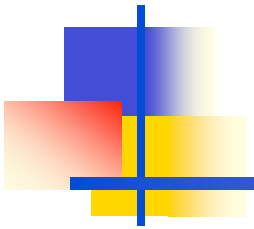


Payload Attribution via Hierarchical Bloom Filters

Kulesh Shanmugasundaram, Hervé Brönnimann
Nasir Memon
http://isis.poly.edu/projects/fornet/

Polytechnic UNIVERSITY

Re: Anubhav's HD - Mozilla Thunderbird

File Edit View Go Message Enigmail Tools Help

Subject: Fwd: Citibank Identity Theft Solutions
From: Nasir Memon <memon@poly.edu>
Reply-To: memon@poly.edu
Date: 03/13/04 00:45
To: 'Kulesh Shanmugasundaram'

Kulesh,

This was a close call; almost fell for it!

- Can we find out who else got this email at Poly?
- And who actually visited the site. (Probably there are lot of other sites as well!)

Bye
Nasir

Recently there have been a large number of identity theft attempts targeting Citibank customers. In order to safeguard your account, we require that you update your Citibank ATM/Debit card PIN.

This update is requested of you as a precautionary measure against fraud. Please note that we have no particular indications that your details have been compromised in any way.

This process is mandatory, and if not completed within the nearest time your account may be subject to temporary suspension.

To securely update your Citibank ATM/Debit card PIN please go to:

https://www.citibank.com/signin/citifi/scripts/login2/update_pin.jsp

Please note that this update applies to your Citibank ATM/Debit card - which is linked directly to your checking account, not Citibank credit cards.

Thank you for your prompt attention to this matter and thank you for using Citibank!

Regards,

Madeline Walter

Head of Citi® Identity Theft Solutions

Copyright © 2004 Citicorp.

10/25/2004

Majority of Home Computers Infected with Spyware

Home > Topics > Security > News > **Another Phishing Hole Found in Google**

Security

Another Phishing Hole Found in Google

By Michael Myser
October 21, 2004

Google Inc. said on Thursday it fixed a flaw in its site that could allow outside
phishing attacks based on Google's familiar interface, and is working on a se
similar vulnerability.

The flaw, which was discovered and posted to Symantec Corp.'s **Bugtraq** sec
Tuesday. The bulletin demonstrated the ability of hackers using JavaScript to

25, 2004
d Daily

Enterprise Linux IT
nt. Linux IT
pen Source
Databases
ux Security
Servers

Linux Security

Red Hat Users Under Phishing Attack

By Kimberly Hill
Enterprise Linux IT
October 25, 2004 12:53PM

A new phishing scam is targeting users of Red Hat Linux. The software company says e-mails that purport to be from Red Hat actually contain a Trojan horse that will compromise system security.

COMPLETE STORY

advertisement

WEB TECHNOLOGY

'Rolex' spam taps into bling-bling culture

By Will Sturgeon Silicon.com October 25, 2004, 9:3

Add your opinion TALKBACK Forward in EMAIL Form

Spam and phishing

Rolex watches have long been seen as a must-h among th high stan being su

In the san

Worldwide Technology

Halo 2 Code Leaked to Internet

By Todd Bishop
October 18, 2004 4:26PM

he latest installment of Microsoft's e, is not expected to affect its release, Downloading the Halo 2 code or s to download will be viewed as theft,

COMPLETE STORY

Search: News SEARCH

'Grand Theft' of intellectual property

By David Becker, Special to ZDNet
22 October 2004

Add your opinion TALK BACK!

