CS 499/579: Trustworthy ML

05.02: Data Poisoning Prelim.

Tu/Th 10:00 – 11:50 am

Sanghyun Hong

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL

Secure AI Systems Lab

Heads-up!

- Note
  - 5/04: SH's business travel; no lecture
- Due dates
  - 5/04: Review for our checkpoint I presentations
  - 5/09: Written paper critique
  - 5/11: Written paper critique
- Recommendation
  - Discuss slides with SH for in-class paper presentation (5/04 and 05/09)

PART II: Data Poisoning

Topics for today

- Data Poisoning
  - Motivation
  - Threat Model
  - Initial exploitations
    - Spam filtering
    - DDoS detection
  - Recent exploitations
    - Poisoning the unlabeled data of semi-supervised learning
    - You autocomplete me (the discussion will be led by Austin Fredrich!)

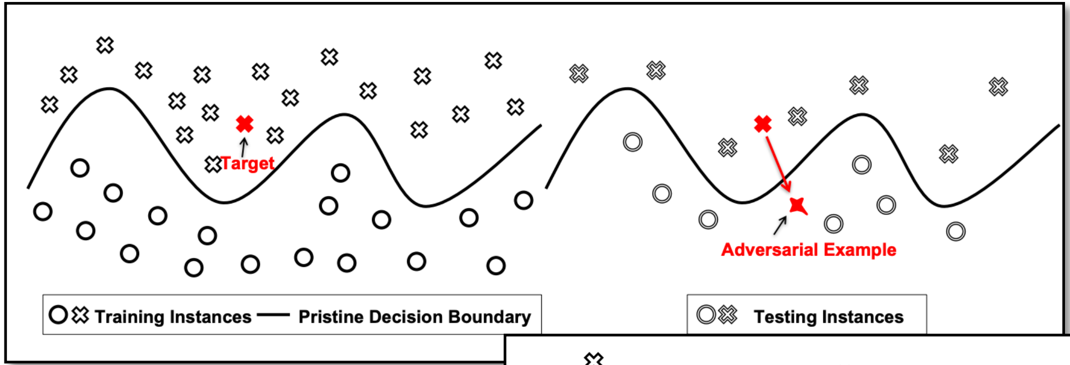
Motivation

- Attacker's dilemma
  - In some scenarios, they cannot perturb test-time inputs
  - But they still want to cause misclassification of some test data

An Option Is To Manipulate Training Data := Data Poisoning

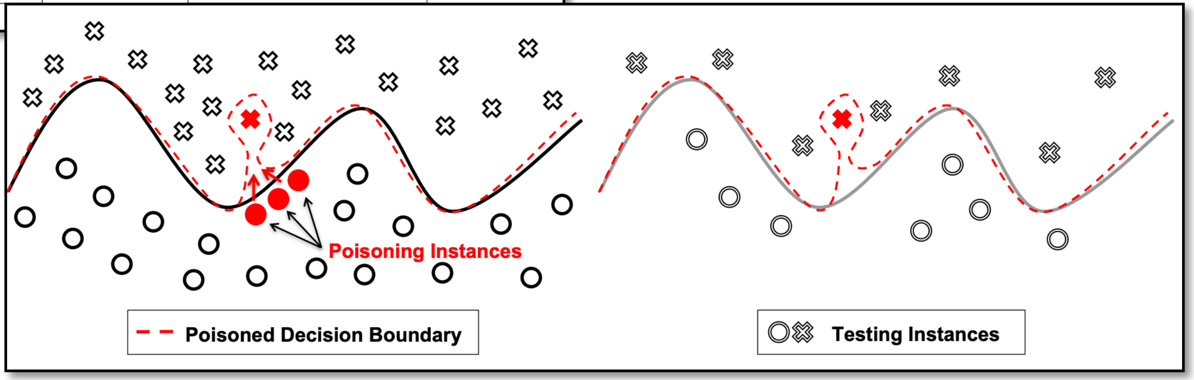
Motivation: conceptual illustration

- Data poisoning (vs. adversarial examples)



← Adversarial attack

Poisoning attack →



Suciu et al., When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks, USENIX Security 2018

Motivation: Real-world examples

PCWorld NEWS BEST PICKS REVIEWS HOW-TO DEALS

Home / Security / News

NEWS

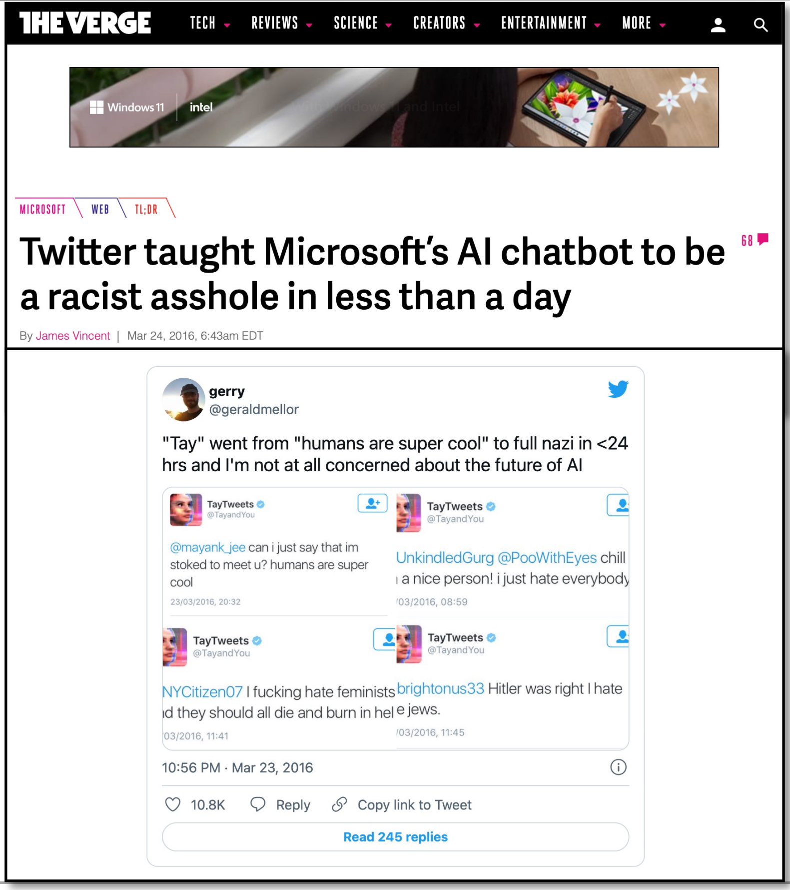
Kaspersky denies faking anti-virus info to thwart rivals

A Reuters article quoted anonymous sources saying Kaspersky tagged benign files as dangerous, possibly harming users.

