Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents

Yang Liu[§], Armin Sarabi[§], Jing Zhang[§], Parinaz Naghizadeh[§]
Manish Karir[♯], Michael Bailey[*], Mingyan Liu[§,♯]

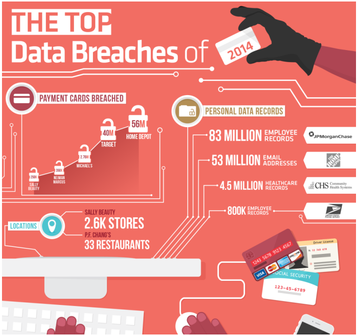
[§] EECS Department, University of Michigan, Ann Arbor
[♯] QuadMetrics, Inc.
[*] ECE Department, University of Illinois, Urbana-Champaign

http://grs.eecs.umich.edu

Motivation

Increasingly frequent and high-impact data breaches

- Target, JP Morgan Chase, Home Depot, to name a few
- Increasing social and economic impact of such cyber incidents



Limitation of current approaches

- Heavily *detection* based
- Fail to detect, or too late by the time a breach is detected
- Not suited for cost/damage control
- Urgent need for more *proactive* measures

Detection

- analogous to diagnosing a patient who may already be ill (e.g., by using biopsy).
- [Qian et al. NDSS14, Wang et al. USENIX Sec14]

Prediction

- predicting whether a presently healthy person may become ill based on a variety of relevant factors.
- [Soska & Christin, USENIX Sec14]

Detection

- analogous to diagnosing a patient who may already be ill (e.g., by using biopsy).
- [Qian et al. NDSS14, Wang et al. USENIX Sec14]

Prediction

- predicting whether a presently healthy person may become ill based on a variety of relevant factors.
- [Soska & Christin, USENIX Sec14]

Our goal:

- *Understand the extent to which one can* forecast *incidents on an organizational level.*

Objective

To develop the ability to *forecast* security incidences

- *Applicability:* we rely solely on *externally* observed data; do not require information on the internal workings of a network or its hosts.

Objective

To develop the ability to *forecast* security incidences

- *Applicability:* we rely solely on *externally* observed data; do not require information on the internal workings of a network or its hosts.
- *Robustness:* we do not have control over or direct knowledge of the error embedded in the data.

Objective

To develop the ability to *forecast* security incidences

- *Applicability:* we rely solely on *externally* observed data; do not require information on the internal workings of a network or its hosts.
- *Robustness:* we do not have control over or direct knowledge of the error embedded in the data.

Key idea:

- tap into a *diverse* set of data that captures different aspects of a network's security posture, ranging from the *explicit* to *latent*.

Why prediction?

Forecast enables entirely new classes of applications which are otherwise not feasible.

Why prediction?

Forecast enables entirely new classes of applications which are otherwise not feasible.

- Prediction allows *proactive* policies and measures to be adopted rather than *reactive* measures following the detection.

Why prediction?

Forecast enables entirely new classes of applications which are otherwise not feasible.

- Prediction allows *proactive* policies and measures to be adopted rather than *reactive* measures following the detection.

Forecast enables effective risk management schemes

Why prediction?

Forecast enables entirely new classes of applications which are otherwise not feasible.

- Prediction allows *proactive* policies and measures to be adopted rather than *reactive* measures following the detection.

Forecast enables effective risk management schemes

- *Internal* to an org.: more informed decisions on resource allocation.

Why prediction?

Forecast enables entirely new classes of applications which are otherwise not feasible.

- Prediction allows *proactive* policies and measures to be adopted rather than *reactive* measures following the detection.

Forecast enables effective risk management schemes

- *Internal* to an org.: more informed decisions on resource allocation.
- *External* to an org.: incentive mechanisms such as cyber insurance.

Outline of the talk

- **Data and Preliminaries**
  - Description of the data
  - Data pre-processing
- Forecasting methods
  - Construction of the predictor
- Forecasting results
  - Main prediction results & analysis

Datasets at a glance

Category	Collection period	Datasets
Mismanagement symptoms	Feb'13 - Jul'13	Open Recursive Resolvers, DNS Source Port, BGP misconfiguration, Untrusted HTTPS, Open SMTP Mail Relays
Malicious activities	May'13 - Dec'14	CBL, SBL, SpamCop, UCEPROTECT, WPBL, SURBL, PhishTank, hpHosts, Darknet scanners list, Dshield, OpenBL
Incident reports	Aug'13 - Dec'14	VERIS Community Database, Hackmageddon, Web Hacking Incidents

- Mismanagement and malicious activities used to extract features.
- Incident reports used to generate labels for training and testing.

