PreSTA: Preventing Malicious Behavior Using Spatio-Temporal Reputation

Andrew G. West
November 4, 2009
ONR-MURI Presentation
**PreSTA: Preventing Malicious Behavior Using Spatio-Temporal Reputation**

**Title and Subtitle**

**Publication Date**

**Permitting Organization**

**Performing Organization**

**Distribution/Availability Statement**

**Supplementary Notes**

**Abstract**

**Subject Terms**

**Security Classification of:**

<table>
<thead>
<tr>
<th>a. REPORT</th>
<th>b. ABSTRACT</th>
<th>c. THIS PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>unclassified</td>
<td>unclassified</td>
<td>unclassified</td>
</tr>
</tbody>
</table>

**Limitation of Abstract**

*Same as Report (SAR)*

**Number of Pages**

23

**Name of Responsible Person**

Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

Approved for public release; distribution unlimited

ONR-MURI Presentation, Nov 2009

Standard Form 298 (Rev. 8-98)
PreSTA: Preventative Spatio-Temporal Aggregation

**PROBLEM**
- Traditional punishment mechanisms (i.e., blacklists) are reactive
- PreSTA: Detect malicious users (i.e., spammers) before harm is done

**SOLUTION**

**HYPO-THESIS:**
- Malicious users are spatially clustered (in any dimension)
- Malicious users are likely to repeat bad behaviors (temporal)

**GIVEN:**
- A historical record of those principals known to be bad, and the timestamp of this observation (feedback)

**PRODUCE:**
- An extended list of principals who are thought to be bad now, based on their past history, and history of those around them
PreSTA Running Example: Spam Detection

- Spatio-temporal properties of spam mail
- Basis for spatial groupings
- Calculating and combining reputations
- Classifier performance

Generalizing PreSTA: Additional Use-Cases for Model

- Malicious editors on Wikipedia
- Applicability to the QuanTM model
- General PreSTA use-case criteria

Conclusions & References
TEMPORAL PROPERTIES

TEMPORAL: Bad Guys Repeat Bad Behaviors

- Spammers want to maximize utilization of available IP addresses, leading to re-use
- Bot-nets will compromise a machine until patched
- Blacklist entries have predictable duration (~6 days), making for trivial recycling
- Most mail servers have static IP addresses, so IP acts as a persistent identifier – though we later discuss DHCP considerations
IP DELEGATION HIERARCHY

(1) Internet Assigned Numbers Auth.: Controls all IP delegation (root of trust)

(2) Regional Internet Registries: Continent-level equivalent of the IANA

(3) Autonomous Systems (ISPs): Broadcast the IPs they control via the Border Gateway Protocol (BGP)

(4) Local routers distribute addresses from some pool (i.e., a /24). Such subnet boundaries are NOT known

(5) Individual IP: Over time a single IP may have multiple inhabitants (due to dynamic nature – DHCP)
**IANA /RIR**
- The IANA and RIR granularity are too broad to be of relevant use

**AS**
- What AS(es) are broadcasting IP?
- An IP may have 0, 1, or 2+ homes

**BLOCK**
- What is /24 (256 IP) membership?
- Valuate that block and two adjacent
- Estimation of subnet membership

**IP**
- Simplest case. Little spatial value.
- Due to DHCP, may have multiple inhabitants over time, though

---

**AS(es)**
1000's IPs

**Subnet-level Block-Heuristic**
768 IPs

**IP-level**
1 IP
SPAM: SPATIAL PROPERTIES


- Some ISPs/AS willing to trade *behavioral leniency* for compensation: McColo Corp. and 3FN
- Some geographical jurisdictions are more lenient than others (and this maps into IP space)
- As IPs become BL'ed, operations must shift to 'fresh' addresses, likely those from the same allocation (*i.e.*, subnets)
PreSTA: SPAM USAGE

PreSTA: Preventative Spatio-Temporal Aggregation

BL Source DBs

BL Source

Spamhaus

Subscription

Blacklist DB

Spatial Analysis

Temporal Analysis

Reputation Engine

PreSTA Server

PreSTA Client

Classifier

Cache DB

SMTP Server

Cache Hit

Incoming Emails

Cache Miss

Decision
PreSTA: SPAM USAGE

PreSTA:
Preventative
Spatio-Temporal
Aggregation

Incoming Emails

Cache Miss

Decision

Cache Hit

SMTP Server

PreSTA Server

Reputation Engine

Spatial Analysis

Temporal Analysis

Classifier

Cache DB

PreSTA Client

BL Source DBs

BL Source

Spamhaus

Subscription

PreSTA:
Preventative
Spatio-Temporal
Aggregation

PreSTA: SPAM USAGE
VALUATION WORKFLOW

Source IP → AS → Time-Decay FN → AS-REP
Source IP → BLOCK → REP ALG → BLK-REP
Source IP → IP → IP-REP

