t)) \cdot D(t) - \theta(t) \cdot S(t) + \boldsymbol{\sigma(t)}$
- $I(t+1) = I(t) + \beta(t) \cdot S(t) \cdot I(t) - \gamma_0 \cdot DET(t)$
- $D(t+1) = \gamma_0 \cdot DET(t)$
- $P(t+1) = P(t) + \delta(t) \cdot D(t) + \theta(t) \cdot S(t)$

δ(t): patching rate

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

γ(t) : # detections in model

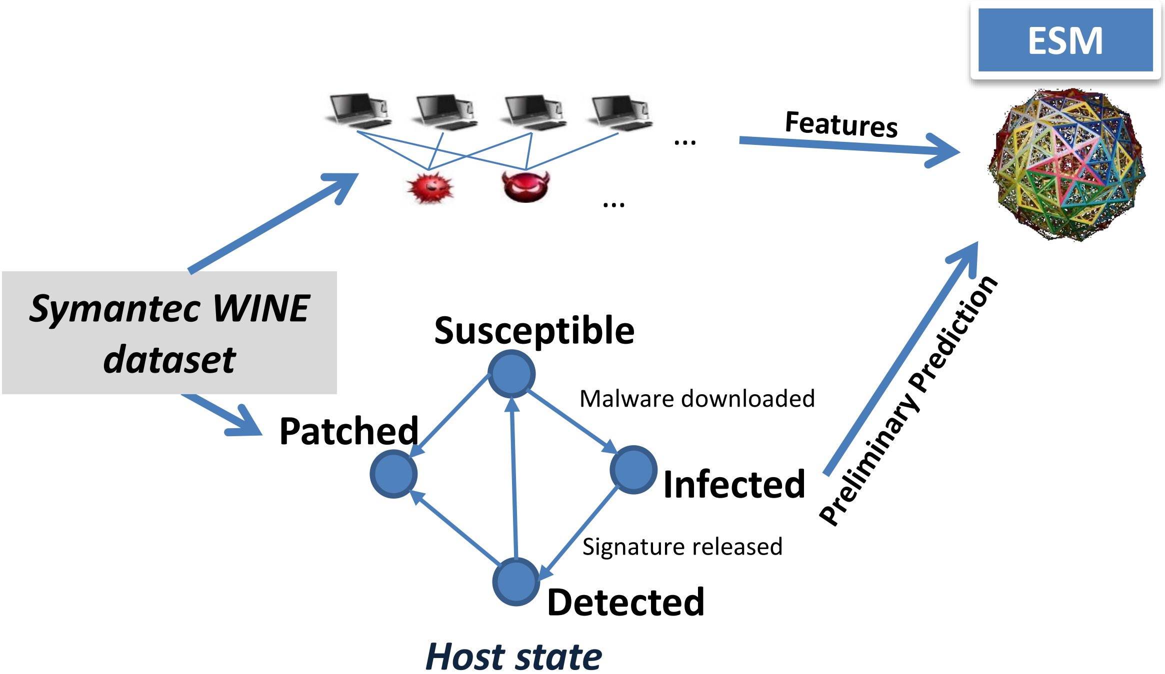
