Quantitative Cybersecurity: Breach Prediction and Incentive Design

Mingyan Liu

Joint work with

Yang Liu, Armin Sarabi, Parinaz Naghizadeh, Michael Bailey, Manish Karir

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 can only be applied at a network/org level.

More specifically

To what extent can we quantify and assess the security posture of a network/organization?

- Enterprise risk management
  - Prioritize resources and take proactive actions
- Third-party/Vendor validation

To what extent can we utilize such assessment to design better incentive mechanisms

- Incentives properly tied to actual security posture and security interdependence

Outline of the talk

- A incident forecasting framework and results
  - As a way to quantify security posture and security risks
  - Data sources and processing
  - A supervised learning approach
- Risk assessment as a form of "public monitoring"
  - Enables inter-temporal incentives in enforcing long-term security information sharing agreements
- Risk assessment as a form of "pre-screening"
  - Enables judicious premium discrimination in cyber insurance to mitigate moral hazard

An incident forecasting framework

Desirable features:

- *Scalability:* we rely solely on *externally* observed data.
- *Robustness:* data will be noisy, incomplete, not all of which is under our control.

An incident forecasting framework

Desirable features:

- *Scalability:* we rely solely on *externally* observed data.
- *Robustness:* data will be noisy, incomplete, not all of which is under our control.

Key steps:

- Tap into a *diverse* set of data that captures different aspects of a network's security posture: source, type (*explicit* vs. *latent*).
- Follow a supervised learning framework.

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.
- Incident reports used to generate labels for training and testing.

Data pre-processing

Conservative processing of incident reports:

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

Data pre-processing

Conservative processing of incident reports:

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

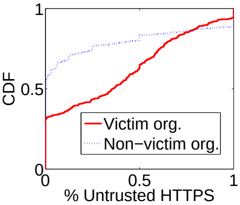
Challenge in data alignment, both in time and in space:

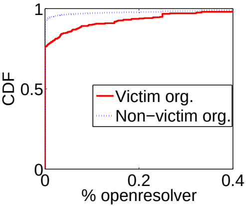
- Security posture records information at the host IP-address level.
- Cyber incident reports associated with an organization.
- Alignment non-trivial: address reallocation, hosting services, etc.

Primary and secondary features

Mismanagement symptoms.

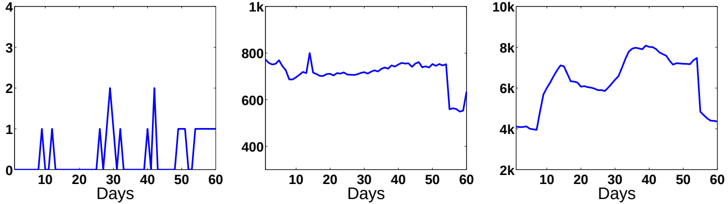
- Five symptoms; each measured as a fraction
- Predictive power of these symptoms.





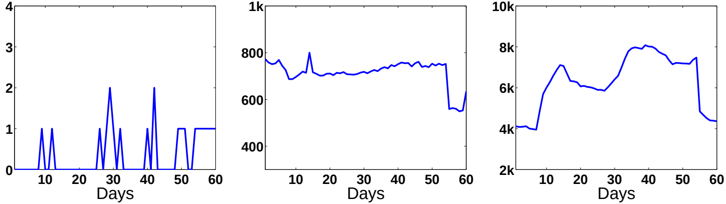
Malicious activity time series.

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



Malicious activity time series.

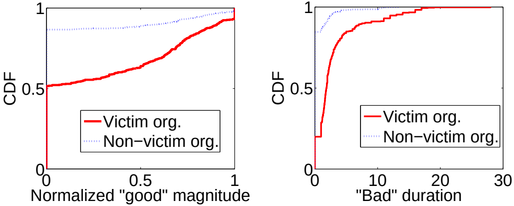
- Three time series over a period: spam, phishing, scan.
- Recent 60 v.s. Recent 14.



Secondary features

- Measuring persistence and responsiveness.

A look at their predictive power:



Training subjects

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

Training subjects

A subset of victim organizations, or incident group.

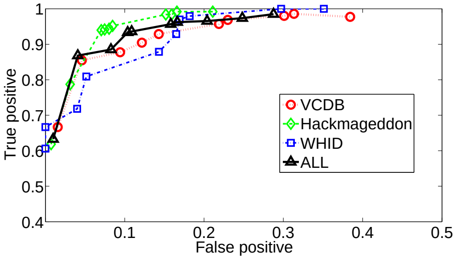
- Training-testing ratio, e.g., **70-30** or **50-50** split .
- Split strictly according to time: use *past* to predict *future*.