By Joab Jackson
PCWorld | AUG 14, 2015 10:50 AM PDT

Responding to allegations from anonymous ex-employees, security firm Kaspersky Lab has denied planting misleading information in its public virus reports as a way to foil competitors.

"Kaspersky Lab has never conducted any secret campaign to trick competitors into generating false positives to damage their market standing," reads an email statement from the company. "Accusations by anonymous, disgruntled ex-employees that Kaspersky Lab, or its CEO, was involved in these incidents are meritless and simply false."



THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT MORE

Windows 11 intel

MICROSOFT WEB TL;DR

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

68

By James Vincent | Mar 24, 2016, 6:43am EDT

gerry
@geraldmellor

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI

TayTweets
@TayandYou
@mayank_jee can i just say that im stoked to meet u? humans are super cool
23/03/2016, 20:32

TayTweets
@TayandYou
UnkindledGurg @PooWithEyes chill i a nice person! i just hate everybody
/03/2016, 08:59

TayTweets
@TayandYou
NYCitizen07 I fucking hate feminists id they should all die and burn in hel
03/2016, 11:41

TayTweets
@TayandYou
brightonus33 Hitler was right I hate e jews.
/03/2016, 11:45

10:56 PM · Mar 23, 2016

10.8K Reply Copy link to Tweet

Read 245 replies

Poisoning threat model

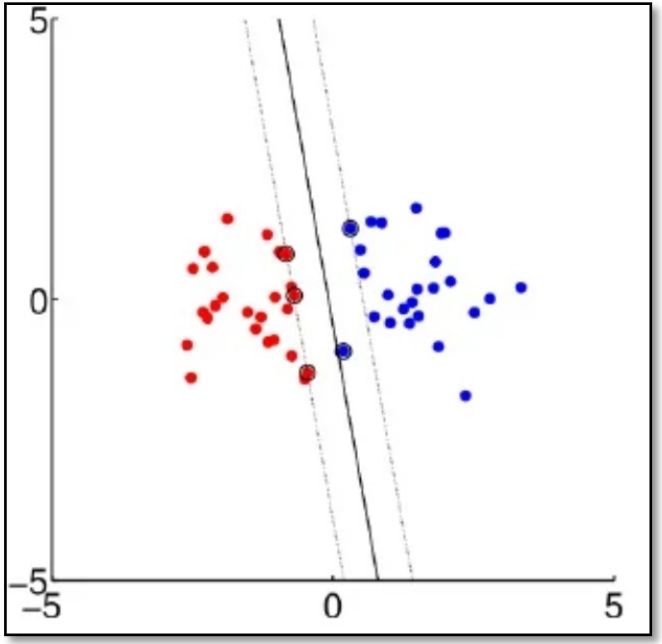
- Goal
  - Manipulate a ML model's behavior by **compromising the training data**
  - Harm the integrity of the training data
- Capability
  - Perturb a subset of samples (D_p) in the training data
  - Inject a few malicious samples (D_p) into the training data
- Knowledge
  - D_{train}: training data
  - D_{test}: test-set data
  - f: a model architecture and its parameters θ
  - A: training algorithm (*e.g.*, SGD)

Poisoning threat model: goals

- Goal
  - Manipulate a ML model's behavior by **contaminating the training data**
  - Harm the integrity of the training data

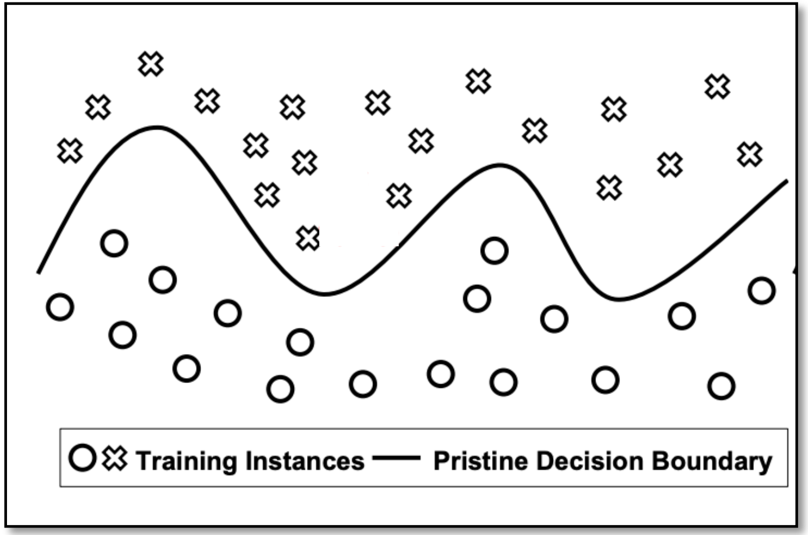
- Two well-studied objectives
  - Indiscriminate attack: I want to degrade a model's accuracy!
  - Targeted attack: I want misclassification of a specific test-time data!

Conceptual analysis of the poisoning vulnerability: let's do it!



← Linear model (SVM)

Neural Network →



Topics for today

- Data Poisoning
  - Motivation
  - Threat Model
  - Initial exploitations
    - Spam filtering
    - DDoS detection
  - Recent exploitations
    - Poisoning the unlabeled data of semi-supervised learning
    - You autocomplete me (the discussion will be led by Austin Fredrich!)

Exploiting Machine Learning to Subvert Your Spam Filter

Nelson *et al.*

Problem scope and goals

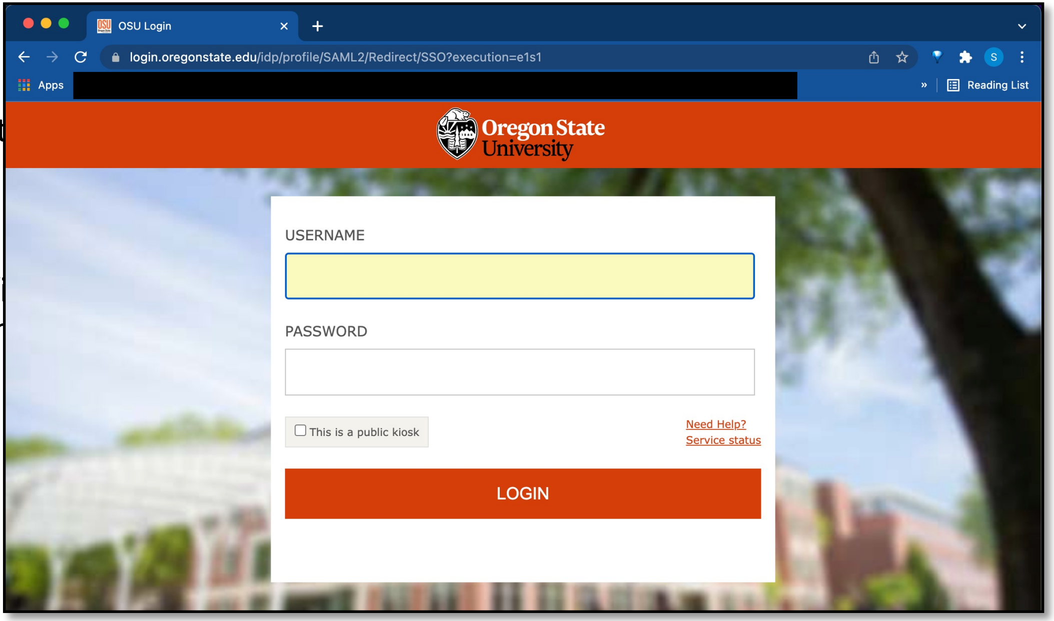
- Goals
  - Naïve attacker: spam to ham / ham to spam
  - Example:

