ForNet: A Distributed Forensics Network

Nasir Memon

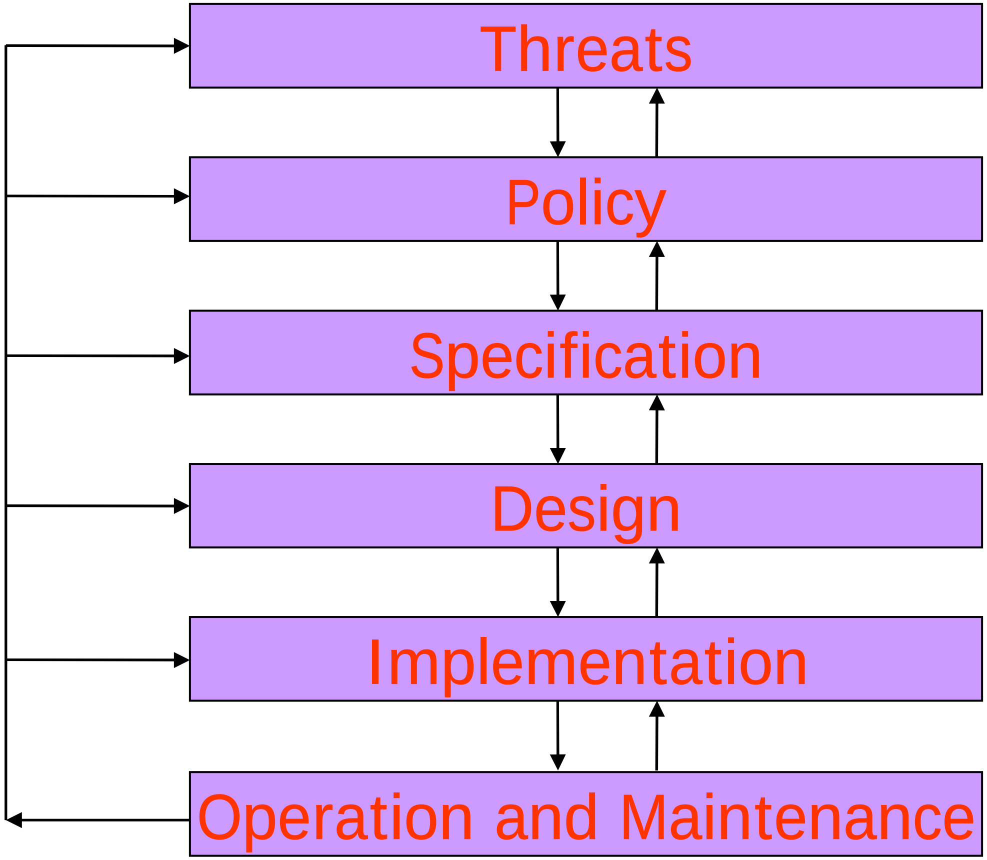
Department of Computer and Information Science

Polytechnic UNIVERSITY

Outline

- Motivation
  - Security fails. How often? Why?
- Infrastructure & Response Model
  - What is our safety net?
  - Is it enough?
  - What do we need?
- Introduction to ForNet
  - What is ForNet?
  - What are synopses?
- Research Challenges & Future Directions

The security life cycle



“To err is human, to really foul things up you need a computer.”

Is security important?

Slammer hits Davis-Besse Nuclear Power Station:

> "[...] But an incident in January at the Davis-Besse Nuclear Power Station, run by the FirstEnergy Corporation outside Toledo, Ohio, showed that this was not always the case. The nuclear plant has not been generating power since early 2002, but a computer system there that was not supposed to be linked to the Internet was invaded by a worm known as Slammer, causing the system to shut down for five hours. The event was not made public until Kevin Poulsen reported it on Aug. 20 on SecurityFocus .com ..." *New York Times, September 7*

Sobig affects Amtrak trains, Air Canada

> "[...] a spokesman for CSX, said the company noticed Wednesday at about 1:15 a.m. that a variant of the Blaster virus was interfering with its train operations and dispatching system."
>
> "[...] A variant of the Blaster virus on Tuesday affected about half of Air Canada's phone-reservation capacity and some of its airport check-in operations, said spokesman John Rebel." *Wall Street Journal, August 21*

AT&T to invest $3 billion in 2003 for global network

> *"AT&T will spend US$3 billion in capital expenditures this year to completely transform its global network from having a voice-based carrier infrastructure into a single Internet Protocol (IP)-based network..." InfoWorld, September 11*

How often does it fail?

- CSI FBI Survey released in 2003
  - 56% of participants reported incidents (lower), 15% don’t know
  - $201M in financial damages (lower). Average reported loss $2.7M.
  - Theft of proprietary information caused greatest loss
  - Computer viruses 82% of attacks
  - Insider abuse 80% of abuses

Why? A lack of effort?

- Why so many incidents? Are we not devoting enough resources to security?
- In the same survey we had…
  - 99% of respondents use anti-virus
  - 98% of respondents use firewalls
  - 93% had access control mechanism
  - 73% of respondents use IDS

Then why so many failures?

- Security is hard. No system is 100% secure.
- Many points of failure
  - Incorrect assessment (or re-evaluation) of threats
  - Incorrect or incomplete specifications
  - Poor design
  - Buggy implementations
  - Sloppy procedures
  - Malicious insiders
  - Lack of education
  - and so on…

What when it fails?

- Incident Response:
  - If problem due to known vulnerability, 93% patched the system
  - 30% reported to enforcement
  - 50% did not report to anyone
  - 21% reported to “legal counsel”
- The vast majority of times, perpetrator of attacks goes unpunished

What happens when security fails in the physical world?

- Various tools developed to help us “solve the crime”
  - Surveillance mechanisms
    - Video cameras
  - Forensics analysis
    - DNA, Fingerprints etc.
  - Intelligence gathering
    - Informers
- Justice system to examine evidence and assign punishment if guilt ascertained
  - Benefit of doubt given to suspect
- More or less the system works

What about the cyber world?

- Most “crimes” go unpunished
  - The source of most viruses never identified
- Poor enforcement infrastructure
  - Laws and international treaties evolving
- If a Trojan Horse were to be smuggled into your network and triggered, would you be able to track it back?
- Would you be able to track it down if it were an inside job?