	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

A random subset of non-victims, or non-incident group.

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

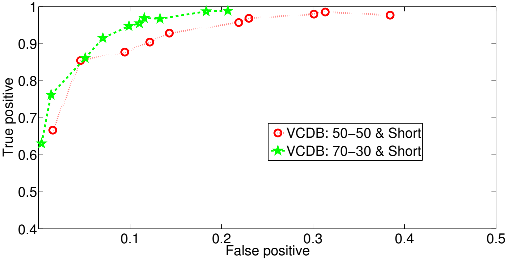
Prediction performance



Example of desirable operating points of the classifier:

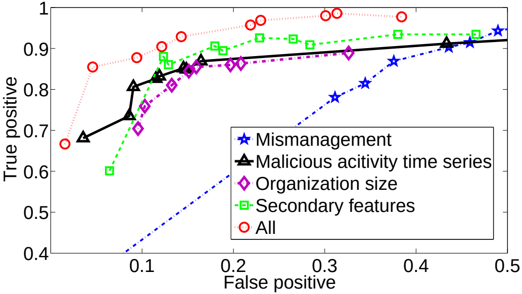
Accuracy	Hackmageddon	VCDB	WHID	All
True Positive (TP)	96%	88%	80%	88%
False Positive (FP)	10%	10%	5%	4%

Split ratio



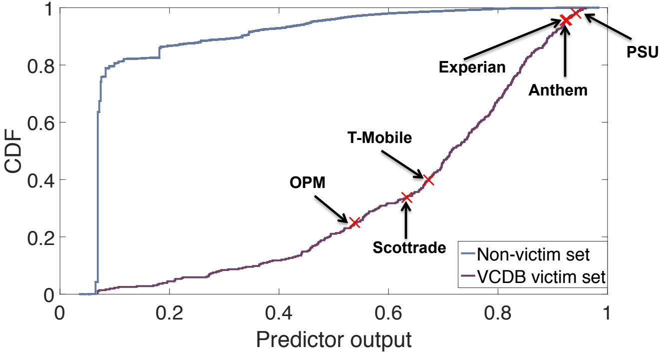
More training data gives better performance.

The power of data diversity



Any single data source does not hold sufficient predictive power

More recent case study: top data breaches of 2015



- Top breaches in 2014: Sony, Ebay, Homedepot, Target, OnlineTech/JP Morgan Chase

Fine-grained prediction

Goal: conditional density estimation

- Perform *conditional prediction*: if an incident should occur, the likelihood of its being of a particular type $\Rightarrow$ *Risk profiles*.

Fine-grained prediction

Goal: conditional density estimation

- Perform *conditional prediction*: if an incident should occur, the likelihood of its being of a particular type $\Rightarrow$ *Risk profiles*.

Shall use VCDB (including non-cyber incidents)

- Details on the incident, actor, action, assets, and the victim.
- Plus information from AWIS: rank (global, regional), rank history, speed, age, locale, category, publicly traded, etc.

Fine-grained prediction

Goal: conditional density estimation

- Perform *conditional prediction*: if an incident should occur, the likelihood of its being of a particular type $\Rightarrow$ *Risk profiles*.

Shall use VCDB (including non-cyber incidents)

- Details on the incident, actor, action, assets, and the victim.
- Plus information from AWIS: rank (global, regional), rank history, speed, age, locale, category, publicly traded, etc.

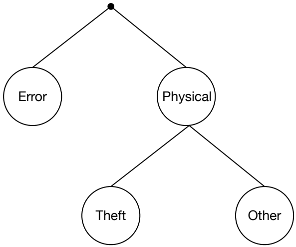
Challenges

- Incomplete labels: the level of details that are available vary for each report.
- Selection bias and rare events.

A layered approach

To address incomplete labels:

- Train multiple binary classifiers, each estimating a portion of the risk
- Chain rule:
  $P(\text{Physical Theft}) = P(\text{Physical}) \times P(\text{Theft} \mid \text{Physical})$



Example risk profiles

Risk profiles for sample organizations and their corresponding industries.

Organization	Error	Hacking: Comp. Cred.	Hacking: Other	Malware	Misuse	Physical: Theft	Physical: Other	Social
Information								
Russian Radio			×					
Verizon			×					
Public Administration								
Macon Bibb County	×							
Internal Revenue Service					×			

- Gray cells signify incident types with high risk;
- Crosses indicate the actual incident.

