Spam Mitigation using Spatio-temporal Reputations from Blacklist History*

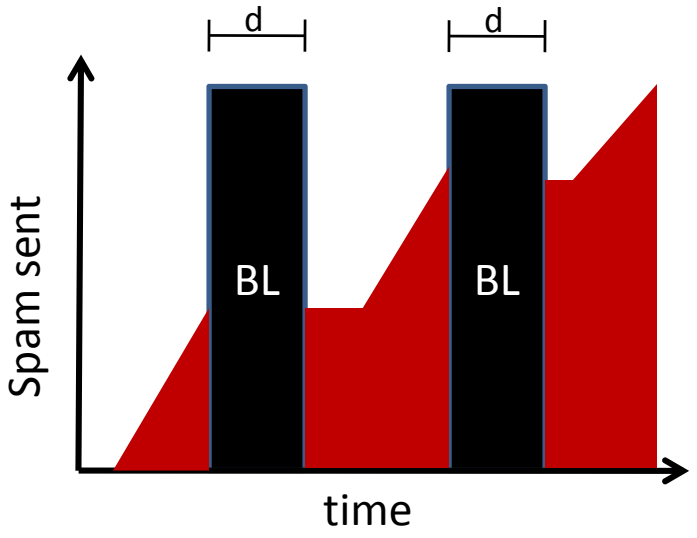
A.G. West, A.J. Aviv, J. Chang, and I. Lee

ACSAC `10 - December 9, 2010

* Note that for conciseness, this version omits the animation and graph annotations that existed in the actual conference version

Motivation

- IP spam blacklists
  - Reactively compiled from major email providers
  - Single IPs can be listed, de-listed, re-listed
  - De-listing policy?
- Blacklist/spam properties
  - High re-listing rates
  - Spam IPs are spatially clustered [5-12]
  - Just 20 ASes account for 42% of all spam [7]



PreSTA: Preventative Spatio-Temporal Aggregation

Penn Engineering

PROBLEM
- Traditional punishment mechanisms are reactive

ASSUM-PTIONS:
- Consistent behaviors (temporal) and spatial clustering

GIVEN:
- User feedback and spatial grouping functions

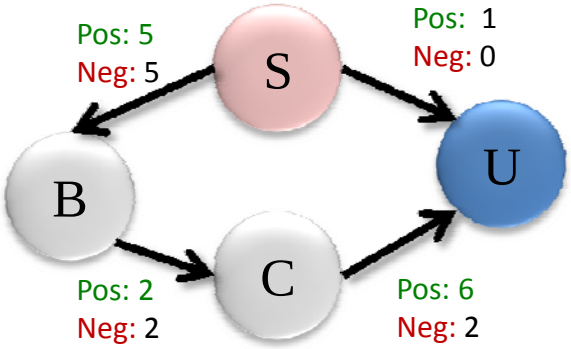
PRODUCE:
- An extended list of principals -- thought to be bad now

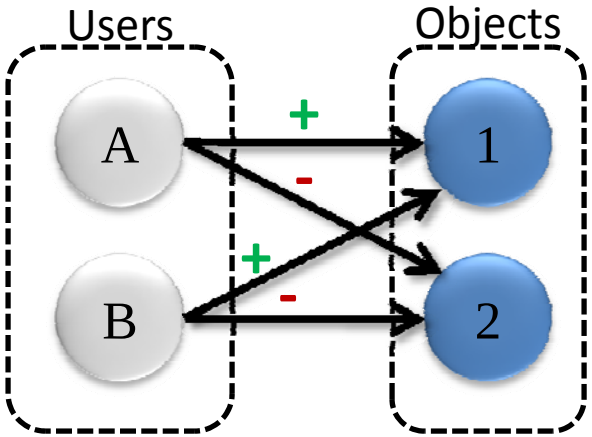
OUTPUT
- The preventative identification of malicious users

PreSTA Model &
Reputation Fundamentals

Reputation Algs.

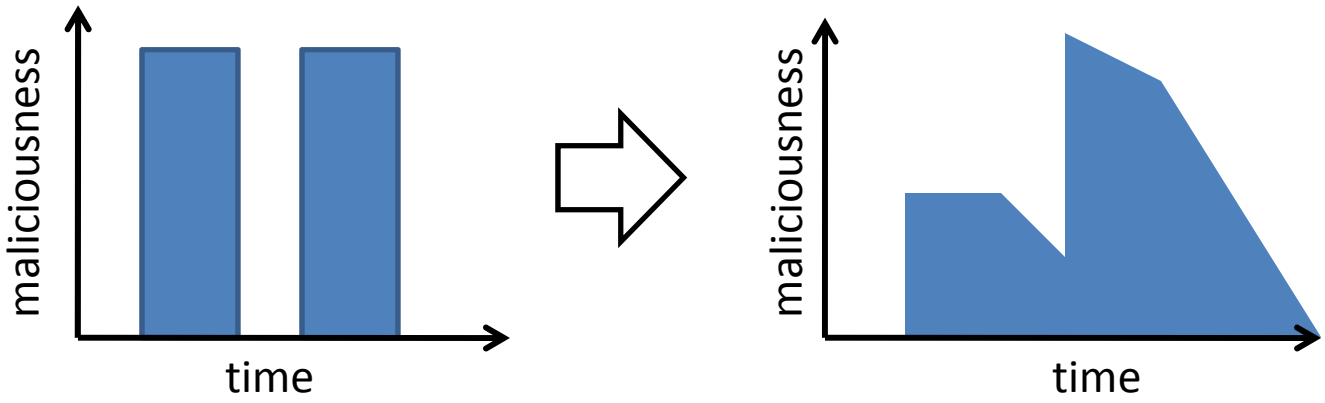
- Reputation systems:
  - EBay, EigenTrust [1], Subjective Logic [2]
  - Use cases: P2P networks, access-control, anti-spam [3]
  - Enum. feedbacks; distributed calculation (transitivity)
- Recommendations: Voting, product suggestions



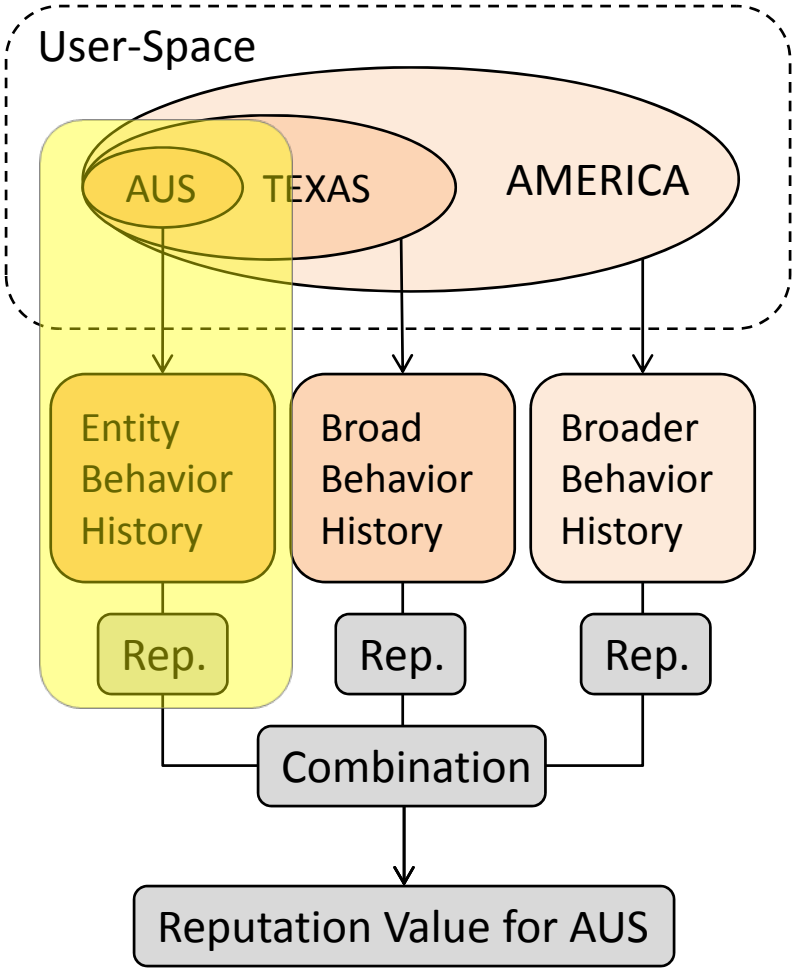


PreSTA Reputation

- PreSTA-style reputation:
  - No positive feedback -- time-decay to heal.
  - Centralized and trusted feedback provider
  - Quantify binary observations into reputations