$$\gamma(t) = \gamma_0 \cdot DET(t)$$

3rd Method

ENSEMBLE PREDICTION MODEL

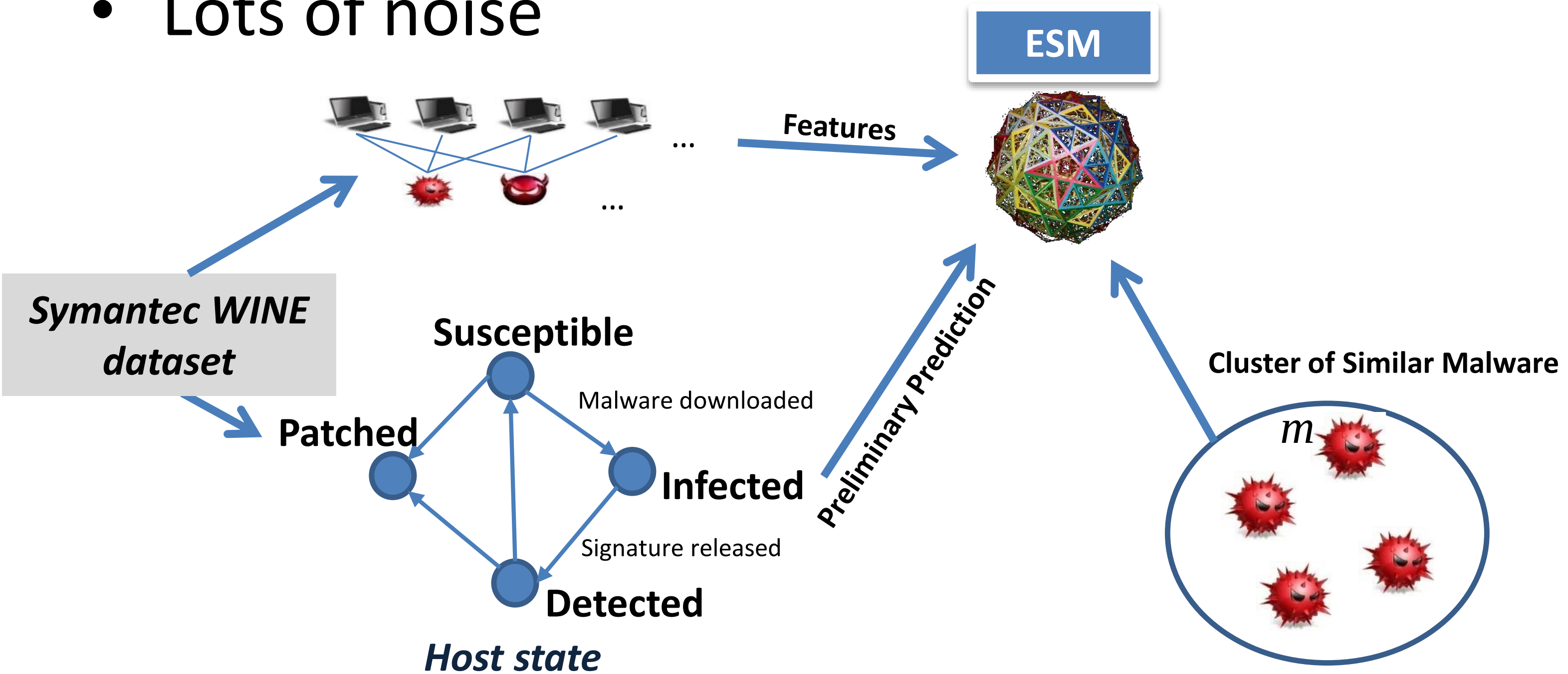
Combine Prediction Models

- 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

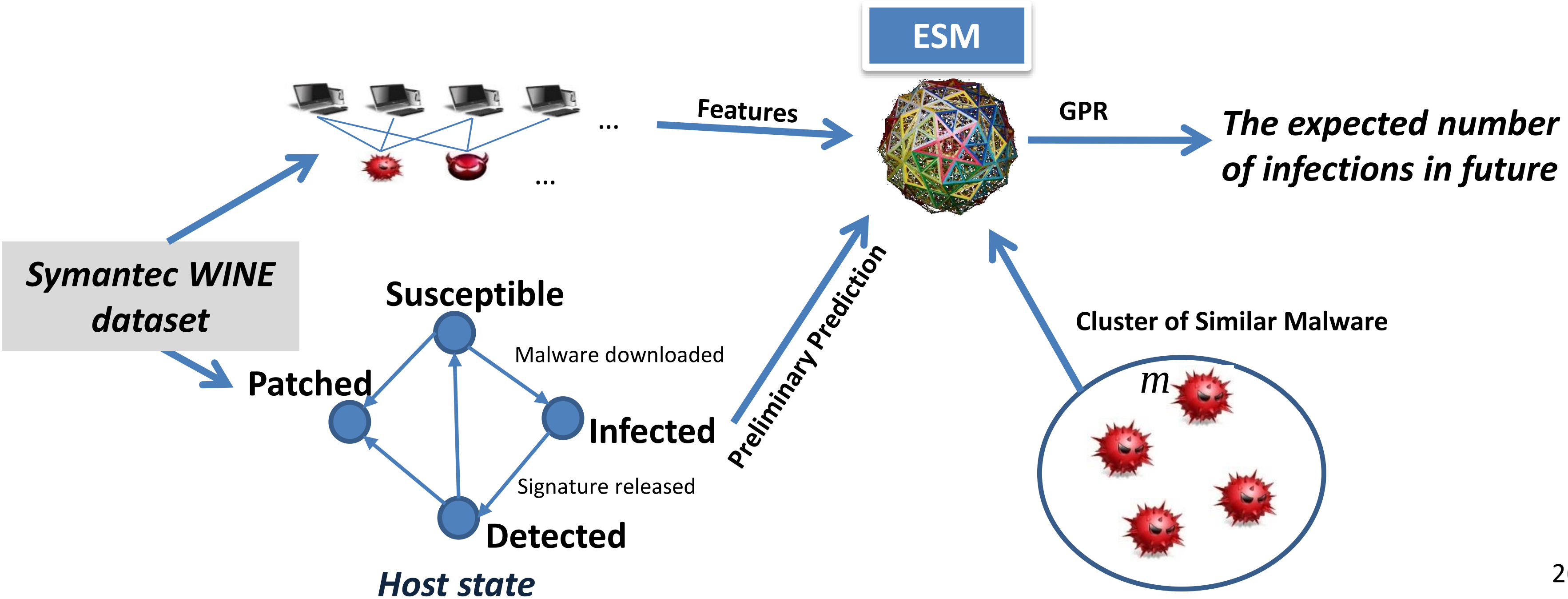
- Need a robust regression
- Gaussian Process Regression
  - Very strong Bayesian regression method
  - Less parametric (Parameters are calculated from data with maximum likelihood.)

Linear Regression	$\hat{y}(w, x) = w_0 + w_1 x_1 + ... + w_p x_p$
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

	Feature #1	Feature #2	DIPS output	……	Infected Host Ratio
80% Training (m)					
80% Training (m1)					
⋮					
20% Test (m)					



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)

