



Big Data for Security: Challenges, Opportunities, and Experiments

Pratyusa K. Manadhat a

HP Labs

Joint work with Stuart Haber, William Horne, Prasad Rao, and Sandeep Yadav

Big Data is everywhere

The New York Times

NEWS ANALYSIS

The Age of Big Data

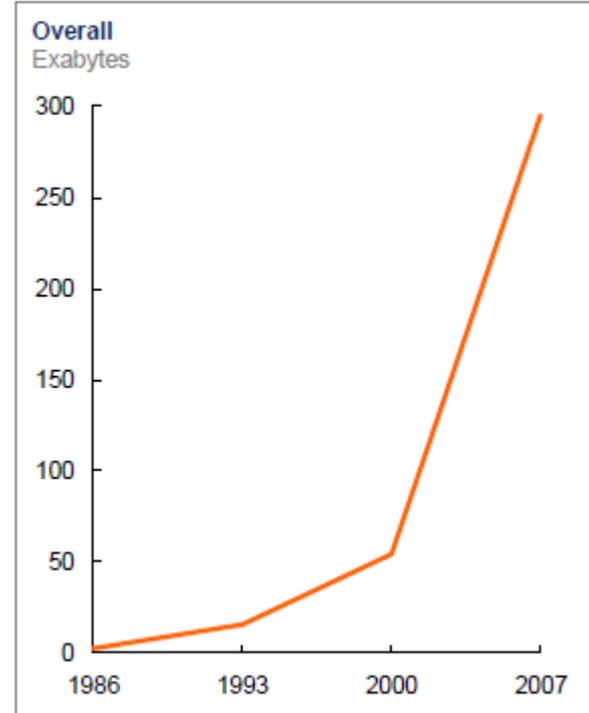
By STEVE LOHR

Published: February 11, 2012



Enterprises collect big data

- Storage is cheaper
- Compliance
 - Audit trails (CFR 11), HIPAA, and SOX



Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011



From 'More is less' to 'More is more'

What can we do with the data?

Algorithms and systems to identify actionable security events from big data.

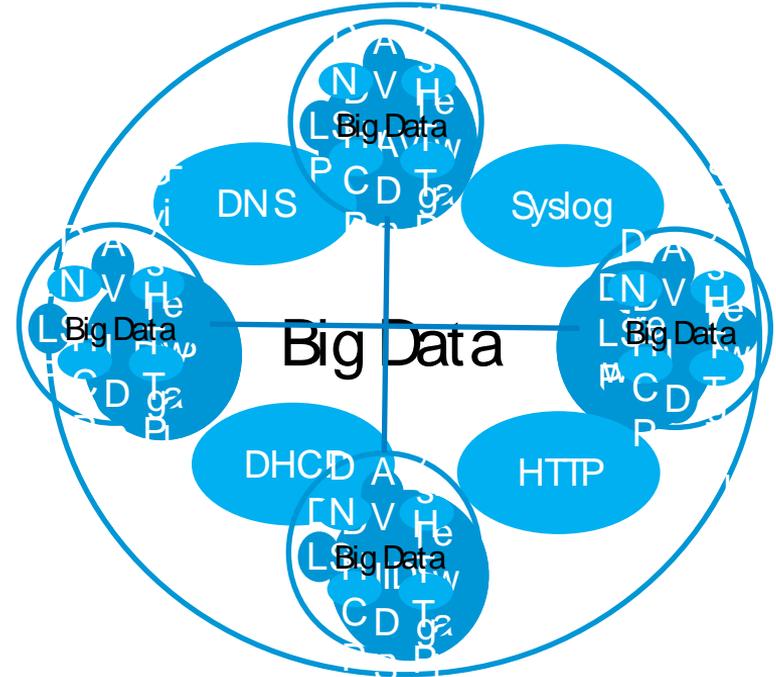


Big data for security

Traditional approach: Point products

Big data: Holistic view of an enterprise

Big data: Global view of enterprises



Challenges

- Data collection and storage – technical, legal, privacy, etc.
- Analysis infrastructure
- Scalable algorithms
- Limitations – what works and what doesn't



Example: Malicious domain detection

Scalable identification of **malware-infected hosts** in an enterprise and of **malicious domains** accessed by the enterprise's hosts



State of the art

- Commercial blacklists
- Traffic analysis
- Machine learning and statistical analysis