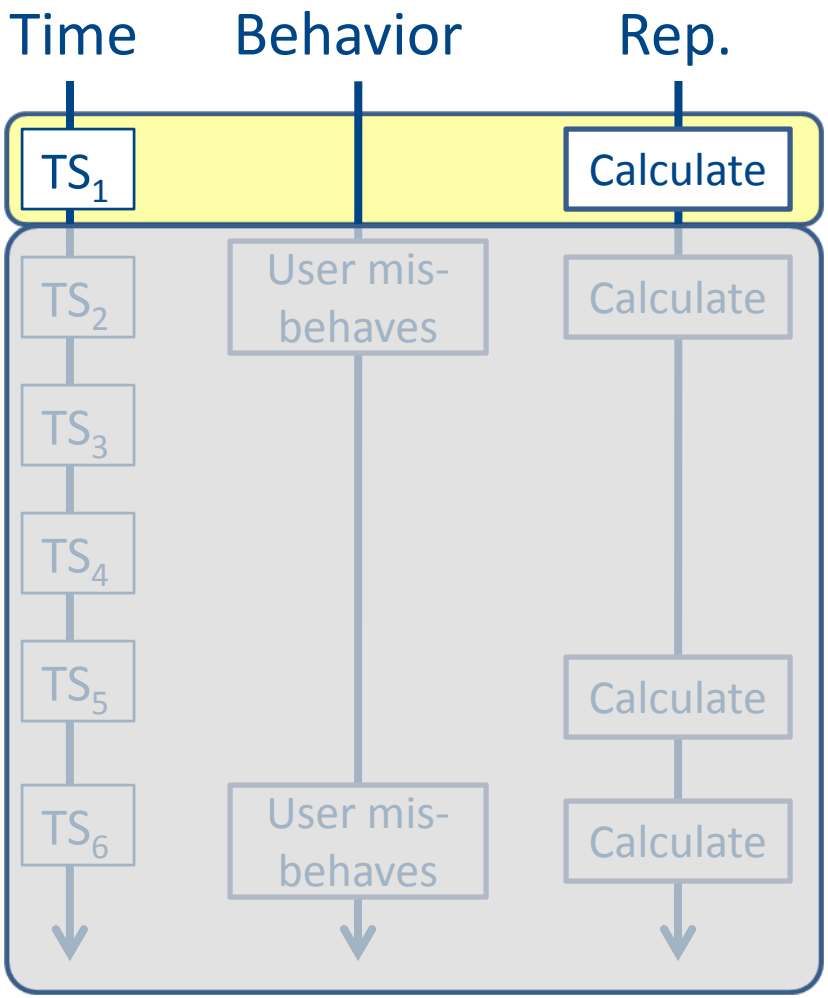


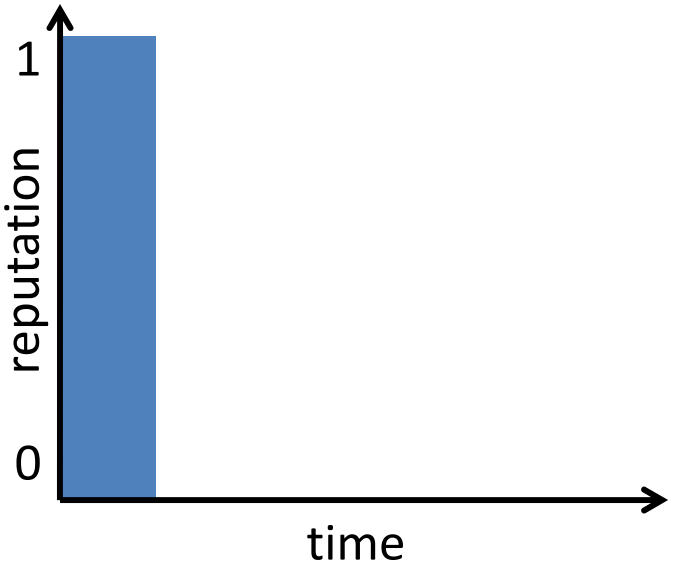
PreSTA Reputation



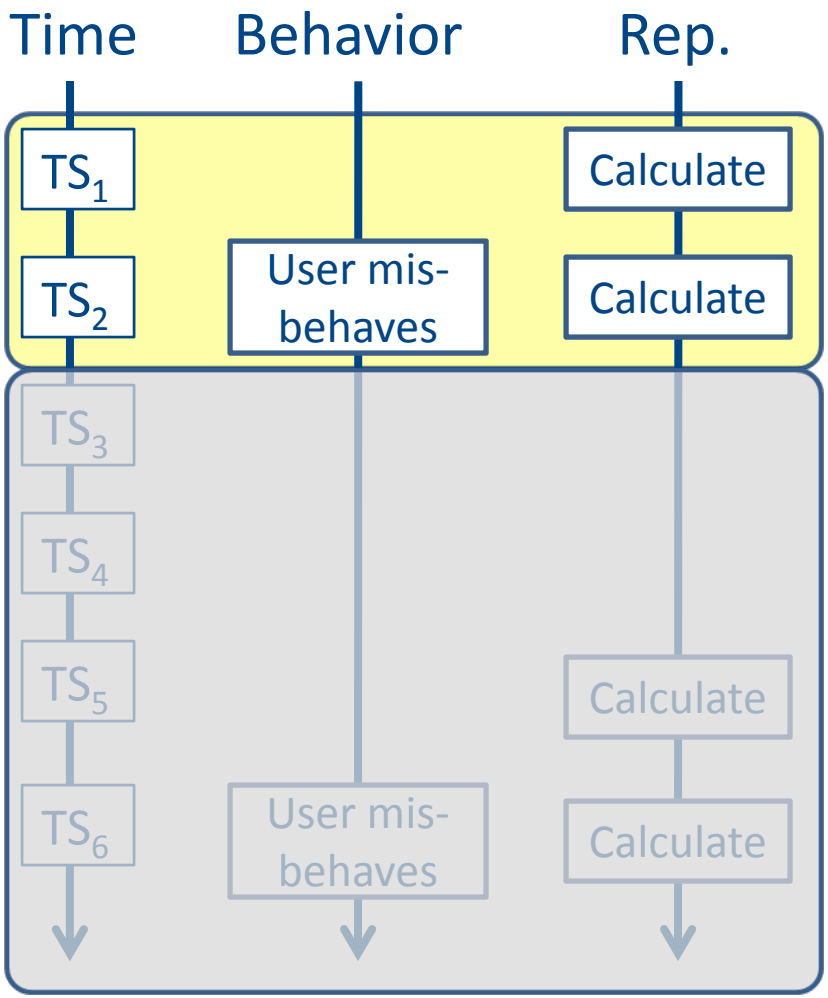
Single-entity calculation and rep. values are status quo (temporal)

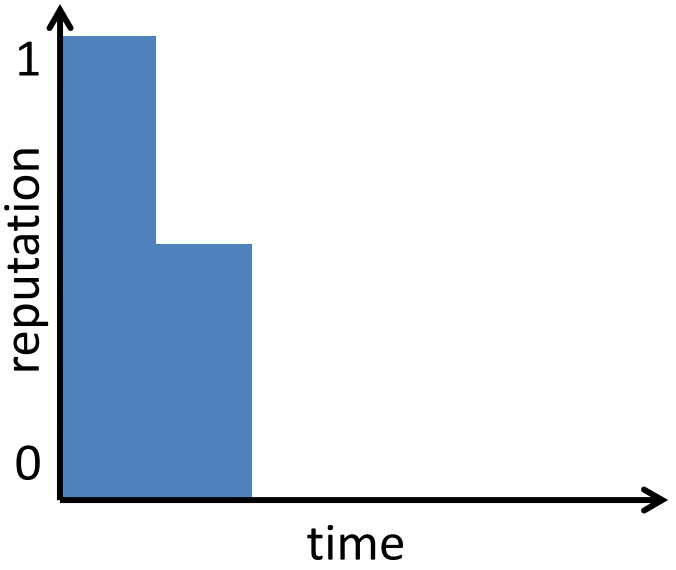
Sample calc. (1)