Security posture data

Mismanagement symptoms

- Deviation from known best practices; indicators of lack of policy or expertise:
  - Misconfigured- HTTPS cert, DNS (resolver+source port), mail server, BGP.
- Collected around mid-2013 (pre-incidnts).

Security posture data

Mismanagement symptoms

- Deviation from known best practices; indicators of lack of policy or expertise:
  - Misconfigured- HTTPS cert, DNS (resolver+source port), mail server, BGP.
- Collected around mid-2013 (pre-incidnts).

Malicious Activity Data: a set of 11 reputation blacklists (RBLs)

- Daily collections of IPs seen engaged in some malicious activity.
- Three malicious activity types: spam, phishing, scan.
- Use data between May 2013 and December 2014.

Security incident Data

Three incident datasets

- Hackmageddon
- Web Hacking Incidents Database (WHID)
- VERIS Community Database (VCDB)

Incident type	SQLi	Hijacking	Defacement	DDoS
Hackmageddon	38	9	97	59
WHID	12	5	16	45
Incident type	Crimeware	Cyber Esp.	Web app.	Else
VCDB	59	16	368	213

Data Pre-processing

Incident cleaning.

- Remove irrelevant cases, e.g., robbery at liquor store, something happened etc.

Data Pre-processing

Incident cleaning.

- Remove irrelevant cases, e.g., robbery at liquor store, something happened etc.

Data diversity presents challenge in alignment in time and space.

- Security posture records information at the host IP-address level.
- Cyber incident reports associated with an organization.
- Such alignment is not travial: reallocation makes boundary unclear.

Data Pre-processing

Incident cleaning.

- Remove irrelevant cases, e.g., robbery at liquor store, something happened etc.

Data diversity presents challenge in alignment in time and space.

- Security posture records information at the host IP-address level.
- Cyber incident reports associated with an organization.
- Such alignment is not travial: reallocation makes boundary unclear.

A mapping process:

- Summarizing owner IDs from RIR databases.
- 4.4 million prefixes listed under 2.6 million owner IDs: finer degree compared to routing table.
- Sample IP from organization + search in above table.

Outline of the talk

- Data and Preliminaries
  - Description of the data
  - Data pre-processing
- **Forecasting methods**
  - Construction of the predictor
- Forecasting results
  - Main prediction results & analysis

Approach at a glance

Feature extraction

- 258 features extracted from the datasets: Primary + Secondary features.

Approach at a glance

Feature extraction

- 258 features extracted from the datasets: Primary + Secondary features.

Label generation

- 1,000+ incident reports from the three incident sets

Approach at a glance

Feature extraction

- 258 features extracted from the datasets: Primary + Secondary features.

Label generation

- 1,000+ incident reports from the three incident sets

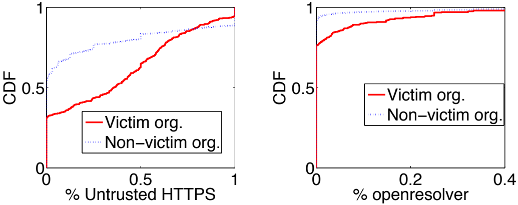
Classifier training and testing

- Random Forest (RF) classifier trained with features and labels.

Primary features: raw data

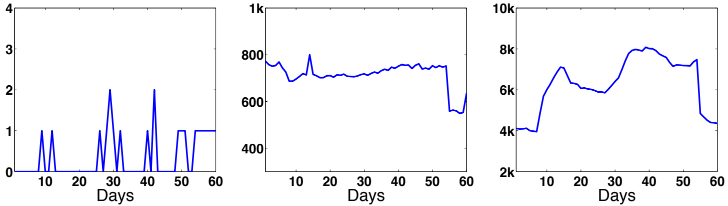
Mismanagement symptoms (5).

- Five symptoms; each measures a fraction
- Predictive power of these symptoms.



