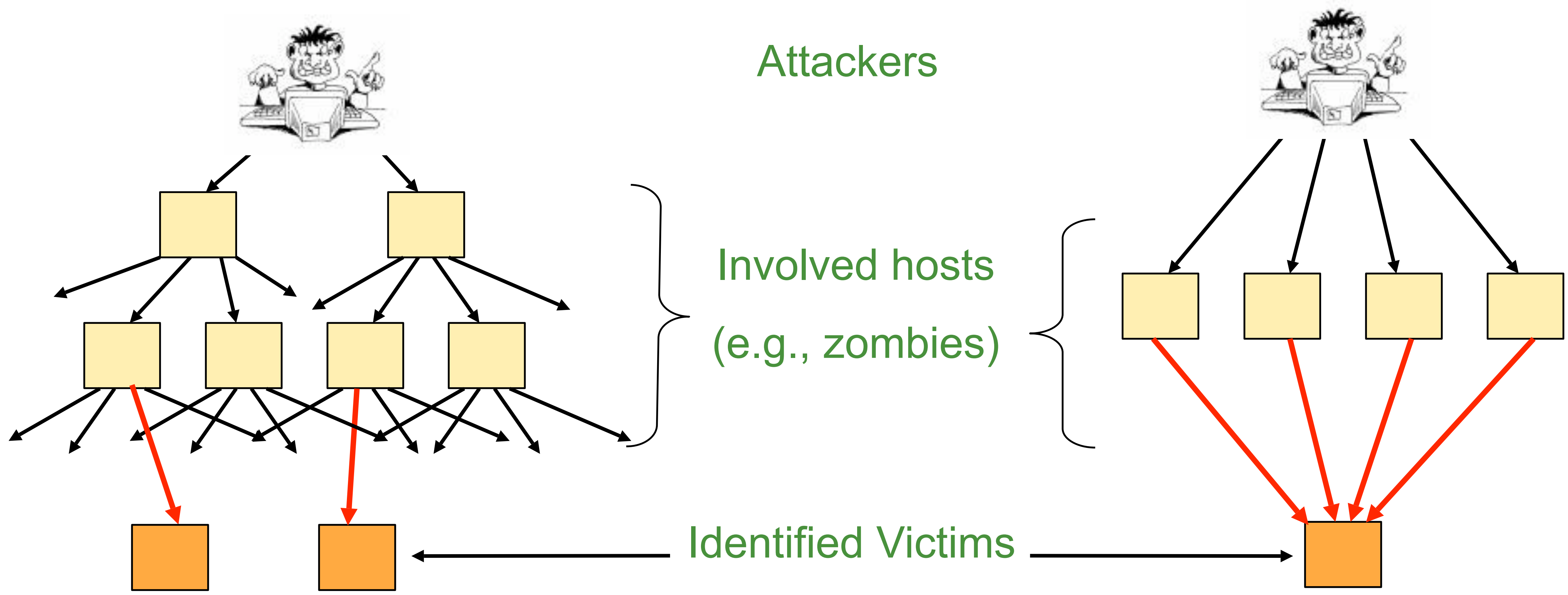
CE 817 - Advanced Network Security
Network Forensics

Lecture 22

Mehdi Kharrazi
Department of Computer Engineering
Sharif University of Technology

Acknowledgments: Some of the slides are fully or partially obtained from other sources. Reference is noted on the bottom of each slide, when the content is fully obtained from another source. Otherwise a full list of references is provided on the last slide.

The Structure of Attacks



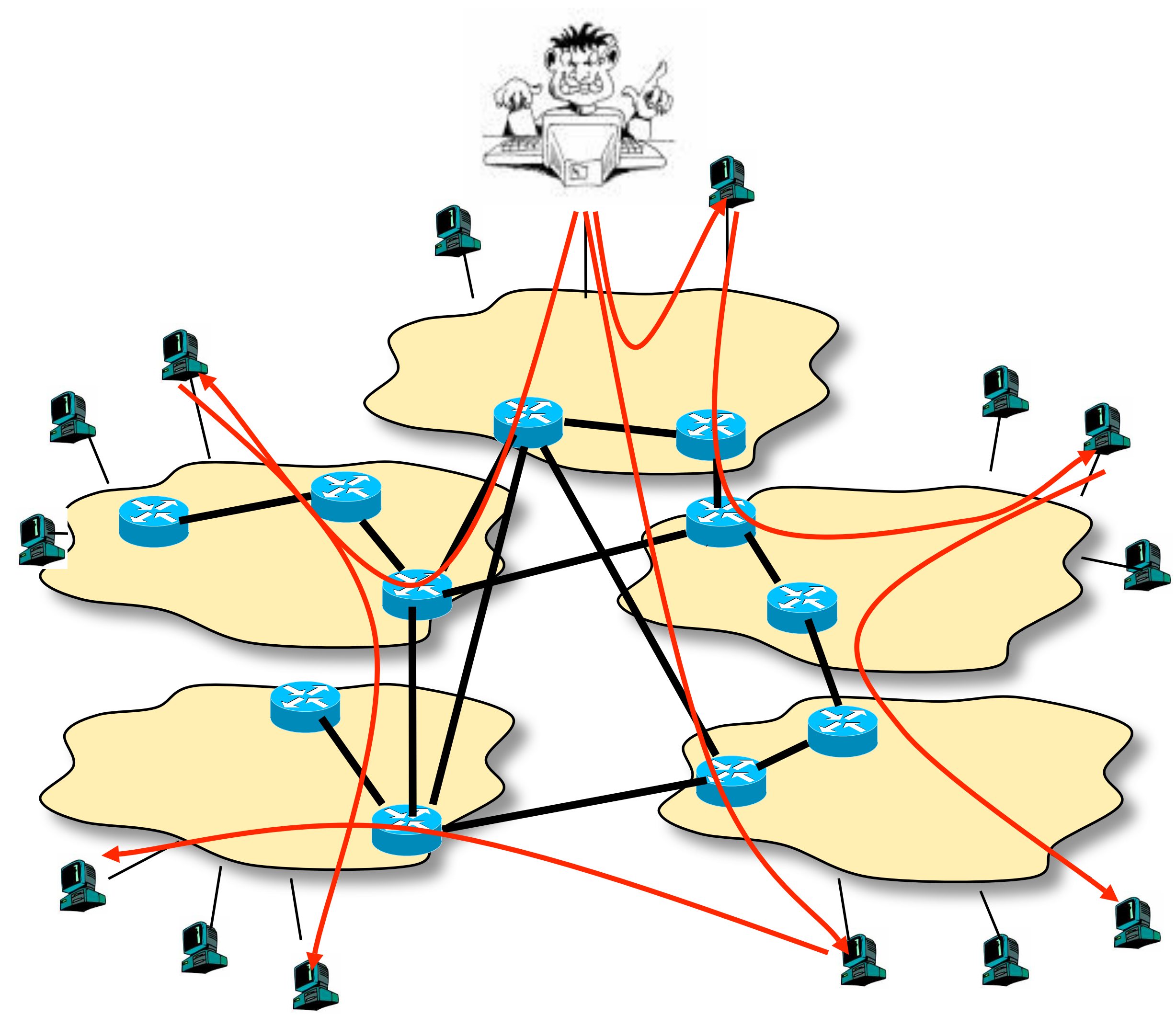
Worm Infection

Distributed DoS

- Modern attacks are multi-level
  - Large scale: difficult to defend
  - Hidden trail: difficult to identify initial launch point