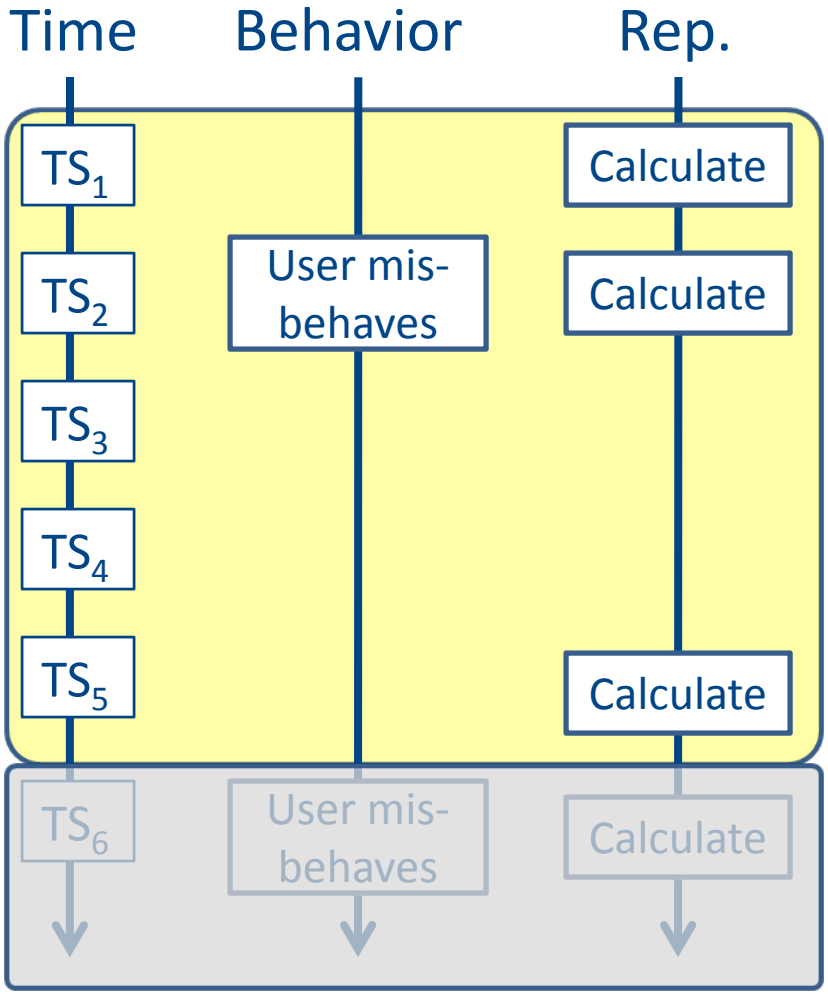


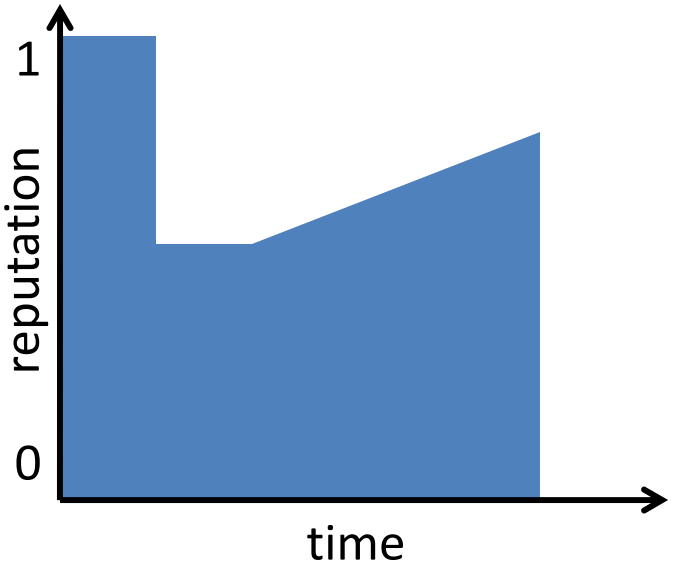
Sample calc. (2)



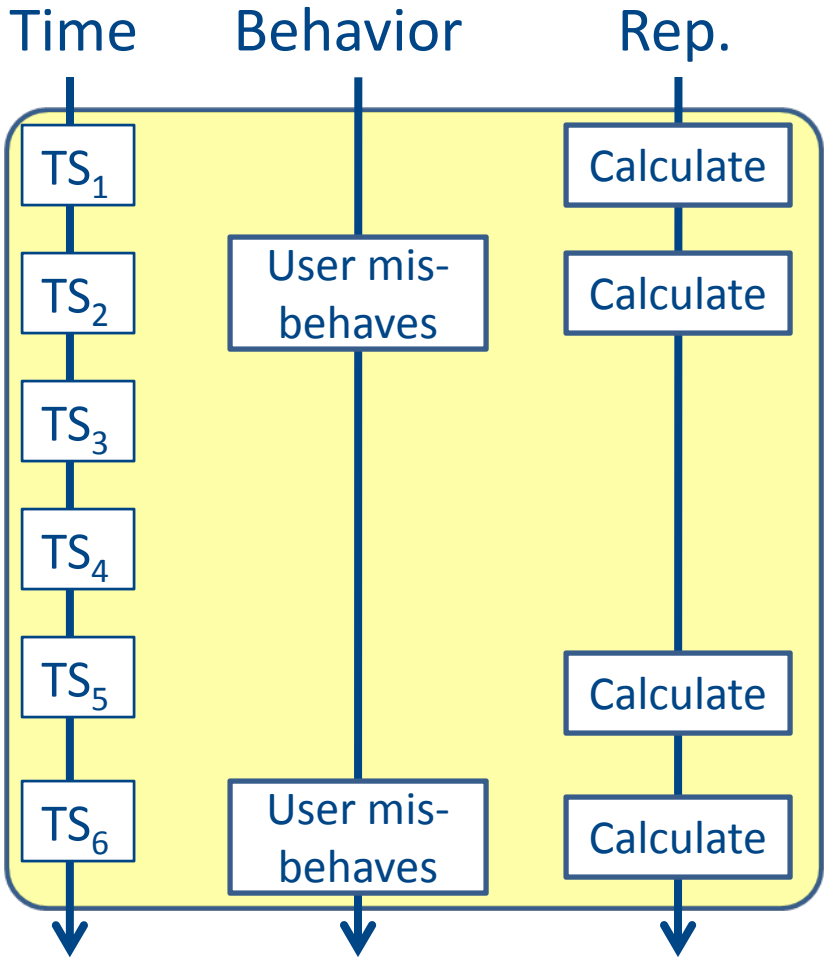


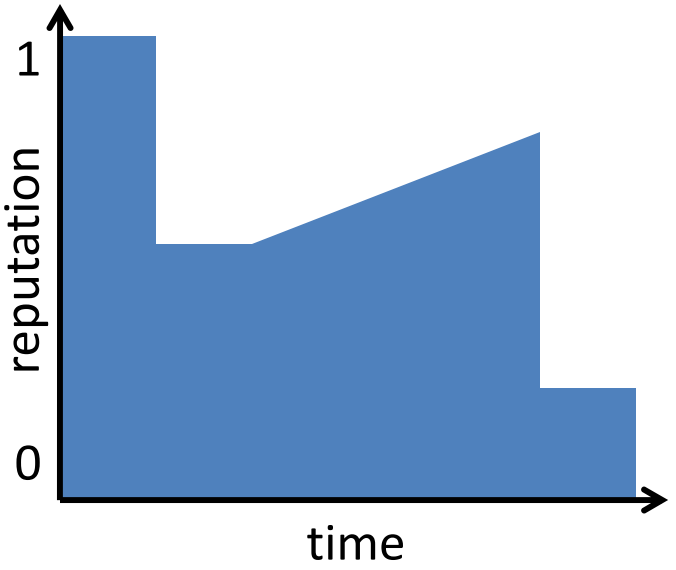
Sample calc. (3)



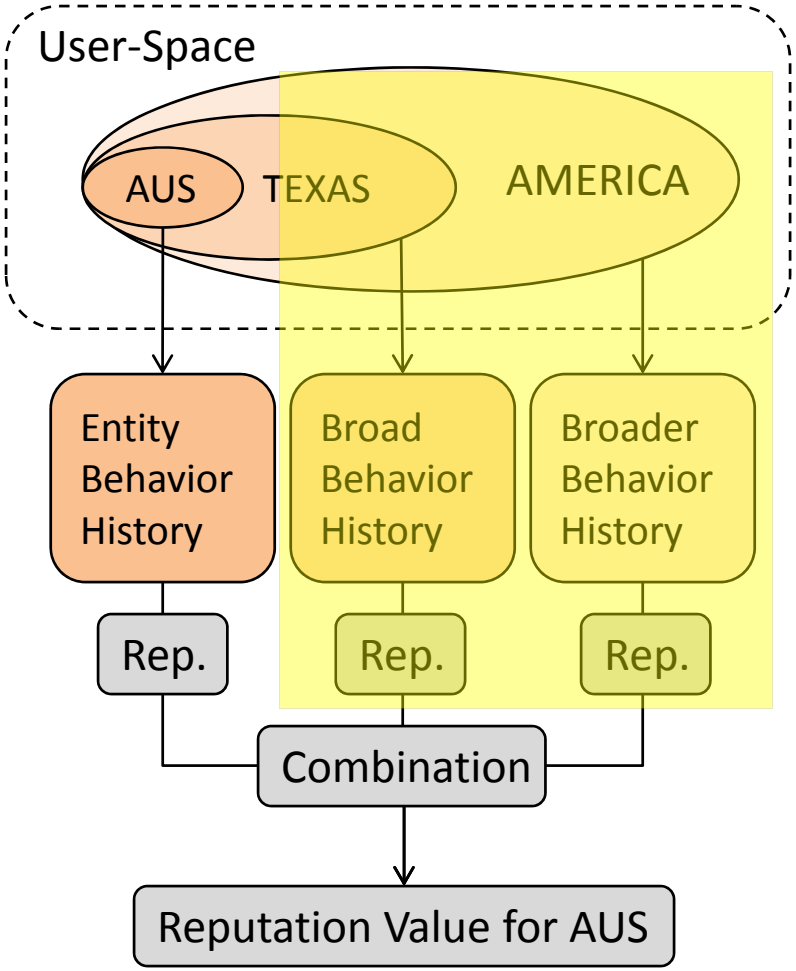


Sample calc. (4)