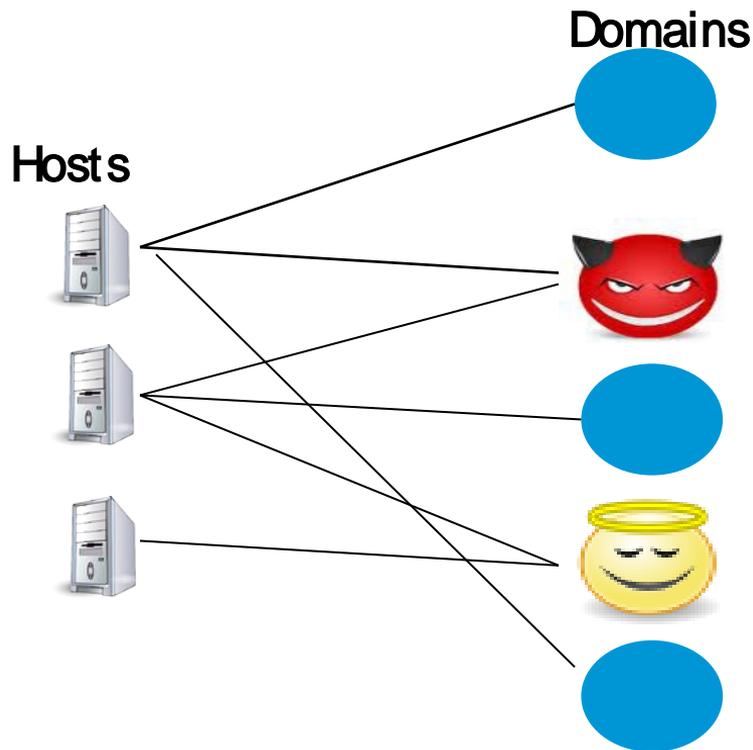


Our approach: Scalable graph inference

Host-Domain graph

Maleasance inference as marginal probability estimation

Minimal ground truth



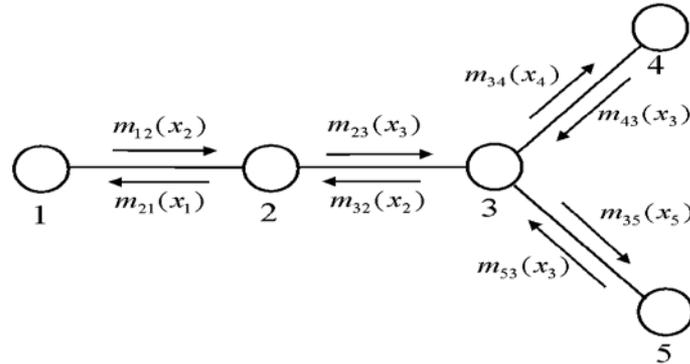
Belief propagation [P82, YFW01]

Marginal probability estimation in graphs

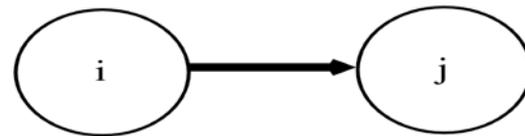
- NP-complete

Belief propagation is fast and approximate

- Iterative message passing



Message passing



Message(*i* → *j*) ∝ (prior, edge potential, incoming messages)

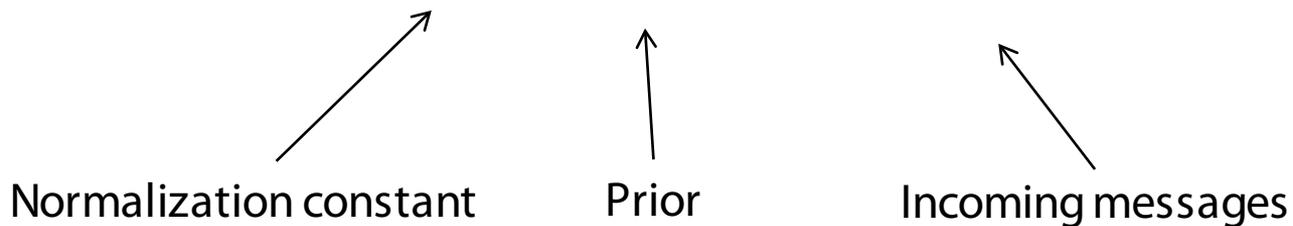
$$m_{ij}(x_j) = \sum_{x_i \in \mathcal{S}} \phi(x_i) \psi(x_i, x_j) \prod_{k \in \mathcal{N}(i) \setminus j} m_{ki}(x_i)$$