Outline of the talk

- A incident forecasting framework and results
    - As a way to quantify security posture and security risks
    - Data sources and processing
    - A supervised learning approach
- Risk assessment as a form of “public monitoring”
    - Enables inter-temporal incentives in enforcing long-term security information sharing agreements
- Risk assessment as a form of “pre-screening”
    - Enables judicious premium discrimination in cyber insurance to mitigate moral hazard

Information sharing agreements among firms

Information sharing agreements among firms

Executive Order 13691 "Promoting Private Sector Cybersecurity Information Sharing"

Information Sharing and Analysis Organizations (ISAOs), Cyber Information Sharing and Collaboration Program (CISCP), Computer Emergency Readiness Team (US-CERT), etc

Information Sharing and Analysis Centers (ISACs)

FINANCIAL SERVICES Information Sharing and Analysis Center

IT ISAC

The disincentive: disclosure costs

Disclosure costs

- Drop in market values following security breach disclosure [Campbell et al. 03][Cavusoglu, Mishra, Raghunathan 04]
- Loss of consumer/partner confidence
- Bureaucratic burden

The disincentive: disclosure costs

Disclosure costs

- Drop in market values following security breach disclosure [Campbell et al. 03][Cavusoglu, Mishra, Raghunathan 04]
- Loss of consumer/partner confidence
- Bureaucratic burden

How to sustain cooperation?

- Audits and sanctions (e.g. by an authority or the government) [Laube and Bohme 15]
- Introducing additional economic incentives (e.g. taxes and rewards for members of ISACs) [Gordon, Loeb, Lucyshyn 03]

The disincentive: disclosure costs

Disclosure costs

- Drop in market values following security breach disclosure [Campbell et al. 03][Cavusoglu, Mishra, Raghunathan 04]
- Loss of consumer/partner confidence
- Bureaucratic burden

How to sustain cooperation?

- Audits and sanctions (e.g. by an authority or the government) [Laube and Bohme 15]
- Introducing additional economic incentives (e.g. taxes and rewards for members of ISACs) [Gordon, Loeb, Lucyshyn 03]
- **Inter-temporal incentives**: conditioning future cooperation on history of past interactions.

Information sharing games: stage game model

- Two firms
- $r_i \in \{0, 1\}$: (partially) concealing and (fully) disclosing
- Gain from other firm's disclosed information G
- Disclosure costs C

	1	0
1	$G - C$, $G - C$	$-C$, G
0	G, $-C$	0, 0

Information sharing games: stage game model

- Two firms
- $r_i \in \{0, 1\}$: (partially) concealing and (fully) disclosing
- Gain from other firm's disclosed information G
- Disclosure costs C

	1	0
1	$G - C$, $G - C$	$-C$, G
0	G, $-C$	0, 0

$\Rightarrow$ Prisoner's dilemma: only equilibrium of one shot game is $(0, 0)$.

Repeated games and monitoring possibilities

- Can we sustain (nearly) *efficient payoffs* in repeated games?
- Depends on whether/how deviations are detected and punished.
- Let b_i denote the *belief* of i about r_j.

Repeated games and monitoring possibilities

- Can we sustain (nearly) *efficient payoffs* in repeated games?
- Depends on whether/how deviations are detected and punished.
- Let b_i denote the *belief* of i about r_j.

Imperfect **Private** Monitoring

$$\pi(b_i|r_j) = \begin{cases} \epsilon, & \text{for } b_i = 0, r_j = 1 \\ 1-\epsilon, & \text{for } b_i = 1, r_j = 1 \\ \alpha, & \text{for } b_i = 0, r_j = 0 \\ 1-\alpha, & \text{for } b_i = 1, r_j = 0 \end{cases}$$

with $\epsilon \in (0, 1/2)$ and $\alpha \in (1/2, 1)$.

Imperfect **Public** Monitoring

$$\hat{\pi}((b_i, b_j)|(r_i, r_j)) := \pi(b_i|r_j)\pi(b_j|r_i)$$

monitoring by an assessment system.

Infinitely repeated games with private monitoring

- Wanted: a *folk theorem* - a full characterization of payoffs that can be achieved in a repeated game if players are sufficiently patient.

Infinitely repeated games with private monitoring

- Wanted: a *folk theorem* - a full characterization of payoffs that can be achieved in a repeated game if players are sufficiently patient.
- No folk theorem for infinitely repeated games with imperfect private monitoring in general.