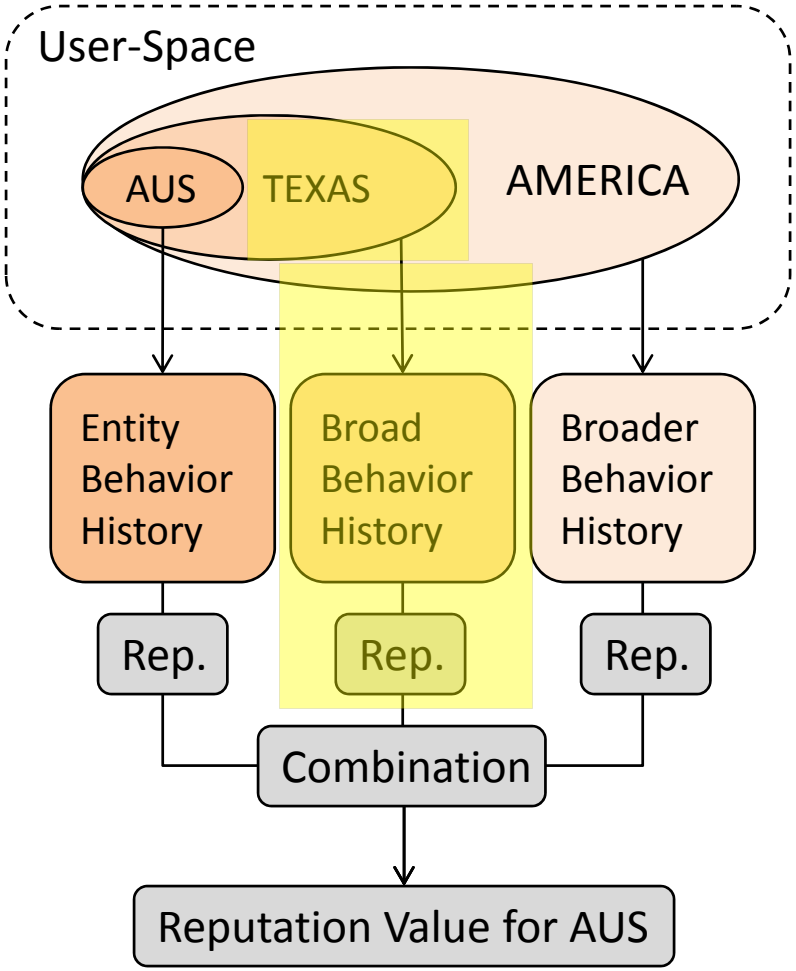


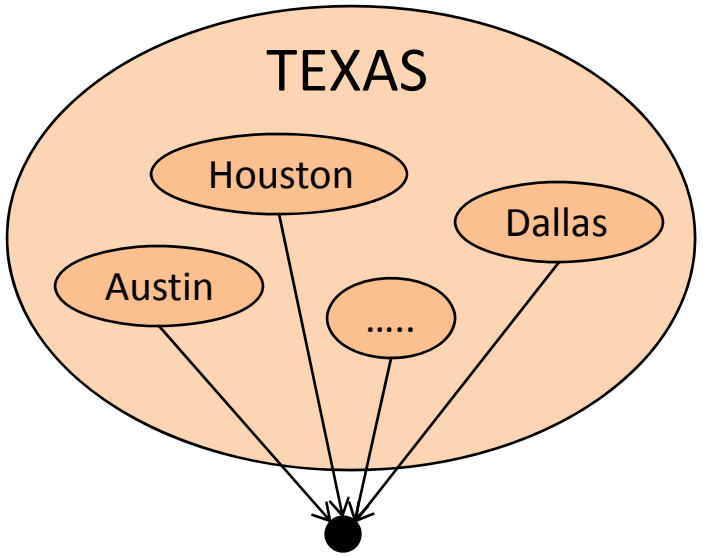
PreSTA (spatial)



- Why spatial reputation:
  - Exploit homophily
  - Overcome the cold-start problem (Sybil [4])
- Grouping functions define group membership
  - Multiple groups/dims.
  - Geo-based/abstract

PreSTA (spatial)





$$rep(group) = \frac{\sum \text{all member rep.}}{\text{group size}}$$

PreSTA + Wikipedia[19]

Penn Engineering

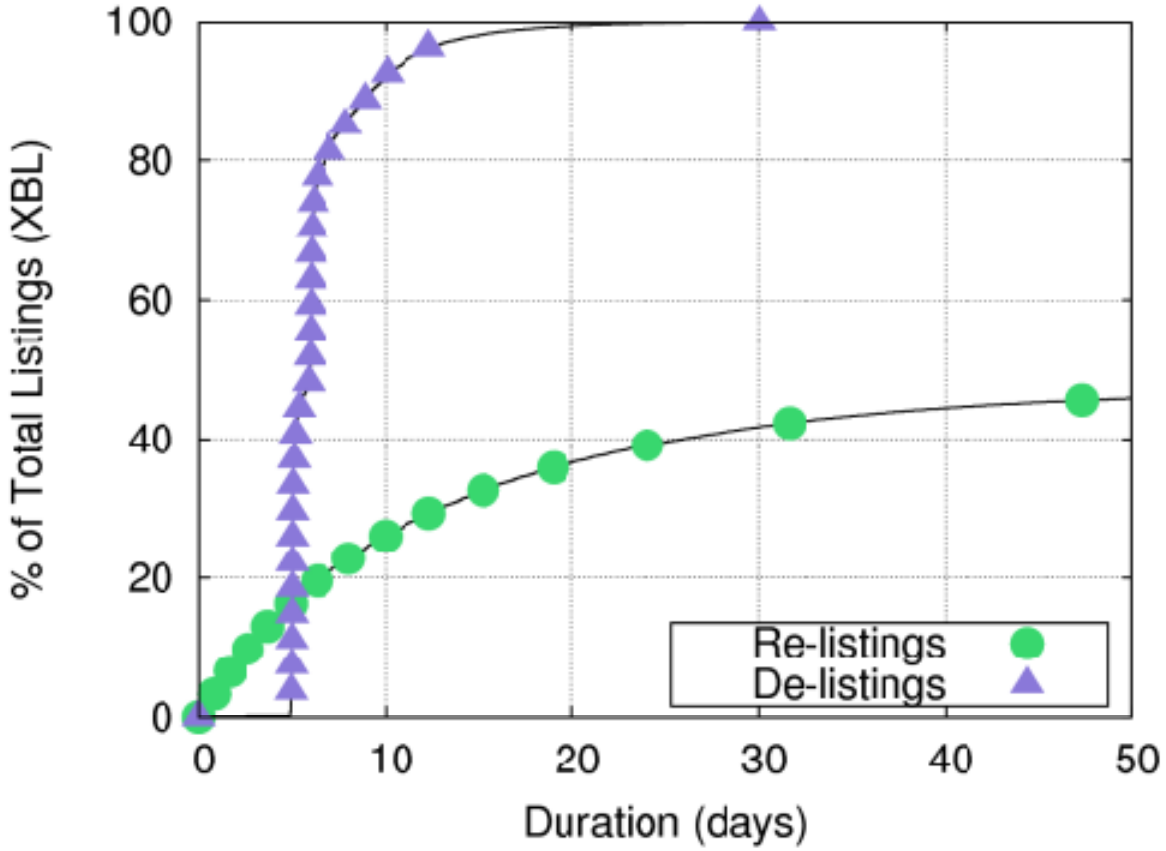
PURPOSE: Detect vandalism edits to Wikipedia

TEMPORAL	• Vandal editors are probably repeat offenders • Frequently vandalized articles may be future targets
SPATIAL	• Group editors by country (geographical space) • Group articles by category (topical space)
FEEDBACK	• Gleamed from administrative “undo” function
SUCCESS	• Live tool -- STiki [20] -- 25,000 vandalisms undone