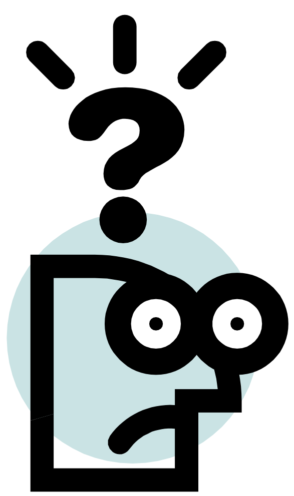
Existing Approaches

- Firewall
- Intrusion detection
    - Identify compromised hosts
- IP traceback
    - Trace to the source of a packet



Question:

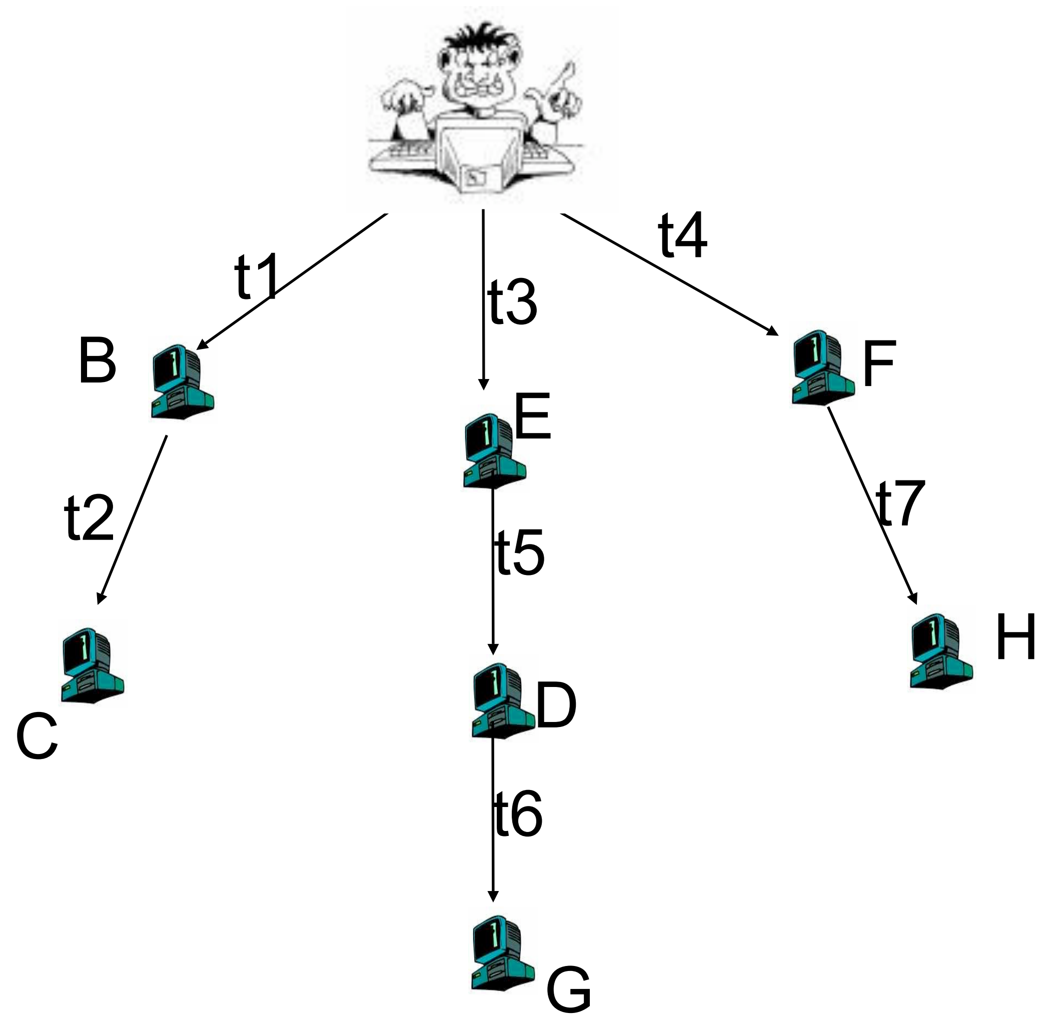
- What fundamental capabilities can be added to the network to achieve better security?



Forensics

- A fundamental capability --- network auditing and forensic analysis
  - Keep communication records
  - Permit post-mortem analysis of patterns across network and time
- Scope: Internet and intranet
  - Correct weak points in a network perimeter
  - Deter future similar attacks

Two Applications



- Attack Reconstruction: infer which communication carry the attack forward
- Attacker Identification: pinpoint the attack source(s)

Key Question 1: Feasibility of Network Auditing

- How much storage is needed for auditing?
  - Directional network flows
  - <source, destination, start-time, end-time>
  - A large ISP with O(100) POPs – 450 GB/hour with compression
  - Intranet requirement will be smaller

Key Question 2: Feasibility of Network Forensic

- Given complete information, can we find needles in the haystack?
  - Host contact graph is large and noisy
  - Algorithms to identify global correlations

Payload Attribution via Hierarchical Bloom Filters,

Kulesh Shanmugasundaram, Hervé Brönnimann, and Nasir Memon. ACM Computer Communications and Security (CCS'04), Washington, DC, 2004.

Payload Attribution

- The problem:
  - Identify the sources and/or the destinations of a bit-string in a network
  - We may only have an arbitrary portion of payload

Challenges Facing Network Forensics

- Lack of Infrastructure:
  - For data collection, archival, and dissemination
- Volume of Data:
  - Prolonged storage, processing, and sharing of raw data infeasible
  - Even a network of ~3000 hosts have a ~1TB/day requirement!
- Process is Manual:
  - Spans multiple administrative domains
  - Response times very long (digital evidence disappears)
- Unreliable Logging Mechanisms:
  - Host logs are usually compromised
  - Growing support for mobility makes it difficult to maintain prudent logging policies on hosts