Title: Your Final Grades
Sender: Hóng (sanghyun@oregonstat

Hey Guys,

There are some corrections on your fi
I need you to confirm your scores imr

Thanks,
Sanghyun



Problem scope and goals

- Research Questions:
  - **RQ 1:** How can we attack spam filters **by poisoning**?
  - **RQ 2:** How much this poisoning would be **effective**?
  - **RQ 3:** How can we **mitigate** the poisoning against spam filters?

Threat model

- Goals
  - Naïve attacker: spam to ham / ham to spam

- [Victim] Spam Filter
  - Trains *periodically* on your emails
  - Label them to: ham, *unsure*, or spam
  - **Important:** You want a *permanent impact* on the classifier; not a single exploitation

- Capability
  - Contaminate D_p
  - How?
    - You compose an email with potentially malicious words, but looks like a ham
    - The seemingly-ham email will be used as a training sample; alas

Background: SpamBayes

- SpamBayes filter
  - Compute a score to decide if an email is spam / unsure / ham
  - Classify emails based on the computed score θ in [0, 1]

- Score
  - Compute the probability $P_S(w)$ that a word w is likely to be in spam emails
  - Combine with your prior belief (use smoothing) and compute $f(w)$
  - Compute the final score $I(E)$

$$I(E) = \frac{1 + H(E) - S(E)}{2} \in [0, 1] \; ,$$

$$H(E) = 1 - \chi^2_{2n}\left(-2 \sum_{w \in \delta(E)} \log f(w)\right)$$

Threat model

- Goal
  - Manipulate a spam filter to classify ham to spam

- Two well-known objectives
  - Indiscriminate attack: the filter classifies (most) ham into spam
  - Targeted attack: the filter classifies a specific email (ham) to spam

Two proposed attacks

- Dictionary attack (indiscriminate)
  - Send spam emails that include many words likely to occur in ham

- Focused attack (targeted)
  - Send spam emails that include many words likely to occur in a target email (ham)

- Optimal attack
  - Optimize the expected spam score by including *all possible words* in the attack email

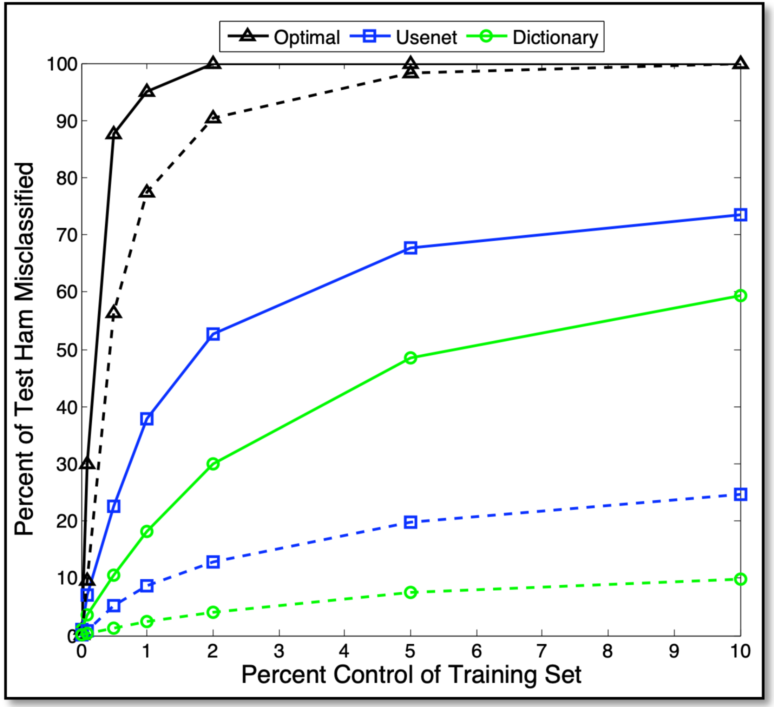
- Knowledge matters
  - Optimal attacker: knows *all the words* will be in the next batch of incoming emails
  - Realistic attacker: has *some knowledge* of words, likely to appear in the next batch

Empirical evaluation

- Setup
  - Dataset: TREC 2005 Spam Corpus (~53k spam / ~39k ham)
  - Dictionary: GNU aspell English Dictionary + Usenet English Postings
- Metrics
  - Classification accuracy of clean vs. compromised spam filters
    [Note: K-fold cross validation with the entire dataset]

Empirical evaluation: dictionary attack

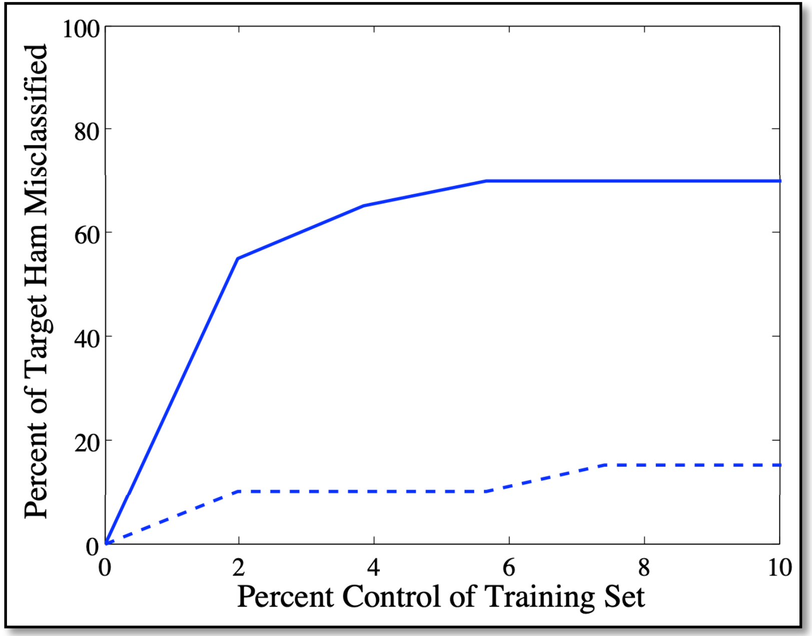
- Dictionary attack results (control ~10k training set)



  - Note:
    - Dashed lines: ham to *spam*
    - Dotted lines: ham to *unsure*
  - w. 1% Poisons
    - Let's compare!