Typical response scenario to an intrusion

1. Adversary violates security policy
2. Some event leads to identification of violation and maybe potential violator
3. Find out where the adversary came from (log files if still there)
4. Picks-up the phone & call the ISP, FBI
5. ISP/FBI notifies the other end
6. Go Back To Step 2!!

What do we need?

- A reliable and efficient mechanism for attribution
  - IP networks are anonymous
- An effective response model
  - Current response model is mostly manual
  - Response time is in days or weeks
  - Digital evidence disappears quickly
- Tools that complement security components
  - Need some safety net to fall back
  - Forensics is that net
- Building support for forensics in networks is a good place to start

The Art and Science of Forensics

The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations.

Crimes that “use” the network

- In addition to investigating attacks on the resources of the network, forensics ability is also needed for crimes that use the network
  - Lowe’s pipe bomber case
  - Child pornography
  - Communications by terrorists
  - And so on …

What forensics mechanisms do we have?

- Logging mechanisms and audit trails
  - Logs from Security Components:
    - Logs only perceived security threats
  - Host Logs:
    - First thing to get disabled during an attack
    - An insider would rather use a host without any logs
    - Mobility, wireless networks create new problems
  - Packet Logs:
    - Usually at the edges hence blinded easily
    - Can’t keep data for long…
    - E.g: Infinistream, NFR, NetWitness

Challenges facing network forensics

- Lack of infrastructure for forensic data collection, storage, and dissemination
- Growth of network traffic outpaces Moore's law making prolonged storage, processing, and sharing of raw network data infeasible
- Most of the process is manual and spans multiple administrative domains making response times undesirably long (digital evidence disappears quickly)
- Inability of current logging mechanisms to help forensic analysts explore networks incrementally
- Unreliable logging mechanisms on hosts. Growing support for mobility makes it difficult to maintain prudent logging policies on hosts

One solution to support network forensics …

- Let the network securely collect, store, disseminate, and process *synopsis* of network traffic
- Give networks the ability to remember network events so that they can answer questions like:
  - Where did a worm appear first?
  - Who sent this (possibly spoofed) packet?
  - Where else was this packet observed on the network?
- Goal of **Project ForNet**: development of tools, techniques, and infrastructure to aid rapid investigation and identification of cyber crimes

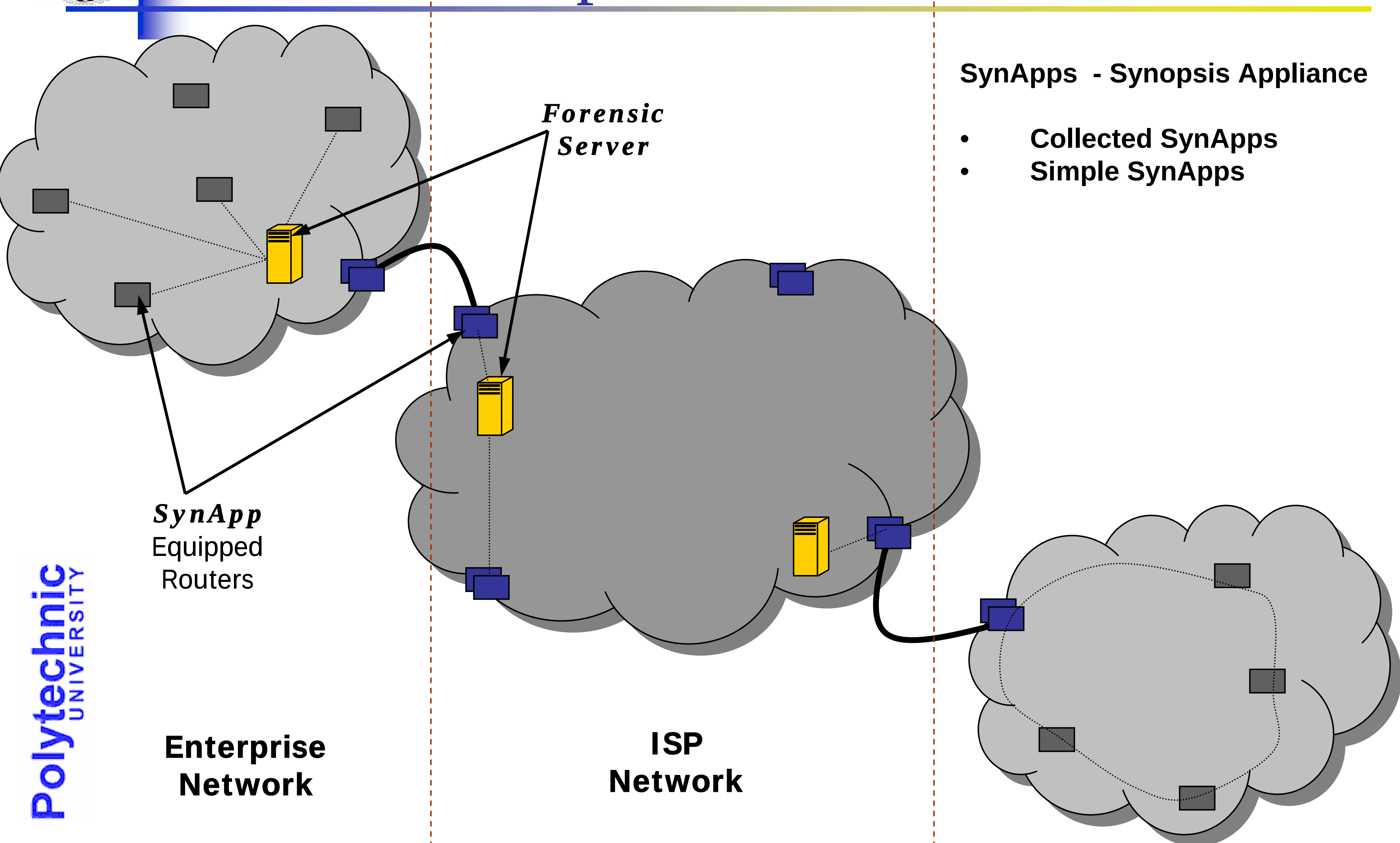
ForNet Design Goals

1. Capture complete & correct evidence
   - To be able to keep up with ever increasing network speeds
2. Longevity & accessibility of evidence
   - Captured evidence must be stored for a prolonged period of time, longer the better
3. Security & privacy of evidence
   - Integrity of collected evidence must be preserved
   - Privacy of users must also be preserved