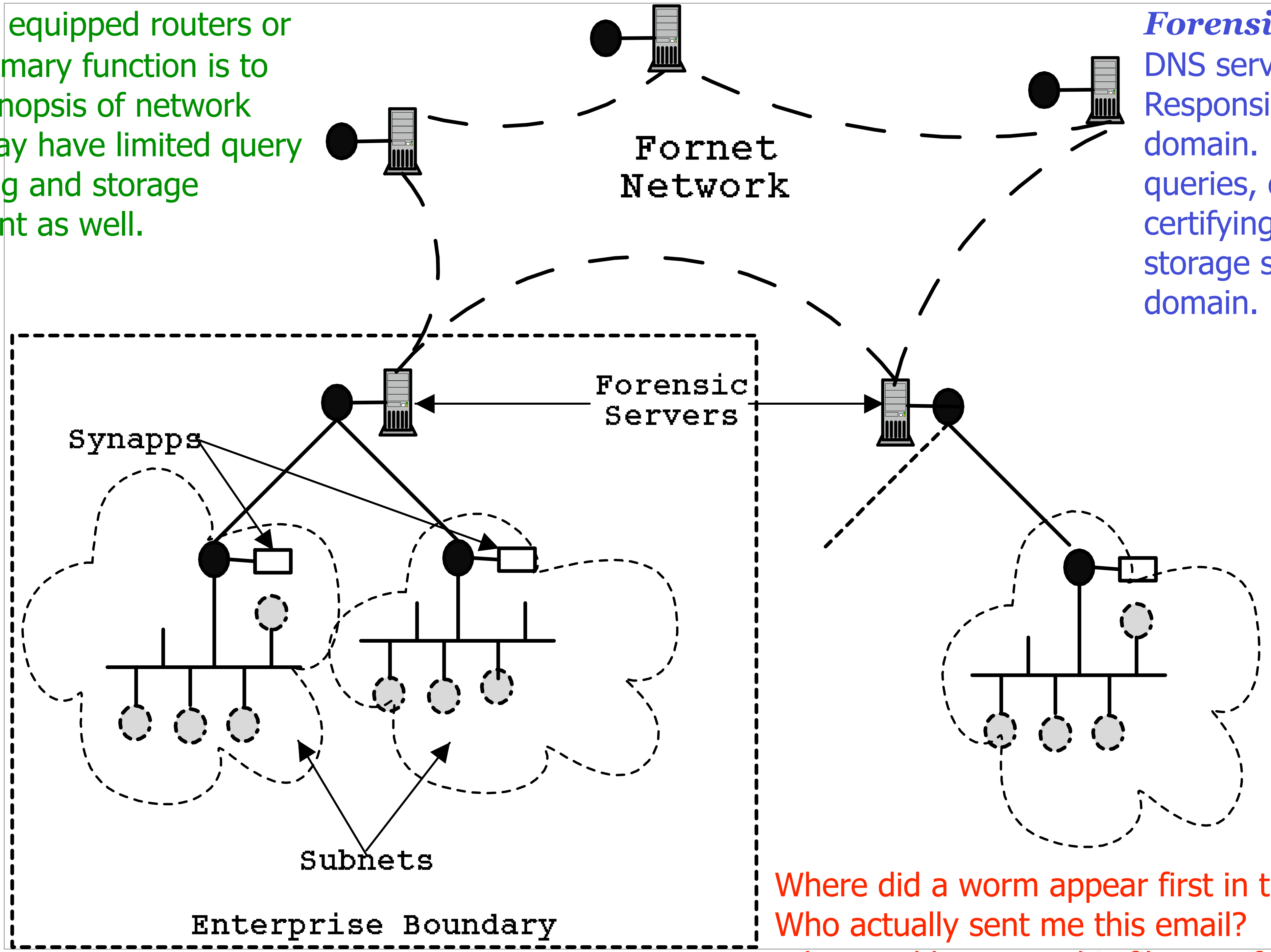
 [Memon]

Our Solution…

- Securely collect, store, disseminate, and process synopsis of network traffic.
- In other words a device analogous to a surveillance camera for the network. Even better – a co-operating network of such surveillance cameras.
- Goal of Project ForNet: development of tools, techniques, and infrastructure to aid rapid investigation and identification of cyber crimes

ForNet Vision - A unified, end-to-end approach for network monitoring

SynApp equipped routers or hosts. Primary function is to create synopsis of network traffic. May have limited query processing and storage component as well.



Forensic Server: Like the DNS servers' of today. Responsible for a network domain. Responsible for routing queries, query processing, certifying results, and long-term storage synopsis for the domain.

Where did a worm appear first in the .edu domain?
Who actually sent me this email?
When and how was this file transferred to this host?
Was this host compromised in the past? When and how?

[Memon]

What is a Synopsis?

Data structures and algorithms for representing a set of elements succinctly with predefined loss in information and has the ability to recall the original set of elements with a preset accuracy.

Properties of a Good Synopsis

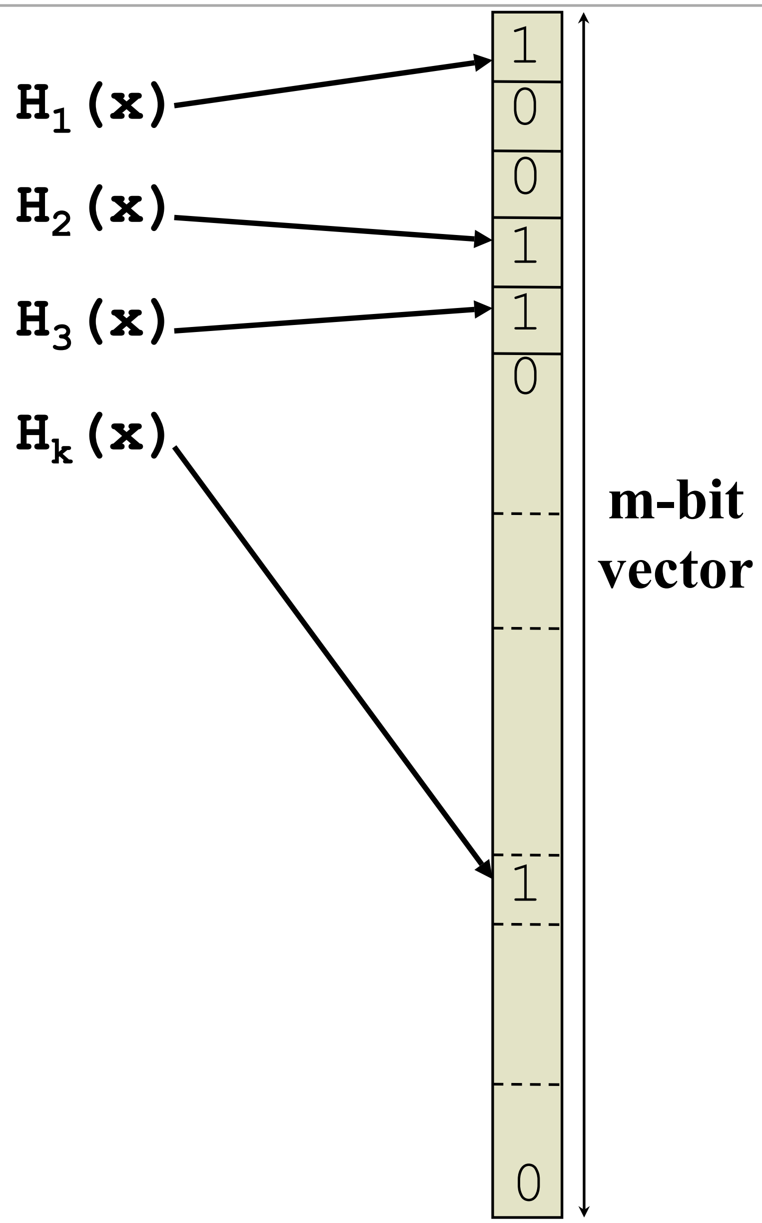
- Contains enough data to answer certain classes of queries
  - Who sent payload “xyz”?
  - What did host bug.poly.edu send?
- Contains enough data to quantify confidence of its answers
  - I’m 99.37% sure bug.poly.edu sent “XYZ”
- Have small memory footprint and easy to update
  - Need 20GB/day to keep 1TB/day of raw network data
  - Need to compute a number of hashes per packet
- Resource requirements are tunable
  - Can only afford 3GB/day, adjust the accuracy to accommodate this.