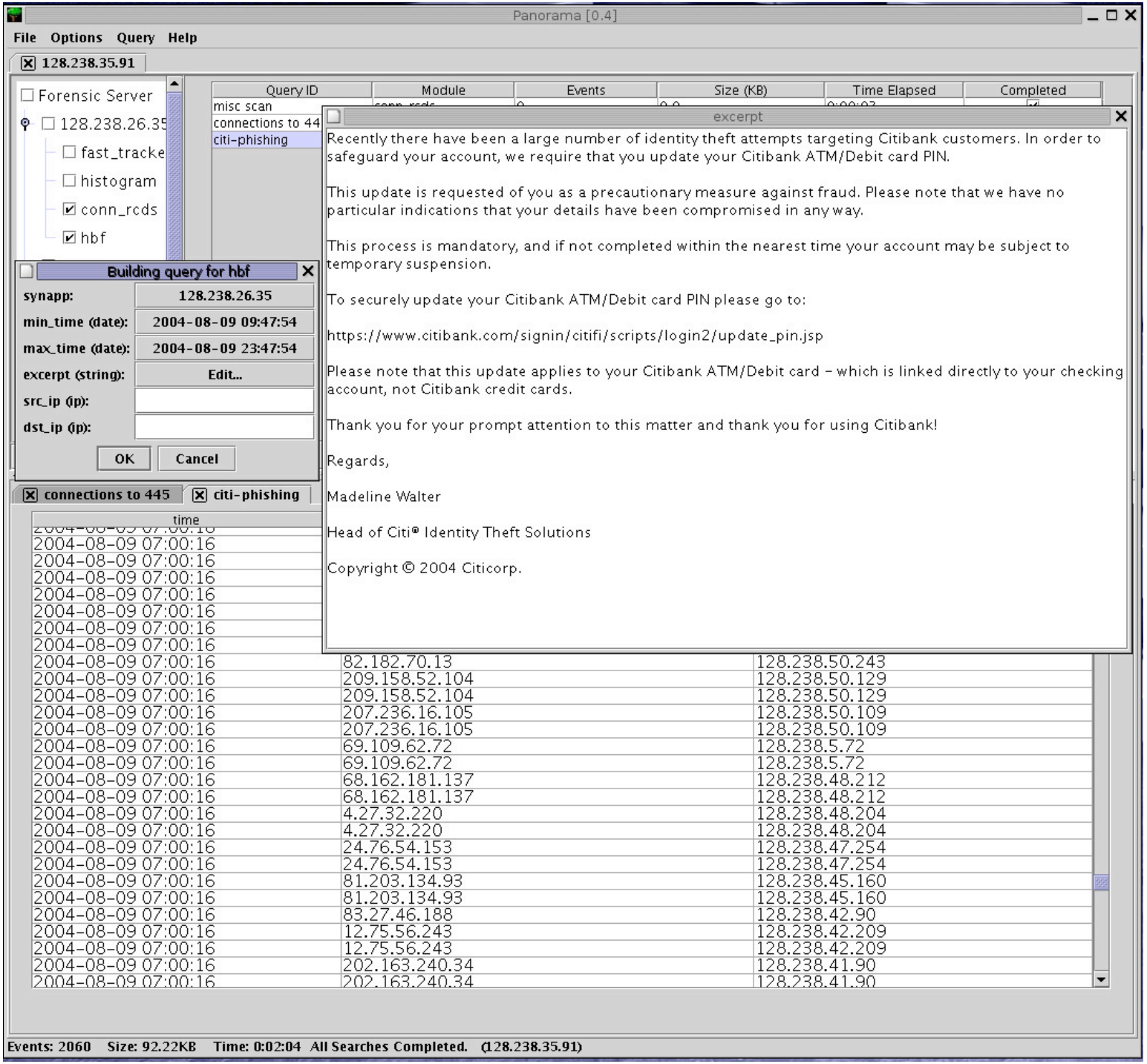
Forward in E-MAIL Format for PRINTER

A stolen copy of the latest sequel in one of the top-selling video game series of all time began circulating on the Web late Wednesday in the US, the second high-profile game theft in a week.

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
- Our contribution:
  - Payload attribution system
    - Currently in alpha-test
  - Not a foolproof system
  - Very useful in most cases