↑ ↑ ↑
Prior Edge potential Incoming messages

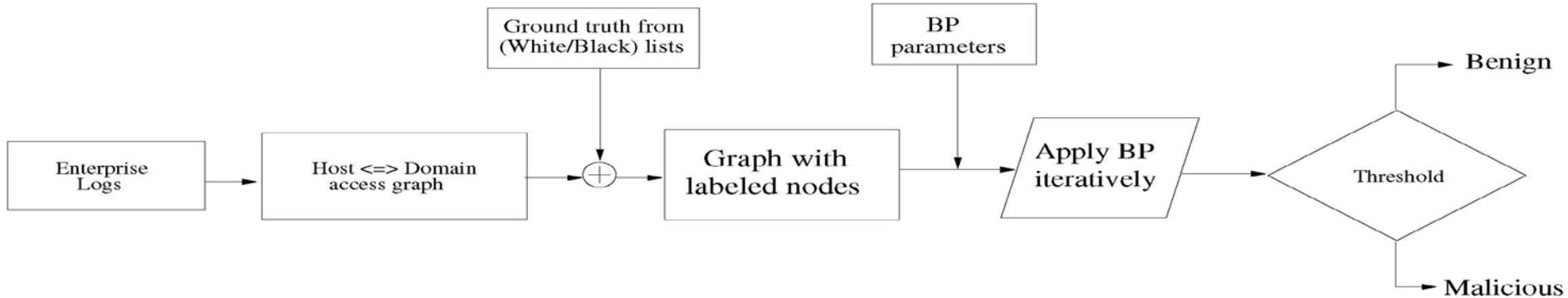
Belief computation

Belief(i) \propto (prior, incoming messages)

$$b_i(x_i) = K \phi(x_i) \prod_{j \in N(i)} m_{ji}(x_i)$$



Our approach



Experimental evaluation

- Graph “completeness”
- Size of ground truth data
- Two class vs multi class classification
- Homogeneous vs heterogeneous data
- One enterprise vs multiple enterprises



HTTP Proxy logs and DHCP logs

Logs from a large enterprise

- 98 HTTP proxy servers and 6 DHCP servers world wide
- 1 day's logs : 2 billion events
- 144K hosts, 1.28M domains, and 12M edges

Priors from ground truth (0.4% nodes)

- 3K known bad domains: 0.99 (TippingPoint)
- 3K known good domains: 0.01 (Alexa)
- Unknown hosts and domains: 0.5