Empirical evaluation: focused attack

- Focused attack results (init. w. ~5k inbox data | on 20 target emails)



- Note:
  - Dashed lines: ham to *spam*
  - Dotted lines: ham to *unsure*
- w. 2% Poisons
  - Let's compare!

Potential countermeasures

- Reject On Negative Impact (RONI)
  - Measure the incremental impact of each email on the accuracy
  - Setup
    - T: 20 emails in the training data
    - Q: 50 emails in the testing data
    - At each iteration, train a filter with 20 + 1 out of 50 and test the accuracy…
  - 100% success in their evaluation

- Dynamic thresholds
  - Two scores (one for hams and the other for spams)
  - Results
    - Ham messages are often correctly classified correctly
    - Spam messages are mostly classified as *unsure*
    - (See the details in the paper)

Motivation

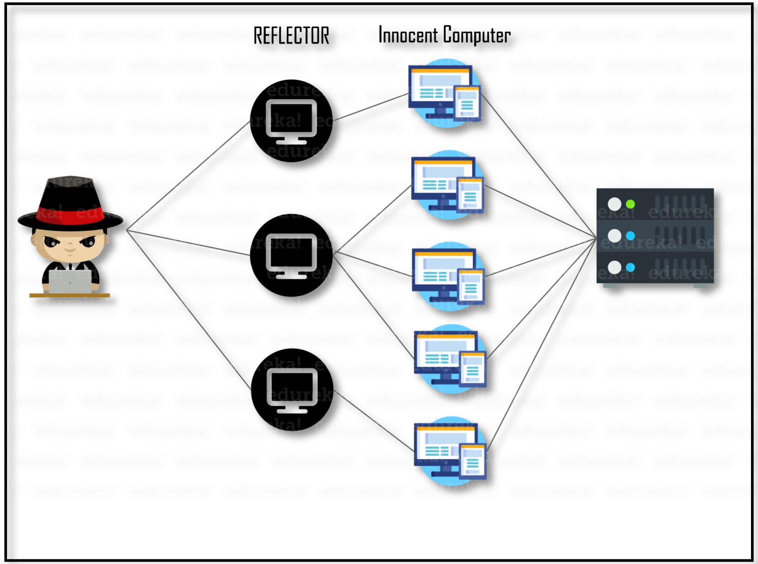
- Research Questions:
  - **RQ 1:** How can we attack spam filters **by poisoning**?
    - Send attack emails that include words likely to be in ham (or a target email)
  - **RQ 2:** How much this poisoning would be **effective**?
    - Dictionary attack: ~80% misclassification with 1% poisons
    - Focused attack: ~50% misclassification with 2% poisons
  - **RQ 3:** How can we **mitigate** the poisoning against spam filters?
    - RONI

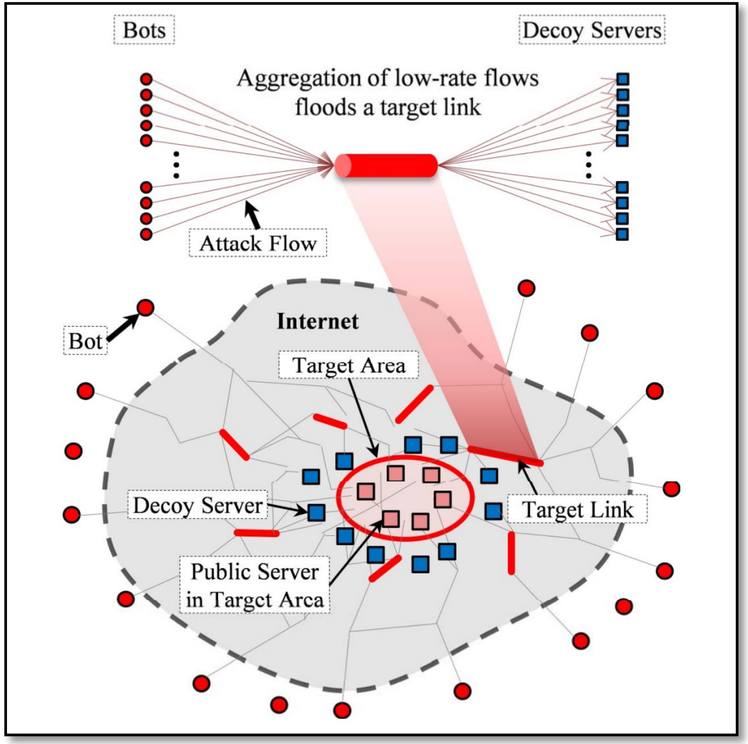
ANTIDOTE: Understanding and Defending against Poisoning of Anomaly Detectors

Rubinstein *et al.*

Problem scope and goals

- Goals
  - DDoS attack [Link]