Advantages of Using Synopses

Can retain potential evidence for months!

- Succinct representation of raw 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
  - Can adopt the size, processing requirements of a Bloom Filter based on various hardware resources and network load

Bloom Filters



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

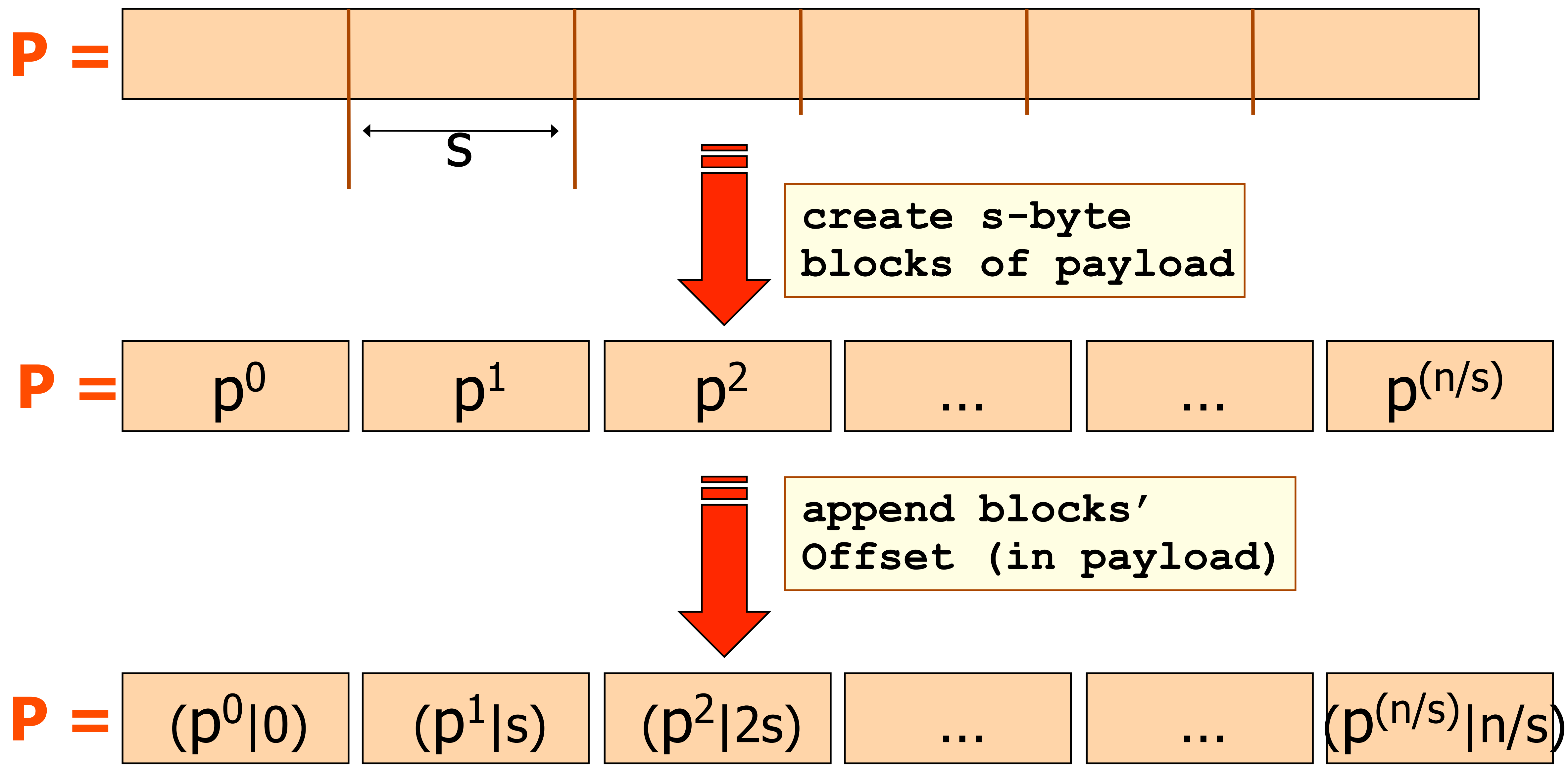
- Bloom Filter:
  - Randomized data structure for representing a set in order to support membership queries.
  - Insert(x):
    - Flip bits H1(x)… Hk(x) to ‘1’
  - IsMember(y):
    - If H1(Y) … Hk(Y) all ‘1’ “yes” otherwise “no”
- Can tradeoff memory (m), compute power (k), and accuracy (FP)
  - m – length of bit vector (range of H(.))
  - k – number of hashes per element
  - n – number of elements in the set

