Building a Global Network Reputation System: Metrics and Data Analytics

Mingyan Liu

Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI

August 2017

Threats to Internet security and availability

From unintentional to intentional, random to financially driven:

- misconfiguration
- mismanagement
- botnets, worms, SPAM, DoS attacks, . . .

Typical countermeasures are *host* based:

- blacklisting malicious hosts; used for filtering/blocking
- installing solutions on individual hosts, e.g., intrusion detection

Also heavily *detection* based:

- even when successful, could be too late
- damage control *post* breach

Our vision

To assess networks as a whole, not individual hosts

- a network is typically governed by consistent policies
    - changes in system administration on a larger time scale
    - changes in resource and expertise on a larger time scale
- consistency (though dynamic) leads to predictability

From a policy perspective:

- leads to *proactive* security policies and enables *incentive mechanisms*, many of which only applicable at an org level.
- enables sensible policies within resource constraints
- facilitates self-inspection by a network using its reputation as feedback

An illustration: host reputation block lists (RBLs)

Commonly used RBLs:

- daily average volume (unique entries) ranging from 146M (BRBL) to 2K (PhishTank)

RBL Type	RBL Name
Spam	BRBL, CBL, SpamCop, WPBL, UCEPROTECT
Phishing/Malware	SURBL, PhishTank, hpHosts
Active attack	Darknet scanners list, Dshield

Strengthen defense:

- filter configuration, blocking mechanisms, etc.

Strengthen security posture:

- get hosts off the list
- install security patches, update software, etc.

Limitations when used at a host level

Host identities can be highly transient:

- dynamic IP address assignment
- reactive policies, leading to significant false positives and misses

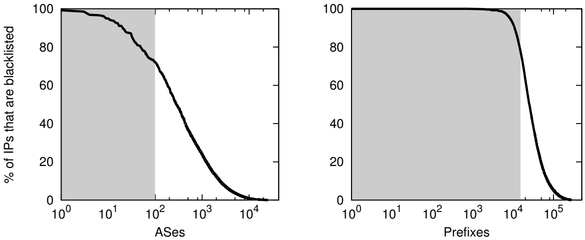
RBLs are application specific:

- a host listed for spamming can initiate a different attack

Lack of standard and transparency in how they are generated

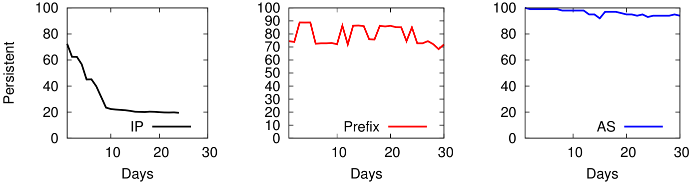
- unknown errors and noises

The power of aggregation: an illustration



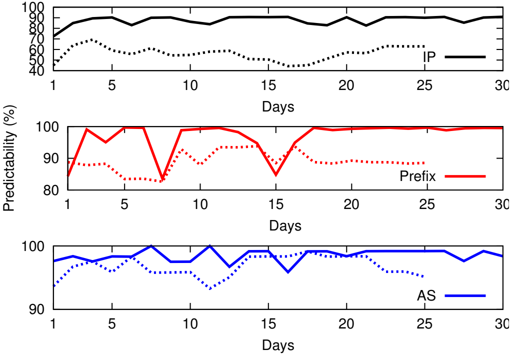
- Taking the union of 12 RBLs
- Right: aggregate at the prefix level (top 15,000-worst prefixes are more than 70% listed; nearly 100% for the worst 9,000 prefixes)
- Left: aggregate at the AS level (top 100-worst ASes are more than 70% listed)

Persistence of maliciousness



- Left: % of IPs listed on the union list on day 1 remain on the list x days later
- Middle: % of the worst set of prefixes on day 1 remain in the worst set x days later
- Right: % of the worst set of ASes on day 1 remain in the worst set x days later

Predictive power



Assume the truth is reflected after a time lag

- Solid: 1-day time lag; Dash: 5-day time lag
- If truth is delayed, how much we see on day x are actually malicious sources

Many applications of such aggregate measures ("reputation")

If it correctly captures the security posture of a network/organization:

- enterprise risk management
  - prioritize resources and take proactive actions
- third-party/vendor validation
- design better incentive mechanisms

How to define and quantify such aggregate measures?

