SAND2013-6752C

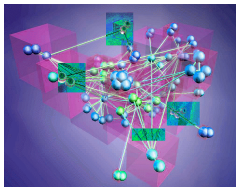
Increasing Cyber Resilience via Predictability-Based Defense

Rich Colbaugh **Kristin Glass** **Dave Zage**

Sandia National Laboratories

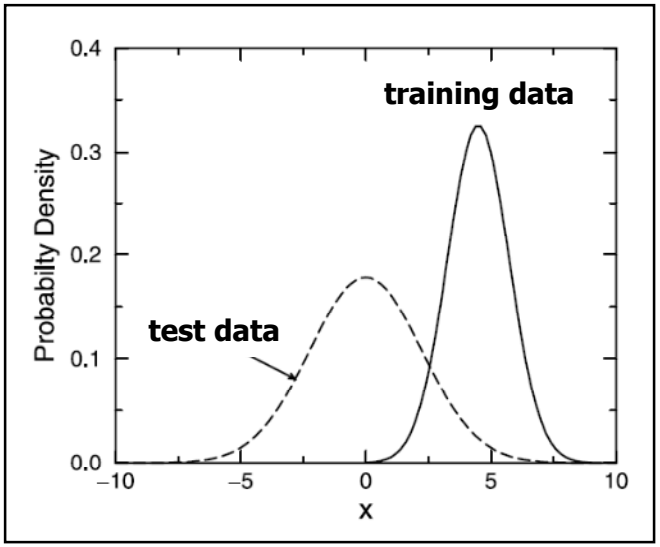
August 2013

Introduction

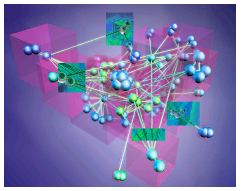


Background

- *Cyber system resilience* – the ability of systems to anticipate/withstand/recover from attacks and failures – is a key objective in cyber security.
- Machine learning (ML) plays a central role in cyber security, but existing ML methods are not resilient to attacks by *adaptive adversaries.*
- In particular, a standard assumption in ML – that training and test (future) data are identically distributed – is violated in adversarial settings. In security domains the test data are generated, in part, by adversaries whose goals conflict with defense and who therefore attempt to adapt the test data to reduce ML effectiveness.



Introduction



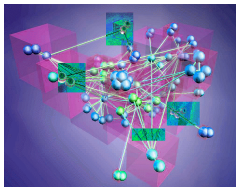
Objective

Increase cyber system resilience via *predictability-based defense design,* for instance as a means of 1.) increasing predictive power of defenses, 2.) reducing ability of adversaries to anticipate defense actions.

Outline

- Problem formulation.
- Predictive defense:
  - game model + ML approach;
  - sample results (Spam, malware).
- Moving target defense:
  - hybrid dynamical system approach;
  - sample results (Spam, malware).

Problem Formulation



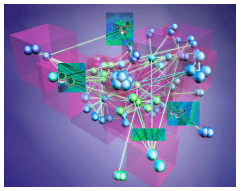
Problem

Task of interest is *behavior classification,* in which innocent and malicious activities are to be distinguished.

Formulation

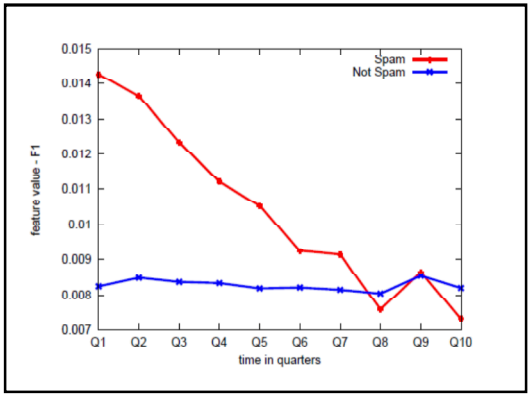
- Defense adopts ML approach, modeling each instance of activity as a vector **x** in feature space (with features $F = \{f_1, \ldots, f_m\}$), and learning $\mathbf{w} \in \Re^{|F|}$ such that the classifier $orient = sign(\mathbf{w}^T\mathbf{x})$ accurately predicts the nature of **x** ($+1 \rightarrow$ malicious, $-1 \rightarrow$ innocent).
- Assume predictability-oriented goals:
  - *predictive defense* – learn classifier which predicts adversary adaptation in order to counter current and (near) future attacks;
  - *moving target defense* – learn dynamic classifier which is difficult for adversary to "reverse-engineer".