ForNet Design Goals

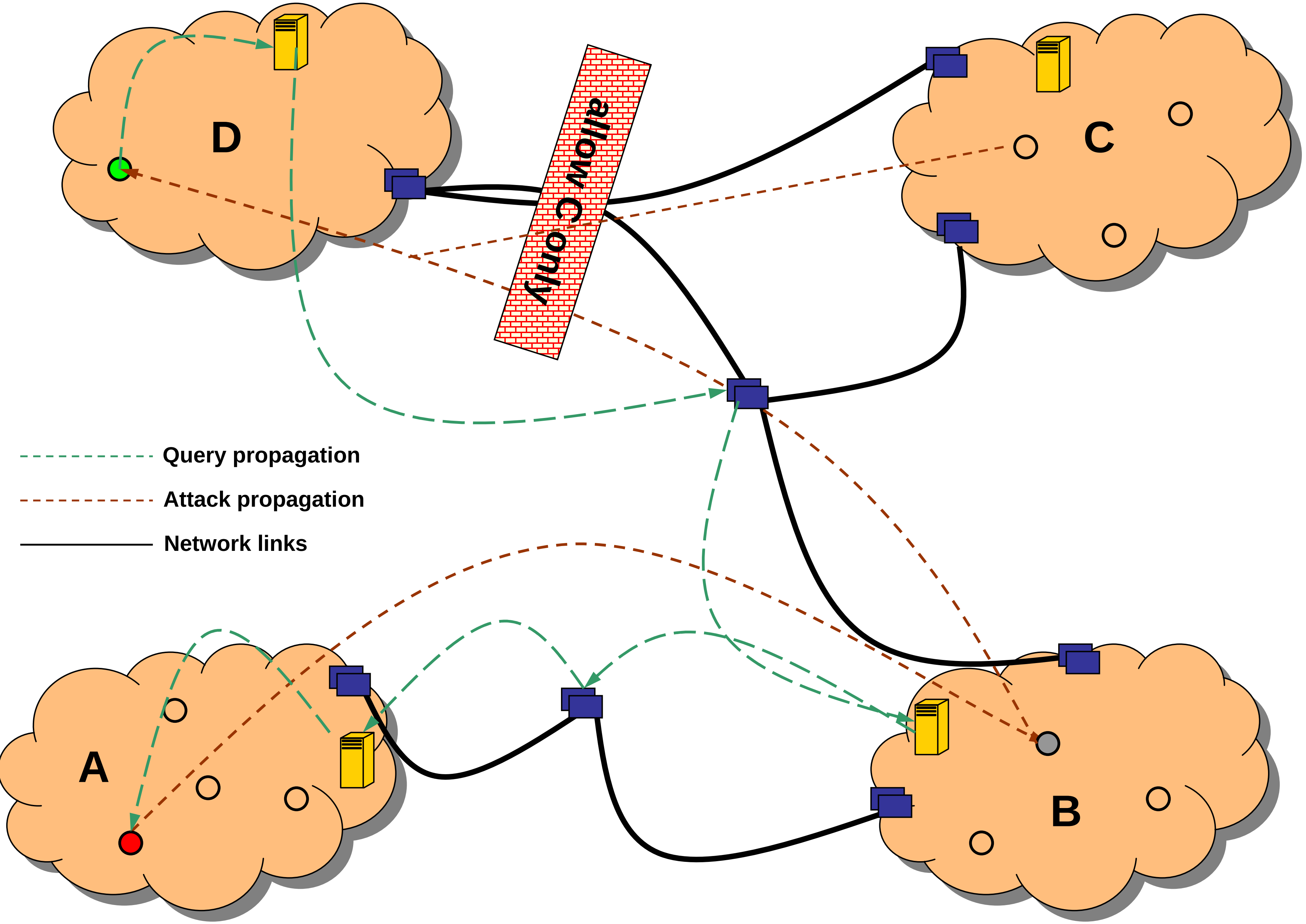
4. Ubiquity & incremental deployment
   - Design should be such that it can be seamlessly integrated into existing network components
   - Because, ubiquitous presence of evidence collection guarantees complete gathering of evidence
5. Modular & scaleable design