RBLs (again)

Commonly used RBLs:

- daily average volume (unique entries) ranging from 146M (BRBL) to 2K (PhishTank)

RBL Type	RBL Name
Spam	BRBL, CBL, SpamCop, WPBL, UCEPROTECT
Phishing/Malware	SURBL, PhishTank, hpHosts
Active attack	Darknet scanners list, Dshield

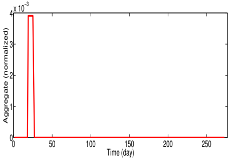
Goal: extract from this dataset information on network-level maliciousness

Data aggregation

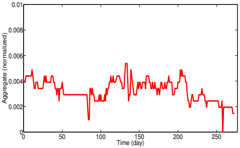
Aggregate the presence on the lists to network level (e.g. /24.)

- Can do this as union of the entire set of RBLs
- or as union of RBLs within a single malicious type.
- apply normalization : fraction of malicious IP addresses.
- $\Rightarrow$ a set of temporal signals, $r_i(t)$

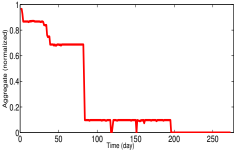
Sample signals



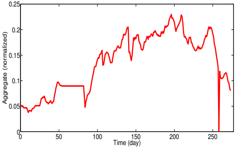
(a) Example /24



(b) Example /21

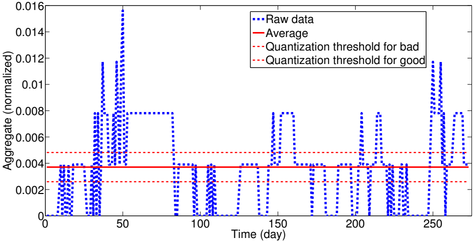


(c) Example /24



(d) Example /20

Feature extraction



- Value-quantize the aggregate signal
- Three regions: good, normal, bad
- Define for each aggregate signal $r_i(t)$, a set of feature vectors λ_i, $\mathbf{d}_i$, $\mathbf{f}_i$: intensity, duration, and frequency vectors.

Why these features?

Hope to capture unique properties in a succinct way

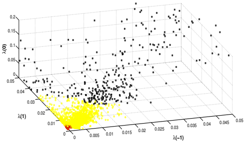
- They allow us to inspect each signal independently and efficiently.
- Large dataset: $N > 360,000$ prefixes.

How to judge whether they are good summaries of the data?

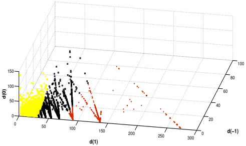
- If we cluster the data using these features (unsupervised), do we get meaningful results?
- If we use these features to train a classifier (supervised), does it make good predictions?

Spectral clustering

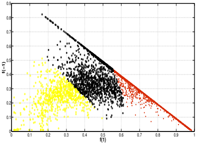
Good: 1; Normal: 0; Bad: -1



(e) Intensity



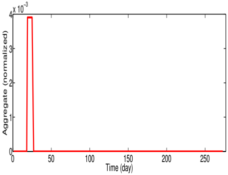
(f) Duration



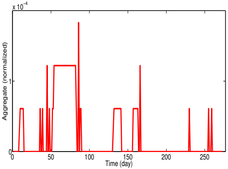
(g) Frequency

Clusters	Intensity	Duration	Frequency
1	low in all 3 elements	long good durations	high good frequency
2	medium in all 3 elements	short bad/good durations	high normal frequency
3	high in all 3 elements	long bad durations	high bad frequency

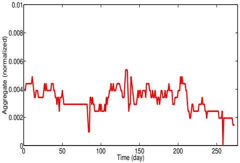
Putting three features together: some examples



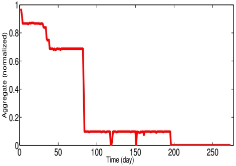
(h) [1,1,1]