Polytechnic UNIVERSITY

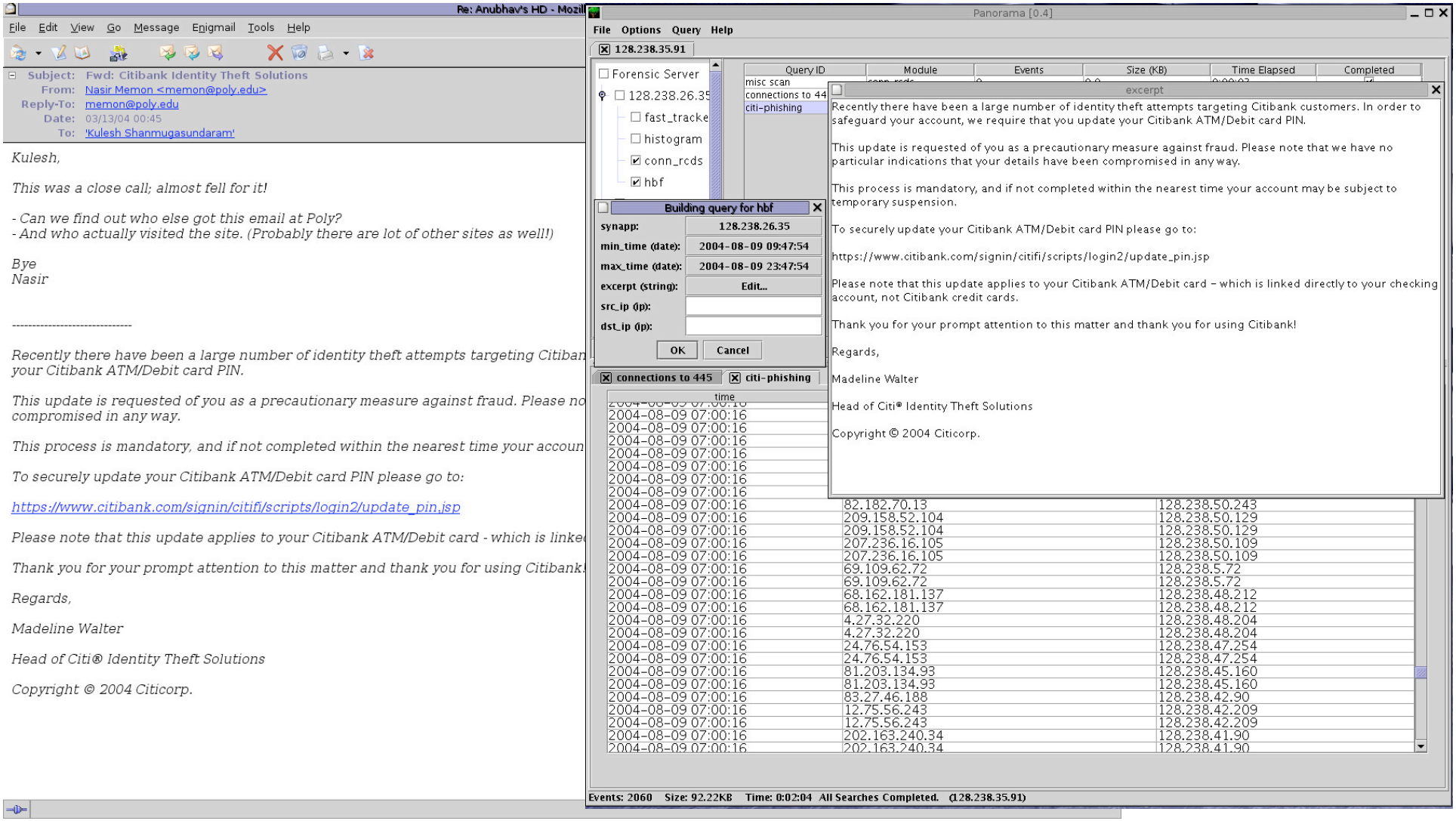

Panorama [0.4]
File Options Query Help
128.238.35.91
Forensic Server
128.238.26.35
fast_tracke
histogram
conn_rcds
hbf
Query ID
Module
Events
Size (KB)
Time Elapsed
Completed
misc scan
connections to 44
citi-phishing
Building query for hbf
synapp:
128.238.26.35
min_time (date):
2004-08-09 09:47:54
max_time (date):
2004-08-09 23:47:54
excerpt (string):
Edit...
src_ip (ip):
dst_ip (ip):
OK
Cancel
excerpt
Recently there have been a large number of identity theft attempts targeting Citibank customers. In order to safeguard your account, we require that you update your Citibank ATM/Debit card PIN.
This update is requested of you as a precautionary measure against fraud. Please note that we have no particular indications that your details have been compromised in any way.
This process is mandatory, and if not completed within the nearest time your account may be subject to temporary suspension.
To securely update your Citibank ATM/Debit card PIN please go to:
https://www.citibank.com/signin/citifi/scripts/login2/update_pin.jsp
Please note that this update applies to your Citibank ATM/Debit card – which is linked directly to your checking account, not Citibank credit cards.
Thank you for your prompt attention to this matter and thank you for using Citibank!
Regards,
Madeline Walter
Head of Citi® Identity Theft Solutions
Copyright © 2004 Citicorp.
connections to 445
citi-phishing
time
2004-08-09 07:00:16 82.182.70.13 128.238.50.243
2004-08-09 07:00:16 209.158.52.104 128.238.50.129
2004-08-09 07:00:16 209.158.52.104 128.238.50.129
2004-08-09 07:00:16 207.236.16.105 128.238.50.109