Spatial Mapping

SPAM or HAM

Plot into 3-D Space

Classify

Mail Body
To calculate reputation for entity $\alpha$:

$$\text{raw_rep}(\alpha) = \sum_{i=1}^{i \leq |\text{BL}(\alpha)|} \frac{\text{time_decay}(\text{BL}(\alpha)_i)}{\text{magnitude}(\alpha)}$$

$$\text{REP}(\alpha) = 1.0 - (\text{raw_rep}(\alpha) \times \phi^{-1})$$

- $\text{time_decay}(*):$ Returns on $[0,1]$, higher weight to more recent events
- $\text{magnitude}(\alpha):$ Number of IPs in grouping $\alpha$
- $\phi$: Normalization constant putting $\text{REP}()$ on $[0,1]$
SVM LEARNING

- Combination strategies
- Support Vector Machine
  - Supervised learning
  - Train over previous email to classify current emails
- Draws surface (threshold) best separating points
  - Can adjust penalty weight to keep false positives low
  - Polynomial, RBF kernels improve on linear performance
SPAM: TESTING DATASETS

BLACKLIST
- Subscribe to Spamhaus provider
- Process diff’s between lists into DB
- Scores 86.2% detection w/0.37% FP

AS-MAP
- Use RouteViews data to map IP->AS

EMAIL
- 10 weeks: 15 mil. UPenn mail headers
- Proofpoint score as definitive spam/ham tag
We capture between 20-50% of spam that gets past current blacklists
- By design our FP-rate is equivalent to BLs: ~0.4%

Total blockage remains near constant: 90%
- Blacklists are reactive, we are predictive. We can cover its slack
- Cat and mouse. Graph should roll over time

Captures up to 50% of mail not caught by traditional blacklists with the same low false-positives
Probable botnet attack which our metric could mitigate via both temporal and spatial means.
SPAM: CONTRIBUTIONS

SNARE [3] (GA-Tech)

- Supervised learning across 13-network level features, including spatio-temporal ones
- Don't need blacklists (but neither do we, only known spamming IPs)

Existing ‘Reputation Systems’ [6]

- Exclusive use of negative feedback
- Existing email reputation systems [5] focus only on sharing classifications

DISTINGUISHING CONTRIBUTIONS

- Formalization of predictive spatio-temporal reputation
- Development of a lightweight mail filter, capable of 500k+ mails/hour
**PURPOSE:** Build a blacklist of user-names/IPs based on the probability they will vandalize.

**TEMPORAL**
- Straightforward, vandals are probably repeat offenders.
- Registered users have IDs indicating when they joined, are new users more likely to vandalize?

**SPATIAL**
- Geographical: Based on user **location** (i.e., Wash. D.C.)
- Topical: A user may vandalize one **topic** (Rush Limbaugh), while properly editing another (Barack Obama).
- Anonymous users: IP address properties.

**FEEDBACK**
- Certain administrators have **rollback** (revert) privileges.
- Comment: “Reverted edit by X to last edition by Y”.
PreSTA may trivially fulfill the reputation component of qualifying QTM systems
- TDG-like hierarchy of IP-delegation
- Spatial groups from credential depth?

General-use case criteria:
- (1) There must be a grouping function to define finite sets of participants
- (2) Observable and dynamic feedback sufficient to construct behavior history
Given a known set of malicious users (and the time at which they mis-behaved)...

...additional malicious users may be identified using...

(1) Temporal histories of principals
(2) w.r.t the space in which they reside

... and such a system is useful for:

(1) Lightweight spam filtering above traditional blacklists
(2) Detecting editors probable of vandalism on Wikipedia
(3) Fulfilling the reputation component of any QTM system

and such a system is useful for:
CONCLUSIONS

Given a known set of malicious users (and the time at which they mis-behaved)...

...additional malicious users may be identified using...

(1) Temporal histories of principals
(2) w.r.t the space in which they reside

... and such a system is useful for:

(1) Lightweight spam filtering above traditional blacklists
(2) Detecting editors probable of vandalism on Wikipedia
(3) Fulfilling the reputation component of any QTM system
CONCLUSIONS

Given a known set of malicious users (and the time at which they mis-behaved)...

...additional malicious users may be identified using...

(1) Temporal histories of principals

(2) w.r.t the space in which they reside

... and such a system is useful for:

(1) Lightweight spam filtering above traditional blacklists

(2) Detecting editors probable of vandalism on Wikipedia

(3) Fulfilling the reputation component of any QTM system
CONCLUSIONS

Given a known set of malicious users (and the time at which they mis-behaved)...

...additional malicious users may be identified using...

(1) Temporal histories of principals  (2) w.r.t the space in which they reside

... and such a system is useful for:

(1) Lightweight spam filtering above traditional blacklists
(2) Detecting editors probable of vandalism on Wikipedia
(3) Fulfilling the reputation component of any QTM system

DONE  IN PROGRESS  FUTURE WORK
REFERENCES