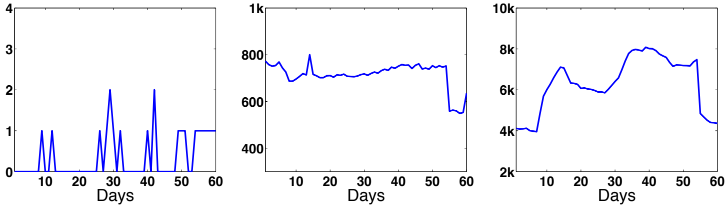
Malicious activity time series (60 $\times$ 3).

- Three time series over a period: spam, phishing, scan.
- Recent 60 v.s. Recent 14.



Malicious activity time series (60 $\times$ 3).

- Three time series over a period: spam, phishing, scan.
- Recent 60 v.s. Recent 14.

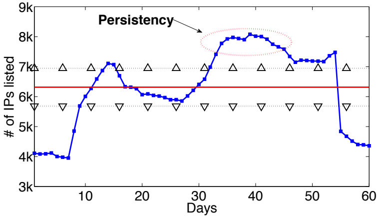


Size: number of IPs in an aggregation unit (1)

- To some extent capture the likelihood of an organization becoming a target of/reproting intentional attacks.

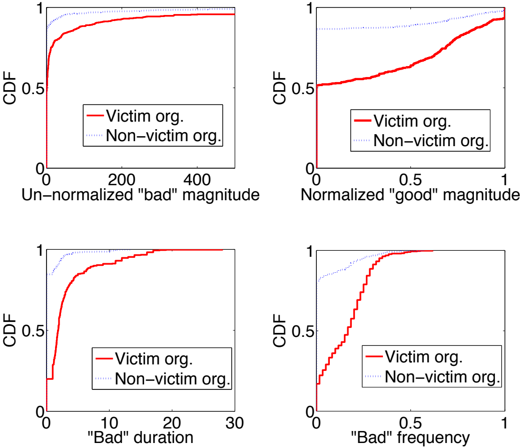
Secondary features

Quantization and feature extraction



- Measure security efforts and responsiveness.
- In each quantized region, measure average magnitude, average duration, and frequency.

A look at their predictive power (using data from Nov-Dec'13):



Training subjects

A subset victim organizations, Group(1) or incident group.

- Training-testing ratio, e.g., **70-30** or **50-50** split .
- Split strictly according to time: use *past* to predict *future*.

	Hackmageddon	VCDB	WHID
Training	Oct 13 – Dec 13	Aug 13 – Dec 13	Jan 14 – Mar 14
Testing	Jan 14 – Feb 14	Jan 14 – Dec 14	Apr 14 – Nov 14

Training subjects

A subset victim organizations, Group(1) or incident group.

- Training-testing ratio, e.g., **70-30** or **50-50** split .
- Split strictly according to time: use *past* to predict *future*.

	Hackmageddon	VCDB	WHID
Training	Oct 13 – Dec 13	Aug 13 – Dec 13	Jan 14 – Mar 14
Testing	Jan 14 – Feb 14	Jan 14 – Dec 14	Apr 14 – Nov 14

A random subset of non-victims, Group (0) or non-incident group.

- Random sub-sampling necessary to avoid imbalance; procedure is repeated over different random subsets.

Outline of the talk

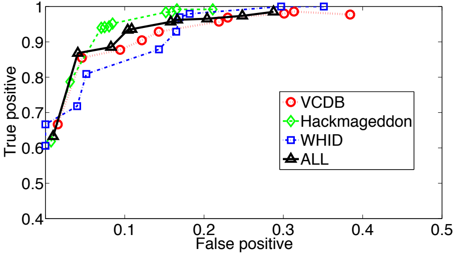
- Data and Preliminaries
  - Description of the data
  - Data pre-processing
- Forecasting methods
  - Construction of the predictor
- **Forecasting results**
  - Main prediction results & analysis

Prediction procedure

Prediction procedure

Prediction procedure

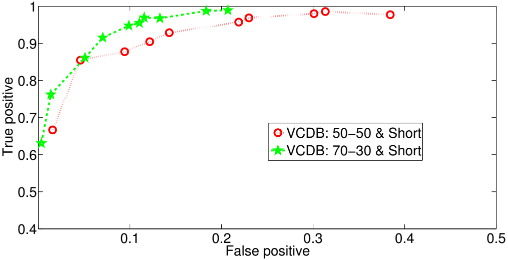
Prediction performance



Example of desirable operating points of the classifier:

Accuracy	Hackmageddon	VCDB	WHID	All
True Positive (TP)	96%	88%	80%	88%
False Positive (FP)	10%	10%	5%	4%
Overall Accuracy	90%	90%	95%	96%

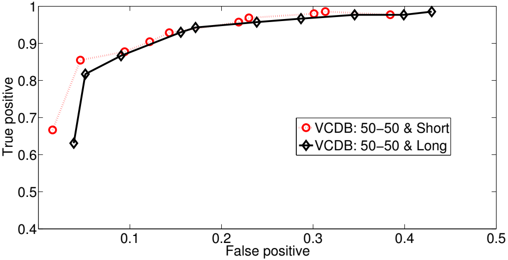
Split ratio



More training data better performance.

Long term prediction

Short term v.s. long term prediction



Temporal features become outdated.

Importance of the Features

Top feature descriptor	Value
Untrusted HTTPS Certificates	0.1531
Frequency	0.1089
Organization size	0.0976
Open recursive resolver	0.0928

- Two mismgmt features rank in top 4.

Importance of the Features

Top feature descriptor	Value
Untrusted HTTPS Certificates	0.1531
Frequency	0.1089
Organization size	0.0976
Open recursive resolver	0.0928

- Two mismgmt features rank in top 4.

Feature category	Normalized importance
Mismanagement	0.3229
Time series data	0.2994
Recent-60 secondary features	0.2602

- Secondary features almost as important as time series data.

Importance of the Features

Top feature descriptor	Value
Untrusted HTTPS Certificates	0.1531
Frequency	0.1089
Organization size	0.0976
Open recursive resolver	0.0928

- Two mismgmt features rank in top 4.

Feature category	Normalized importance
Mismanagement	0.3229
Time series data	0.2994
Recent-60 secondary features	0.2602

- Secondary features almost as important as time series data.
- Dynamic features > static features.

Importance of the Features

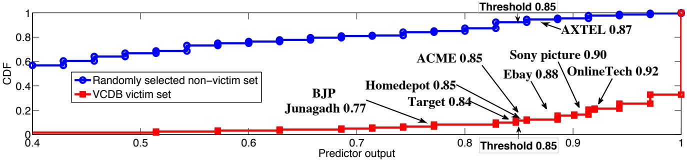
Top feature descriptor	Value
Untrusted HTTPS Certificates	0.1531
Frequency	0.1089
Organization size	0.0976
Open recursive resolver	0.0928

- Two mismgmt features rank in top 4.

Feature category	Normalized importance
Mismanagement	0.3229
Time series data	0.2994
Recent-60 secondary features	0.2602

- Secondary features almost as important as time series data.
- Dynamic features > static features.
- Separate data does NOT achieve comparable results.

Case study: Data Breaches of 2014



- High profile data breaches from 2014: Sony (0.9), Ebay (0.88), Homedepot (0.85), Target (0.84), OnlineTech/JP Morgan Chase (0.92)

Discussions

Errors in the data.

Discussions

Errors in the data.

Robustness against advasarial data.

Discussions

Errors in the data.

Robustness against advasarial data.

Prediction by incident type.

- *O. Thonnard, L. Bilge, A. Kashyap, and M.Lee, Are You At Risk? Profiling Organizations and Individuals Subject to Targeted Attacks. Financial Cryptography and Data Security 2015.*
- *A. Sarabi, P. Naghizadeh, Y. Liu and M. Liu, Prioritizing Security Spending: A Quantitative Analysis of Risk Distributions for Different Business Profiles, WEIS 2015.*

Discussions

Errors in the data.

Robustness against advasarial data.

Prediction by incident type.

- *O. Thonnard, L. Bilge, A. Kashyap, and M.Lee, Are You At Risk? Profiling Organizations and Individuals Subject to Targeted Attacks. Financial Cryptography and Data Security 2015.*
- *A. Sarabi, P. Naghizadeh, Y. Liu and M. Liu, Prioritizing Security Spending: A Quantitative Analysis of Risk Distributions for Different Business Profiles, WEIS 2015.*

Quality of reported data.

- Part of our data can be downladed here: `http://grs.eecs.umich.edu`.

Q & A

Acknowledgement

- We thank NSF and DHS for fundings.

Project webpage (part of data being available)

- http://grs.eecs.umich.edu
- http://www.umich.edu/~youngliu