ForNet Blueprint

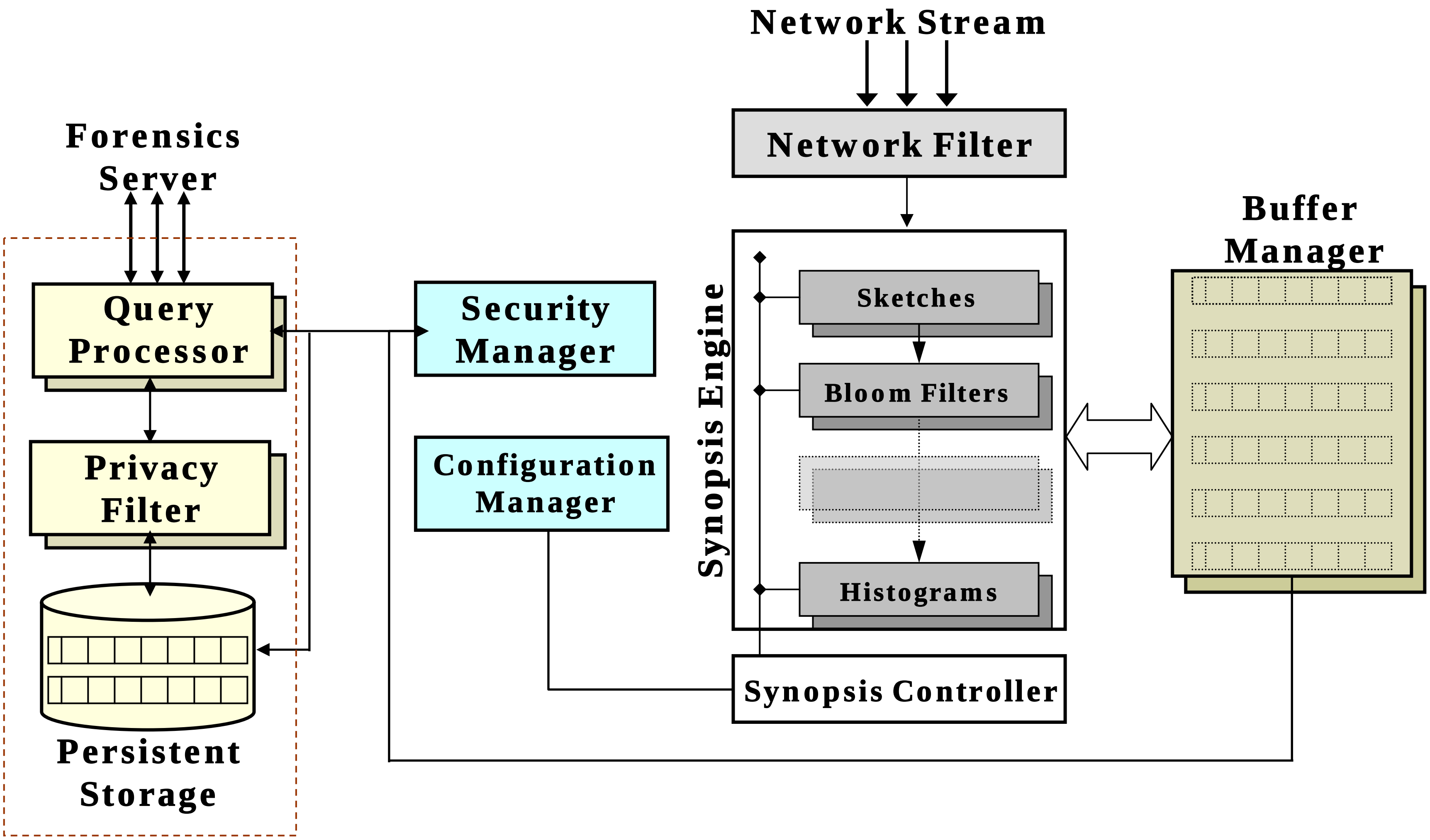


SynApps - Synopsis Appliance

- Collected SynApps
- Simple SynApps


D
allow C only
C
Query propagation
Attack propagation
Network links
A
B

Architecture of SynApp



What is a Synopsis?

- Properties of a good synopsis:
  - Contains enough data to answer *certain classes of queries*
  - Contains enough data to *quantify confidence of its answers*
  - Have *small memory footprint* and easy to update
  - Resource requirements are *tunable*

Advantages of Using Synopses

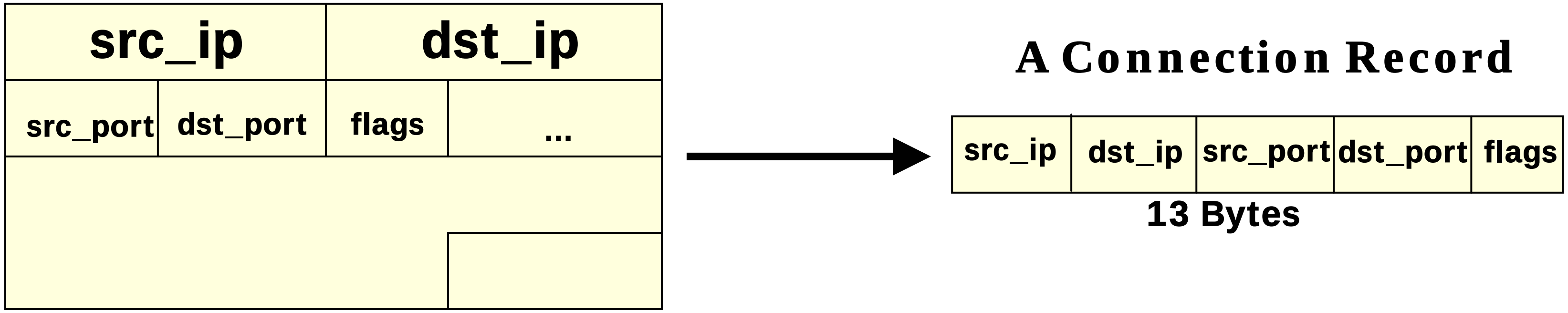
- Succinct representation of base data makes it possible to transfer network data to disks
- Sharing/transferring raw data over network is impossible but synopsis can be moved to remote sites
- Query processing would be expensive with raw data
  - What’s the frequency of traffic to port 80 in the past week? (raw data vs. a histogram)
- Easily adaptable to various resource requirements
  - For example, can adopt the size, processing requirements of a Bloom Filter based on various hardware resources and network load
- Allows for cascading different techniques in the network hierarchy

Examples of synopsis techniques

- Connection Records
- Bloom Filters
- Sampling
- Histograms
- Decision Trees/Clusters
- Wavelets

An Illustrative Example

Connection records



40 Bytes

A Connection Record

src_ip	dst_ip	src_port	dst_port	flags

13 Bytes

- Unfortunately, this is a less flexible data structure
- Does not tell us what happened during a connection