Packet Digests & Bloom Filters

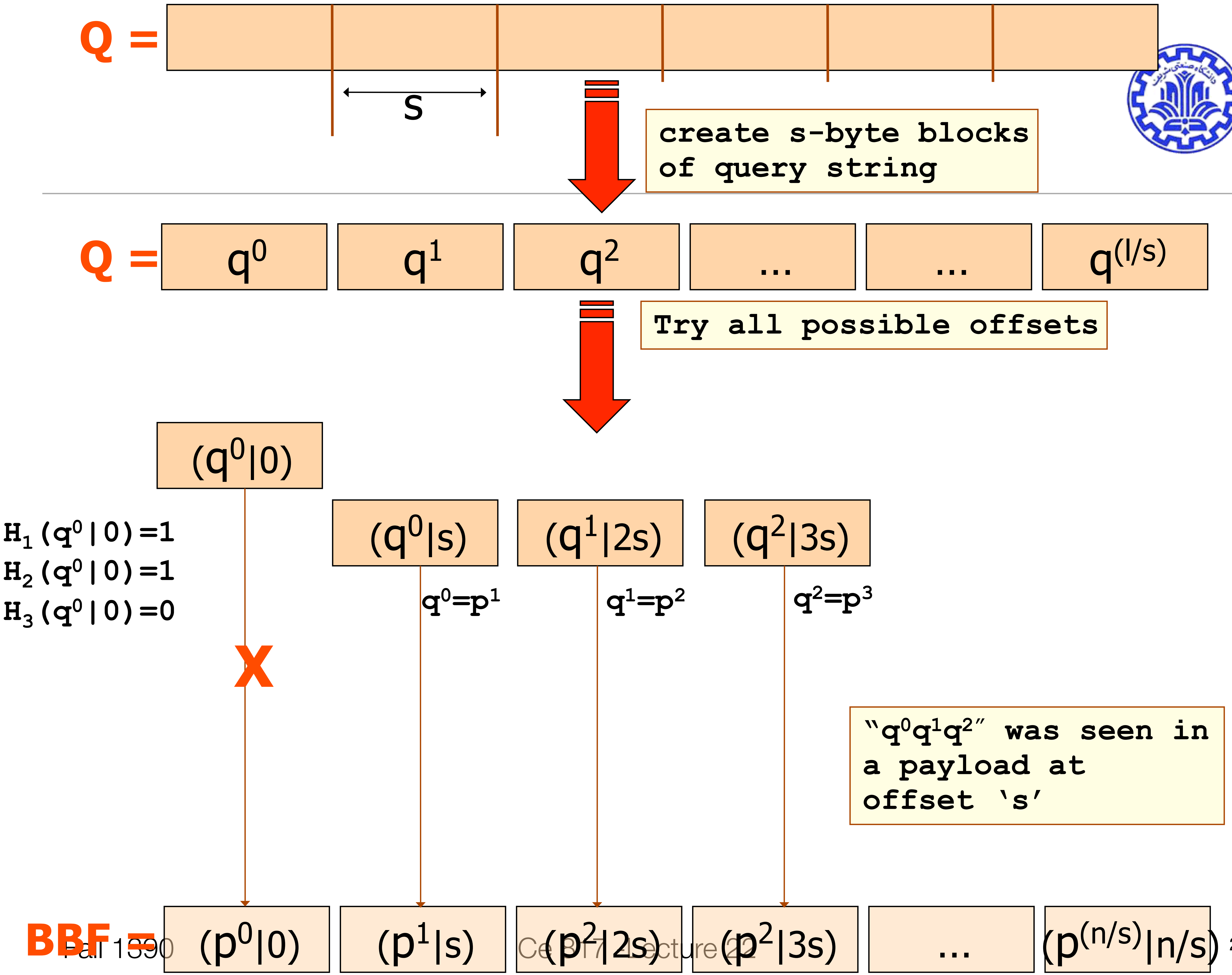
- Space Efficient:
  - m/n=16 and 8 hashes (k=8)
    false positive (FP) = 5.74 x 10-4
  - No false negatives!
- However, suppose we don't have packets.
  - We only have some excerpts of payload
  - Don't know where the excerpt was aligned in the packet

Extend Bloom Filters to support excerpt/substring matching

Block-based Bloom Filter



Insert each block into a Bloom Filter


Q =
s
create s-byte blocks
of query string
Q =
q0
q1
q2
...
...
q(l/s)
Try all possible offsets
(q0|0)
H1(q0|0)=1
H2(q0|0)=1
H3(q0|0)=0
X
(q0|s)
(q1|2s)
(q2|3s)
q0=p1
q1=p2
q2=p3
"q0q1q2" was seen in
a payload at
offset 's'
BBF =
(p0|0)
(p1|s)
(p2|2s)
(p2|3s)
...
(p(n/s)|n/s)

[Memon]

P1 = | A | B | R | A | C | A |

P2 = | C | D | A | B | R | A |

BBF =

(A|0)	(B|s)	(R|2s)	(A|3s)	(C|4s)	(A|5s)
(C|0)	(D|s)	(A|2s)	(B|3s)	(R|4s)	(A|5s)

"Offset Collisions"

(A|0)	(B|s)	(R|2s)	(A|3s)	(C|4s)	(A|5s)
(C|0)	(D|s)	(A|2s)	(B|3s)	(R|4s)	(A|5s)

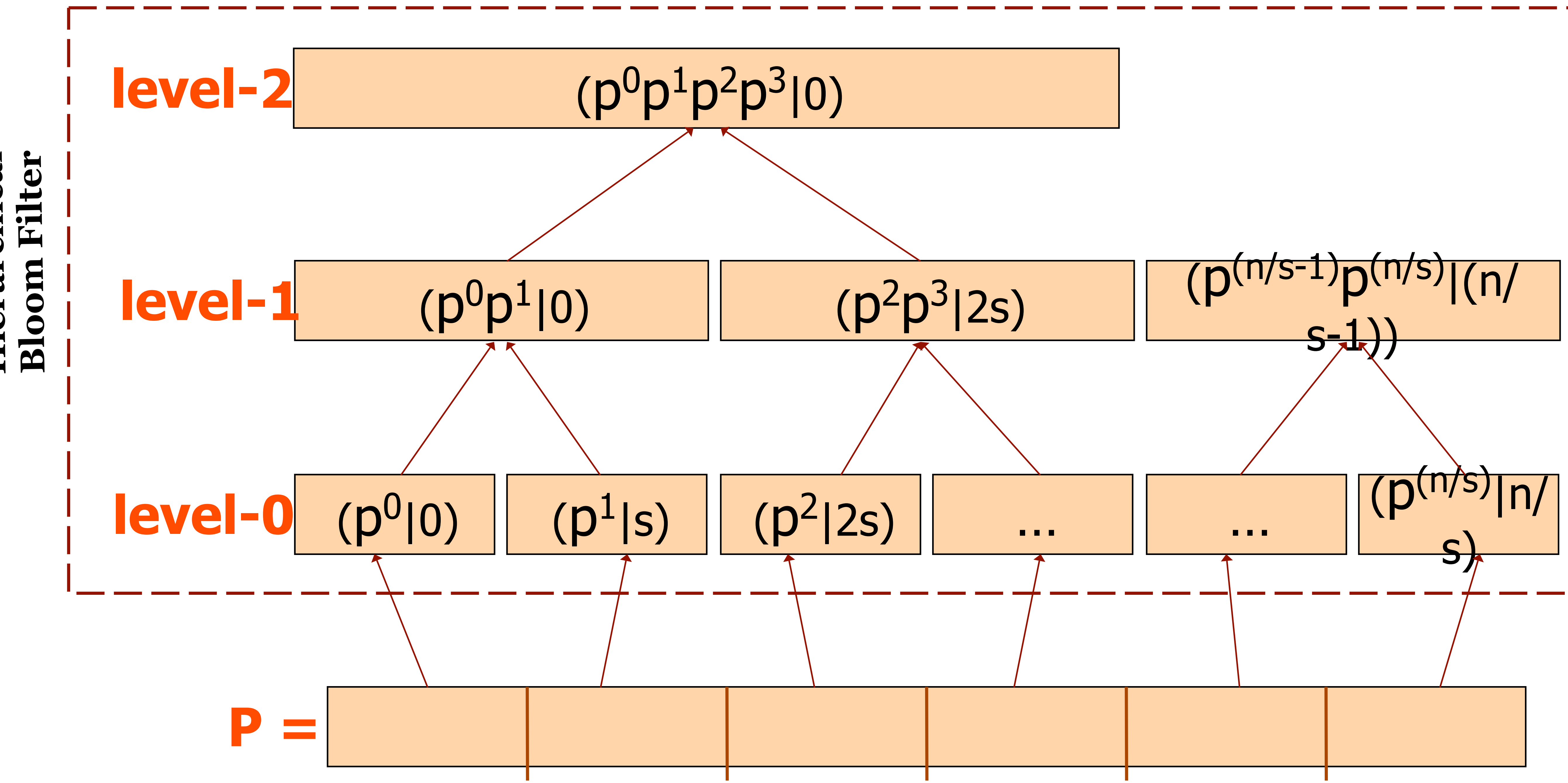
For query strings: "AD", "CB", "DR", "AA" etc. BBF falsely identifies them as seen in the payload!