Applying PreSTA to spam mitigation

Spatio-temporal
props. of spam email

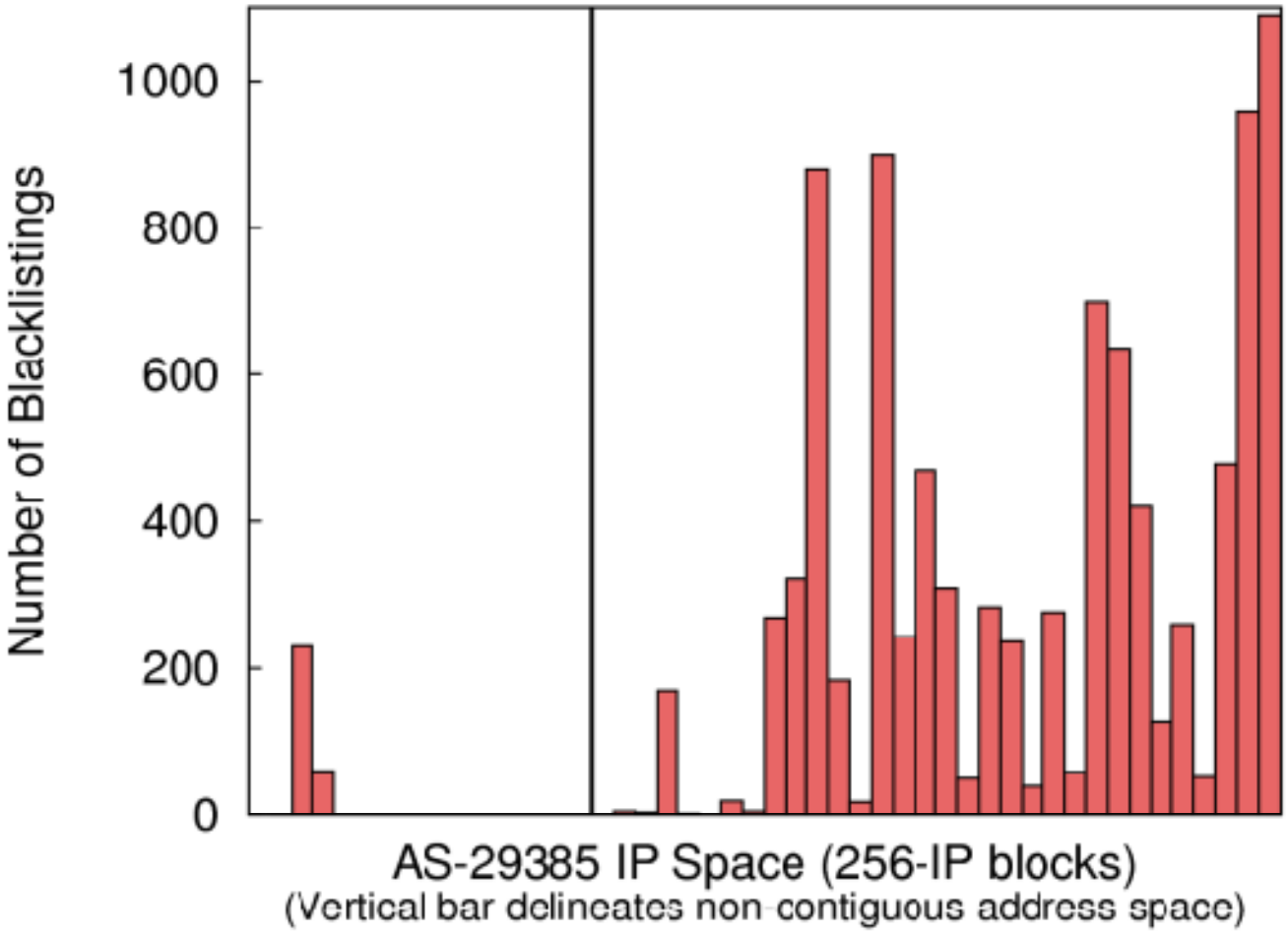
Temporal Props.



Of IPs removed from a popular blacklist, 26% are re-listed within 10 days, and 47% are relisted within ten weeks.

Consistent listing length permits normalization

Spatial Props.



Applying PreSTA to spam mitigation

Implementing the model

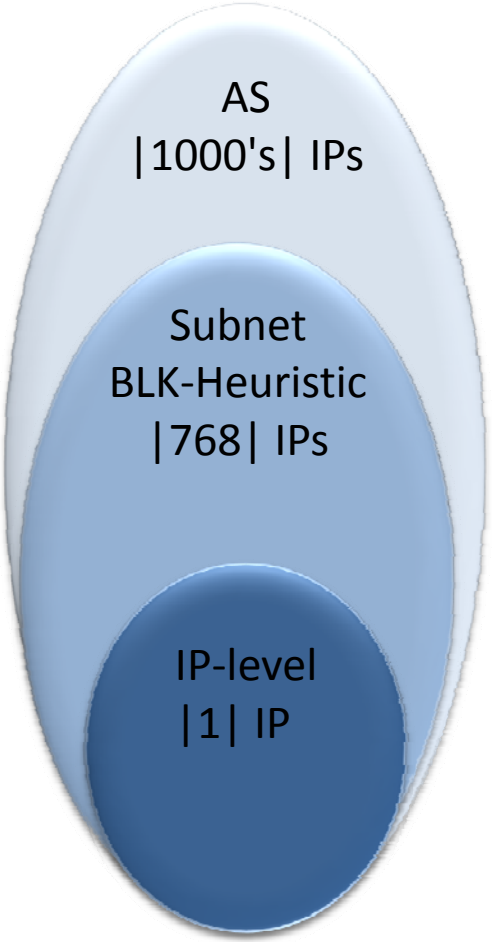
Grouping Functions

~~IANA /RIR~~
- The IANA and RIR granularity are too broad to be of relevant use

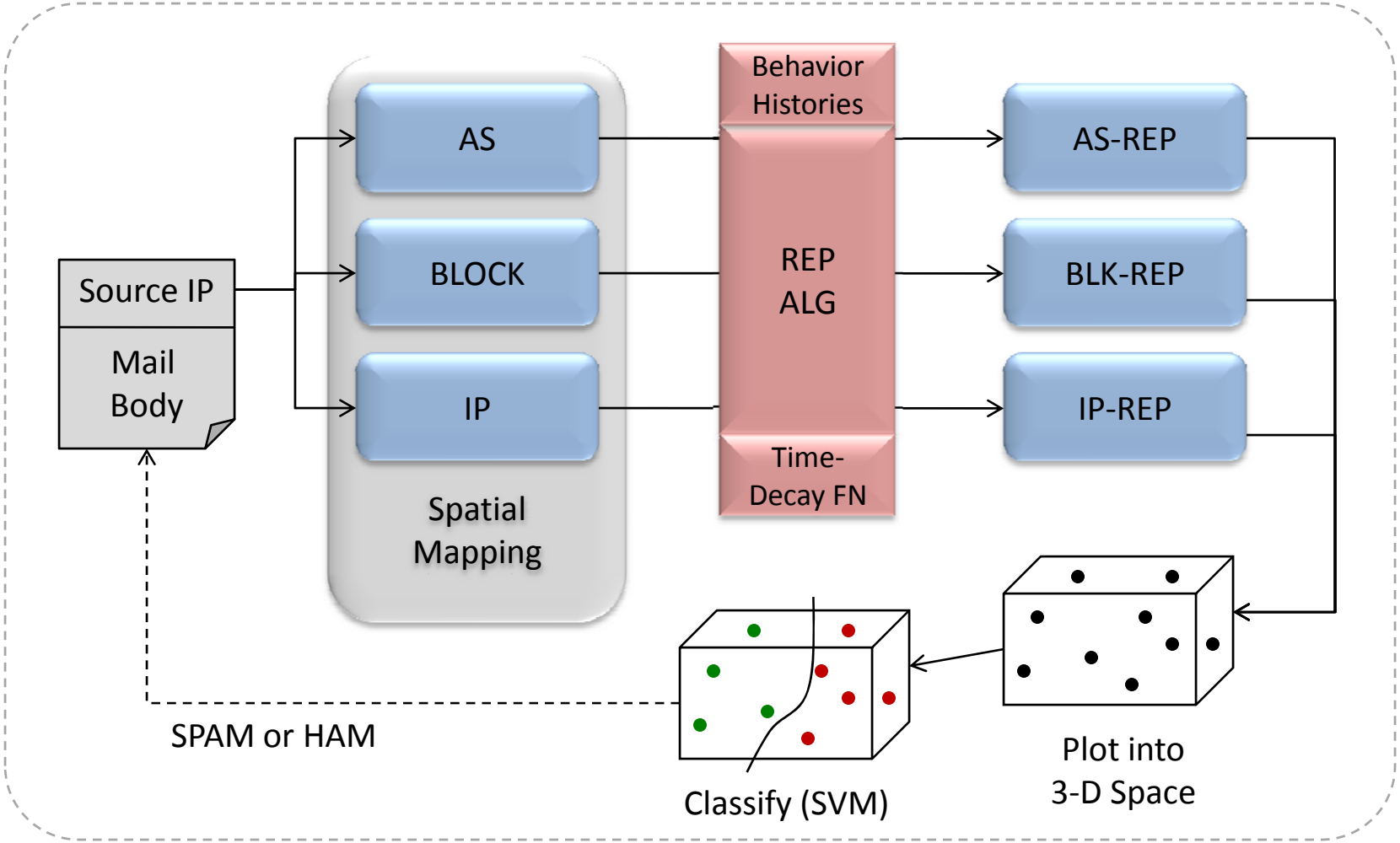
AS
- What AS(es) are broadcasting IP?
- An IP may have 0, 1, or 2+ homes