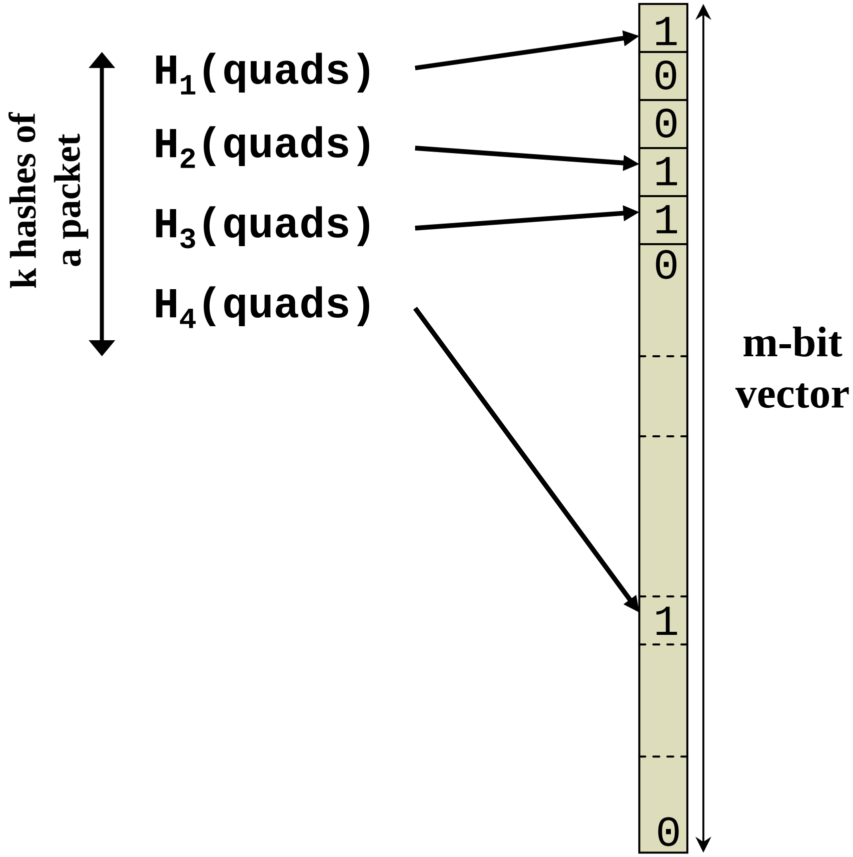
Importance of payload attribution

- Tracing worms and Spam
  - Only payload distinguishes a worm from a benign packet
  - Spam is harder to trace thanks to falsified mail-headers
  - Can we use certain unique portions of Spam (like Message-ID or keywords in the body) to trace back?
- Tracing Honey Tokens
  - Honey tokens are simple watermarks on files or databases
  - E.g. an invalid credit card number in a database
  - Helpful in tracing stolen intellectual property
- And many more …

Simple methods for payload attribution

- Storing Raw Packets
  - Not feasible, too much data to store, privacy issues
- Storing Hashes of Packets
  - Not flexible enough for forensics
  - Flipping a bit causes mismatches and makes the method useless
    - Like adding a space or escape character in a packet

Bloom Filters



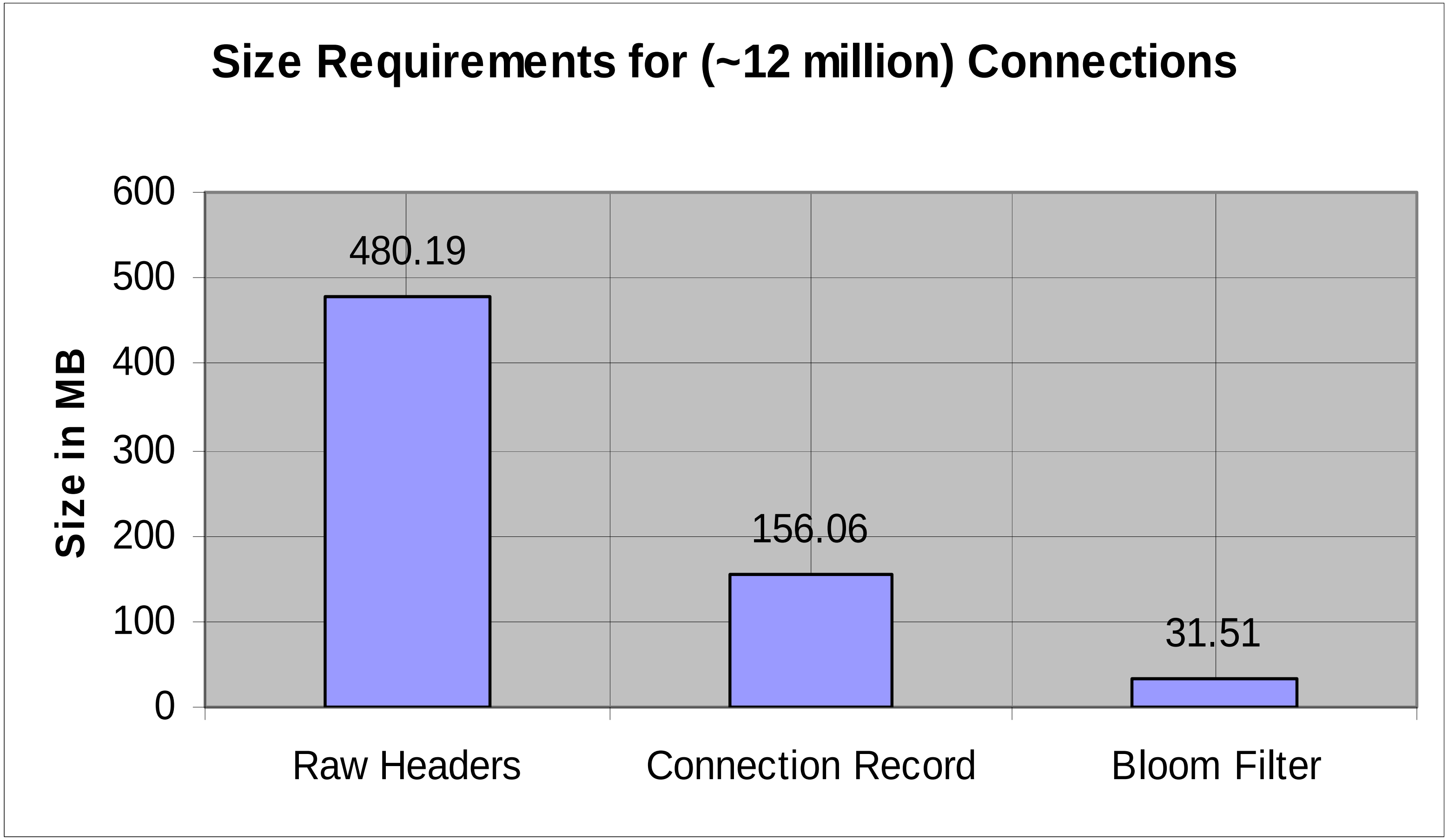
$$FP = \left(1 - (1 - 1/m)^{kn}\right)^{k}$$