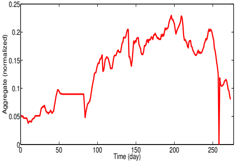
(i) [1,2,1]



(j) [1,2,2]



(k) [3,1,1]



(l) [3,3,3]

Some observations of prefix distribution

Combining the worst patterns (6.8K between [3,3,3] and [3,2,2]):

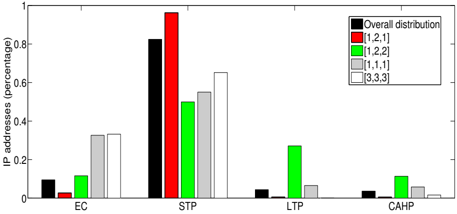
- 1.65K from India,
- 587 from Vietnam,
- 388 from Iran,
- 366 from Peru, and
- 340 from Kazakhstan.

By contrast, of the almost 75K prefixes in [1,1,1]:

- one-third comes from the US,
- 5.8K from UK,
- 4.6K from Brazil,
- 3.1K from China and
- 2.7K from Russia.

ASes categorized into four types:

- Enterprise Customers (ECs),
- Small Transit Providers (STPs),
- Large Transit Providers (LTSs), and
- Content/ Access/ Hosting Providers (CAHPs).



Can similar features be used to train a classifier?

Follow a supervised learning framework:

- features: capturing security posture of an entity
- labels: ground truth data on whether an entity has had a cybersecurity incident

Both datasets are noisy and incomplete

- Tap into a larger set of data that captures different aspects of a network's security posture: *explicit* as well as *latent*.

Security posture data

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.

Security posture data

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.

Mismanagement symptoms

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

Cyber 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

Datasets at a glance

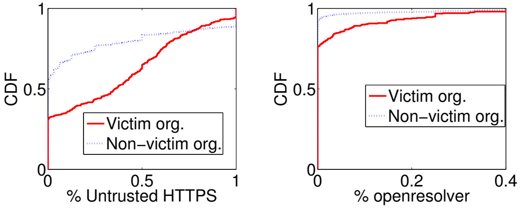
Category	Collection period	Datasets
Mismanagement symptoms	Feb'13 - Jul'13	Open Recursive Resolvers, DNS Source Port, BGP misconfiguration, Untrusted HTTPS, Open SMTP Mail Relays
Malicious activities	May'13 - Dec'14	CBL, SBL, SpamCop, UCEPROTECT, WPBL, SURBL, PhishTank, hpHosts, Darknet scanners list,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

- Mismanagement and malicious activities used to extract features:
  - aggregation now at the org/entity level.
- Incident reports used to generate labels for training and testing.

Primary and secondary features

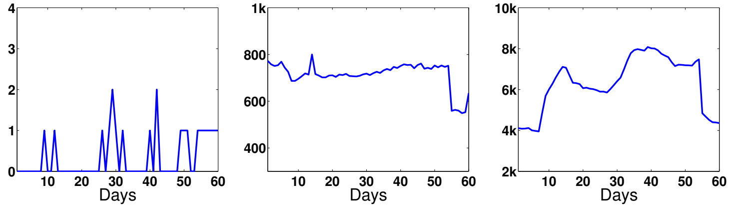
Mismanagement symptoms.

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



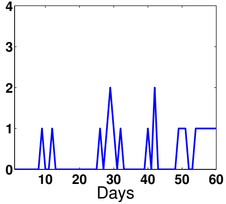
Malicious activity time series.

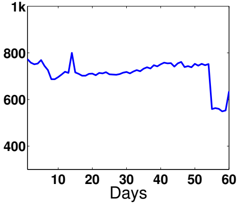
- Three time series over a period: spam, phishing, scan.
- Recent 60 v.s. Recent 14.

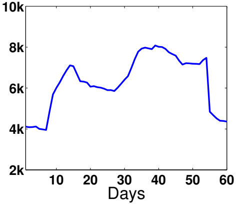


Malicious activity time series.