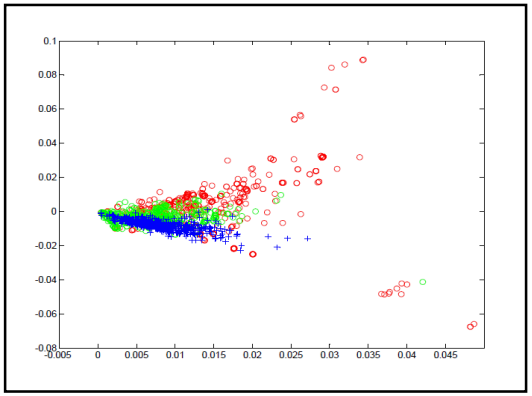
Predictive Defense

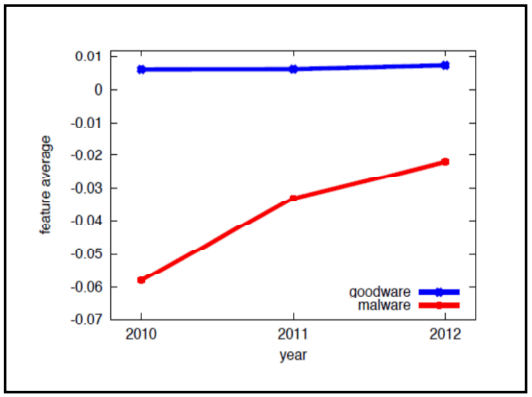


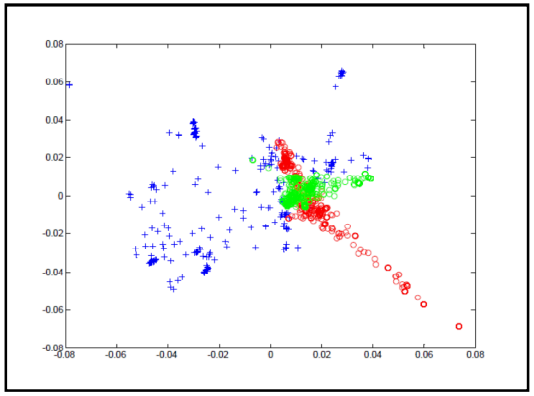
Preliminaries

Analysis of Spam, phishing, and malware datasets indicates adversary adaptation is often "sensible" and regular.





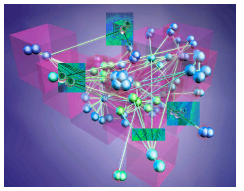




Sensibility Test

We studied the five most "significant" features in six different longitudinal datasets (Spam1 and 2, phishing 1 and 2, malware 1 and 2), yielding 30 observations of adversary adaptation. In 28 of the 30 cases (93%), the observed adaptation is "rational": the change in feature value for the malicious instances is toward the benign instance feature value.

Predictive Defense



Game model + machine learning (GM+ML) approach

Two challenges and proposed approaches

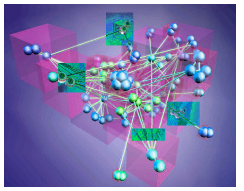
- C1: Difficult to develop realistic models for innocent/malicious activity "from scratch" [e.g., Roy et al. 2010].

  A1: Model the way an ecology of attackers *transforms* data, rather than the way a set of attackers generates data.

- C2: Space of possible attacker actions is very large, so that realistic GM are typically intractable [e.g., Laskov/Lippmann 2010, Sandholm 2011].

  A2: Incorporate GM directly in reduced ML feature space, yielding an aggressive yet lossless information abstraction.

Predictive Defense



GM+ML approach (cont'd)

- Standard linear classifier:

$$\min_{w} \left[\beta \|w\|^2 + \sum_i \text{loss}\left(y_i, w^T x_i\right) \right]$$

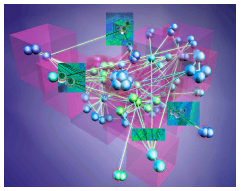
$$\text{class} = \text{sign}(x^T w^*).$$

- Sequential game-informed classifier:

$$\min_{w} \max_{a} \left[-\alpha \|a\|^2 + \beta \|w\|^2 + \sum_i \text{loss}\left(y_i, w^T (x_i + a)\right) \right]$$