Infinitely repeated games with private monitoring

- Wanted: a *folk theorem* - a full characterization of payoffs that can be achieved in a repeated game if players are sufficiently patient.
- No folk theorem for infinitely repeated games with imperfect private monitoring in general.
    - They exist for some modifications/subclasses:
        - Communication (cheap talk) [Compte 98, Kandori and Matsushima 98].
        - Pubic actions, e.g., announcing sanctions [Park 11].
        - Sufficiently correlated private signals [Mailath and Morris 02].

Imperfect public monitoring: A folk theorem

[Fudenberg, Levine, and Maskin 1994]

If the imperfect public monitoring is *sufficiently informative*, s.t.:

- individual full rank: deviations by an individual player are statistically distinguishable.
- pairwise full rank: deviations by players i and j are distinct, i.e., induce different distributions over public outcomes.

Imperfect public monitoring: A folk theorem

[Fudenberg, Levine, and Maskin 1994]

If the imperfect public monitoring is *sufficiently informative*, s.t.:

- individual full rank: deviations by an individual player are statistically distinguishable.
- pairwise full rank: deviations by players i and j are distinct, i.e., induce different distributions over public outcomes.

then there exists a discount factor $\underline{\delta} < 1$, such that for all $\delta \in (\underline{\delta}, 1)$, any feasible and strictly individually rational payoff profile can be sustained by public strategies.

Our monitoring mechanism is informative

- It can be verified that our public monitoring model satisfies these two conditions.
- The folk theorem holds with the **same monitoring technology** of that of individual firms $\Rightarrow$ the rating/assessment system facilitates coordination.
- Conclusions hold with countably finite disclosure decisions and discrete ratings by the monitoring system.
- The monitoring model captures the predictive framework presented earlier: binary outcome, imperfect but sufficiently accurate.

Outline of the talk

- A incident forecasting framework and results
  - As a way to quantify security posture and security risks
  - Data sources and processing
  - A supervised learning approach
- Risk assessment as a form of “public monitoring”
  - Enables inter-temporal incentives in enforcing long-term security information sharing agreements
- Risk assessment as a form of “pre-screening”
  - Enables judicious premium discrimination in cyber insurance to mitigate moral hazard

Cyber Insurance as a risk management tool

Risk transfer rather than risk reduction:

- Inherits typical issues: adverse selection and moral hazard
- Has the effect of lowering the effort exerted by the client

Lack actuarial data in cyber security compared to traditional products

- Lack of understanding on both sides
- Policy underwriting driven by regulation rather than by security concerns

Cyber security in a fast changing threat landscape

- compared to more predictable or deterministic conditions: home, life, auto, flood, etc.

Current state of practice

Prospective client taking a survey:

- questions on IT systems: products in place, etc.
- questions on practice: software/system update, policy
- questions on users: number, access, etc.

Followed by some estimates on value at risk (VaR)
Extensive exclusions

- Generally covers only legal fees and crisis management
- Clients seek to self-insure to lower the premium
- Structured as catastrophe protection but grossly insufficient coverage

Literature on cyber insurance

as an incentive mechanism for risk reduction

In a competitive cyber insurance market:

- Pal, Glubchik, Psounis, Hui 2014; Shetty, Schwartz, Felegyhazi, Walrand 2010
- contracts designed to attract clients; not optimized to induce better security behavior;
- introduction of cyber insurance deteriorates network security;
- insurers make no profit.

With a monopolistic and profit-neutral insurer aiming for maximum social welfare:

- Bolot, Legarge 2008
- use premium discrimination: higher premium to those with worse types/lower efforts;
- insurance contracts can lead to better efforts and improved security;
- non-negative profit for the insurer;
- however, client participation is mandated and insurer does not seek to maximize profit.

Our own work on a monopolistic insurer seeking max social welfare:

- it is generally impossible to simultaneously achieve social welfare maximization, weak budget balance (non-negative profit), and voluntary participation.