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



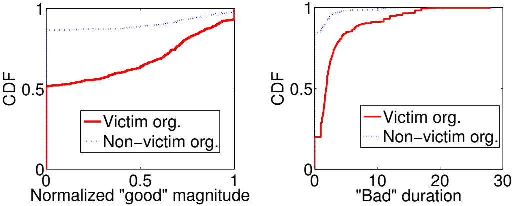




Secondary features: discussed earlier

- Measuring persistence and responsiveness.

A look at their predictive power:



Training and testing procedure

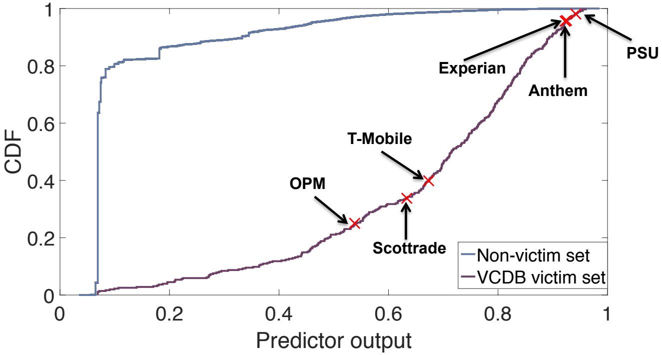
A subset of victim organizations, or incident group.

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

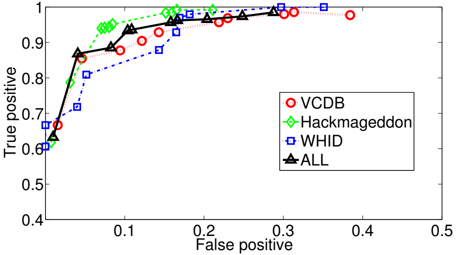
	Hackmageddon	VCDB	WHID
Training	Oct 13 – Dec 13	Aug 13 – Dec 13	Jan 14 – Mar 14
Testing	Jan 14 – Feb 14	Jan 14 – Dec 14	Apr 14 – Nov 14

Examples: top data breaches of 2015

Distribution of predictor output



Overall 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%

Conclusion & Discussion

A macroscopic view of security posture: network reputation

- as a way of holistic assessment
- defined possible metrics and demonstrated their utility
  - feature extraction and clustering
  - classifier training and breach prediction at an org level

Conclusion & Discussion

A macroscopic view of security posture: network reputation

- as a way of holistic assessment
- defined possible metrics and demonstrated their utility
  - feature extraction and clustering
  - classifier training and breach prediction at an org level

Transition to practice

- a global enterprise cybersecurity ratings system
- QuadMetrics, Inc. $\Rightarrow$ FICO.

Conclusion & Discussion

A macroscopic view of security posture: network reputation

- as a way of holistic assessment
- defined possible metrics and demonstrated their utility
  - feature extraction and clustering
  - classifier training and breach prediction at an org level

Transition to practice

- a global enterprise cybersecurity ratings system
- QuadMetrics, Inc. $\Rightarrow$ FICO.

Other applications to be explored:

- deep packet inspection
- peering policies

Acknowledgement

Work supported by the NSF and the DHS

- Y. Liu, A. Sarabi, J. Zhang, P. Naghizadeh, M. Karir, M. Bailey and M. Liu, "Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents", *USENIX Security*, August 2015, Washington, D. C.
- A. Sarabi, P. Naghizadeh, Y. Liu and M. Liu, "Prioritizing Security Spending: A Quantitative Analysis of Risk Distributions for Different Business Profiles", *WEIS*, June 2015, Delft University, The Netherlands.
- P. Naghizadeh and M. Liu, "Inter-Temporal Incentives in Security Information Sharing Agreements", *ITA*, February 2016, San Diego, CA.
- P. Naghizadeh and M. Liu, "Voluntary participation in cyber-insurance markets," WEIS, June 2014, PSU.
- J. Zhang, Z. Durumeric, M. Bailey, M. Karir, and M. Liu, "On the Mismanagement and Maliciousness of Networks," *Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, February 2014.
- J. Zhang, A. Chivukula, M. Bailey, M. Karir, and M. Liu, "Characterization of Blacklists and Tainted Network Traffic," *14th Passive and Active Measurement Conference (PAM)*, Hong Kong, March 2013.