# Project Bloom: Empowering the Security Research Community Through Data Products and Computing

Minaxi Gupta — School of Informatics and Computing
Greg Travis, David Ripley — ANML
Doug D. Pearson — REN-ISAC

Indiana University, Bloomington

# The Security Landscape

- Cybercrime involves a thriving multi-billion dollar underground economy
- Websites connected to phishing, malware, and scams spring up by the millions every day
- Creative spam, social engineering, and search-engine optimization techniques lure users to malicious websites
- Bots are a major enabler of cybercrime
- Security researchers need good data to study trends and devise effective defenses

# Impediments to Data Availability

- Good data sets are either not available or available only within a closed group
- This hurts new progress and verification of previous results
- Even when data is available, it is rarely documented well or available for long durations
- Data procurement process is long or riddled with legal hassles
- Even when data is available, storing and computing on it, and indexing it for the future, requires resources rarely available at a single institution

# Goals of Project Bloom

- Goal 1: to provide well-curated, long-term, raw data sets to the security research community in an efficient manner using an established federation that can make the data acquisition process almost instantaneous
- Goal 2: to provide researchers compute power close to the data sources so they can avoid moving large data sets across the Internet and can focus only on taking back the key derived results
- Goal 3: to offer rich data products derived from raw data so researchers can bypass common data-processing hurdles and be more productive

# Partnerships

- Initial focus on three commonly-used data types: NetFlow, Darknet, Passive DNS
- To ensure quality and long-term availability, they are not community driven as of now
- Raw data will be kept for 10 years on a rolling basis
- Partners:
  - Indiana University:
    - REN-ISAC: 266 higher-education member institutions, some offer their darknet data
  - Internet2: NetFlow data from 300 member institutions
  - ISC/SIE: Passive DNS data from 15 ISPs around the world

# Data Set 1: NetFlow

- Information contained in each NetFlow record: IPs, ports, protocol, bytes, packets of a *flow*, TCP flags, neighboring ASNs, routing information
- Widely used for traffic engineering, network provisioning, and for identifying security and performance problems
- A commonly-used source of high-quality NetFlow data is the Internet2 Observatory
- Limitations of the Observatory's offering:
  - Lengthy proposal process required to access data
  - Data is masked due to privacy concerns
  - Data is sampled
  - No data products or computing facility to enhance utility

# NetFlow Under Bloom

- Data sources:
  - Internet2 Observatory
  - Indiana University
- A comprehensive technical, policy, and federation framework to cut down access times
- Time-limited unmasked data so conclusions about individual IPs can be drawn by trusted researchers
- Unsampled flow records from Indiana University
- Rich data products and parallel-computing facilities to process data

# Data Set 2: Darknet

- Darknet: Allocated, advertised, but unused IP address space
- Packets coming to a darknet are unsolicited and may indicate malicious activity. Ex: Backscatter, scanning packets
- State of data availability:
  - IMS: not available publicly
  - Team Cymru: offers software to allow organizations to monitor data
  - UCSD's /8: CAIDA plans to offer it under PREDICT, offers limited view
  - Others: not widely known or available

# Darknet Under Bloom

- Data sources:
  - Midwest: six /24s
  - East coast: 2 /24s
  - South: four /24s
  - West coast: UCSD's /8
  - Australian subcontinent: two /24s
  - More to be added
- Rich data products and parallel-computing facilities to process data

# Data Set 3: Passive DNS

- Passive DNS: Responses obtained by the local DNS servers from (authoritative) DNS servers around the world
- When collected globally, passive DNS data offers a window into access patterns for hosts and domains around the world, including malicious ones
- State of data availability:
  - ISC/SIE: lengthy data procurement process, expensive
  - Others: mostly local, do not offer a global view

# Passive DNS Under Bloom

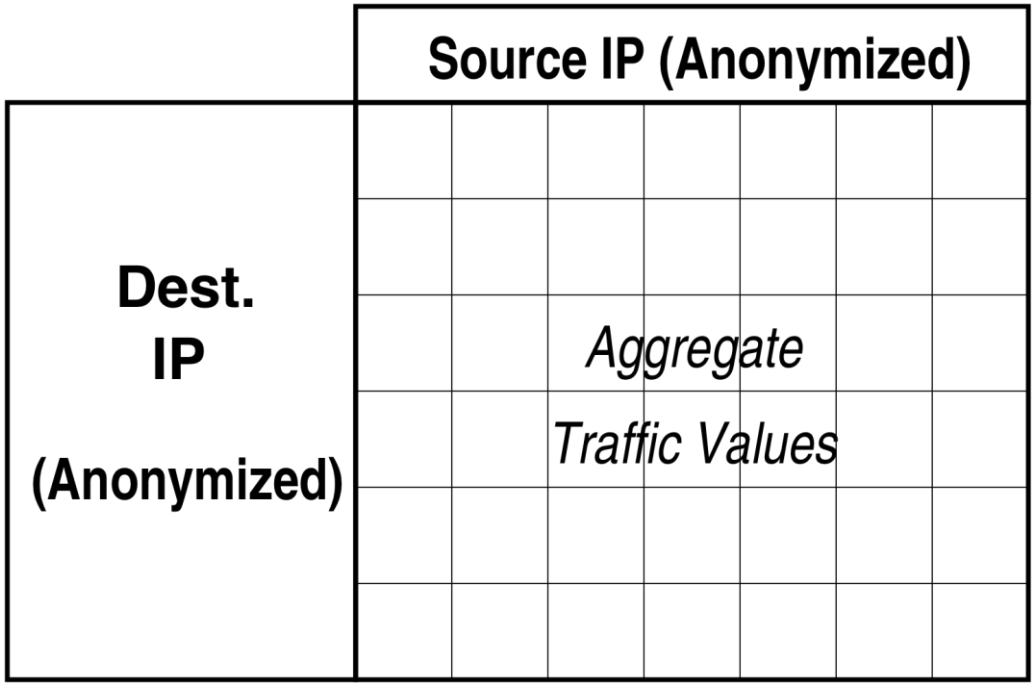
- Data sources:
  - ISC/SIE: collects passive DNS data from 15 large ISPs and commercial DNS service providers around the world, including a Tier1 ISP, two US Cable/DSL providers, and four US-based universities
  - DNS History Database Project (DHDB): collects passive DNS data from 8 data sources, including University of Auckland (New Zealand), France, and Norway
- Rich data products and parallel-computing facilities to process data