Introducing credible pre-screening

Utilizing our risk assessment framework:

- As a signal that enables premium discrimination prior to entering the contract
- As a monitoring tool that reduces information asymmetry and enhances transparency

Basic (principle-agent) model:

- a single profit-maximizing insurer
- one or more risk-averse clients, who may not voluntarily participate (contracts must be individually rational (IR))
- insurer seeks to maximize its utility, subject to incentive compatibility (IC)
- clients' security inter-dependent: correlated losses

One insurer, one risk-averse client

Insurer's utility:

$$V(p, \alpha, \beta, e) = p - \alpha S_e - \beta L_e$$

- e: effort exerted by the client
- S_e: signals observed by both, effort plus noise
- L_e: realized loss
- p: base premium
- α: discount factor
- β: coverage factor, $0 \leq \beta \leq 1$

Client's payoff without contract:

$$U(e) = -e^{-\gamma(-L_e - ce)}$$

- γ: risk attitude; higher γ means more risk aversion; assumed known to the insurer
- $U^o := \max_e \overline{U}(e)$.

Client's payoff with contract:

$$U^c(p, \alpha, \beta, e) = -e^{-\gamma(-p + \alpha S_e - L_e + \beta L_e - ce)}$$

Insurer's problem

$$\begin{aligned}
\max_{p,\alpha,\beta,e\geq 0} \quad & \overline{V}(p,\alpha,\beta,e) \\
\text{s.t.} \quad & \overline{U}^c(p,\alpha,\beta,e) \geq U^o \qquad \text{(IR)} \\
& e \in \arg\max_{e'\geq 0} \overline{U}^c(p,\alpha,\beta,e') \qquad \text{(IC)}
\end{aligned}$$

Key results

Policies can be designed to offer non-negative profit for the insurer and incentive for the client to participate (increased utility)

Risk transfer

- State of security worsens compared to no-insurance scenario
- Risk-averse agent transfers part of the risk to the insurer and reduce its effort

Credible pre-screening can improve the state of security

- also leads to higher profit for the insurer
- the higher the quality of the screening the more significant the impact

One insurer, two risk-averse clients

Consider three cases:

- neither enters a contract
- one enters a contract, the other opts out
- both purchase a contract

Each case results in a game between the two clients

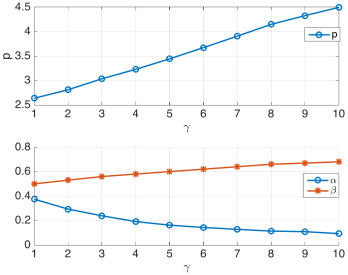
Risk inter-dependence:

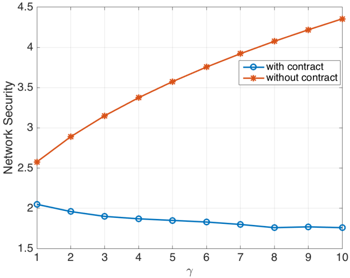
$$L^{(i)}_{e_1,e_2} \sim \mathcal{N}(\mu(e_i + xe_{-i}), \lambda(e_i + xe_{-i}))$$

Numerical results: single agent

assuming $\mu(e) = \frac{10}{e+1}, \lambda(e) = \frac{10}{(e+1)^2}, c = 1$

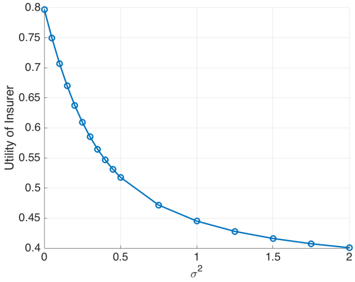
The effect of risk aversion; fix $\sigma^2 = 1$

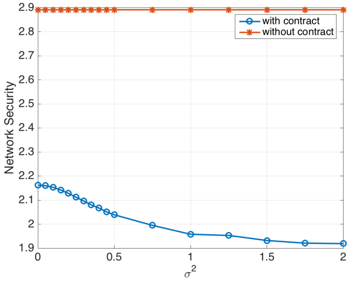




The impact of pre-screening; fix $\gamma = 2$

- Increasing σ^2: less informative pre-screening
- p, α, β all decrease with increasing σ^2





Conclusion

A prediction framework for forecasting cybersecurity incidents

- Data sources, pre-processing, features, and training.

Conclusion

A prediction framework for forecasting cybersecurity incidents

- Data sources, pre-processing, features, and training.

Its role in encouraging better information sharing

- As a form of public monitoring to induce inter-temporal incentives to sustain cooperation.

Conclusion

A prediction framework for forecasting cybersecurity incidents

- Data sources, pre-processing, features, and training.

Its role in encouraging better information sharing

- As a form of public monitoring to induce inter-temporal incentives to sustain cooperation.

Its role in enabling better cyber insurance policies

- Steering insurance toward risk reduction in addition to risk transfer.

Acknowledgement

Work supported by the NSF and the DHS

References:

- 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.