k: Number of hash functions
m: Size of bit vector
n: Number of elements stored

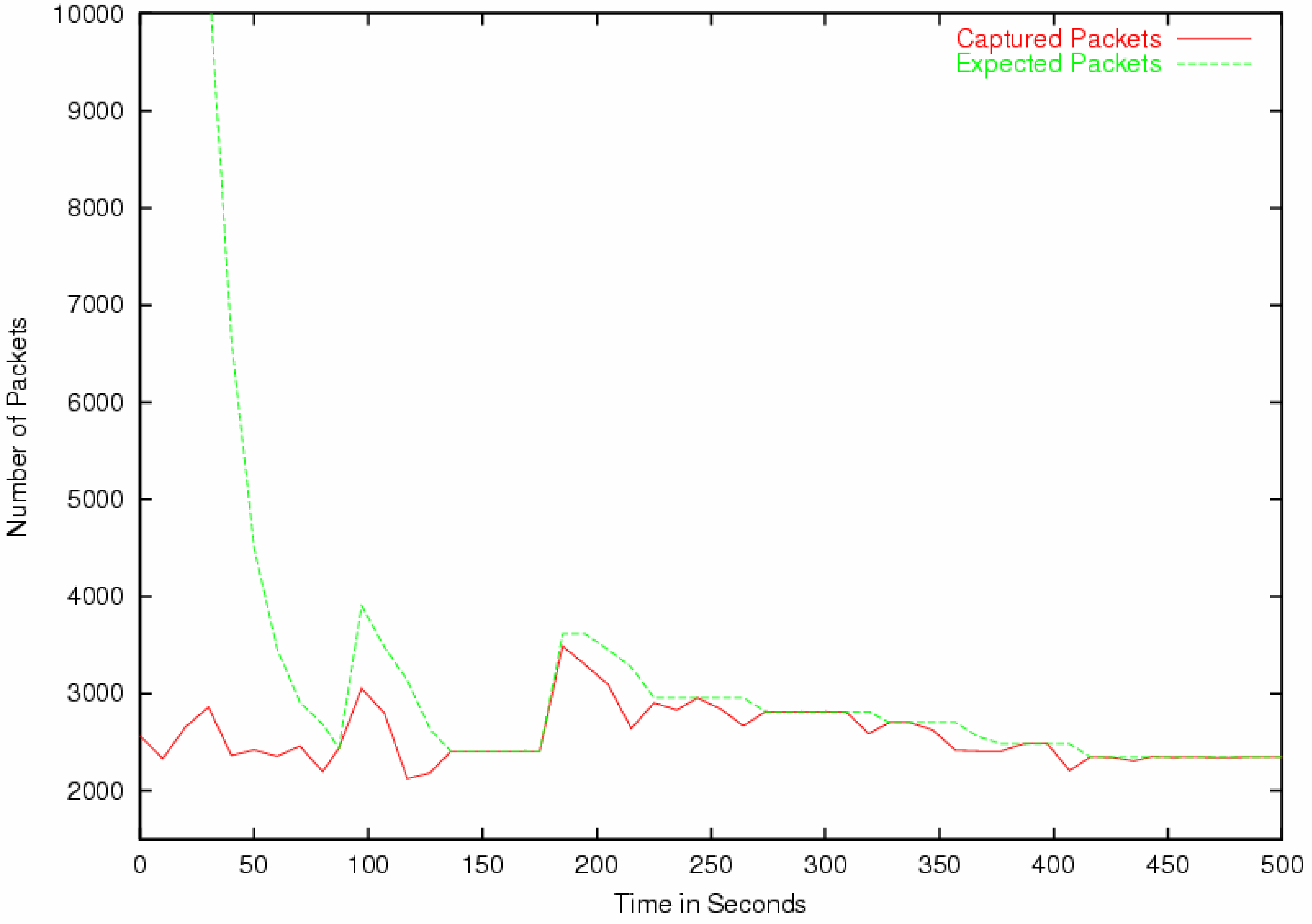
Can tradeoff memory (m) and computing power (k) for accuracy (FP)

- For 1TB of data, filter is ~2.4GB
- 1TB is 100Mbps per day
- How much traffic are we talking about?
  - 100GB of headers (68-bytes) in a day