Edge potential

	Benign	Malicious
Benign	0.51	0.49
Malicious	0.49	0.51



BP scales to enterprise settings

Java implementation of BP

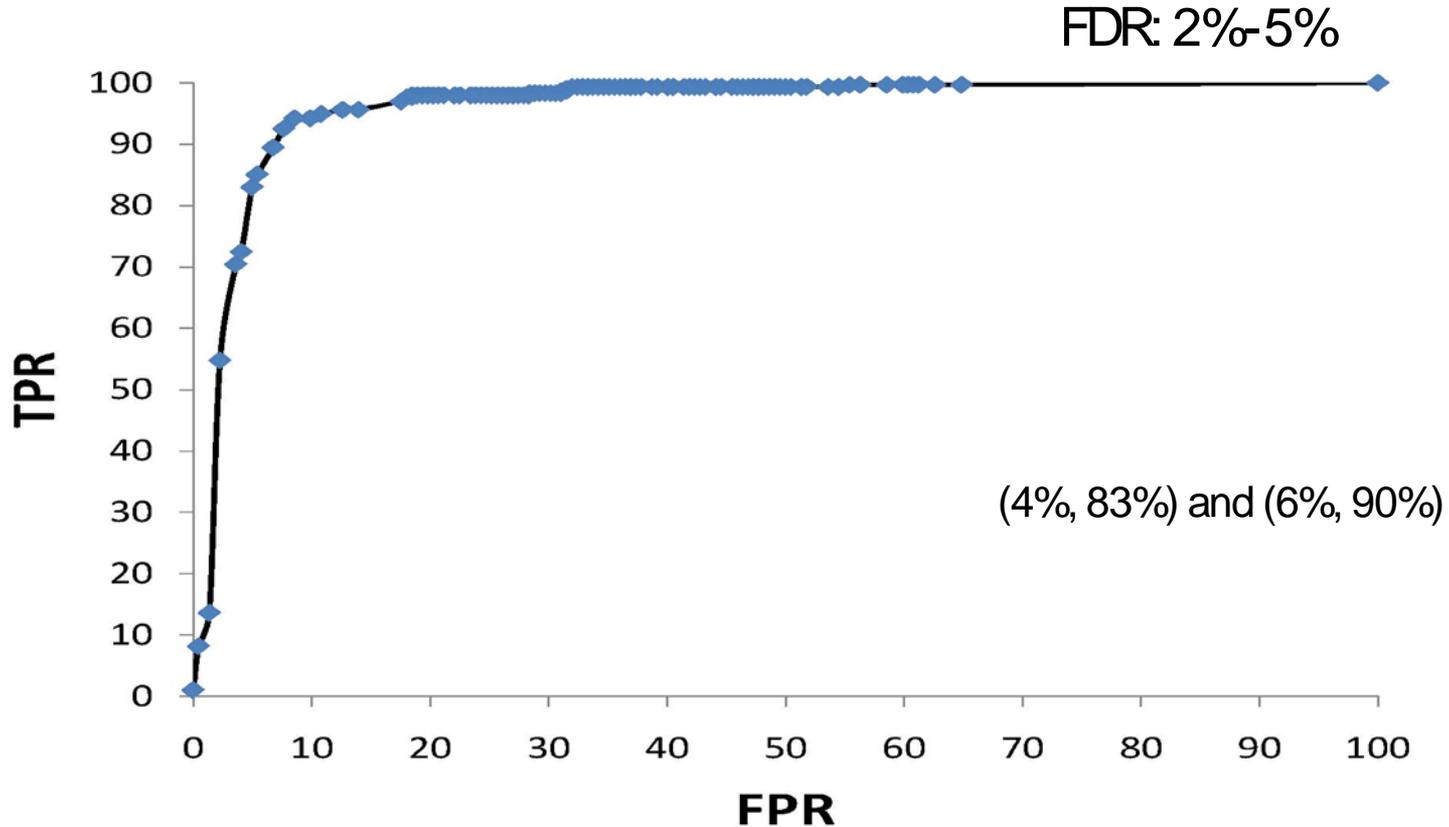
12 core 2.67GHz desktop with 48GB RAM

Each iteration takes 2-4 minutes

Converges in 13 iterations



Malicious domain detection ROCPlot



DNSrequest logs

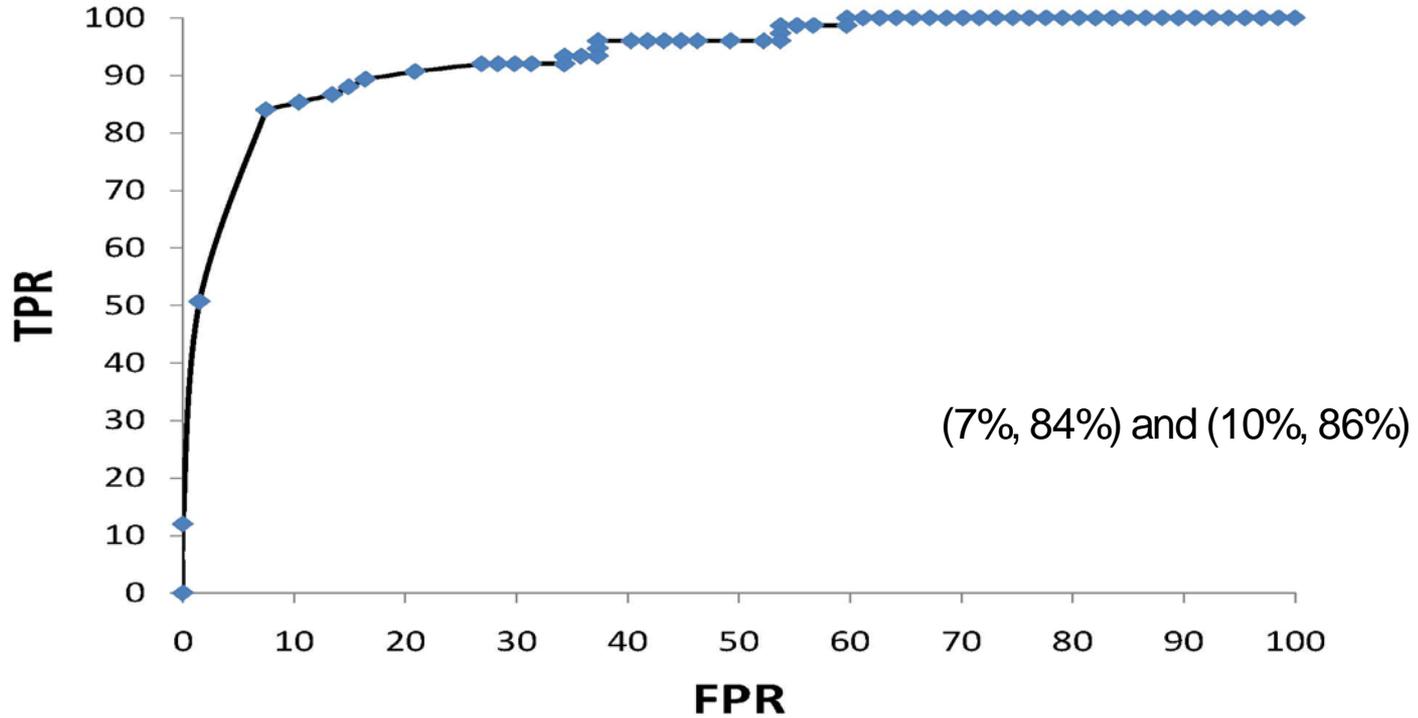
Collected from a medium sized ISP

- 2 Gbps packet captures
- 1 week's data: 1.1 billion DNS requests
- 927K hosts, 1.32M domains, and 12M edges

Priors and edge potential similar to HTTP data



DNS ROCplot



IDS alerts logs

Collected from 916 enterprises worldwide

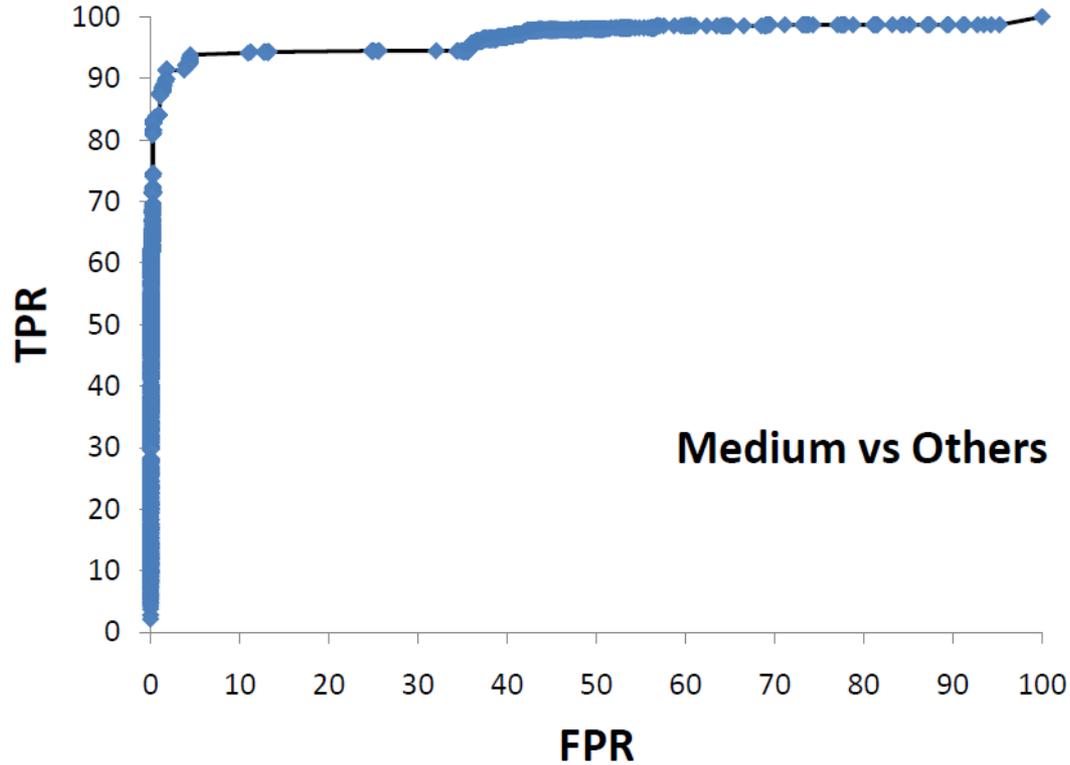
- 5 years' data: 15.5 billion alerts
- 3.1M internal nodes, 3.69M external nodes, and 21.4M edges

Classify nodes into 4 classes

- IDS expert annotates 400 IDS signatures
- Assign priors according to alert classes (6.6% nodes in ground truth)
- Edge potential according to homophilic relationship



IDS logs ROCplot



Summary of results

Works well when the graph is complete (IDS), does poorly when the graph is incomplete (DNS)

Requires minimal ground truth (HTTP), but more ground truth data is better (IDS)

Works in multiclass settings (IDS)

Can handle heterogeneous data (DNS) and from multiple enterprises (IDS)

Discovers genuine anomalies, manually verified (HTTP)



Future work

Combine different data sources

e.g., use IDS logs as ground truth for DNS logs

Root cause analysis



Thank you



manadhat@hp.com