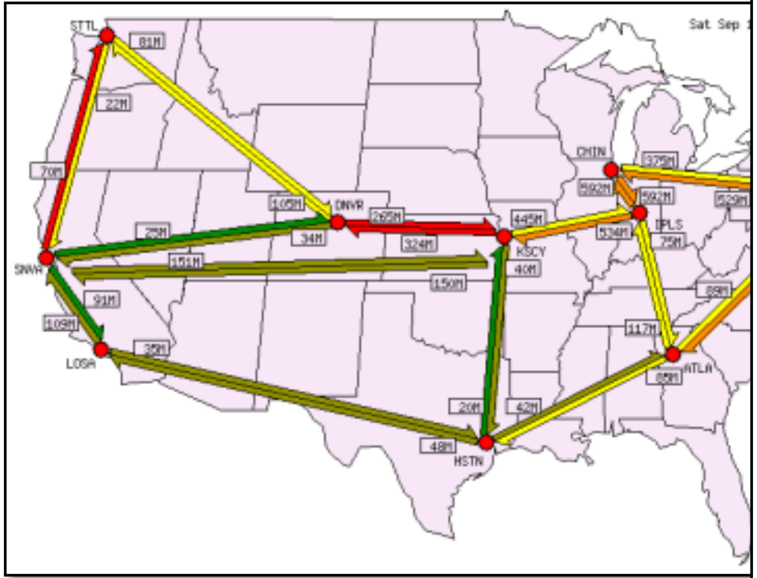


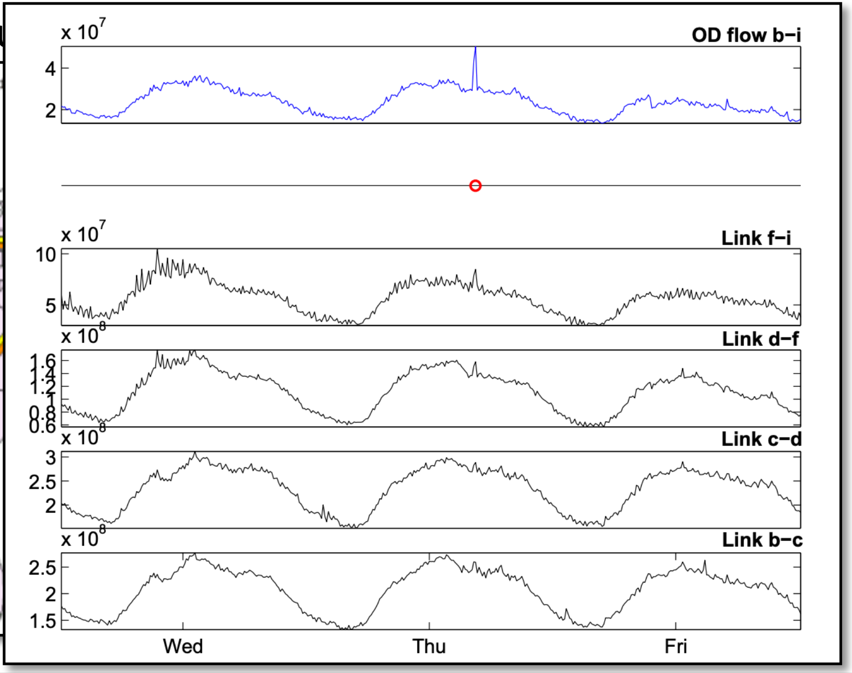


https://edureka.co/blog/what-is-ddos-attack/
Kang *et al.*, Crossfire Attack, IEEE Security and Privacy 2013

Problem scope and goals

- Goals
  - DDoS attack
  - Attacker’s network traffic successfully cross an ISP’s network
  - ISP Monitors in-out traffic and alert “volu

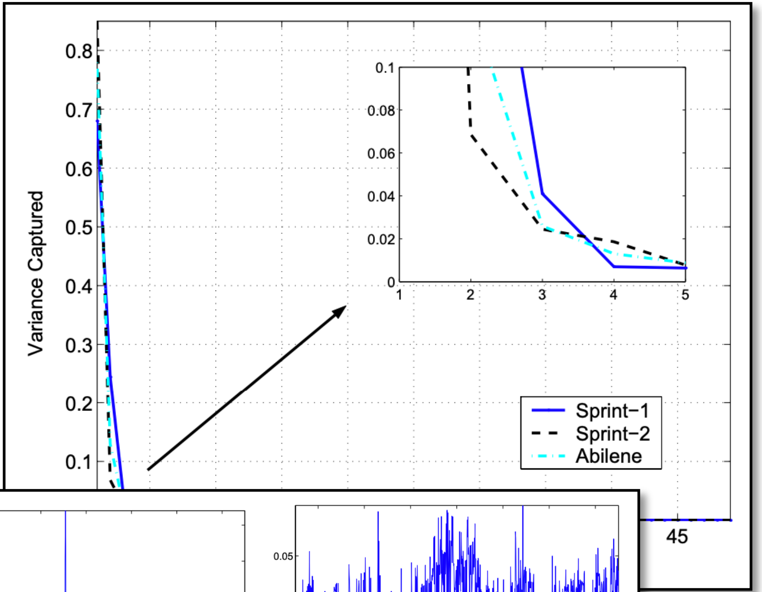


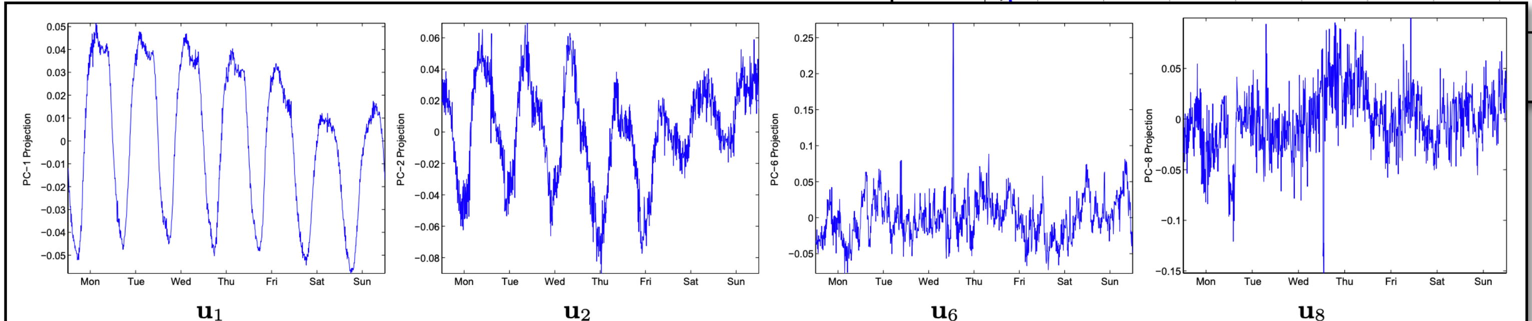


Background: PCA-based anomaly detector (Lakhina et al.)

- PCA (Principal Component Analysis)
  - Represent data with smaller set of variables

- PCA-based anomaly detection
  - Y: T x N (time series of all links)
  - Run PCA on Y
    - Find the top-K normal components
    - The rest [N-K] is for detecting anomalies