Because BBF cannot distinguish between P1 and P2

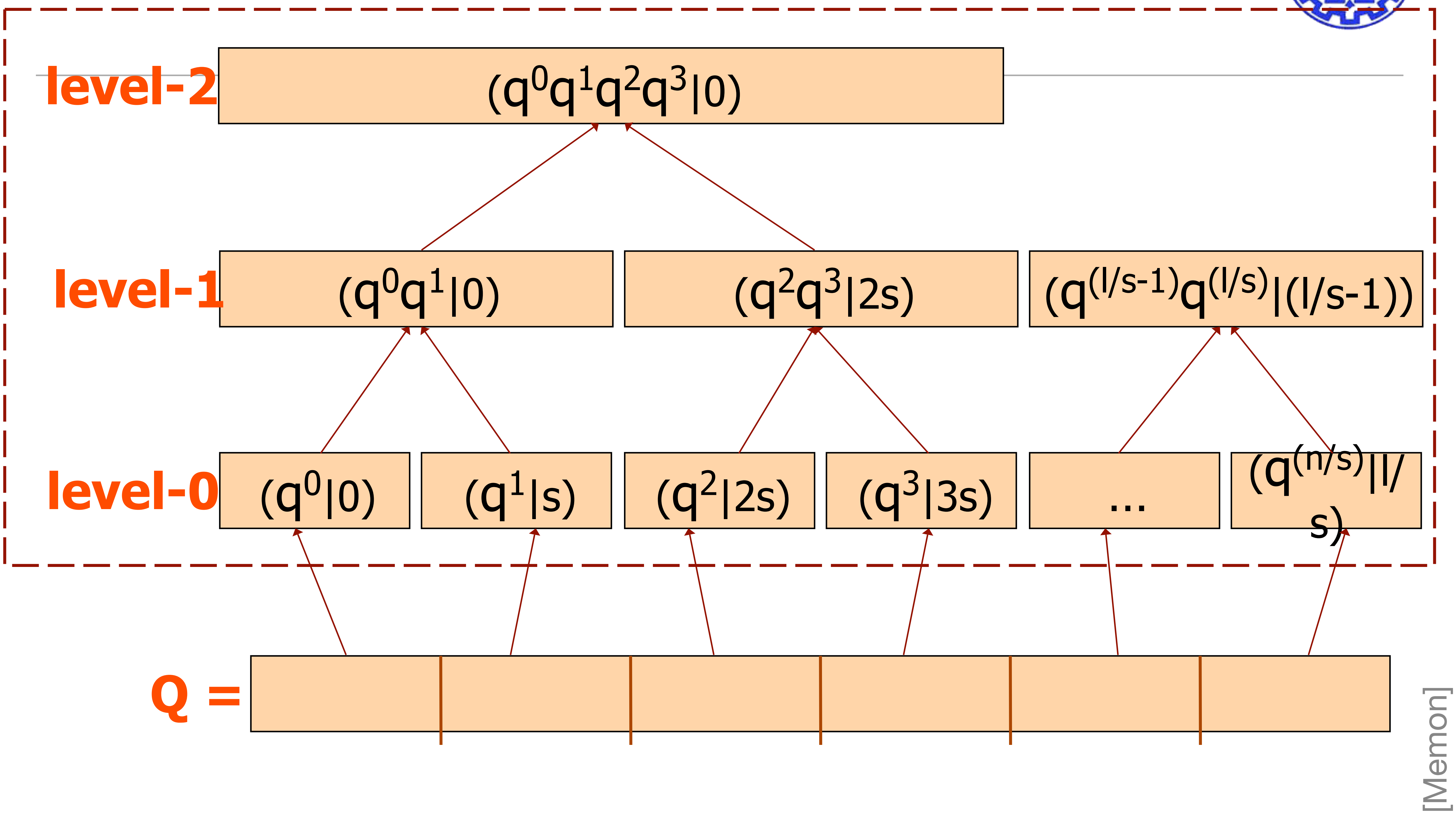
[Memon]

Hierarchical Bloom Filter

- **An HBF is basically a set of BBF for geometrically increasing sizes of blocks.**