BLOCK
- What is /24 (256 IP) membership?
- Estimation of subnet

IP
- Static IP addresses
- Due to DHCP; multiple inhabitants



PreSTA Workflow



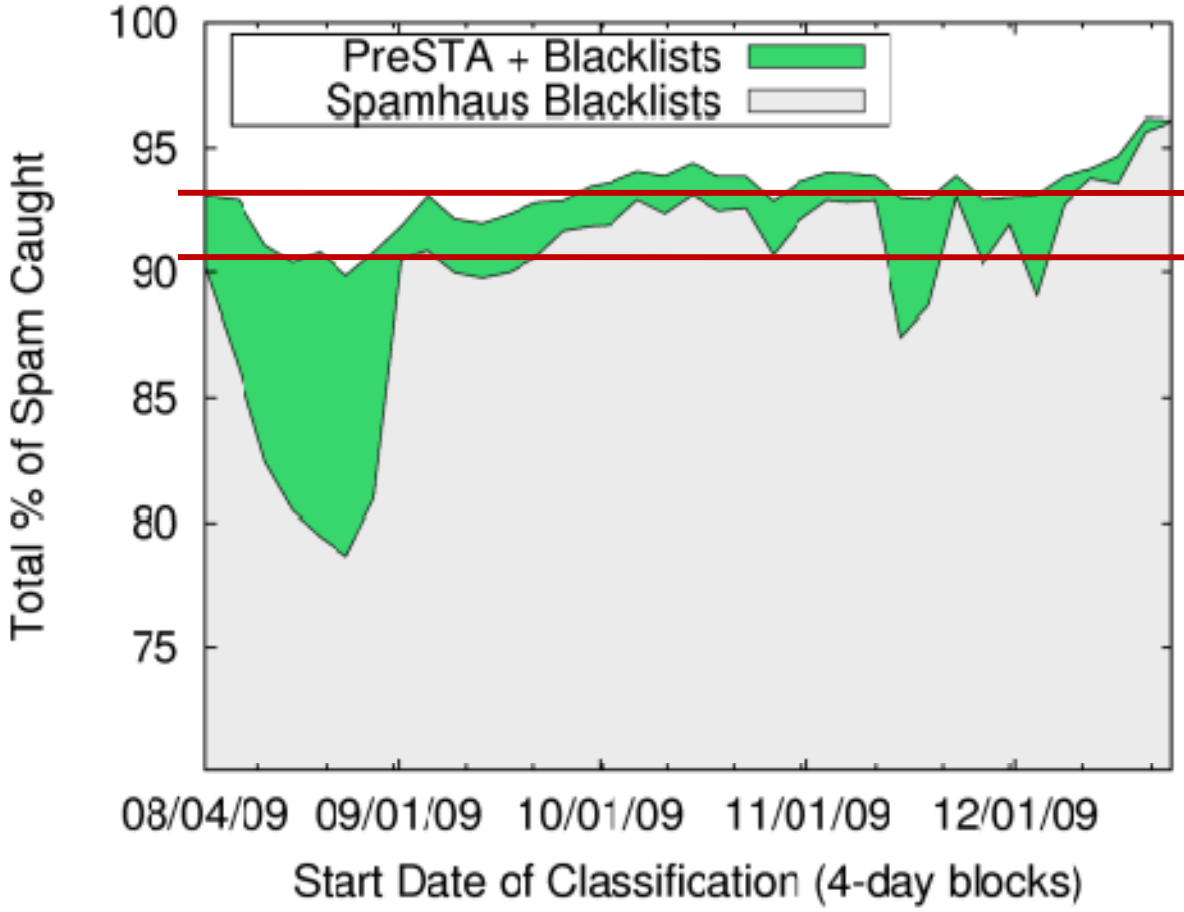
Data Sources

FEEDBACK	• Subscribe to Spamhaus [13] provider • Process `diff` between versions into DB
AS-MAP	• Use RouteViews [14] data to map IP→AS
EMAIL	• 5 months: 31 mil. UPenn mail headers • Proofpoint [15] for ground truth

Applying PreSTA to spam mitigation

Results

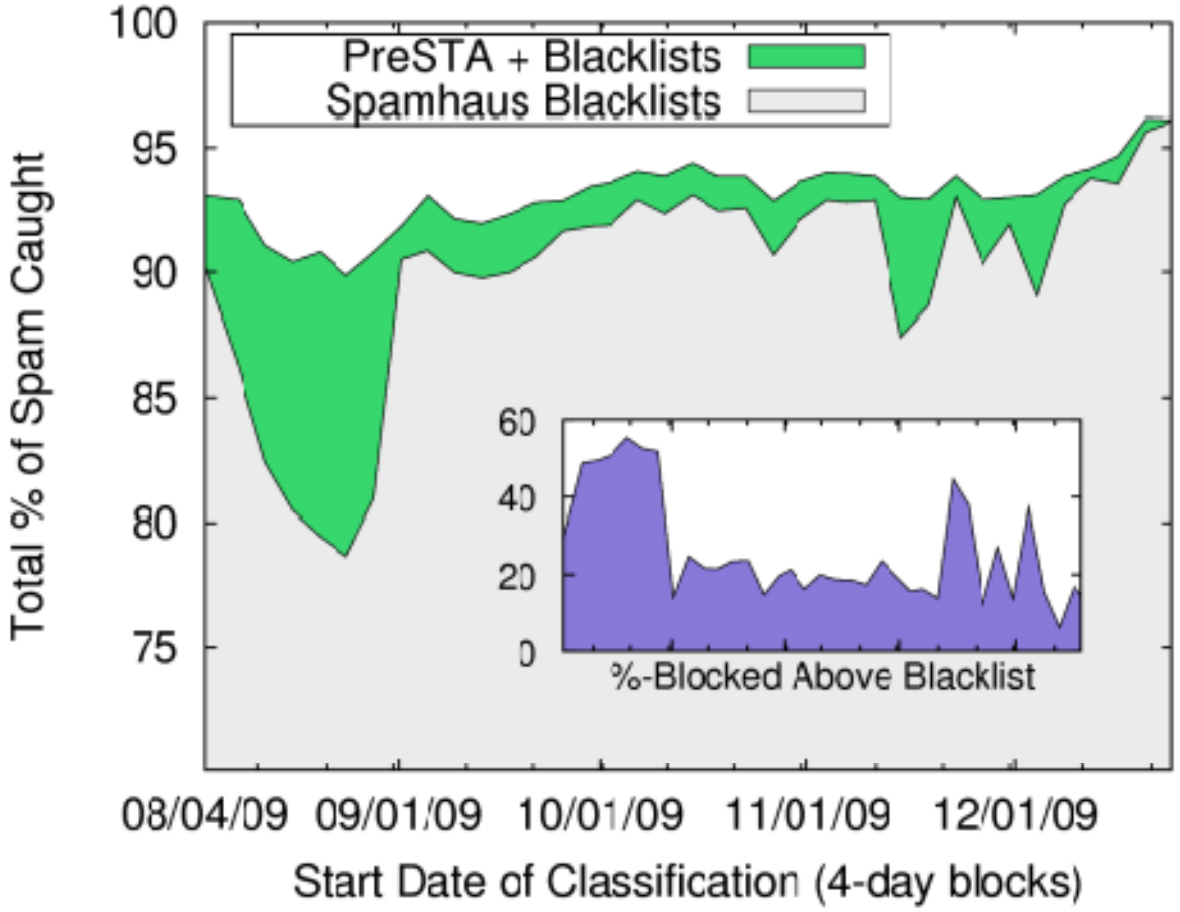
Big-Picture Result



PreSTA + Blacklist:
94% avg.

Blacklist alone:
91% avg.

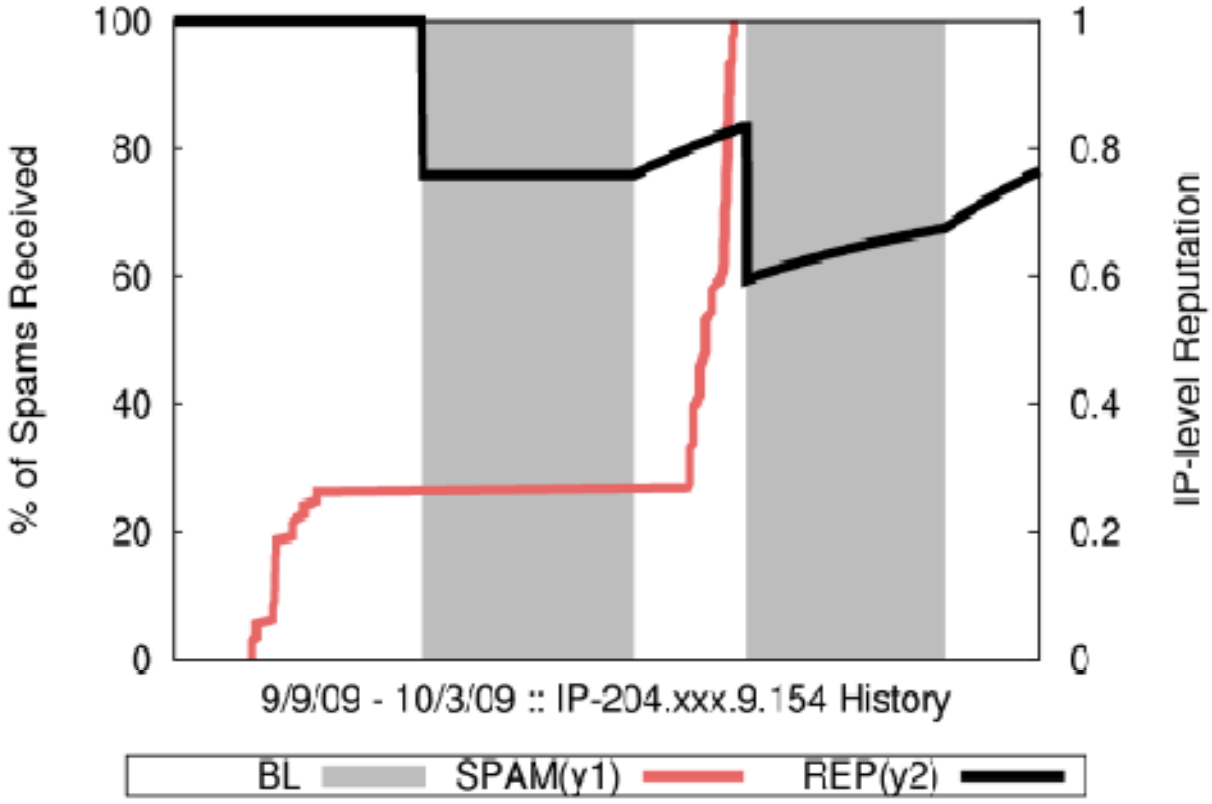
Big-Picture Result



Captures up to 50% of spam mails not caught by blacklist

Would have blocked an addl. 650k spam emails

Case Studies (1)