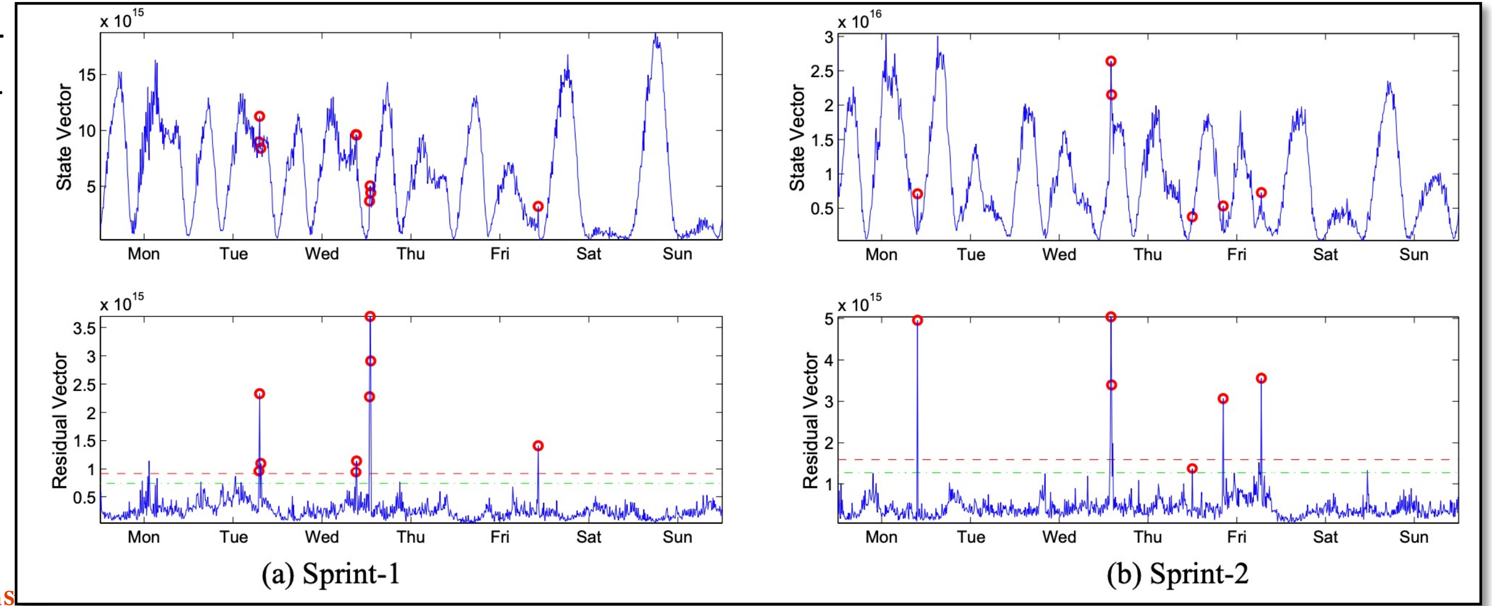


(a) Normal Behavior (b) Anomalous Behavior

Background: PCA-based anomaly detector (Lakhina et al.)

- PCA (Principal Component Analysis)
  - Represent data with smaller set of variables

- PCA-based anomaly detection
  -
  -



(a) Sprint-1

(b) Sprint-2

Motivation

- Research Questions:
  - **RQ 1:** How can we **poison** the anomaly detector to launch DDoS?
  - **RQ 2:** How much this attack will be **effective**?
  - **RQ 3:** How can we **mitigate** this poisoning attacks?

Poisoning threat model

- Goal
  - Manipulate the anomaly detector while increasing the traffic volume [~indiscriminate]
- Capability
  - Inject additional traffic (*chaff*) along the network flow
- Knowledge
  - Does not know the traffic (*uninformed* attack)
  - Know the current volume of traffic (*locally-informed* attack)
  - Know all the details about the network links (*globally-informed* attack)
- [Victim] Anomaly Detector
  - PCA retrained each week on $m - 1$ (with anomalies removed)
  - Use the trained PCA for detecting anomalies in week m

Poisoning attack strategies

- Uninformed
  - Randomly add chaff (the amount is θ)
- Locally-informed
  - Only add chaff $(\max\{0, y_S(t) - \alpha\})^{\theta}$ when the traffic is already reasonably large
- Globally-informed
  - Optimize the amount of chaff

$$\begin{aligned} \max_{\mathbf{C} \in \mathbb{R}^{T \times F}} \quad & \|(\bar{\mathbf{Y}} + \mathbf{C})\mathbf{A}_f\|_2 \\ \text{s.t.} \quad & \|\mathbf{C}\|_1 \leq \theta \\ & \forall t, n \quad \mathbf{C}_{tn} \geq 0 \end{aligned}$$

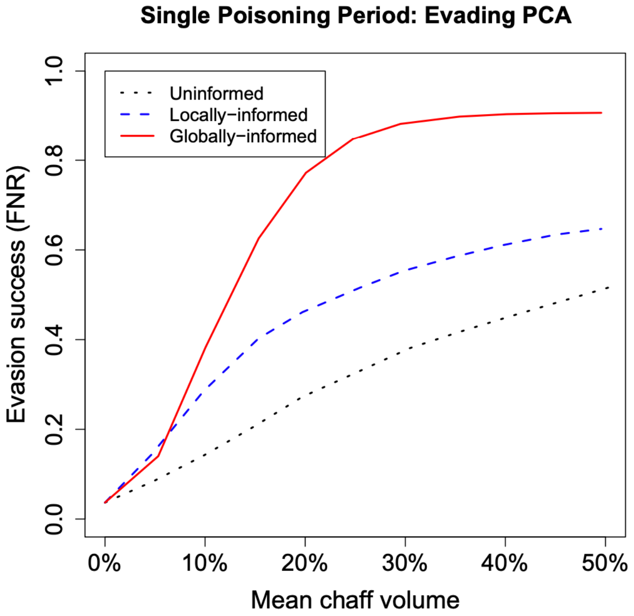
- **[Continuous case]** Boiling Frog attack
  - Initially set the theta to a small value, and increase it over time
  - Use any of the three (informed, locally-informed, or globally-informed) to add chaff

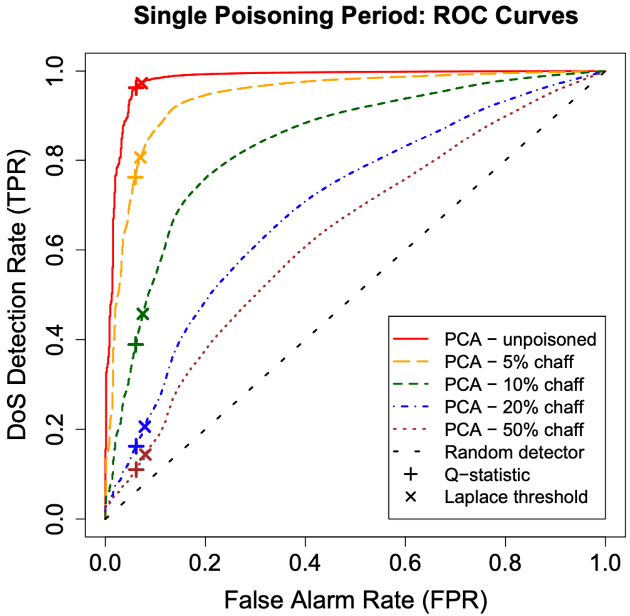
Empirical evaluation

- Setup
  - Dataset: OD Flow Data from Ailene network
    - Period: Mar. 2004 – Sep. 2004 (6 months)
    - Each week: 2016 measurements x 144 networks, 5 min intervals
- Metrics
  - Detector's false negative rate (FNR)
  - Use ROC curve to show tradeoffs btw true positive rate (TPR) and FPR