Hierarchical Bloom Filter

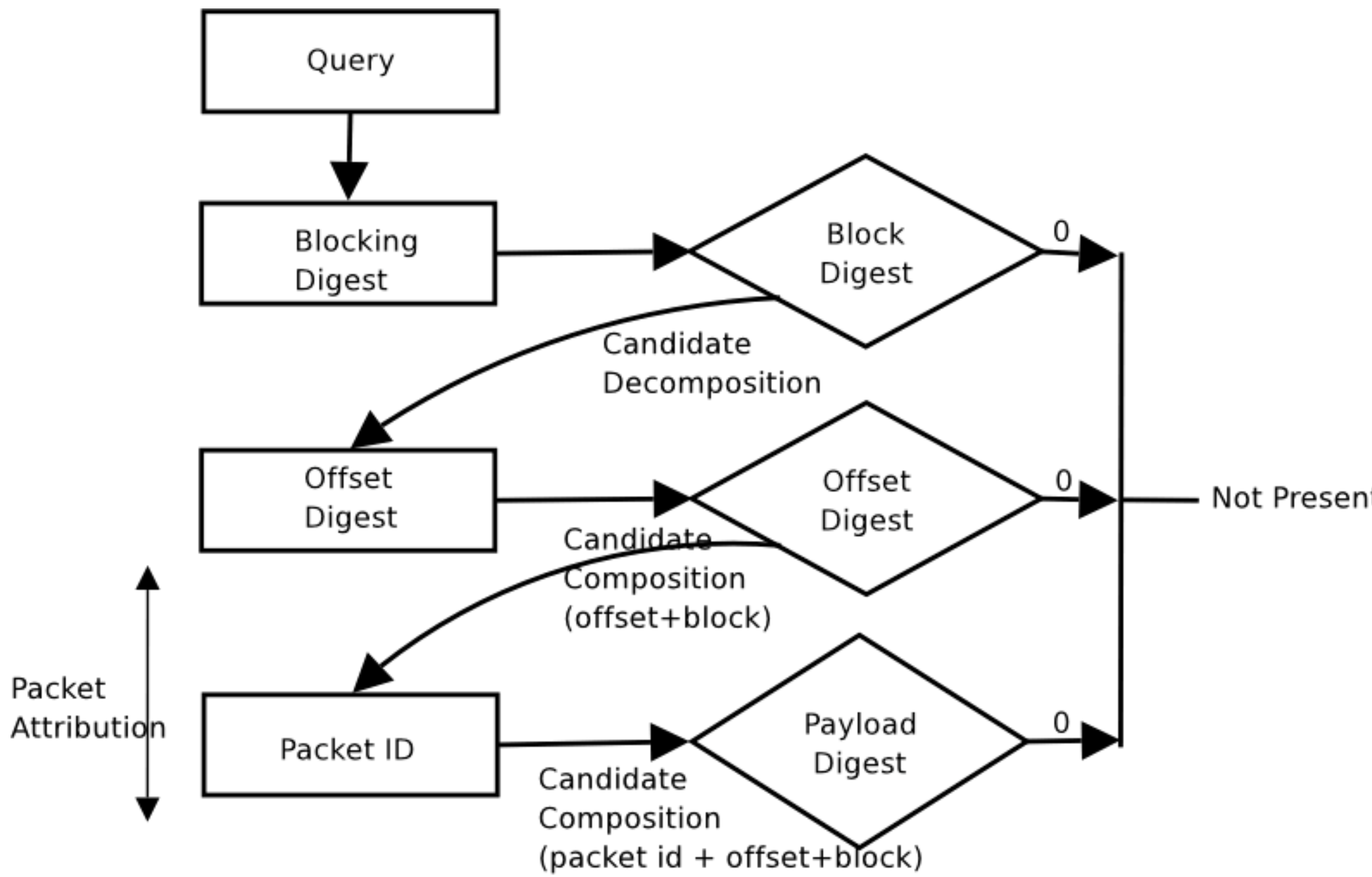


- Querying is similar to BBF.
- Matches at each level can be confirmed a level above.

Adapting an HBF for ForNet

- So far an HBF can attest for the presence of a bit-string in payloads
- We need to tie this bit-string to a source and/or destination hosts
- Our Approach:
  - Similar to tying an offset to a block/bit-string
  - In addition to inserting (block||offset) also insert (block||offset||hostid)
  - Hostid could be (srcIP||dstIP)

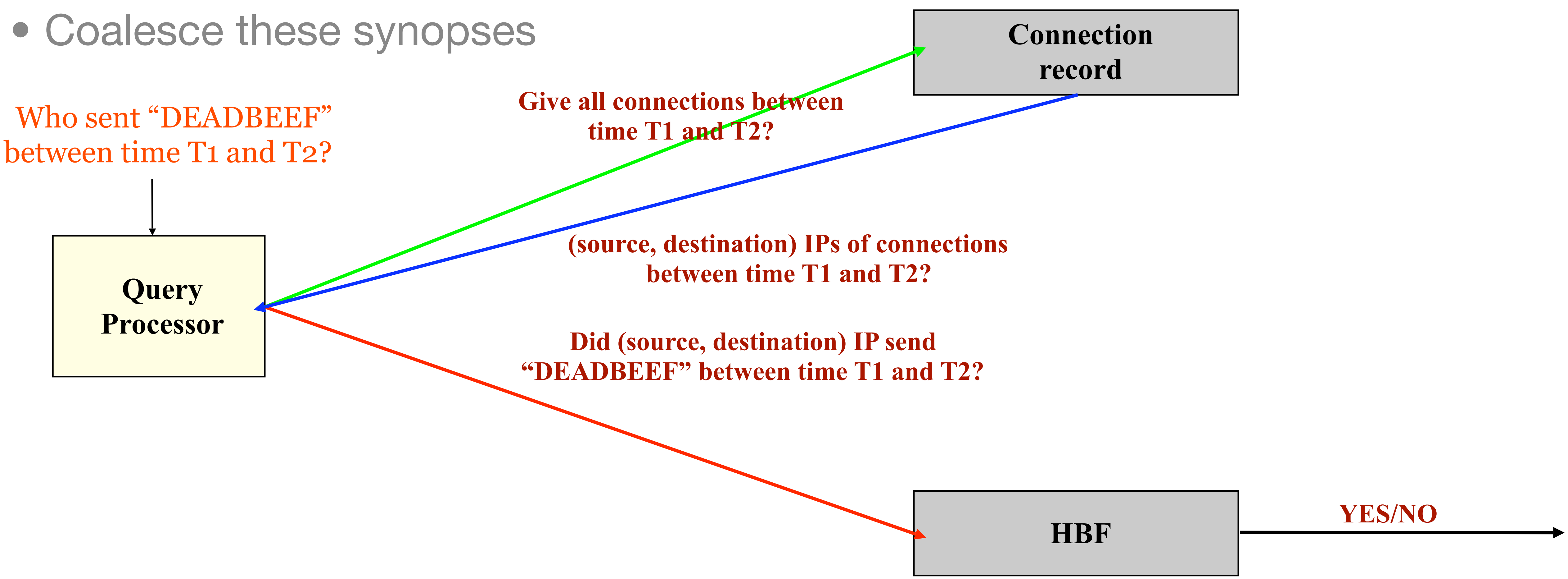
How to run a query?



 [Memon]

Fusion of Synopses

- HBF requires:
    - Source IP, destination IP, excerpt
    - But where do we get Source IP, Destination IP
- Connection Record
    - Given two time intervals can give us list of source, destination IPs
- Coalesce these synopses



ForNet in Intranet Usage

- Determine victims of worm, trojans and other malware.
- Detection of potential victims of phising and spyware

ForNet Deployed on Internet

- Traceback based on partial content of single packets
- Source of malware, worms, etc.