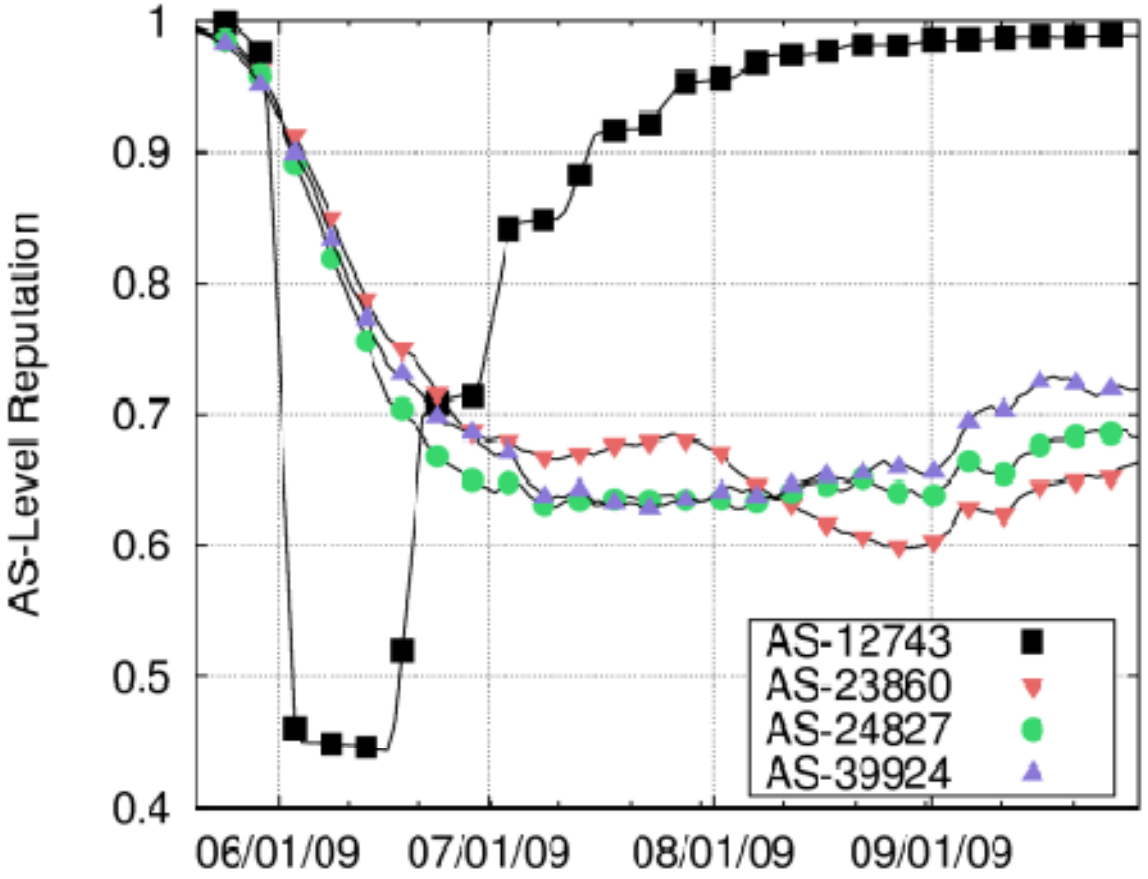
Temporal (single IP) example.

Offender sent 150 spam emails and likely monitored own BL status [16].

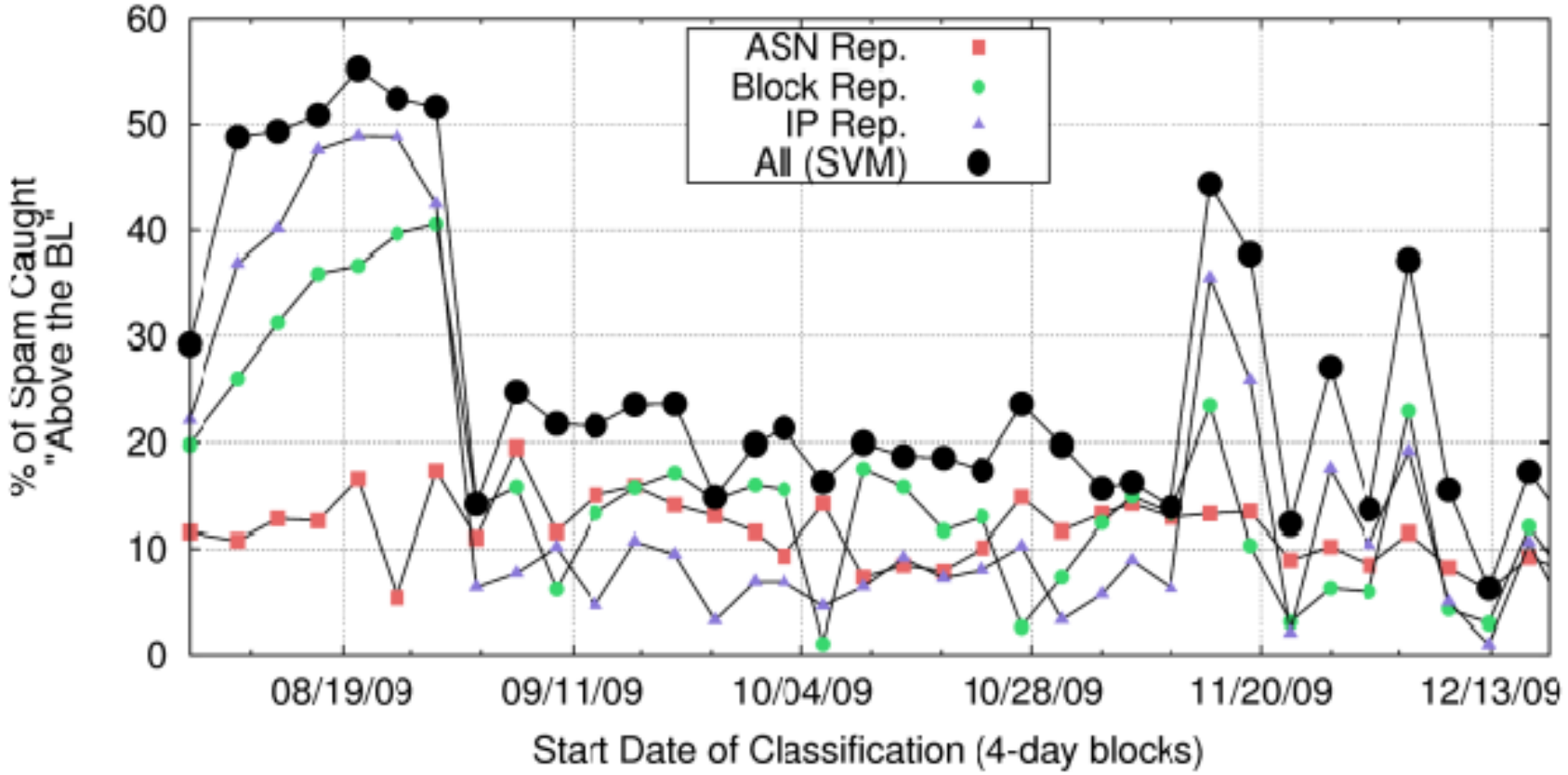
Case Studies (2)

Temporal and spatial example (AS granularity).