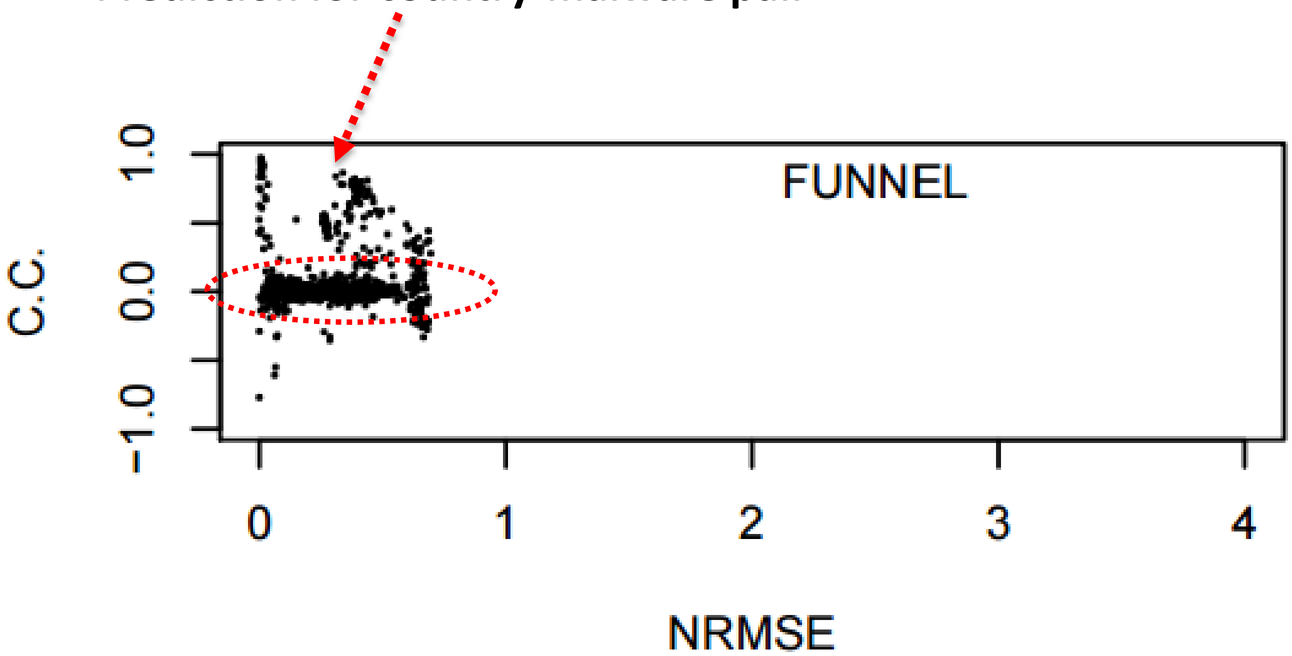
Measurements

- MAE*=|true infections - predicted infections|
- MSE = (true infection ratio - predicted infection ratio)^2
- RMSE = sqrt(MSE)
- NRMSE
- Pearson Correlation Coefficient

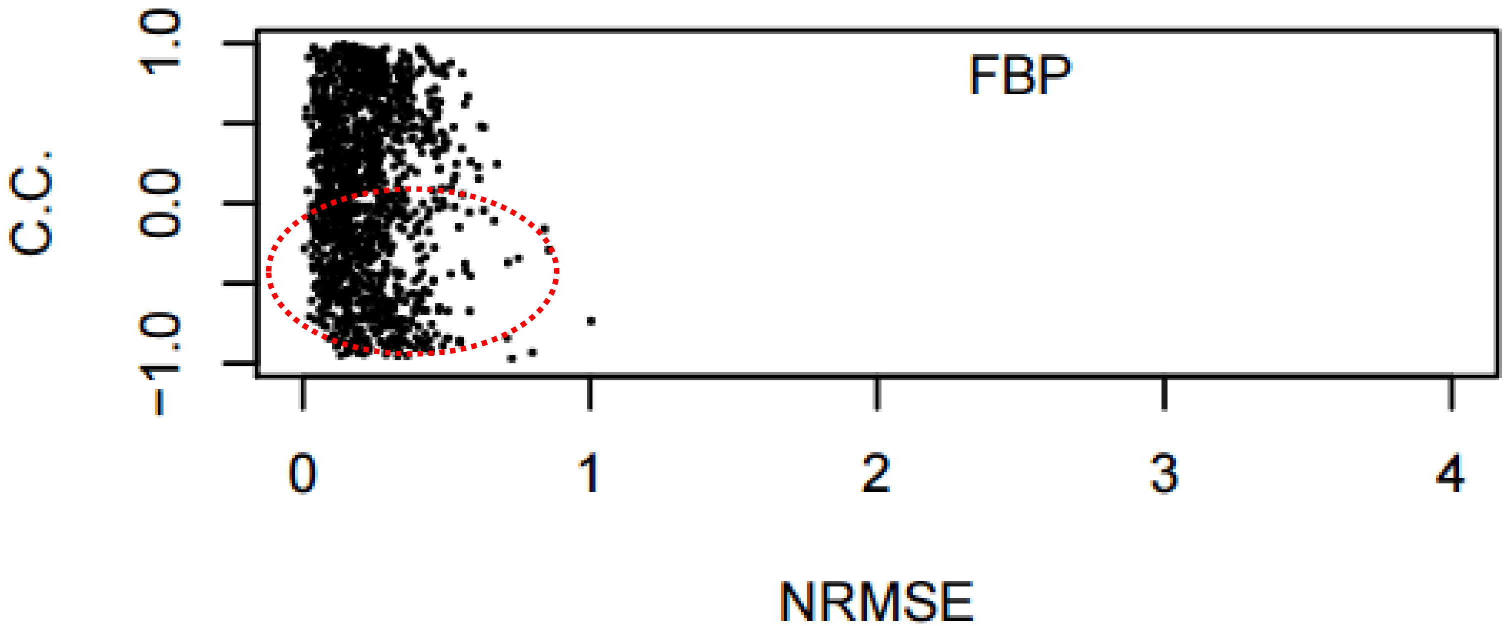
FUNNEL (prior art)

- State of the art epidemic model for human disease
- C.C. between truths and predictions are very bad.