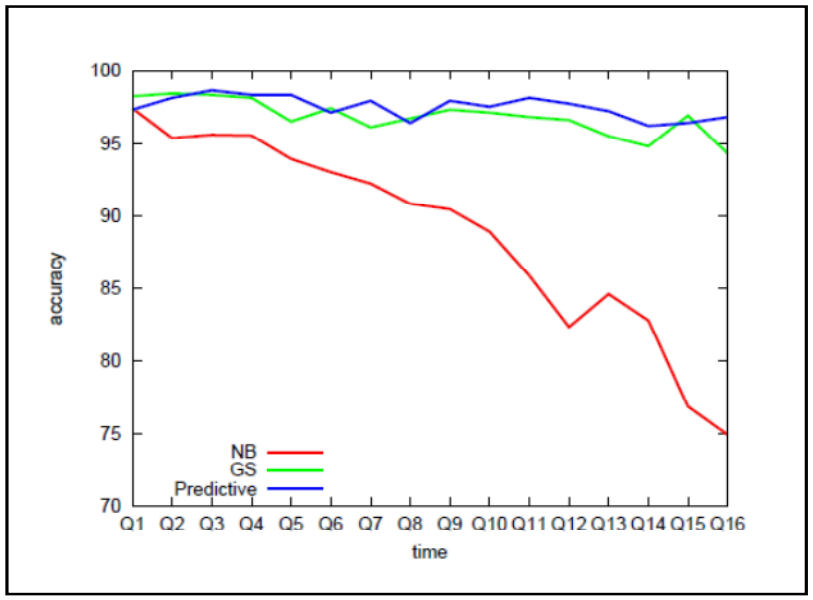
$$\text{class} = \text{sign}(x^T w^*).$$

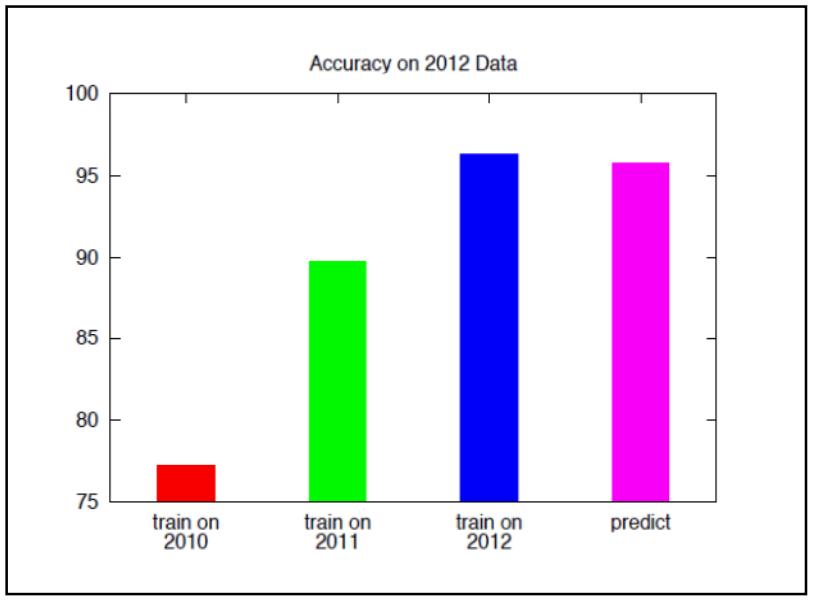
Predictive Defense



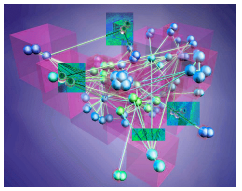
Sample results

GM+ML predictive learning for Spam filtering (left) and malware detection (right).