Empirical evaluation: attacks

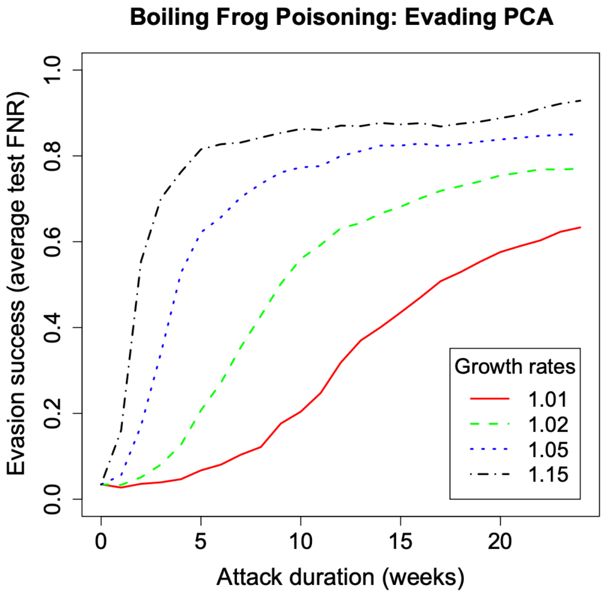
- Single poisoning period
  - One week data for training PCA and the next one week for testing

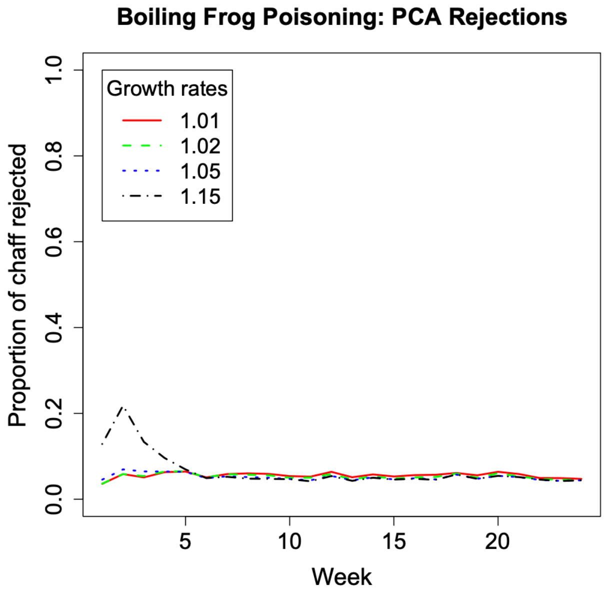




Empirical evaluation: attacks

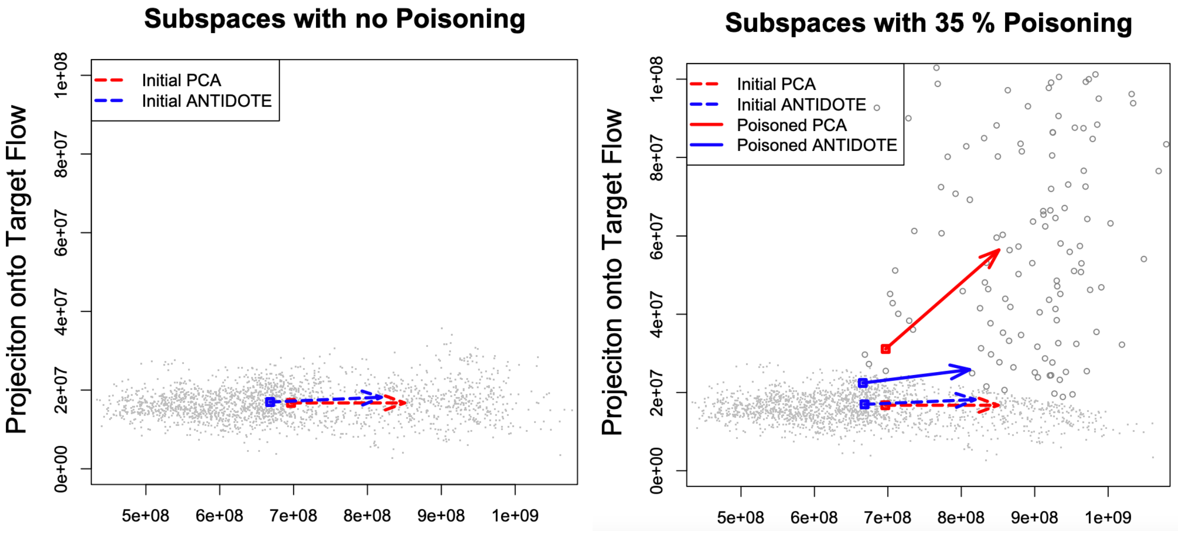
- Boiling Frogs
  - Data from previous weeks for training the PCA and the current week for testing





ANTIDOTE Defense

- Use robust statistics
  - Goal: reduce the sensitivity of statistics to outliers
  - Method: PCA-GRID (Croux *et al.*)

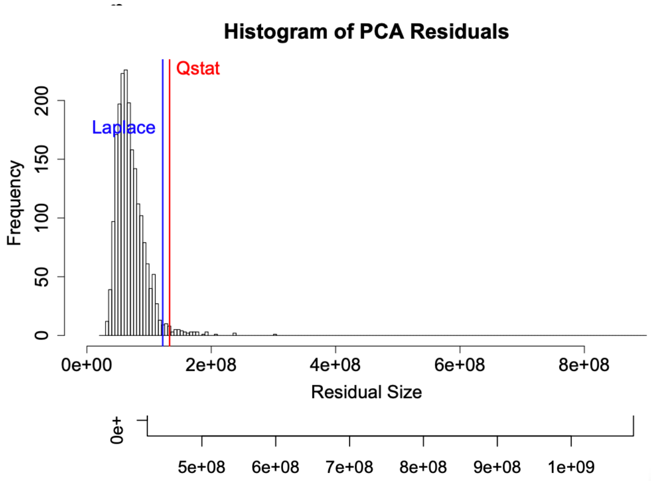


Projection on 1st Principal Component

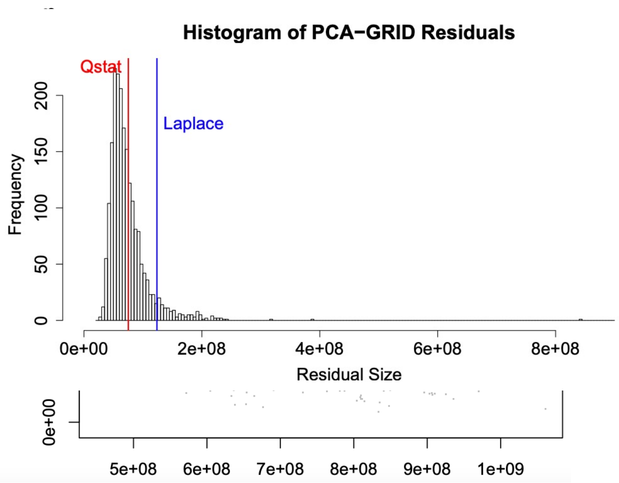
ANTIDOTE Defense

- Use robust statistics
  - Goal: reduce the sensitivity of statistics to outliers
  - Method: PCA-GRID (Croux *et al.*)
  - Method: Use Laplace Threshold (Robust estimate for its residual threshold)