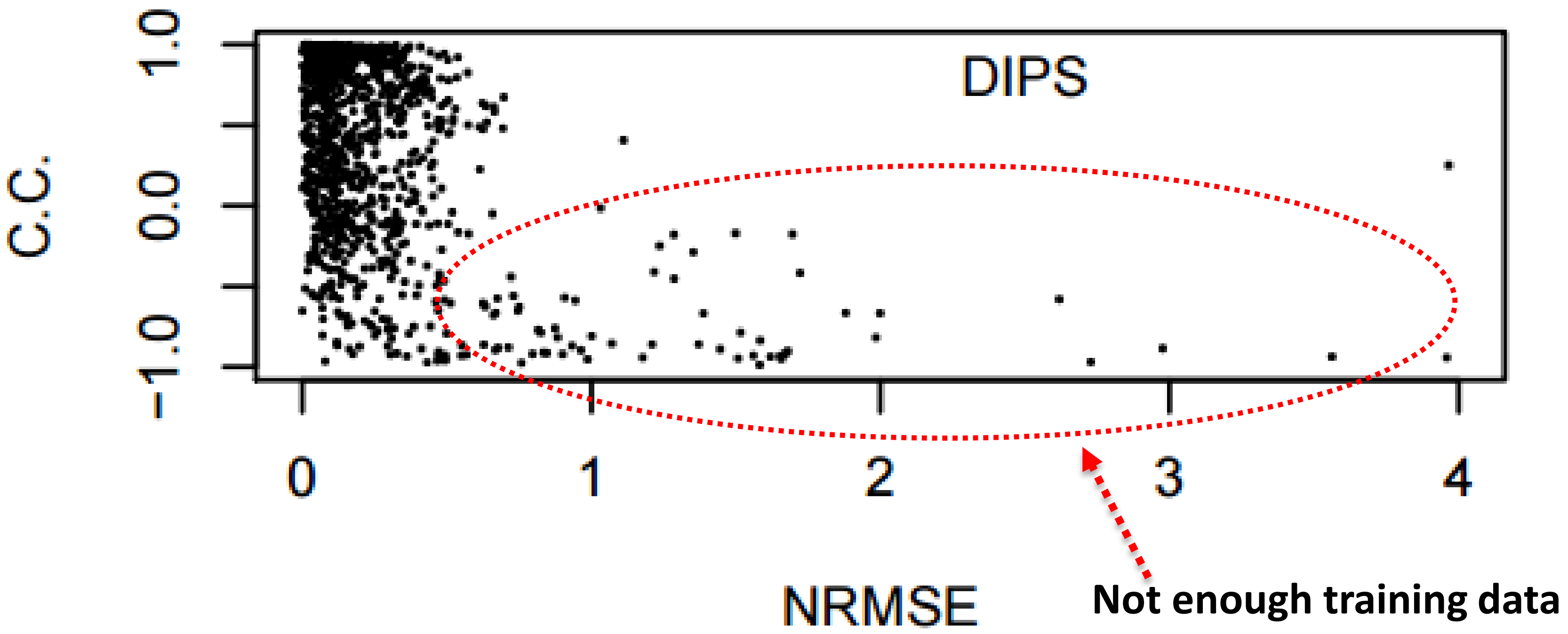
Prediction for country-malware pair



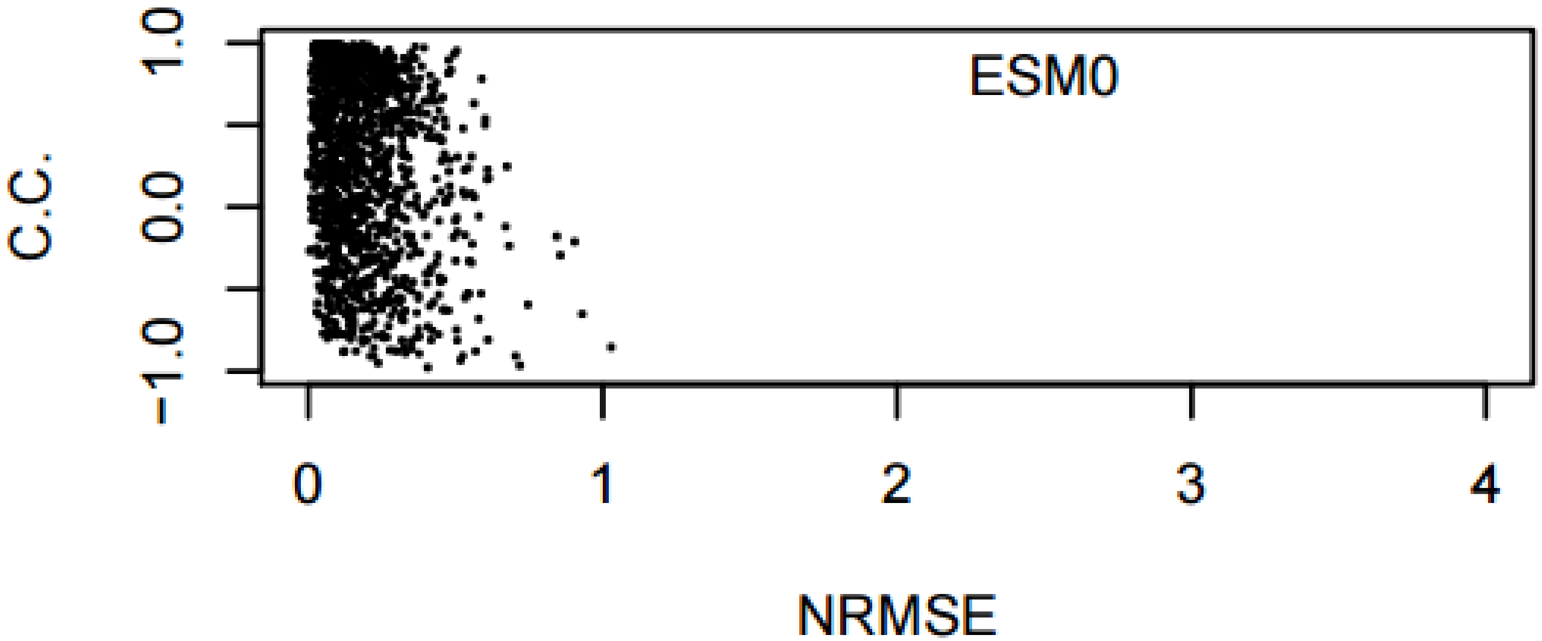
Feature Based Prediction



DIPS

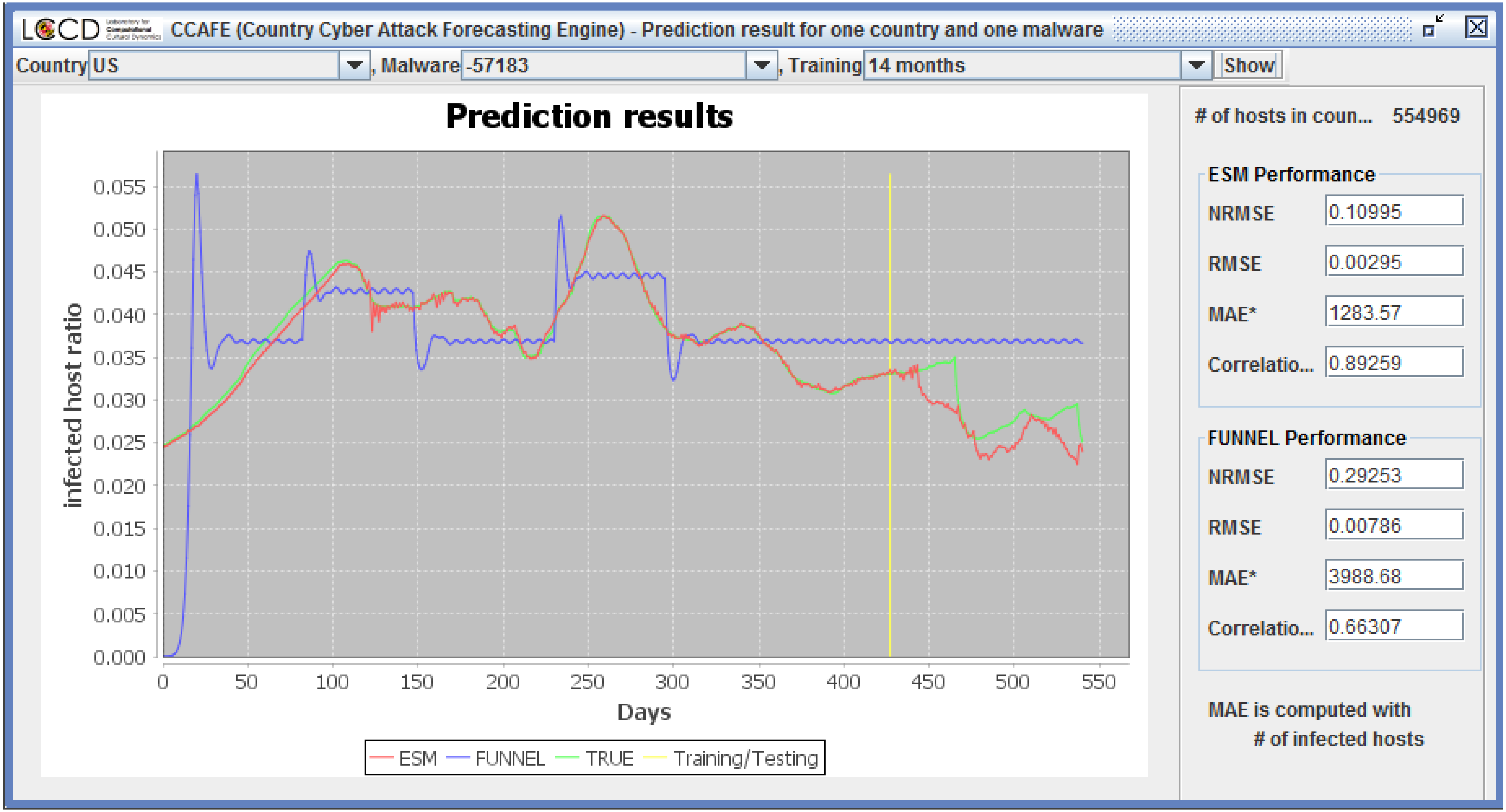


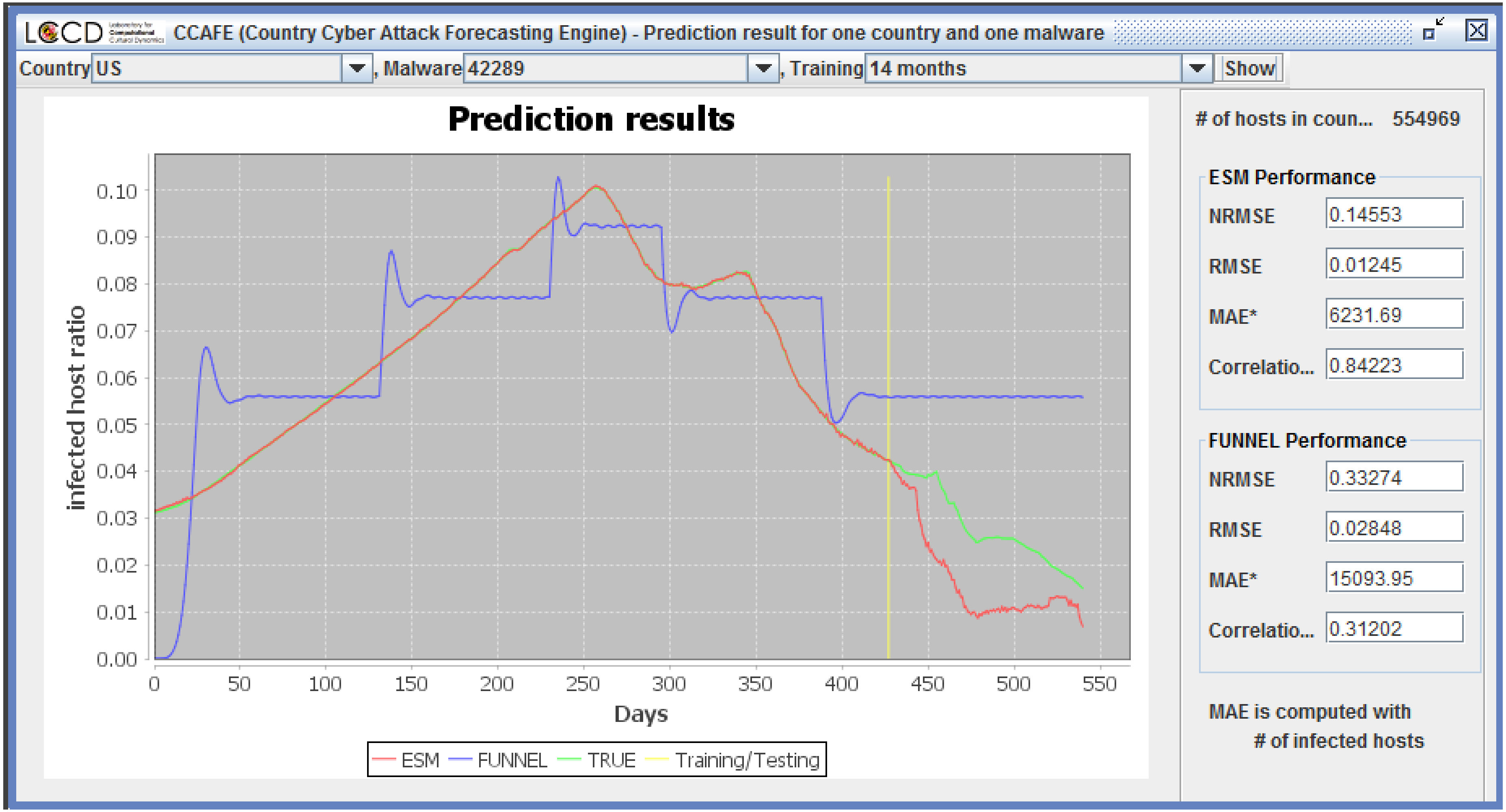
ESM0



Error Values

Model	MAE*	RMSE	NRMSE
FBP	73.74	0.00170	0.179
FUNNEL	127.83	0.00269	0.226
DIPS	32.36	0.00083	0.165
DIPS-EXP	36.56	0.00096	0.223
ESM_0	39.41	0.00115	0.150
ESM_1	41.84	0.00118	0.151
FBP^{+}_{Funnel}	79.01	0.00189	0.179


CCAFE (Country Cyber Attack Forecasting Engine) - Prediction result for one country and one malware
Country US
, Malware -57183
, Training 14 months
Show
Prediction results
infected host ratio
Days
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


CCAFE (Country Cyber Attack Forecasting Engine) - Prediction result for one country and one malware
Country US , Malware 42289 , Training 14 months Show
Prediction results
infected host ratio
Days
ESM FUNNEL TRUE Training/Testing
# of hosts in coun... 554969
ESM Performance
NRMSE 0.14553
RMSE 0.01245
MAE* 6231.69
Correlatio... 0.84223
FUNNEL Performance
NRMSE 0.33274
RMSE 0.02848
MAE* 15093.95
Correlatio... 0.31202
MAE is computed with
# of infected hosts

References

C. Kang, N. Park, B.A. Prakash, E. Serra, V.S. Subrahmanian. Ensemble Models for Data-Driven Prediction of Malware Infections, *Proc. 2016 ACM Web Science & Data Mining Conference (WSDM)*, Feb 2016.

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/