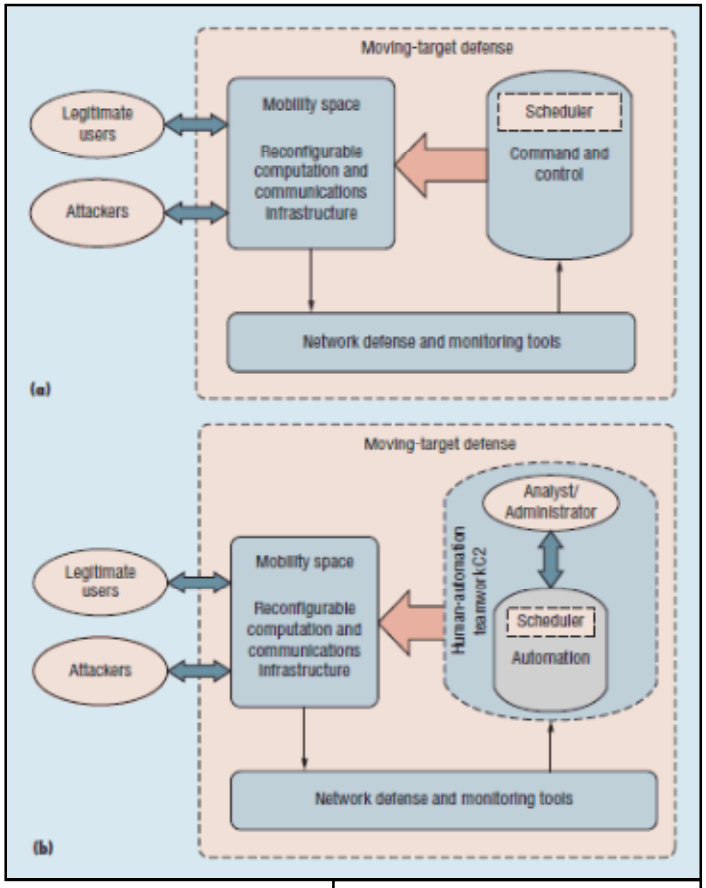


MT Defense



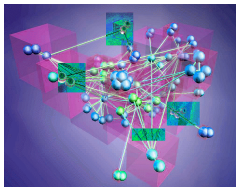
Preliminaries

- Goal: develop learners which make it difficult for attackers to transform test data in an informed way.
- Recently the "moving target" (MT) concept has been proposed as a way to increase the difficulty of adversary's reverse-engineering task [NSTC 2011, MTR 2012], and we adopt this approach here.



Carvalho et al. 2012

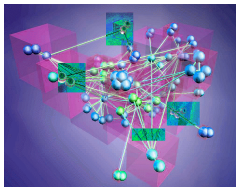
MT Defense



Basic idea

- The proposed approach to MT learning consists of three steps:
  - randomly divide the feature set F into subsets $\{F_1, \ldots, F_K\}$;
  - train one defense system for each F_i, yielding $W = \{\mathbf{w}_1, \ldots, \mathbf{w}_K\}$;
  - appropriately schedule which defense $\mathbf{w}_i$ is "active" at any time.
- Key question: how should we schedule the defenses?
- Remark: the scheduling problem can be subtle, as revealed for instance in the study of repeated incomplete information games [Sandholm 2011].

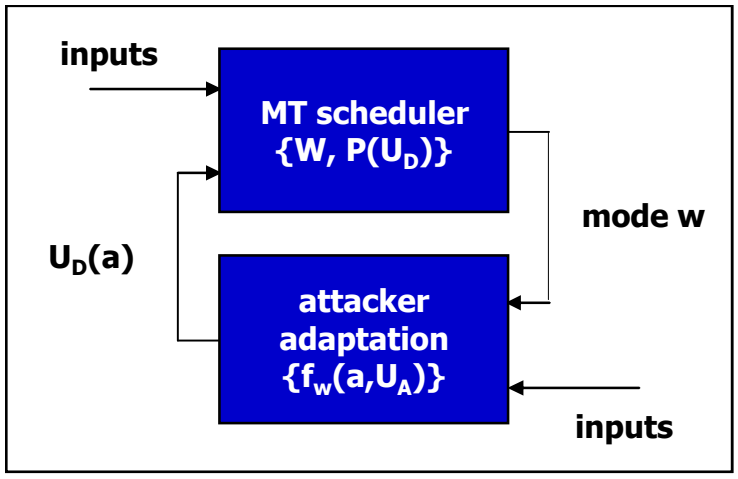
MT Defense



HM-HDS approach

Model: represent the attack-defend incomplete information game as a *hidden-mode hybrid system* (HM-HDS) [Verma 2011, RC/KG 2012]:

- *continuous system* captures attacker adaptation;
- *discrete system* is MT defense scheduler;
- *hidden mode* is defense $\mathbf{w}_i$ that is currently active.



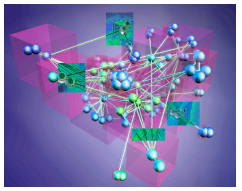
attackers: $U_A = \text{loss}(y, \mathbf{w}^T(\mathbf{x}+\mathbf{a})) - R(\mathbf{a})$

$d\mathbf{a}/dt = \mathbf{f}_w(\mathbf{a}, U_A)$

defense: $U_D = -\text{loss}(y, \mathbf{w}^T(\mathbf{x}+\mathbf{a}))$

$\mathbf{w}^+ = \mathbf{g}(\mathbf{w}, U_D)$

MT Defense



HM-HDS approach (cont'd)

Theorem: for HM-HDS scheduler:

- if either i.) attacker adaptation is good or ii.) defenses $\mathbf{w}_i$ are all equally good then the optimal strategy is to switch defenses uniformly at random;
- as i.) and ii.) are relaxed the above strategy remains nearly optimal.

Empirical study:

MT defense against strong attacks in Spam filtering (top) and malware detection (bottom) tasks.