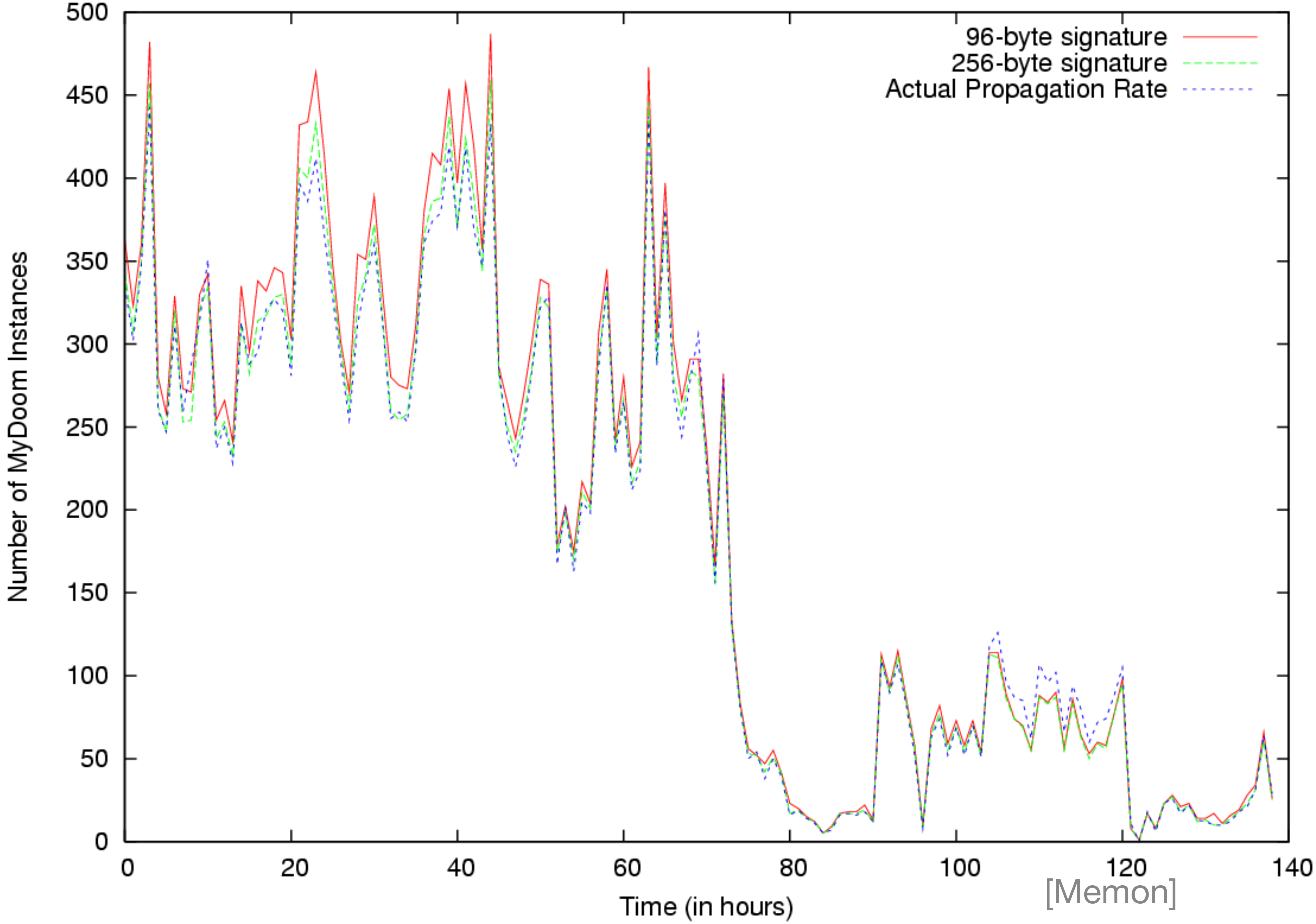
Current Status

- Implemented a PC based SynApp device for placement within an intranet.
- Implemented Forensics Server with simple querying capabilities.
  - Current Forensic Server has 1.3TB of storage with over 3 months worth of data from the edge-router and two subnets
  - Normal bandwidth consumption of network is about a 1 – 2 TB/day
  - Synopses reduces this traffic to about 20GB/day
- Implemented Panorama (GUI Client)

Tracking MyDoom

- Recorded all email traffic for a week
  - Using HBF and raw traffic
  - Was not aware of MyDoom during this collection
- When signatures became available we used them to query the system
  - To find hosts that are infected in our network
  - How the hosts were infected
- Some statistics:
  - 679 hosts originated at least one copy of the virus
    - 52 of which were in our network
  - These hosts sent out copies of the virus to 2011 hosts outside our network boundary

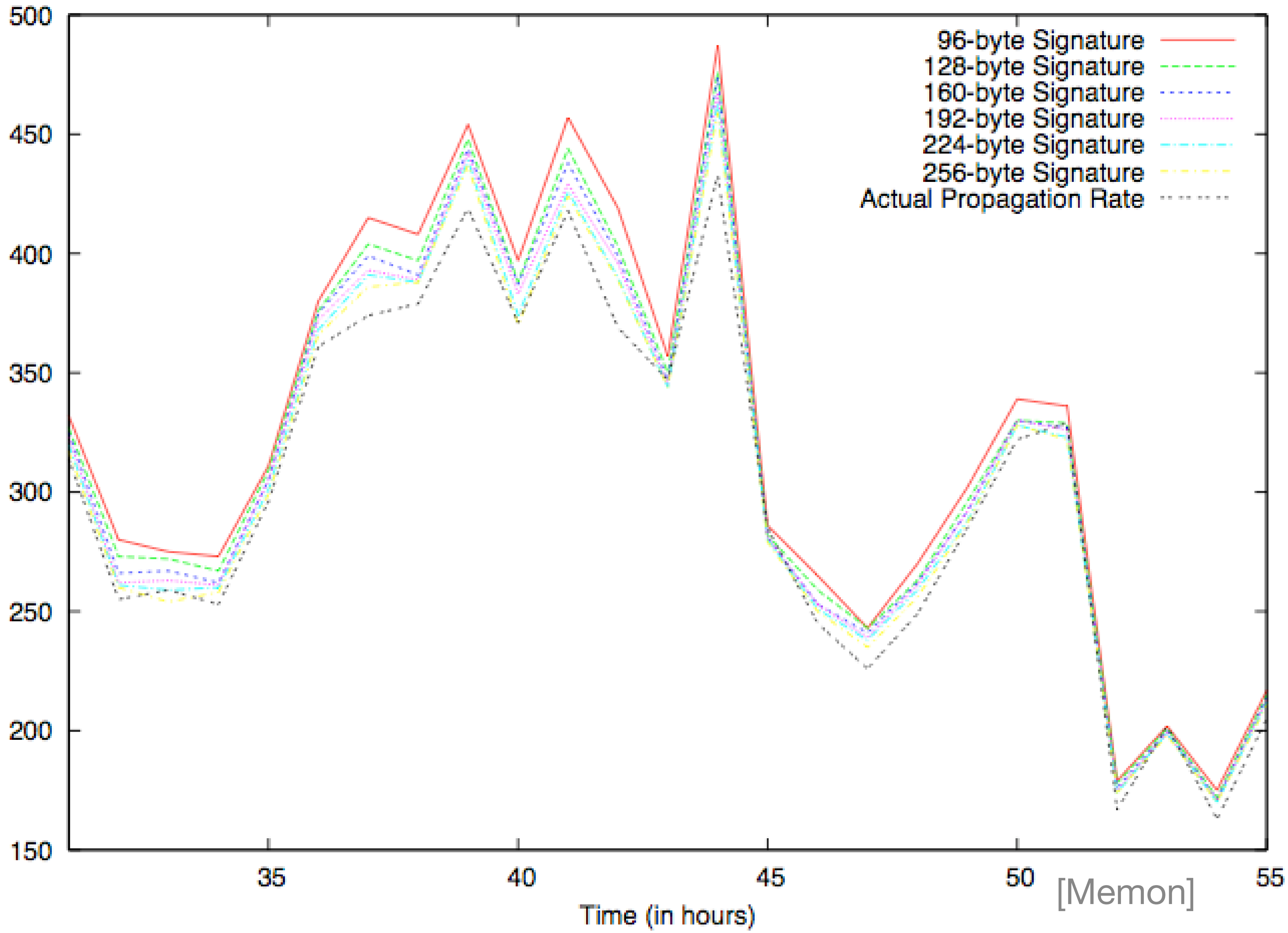
Analyzing MyDoom Infections…



F

[Memon]

MyDoom’s One Day Zoom



[Memon]

Acknowledgments/References

- [Xie] Toward a Framework for Internet Forensic Analysis, Yinglian Xie, presenting (Toward a Framework for Internet Forensic Analysis, V. Sekar, Y. Xie, D. Maltz, M. Reiter, H. Zhang, HotNets-III, 2004.) at the 100x100 clean slate prject presentatio in Pittsburgh, December 2004.
- [Memon] From slides prepared by Kulesh Shanmugasundaram, Hervé Brönnimann, and Nasir Memon, presented at various talks/lectures.