Space requirements …



Predicting space requirements

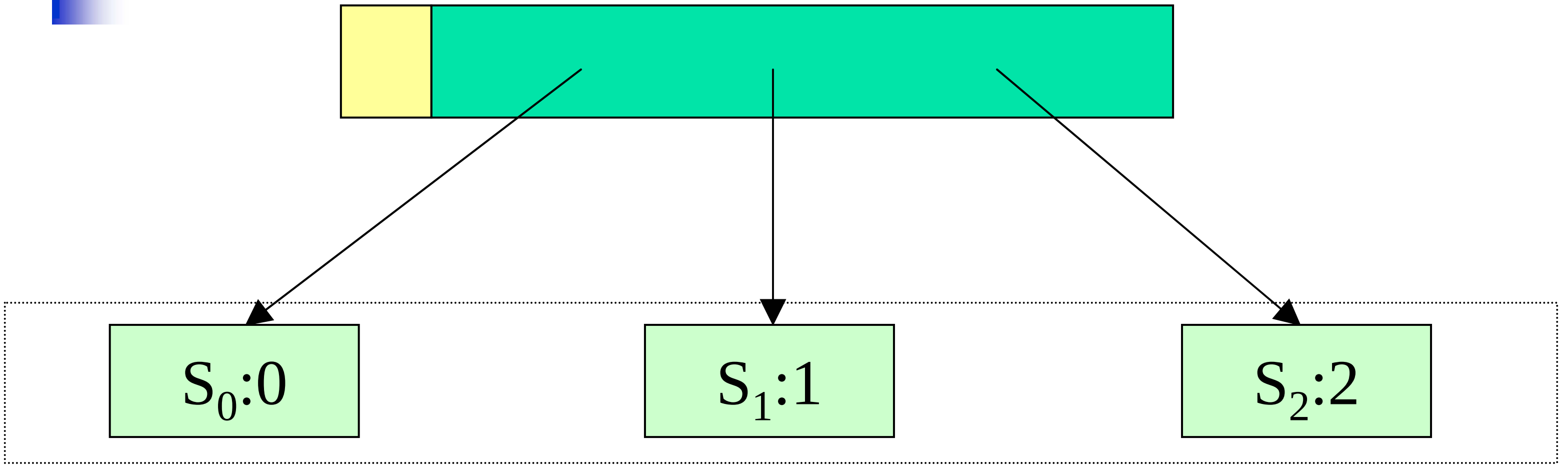


Polytechnic UNIVERSITY

Limitations of Bloom Filters

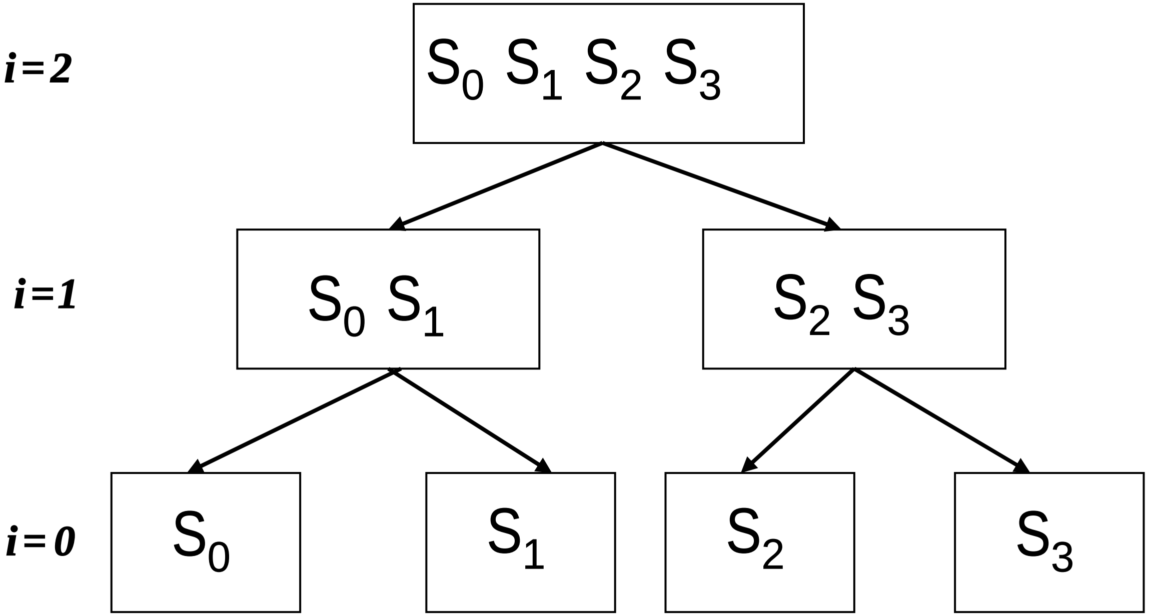
- Need to know what you are looking for
  - Was “xyz” seen on the network?
- Often exact payload is not available
- However we have certainty about a portion of the payload
- We need to be able to make queries based on portion of the payload

Sliced Bloom Filter



- Simply split the payload into *k*-byte blocks
- Append their position in the payload and insert them into a Bloom Filter along with packet identifier (address, port 4-tuple)
- Disadvantage: Cannot guarantee two sub-strings appeared in a payload subsequently

Hierarchical Bloom Filters (HBF)

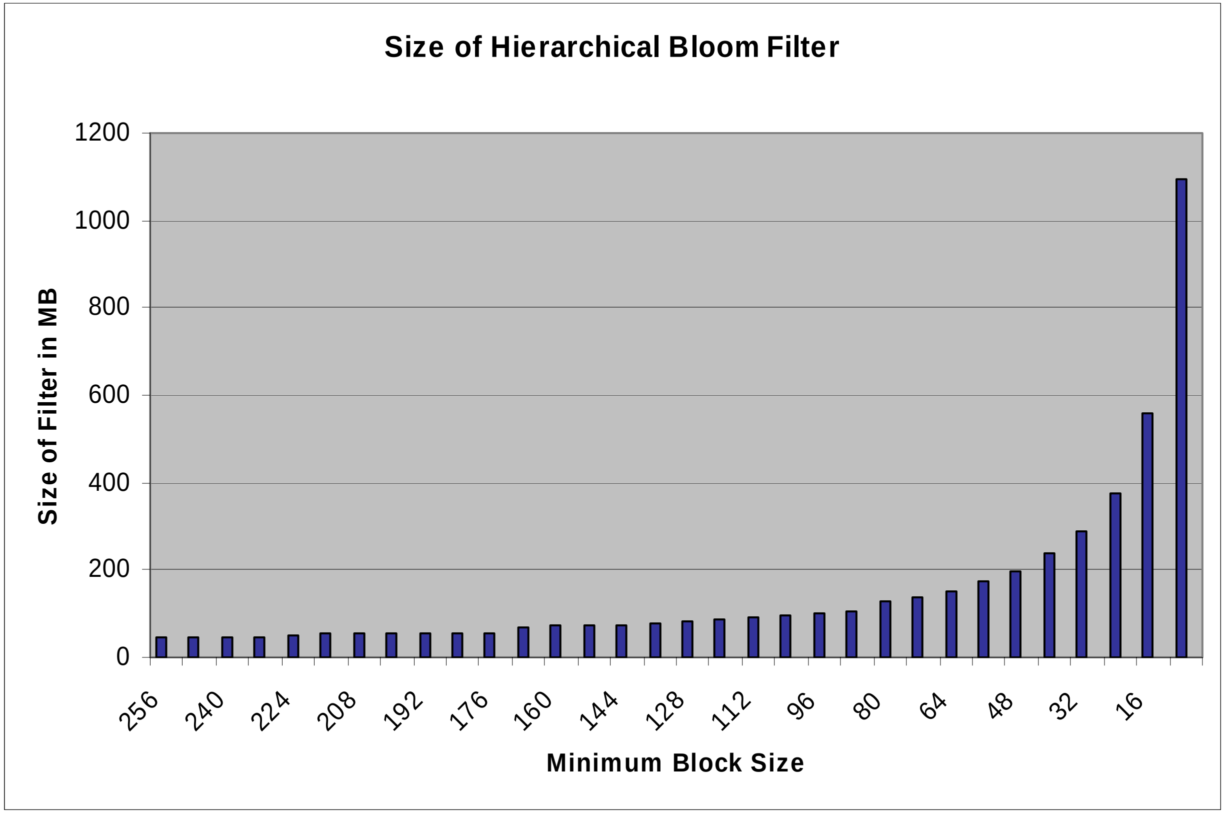


- Split payload into blocks of size (2^i*k) at level *i*
- Insert the block into a Bloom Filter concatenated with packet identifier