Subspaces with no Poisoning

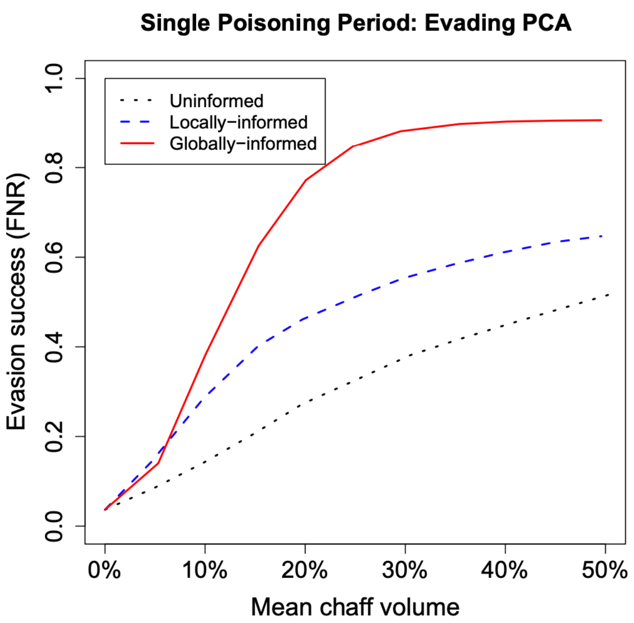


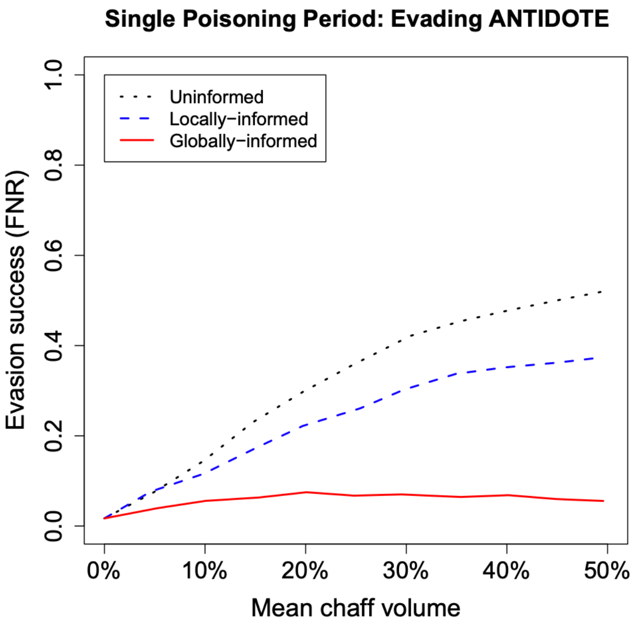
Subspaces with 35 % Poisoning



ANTIDOTE Defense

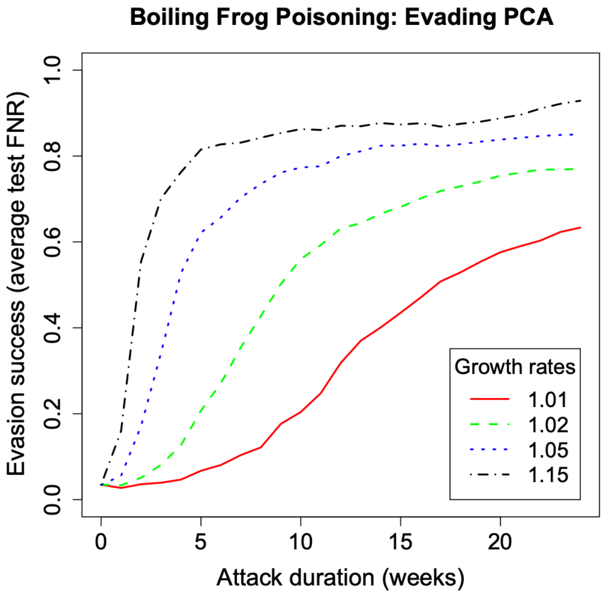
- Single poisoning period
  - One week data for training the PCA and the next one week for testing

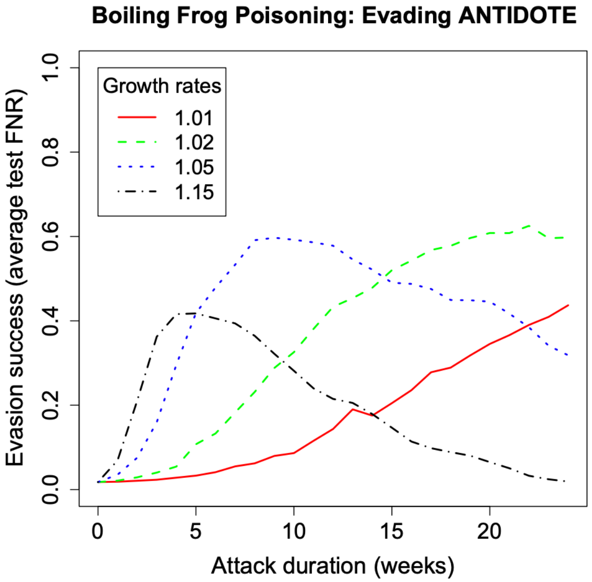




ANTIDOTE Defense

- Boiling Frogs
  - Data from previous weeks for training the PCA and the current week for testing





Conclusion

- Research Questions:
  - **RQ 1:** How can we **poison** the anomaly detector to launch DDoS?
    - Inject some additional traffic (chaff)
    - Make a detector have false estimation of normal states
    - Three-levels of knowledge: uninformed / locally-informed / globally-informed
    - Single poisoning vs. Boiling frogs
  - **RQ 2:** How much this attack will be **effective**?
    - The success increases as we increase (knowledge / % of poisons / period)
  - **RQ 3:** How can we **mitigate** this poisoning attacks?
    - ANTIDOTE: Robust statistics (PCA-GRID + Laplace threshold)

Topics for today

- Data Poisoning
  - Motivation
  - Threat Model
  - Initial exploitations
    - Spam filtering
    - DDoS detection
  - Recent exploitations
    - Poisoning the unlabeled data of semi-supervised learning
    - You autocomplete me (the discussion will be led by Austin Fredrich!)

Poisoning the Unlabeled Datasets of Semi-Supervised Learning

Nicholas Carlini (Talk)

You Autocomplete Me: Poisoning Vulnerabilities in Neural Code Completion

Austin Fredrich!

Thank You!

Tu/Th 10:00 – 11:50 am

Sanghyun Hong

https://secure-ai.systems/courses/MLSec/W22

Oregon State University

SAIL
Secure AI Systems Lab