# Data Products

- Fall in two categories:
  - based on individual data types
  - derived by combining data types
- Initial offering is composed of 12 products
- To derive them, we surveyed:
  - how researchers have used NetFlow data available through the Internet2 Observatory
  - CAIDA's offering of various data sets
  - research papers using darknet and passive DNS data
  - DHS, I3P, and Internet2 workshops that discussed cybersecurity data needs

# Data Products: Example 1

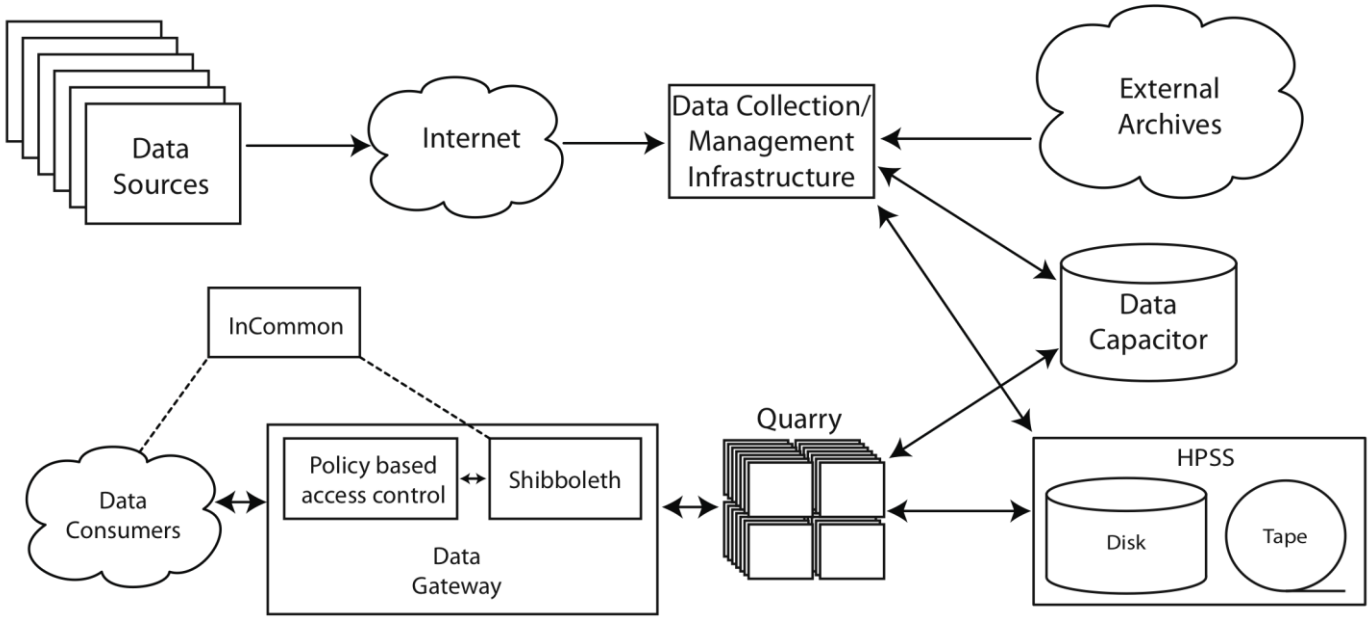
- Aggregate TCP traffic between each source and destination IP at a 15-minute granularity
- Can be combined to infer aggregate traffic for an IP address or BGP prefix at a daily, weekly, or monthly granularity
- 30 days worth of matrices would require 3TB storage



# Data Products: Example 2

- Internet domains fall under gTLDs (.com, .net, etc.) and ccTLDs (.kr, .cn, etc.)
- Zone files enumerating domains within a TLD are a useful tool to investigate DNS aspects of malicious websites
- Prominent gTLD zone files are available but they cover only about 50% of Internet domains
- ccTLD zone files are not available, preventing researchers from obtaining a global view of malicious domains
- Using passive DNS and darknet data, we will generate nightly zone files for all TLDs, including ccTLDs, and archive them for 10 years on a rolling basis
- This data product would require 100TB of storage!

# Architecture and Indiana University Resources




- Quarry: IBM e1350 distributed shared-memory supercomputer cluster with 1120 processor cores sharing 1.2TB of RAM
- Data Capacitor: 1PB of inline mass storage connected
- Massive Data Storage Service (MDSS): 2.8PB of magnetic tape and disk storage
- All connections are 20Gb/s Ethernet
- Access to Bloom will be through a Shibboleth instantiation, *InCommon*, already in use by Internet2

# Conclusions

- Raw data and data products combined warrant 1PB of storage and this will increase with Internet traffic
- Access to Indiana University resources is not exclusive
- Financial support is needed to
  - provision adequate compute and storage for the research community
  - instantiate and maintain Bloom infrastructure
  - program and offer data products
  - evolve data products based on community feedback
  - administrative maintenance
- Your feedback and support is crucial!

# Questions/Thoughts?

Contact Gregory Travis at:

greg@iu.edu

812 855 5091