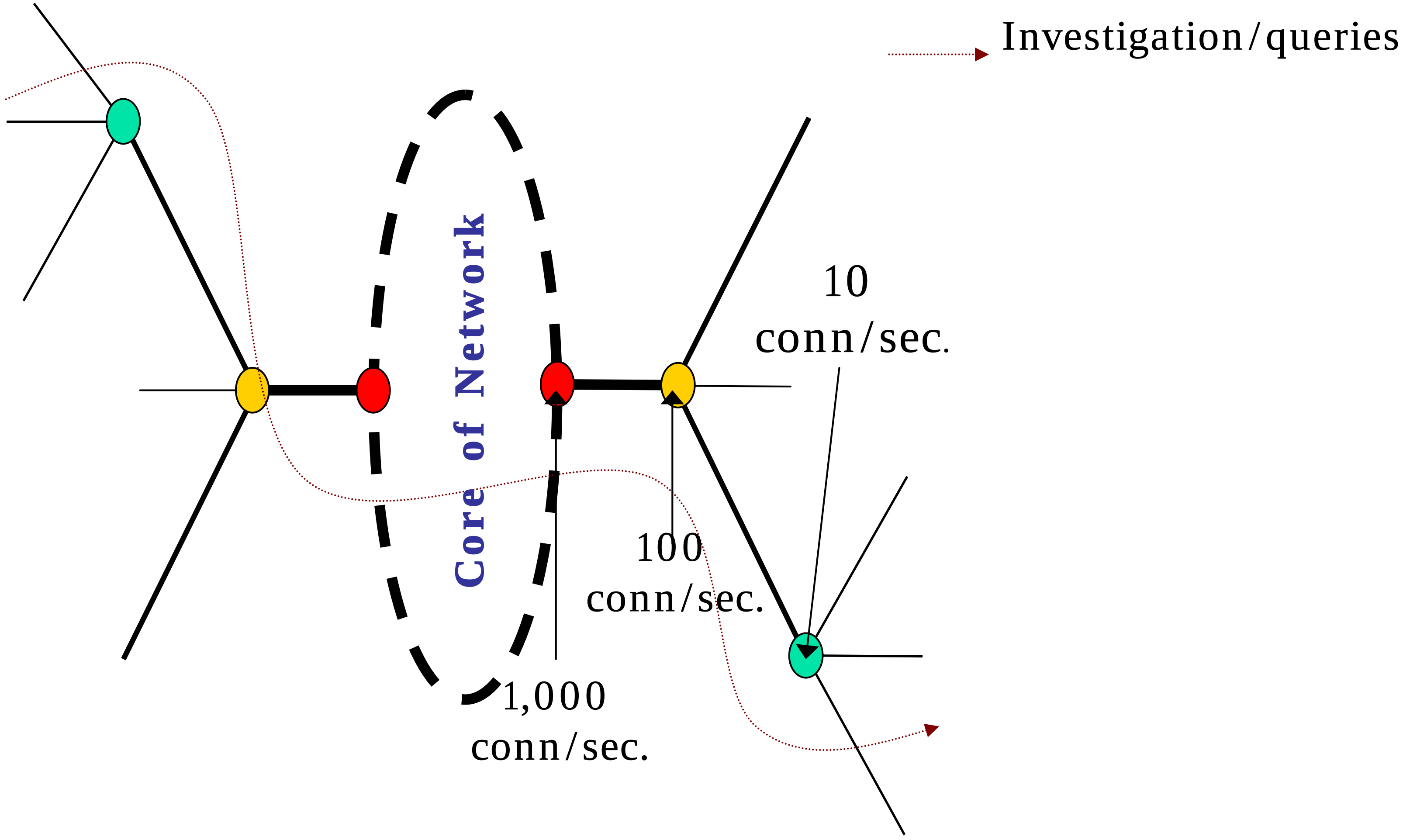
Querying with HBF

- Move a sliding window of size k and check HBF. For any positive
  - Check all subsequent blocks of size k until first negative.
  - You now have a string of length t_0k for which HBF has given positive response.
  - Form blocks of length 2k in this substring and check HBF
  - You will have substring of length t_1k for which HBF has given positive response.
  - Proceed in similar manner climbing tree until you finally have substring of length t_nk for which HBF has given positive response.
- HBF reduces the chances of false positives
- Preserves privacy to a certain extent
- Can attribute content to a packet when a packet-id is incorporated

Storage resources needed for HBF



Simple Network Hierarchy



Cascading of Synopsis on Networks

- Traffic volumes decrease as we move down the network hierarchy
- So use different synopses technique at every level
- For example: without the quadtuples a Bloom filter is useless, on the other hand it is not possible to store connection records at higher network loads
- Therefore, use connection records at the subnets and Bloom filters at enterprise and edge routers (where traffic rate is high)
- When an analyst wants to find out who established a connection:
  - First ask the subnet router and get all connection records
  - Now verify whether the packets were spoofed or not by querying enterprise router

Research Challenges

- Identification of useful network events
  - Network is the virtual crime scene that holds evidence in the form of network events
- Developing efficient synopses
  - Handling connection oriented & connectionless traffic
  - Techniques for looking into payload
  - Cascading of various synopsis techniques
- Identification of various query types
  - Selection queries
  - Neighbor queries
  - Temporal queries
  - Similarity queries
  - Aggregate queries
  - Spatio-temporal joins

Research Challenges

- Integration of information from synopses across networks
  - Real power of ForNet is realized when information from SynApps is fused to answer queries
  - Development of a protocol for secure communication of various ForNet components
- Query processing and storage of synopses
  - A query language transparent of various underlying synopsis techniques
  - A query manage system to interpret the language for the underlying database
  - Various storage and garbage collection strategies for collected-SynApps
  - Storage and query processing infrastructure for Forensics Servers

Questions