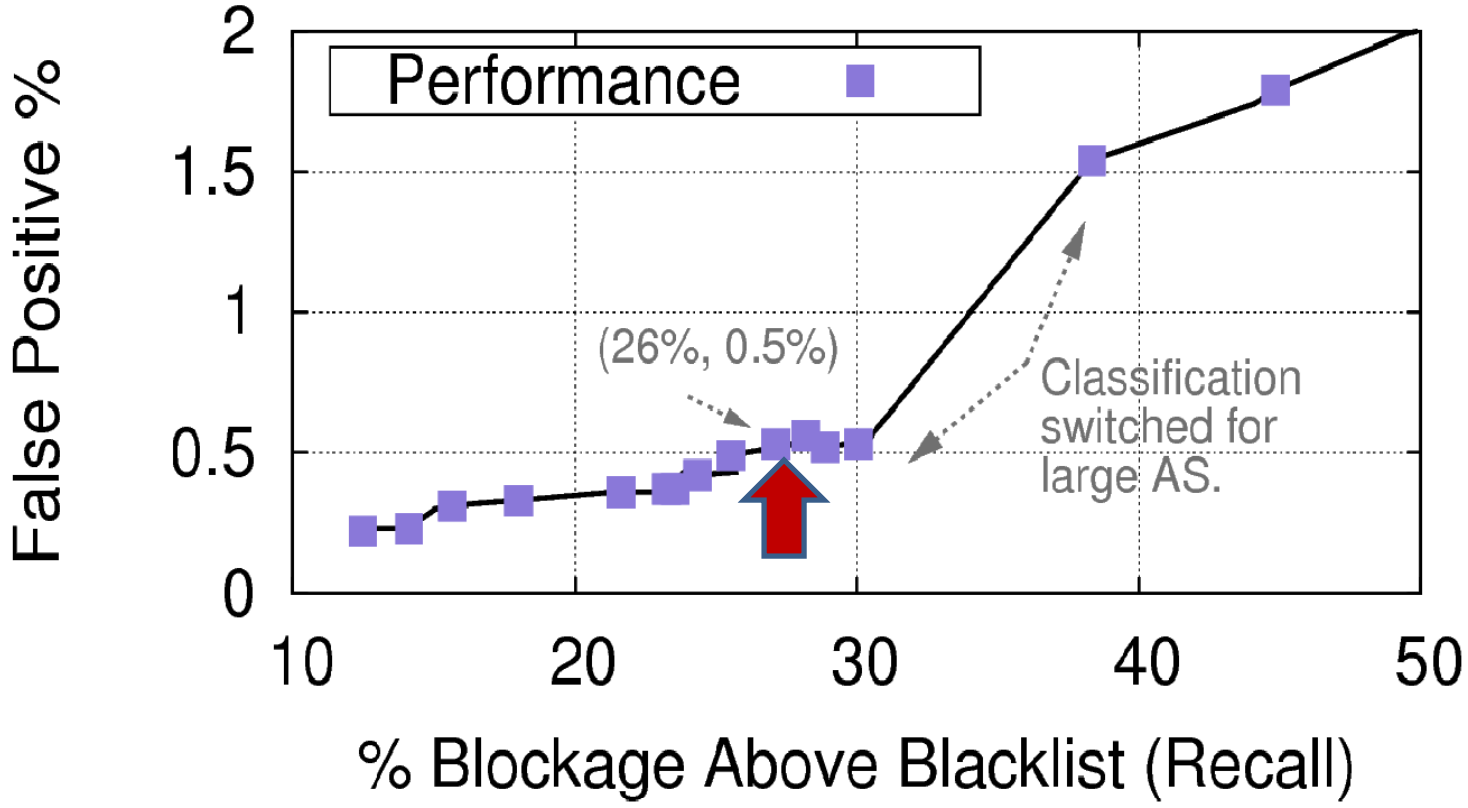
Spam campaign involving 4,500 IP addresses



Other Results (1)



Other Results (2)



Scalability

- Intended Purpose
  - NLP is superior, but computationally expensive
  - Initial and lightweight filter
- Scalability
  - Heavy caching
    - *All* AS-level reputations are cached offline
    - 43% cache hit rate for IPS, 57% for blocks
  - Handles 500,000+ emails/hour (commodity)
  - One month's BL history = 1 GB

Gamesmanship

- Avoid FEEDBACK in the first place
- Temporal evasion? Nothing but patience
- Spatial evasion
  - Move around (reduces IP utility; increases cost)
  - Prefix-hijacking. Fortunately, mostly seen from bogon space [7]. Assign to a special "bad AS"
- Don't let individuals control group size (*sizing attack*) and maintain persistent IDs (Sybil [4])

Contributions

- Formalization of a predictive spatio-temporal reputation model (PreSTA)
  - A dynamic access-control solution for: email, Wikipedia, web-service mash-ups, BGP routing
- Implementation of PreSTA for use as a lightweight initial filter for spam email
  - Blocking up to 50% of spam evading blacklists
  - Extremely consistent blockage rates
  - Scalability of 500k+ emails/hour

References

[1] Kamvar, S.D. *et al*. The EigenTrust Algorithm for Reputation Management in P2P Systems. In *WWW,* 2003.

[2] Jøsang, A. *et al.* Trust Network Analysis with Subjective Logic. In *29th Australasian Computer Science Conference*, 2006.

[3] Alperovitch, D. *et al*. Taxonomy of Email Reputation Systems. In *Distributed Computing Systems Workshops,* 2007.

[4] Douceur, J. The Sybil Attack. In *1st IFTPS*, March 2002.

[5] Krebs, B. Host of Internet Spam Groups is Cut Off. [online] *http://www.washingtonpost.com/wp-dyn/content/article/2008/11/12/AR2008111200658.html*, November 2008 (McColo shut-down).

[6] Krebs, B. FTC Sues, Shuts Down N. California Web Hosting Firm. [online] *http://voices.washingtonpost.com/securityfix/2009/06/ftc_sues_shuts_down_n_calif_we.html*, June 2009 (3FN shut-down).

[7] Hao, S. *et al*. Detecting Spammers with SNARE: Spatio-temporal Network…. In *USENIX Security Symposium*, 2009.

[8] Ramachandran, A. *et al.* Understanding the Network-level Behavior of Spammers. In *SIGCOMM*, 2006.

[9] Ramachandran, A. *et al.* Filtering Spam with Behavioral Blacklisting. In *CCS*, 2007.

[10] Qian, Z. *et al*. On Network-level Clusters for Spam Detection. In *NDSS*, 2010.

[11] Venkataraman, S. *et al*. Tracking Dynamic Sources of Malicious Activity at Internet Scale. In *NIPS*, 2009.

[12] Venkataraman, S. *et al*. Exploiting Network Structure for Proactive Spam Mitigation. In *USENIX Security Symposium*, 2007.

[13] Spamhaus Project. [online] *http://www.spamhaus.org*

[14] Univ. of Oregon Route Views. [online] *http://www.routeviews.org*

[15] Proofpoint, Inc. [online] *http://www.proofpoint.com*

[16] Ramachandran, A. *et al.* Revealing bot-net membership using DNSBL Counter-Intelligence. In *USENIX Security Workshops: Steps to Reducing Unwanted Traffic on the Internet*, 2006.

[17] IronPort Systems Inc.. Reputation-based Mail Flow Control. *White paper for the SenderBase system*, 2002.

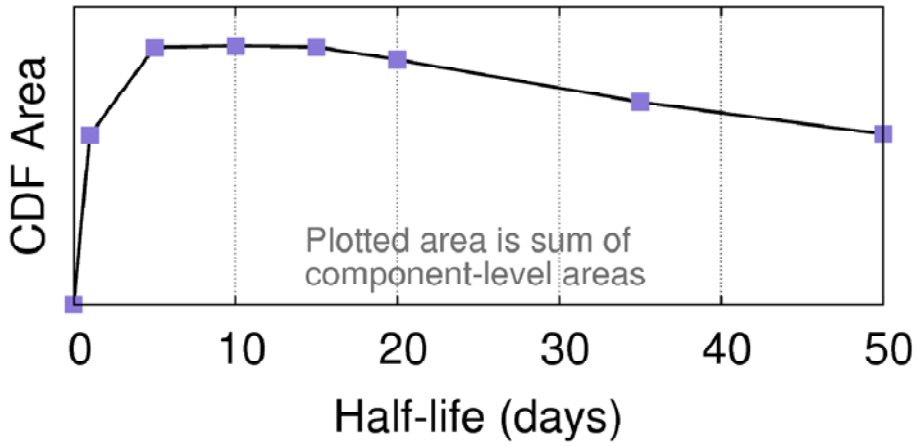
[18] Symantec Corporation. IP Reputation Investigation. [online] *http://ipremoval.sms.symantec.com*

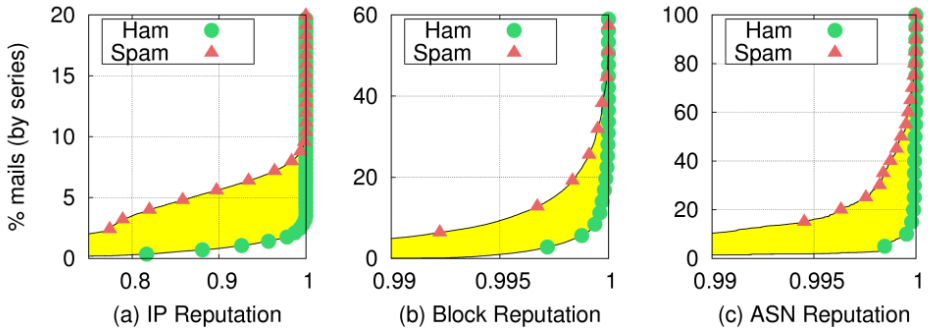
[19] West, A. *et al*.. Detecting Wikipedia Vandalism via Spatio-temporal Analysis of Revision Metadata. In *EUROSEC*, 2010.

[20] West, A. STiki: An Anti-vandalism Tool for Wikipedia. [online] *http://en.wikipedia.org/wiki/Wikipedia:STiki*

Additional slides

Decay Function





- Half-life function is straightforward exponential decay using 10-day half life.
- Half-life was arrived at empirically (sum of CDF areas).
- **TAKEAWAY**: Long punishments lead to few false-positives. Don't led bad guys off the hook too easily!

Related Work

- SNARE (GA-Tech, Hao *et al.* [7])
  - Identifies 13 spatio-temporal metrics → ML classifier
    - Temporally weak aggregation (*i.e.*, mean and variance)
    - "Doesn't need blacklists" → Neither does PreSTA
  - Not scalable. PreSTA uses just 1 metric.
- Similar commercial services
  - Symantec [17] and Ironport SenderBase [18]
  - Closed source, but binary APIs indicate correlation