2004-08-09 07:00:16 207.236.16.105 128.238.50.109
2004-08-09 07:00:16 69.109.62.72 128.238.5.72
2004-08-09 07:00:16 69.109.62.72 128.238.5.72
2004-08-09 07:00:16 68.162.181.137 128.238.48.212
2004-08-09 07:00:16 68.162.181.137 128.238.48.212
2004-08-09 07:00:16 4.27.32.220 128.238.48.204
2004-08-09 07:00:16 4.27.32.220 128.238.48.204
2004-08-09 07:00:16 24.76.54.153 128.238.47.254
2004-08-09 07:00:16 24.76.54.153 128.238.47.254
2004-08-09 07:00:16 81.203.134.93 128.238.45.160
2004-08-09 07:00:16 81.203.134.93 128.238.45.160
2004-08-09 07:00:16 83.27.46.188 128.238.42.90
2004-08-09 07:00:16 12.75.56.243 128.238.42.209
2004-08-09 07:00:16 12.75.56.243 128.238.42.209
2004-08-09 07:00:16 202.163.240.34 128.238.41.90
Events: 2060 Size: 92.22KB Time: 0:02:04 All Searches Completed. (128.238.35.91)

Have Problem. Will Solve.

Packet-loggers:
- Need ~1TB/day
- IS is not going to like this

Hash-packets:
- Storage is better
- Need packets for attribution!
  - But we only have emails/excerpts!

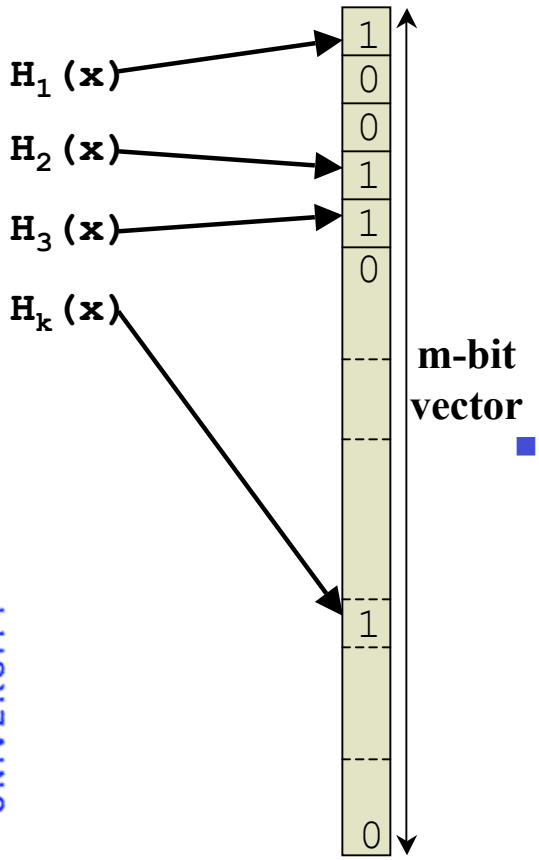
Polytechnic UNIVERSITY



- The Problem: *Identify the sources/ destinations of a bit-string in a network*

- Our Solution: *Create digests of payload s.t we can:*
  - Attribute excerpts of payload
  - Reduce storage requirements

Bloom Filters



- Bloom Filter:
  - Randomized data structure for representing a set in order to support membership queries.
  - **`Insert(x)`:**
    - Flip bits $H_1(x)$... $H_k(x)$ to '1'
  - **`IsMember(y)`:**
    - If $H_1(Y)$ ... $H_k(Y)$ *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

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

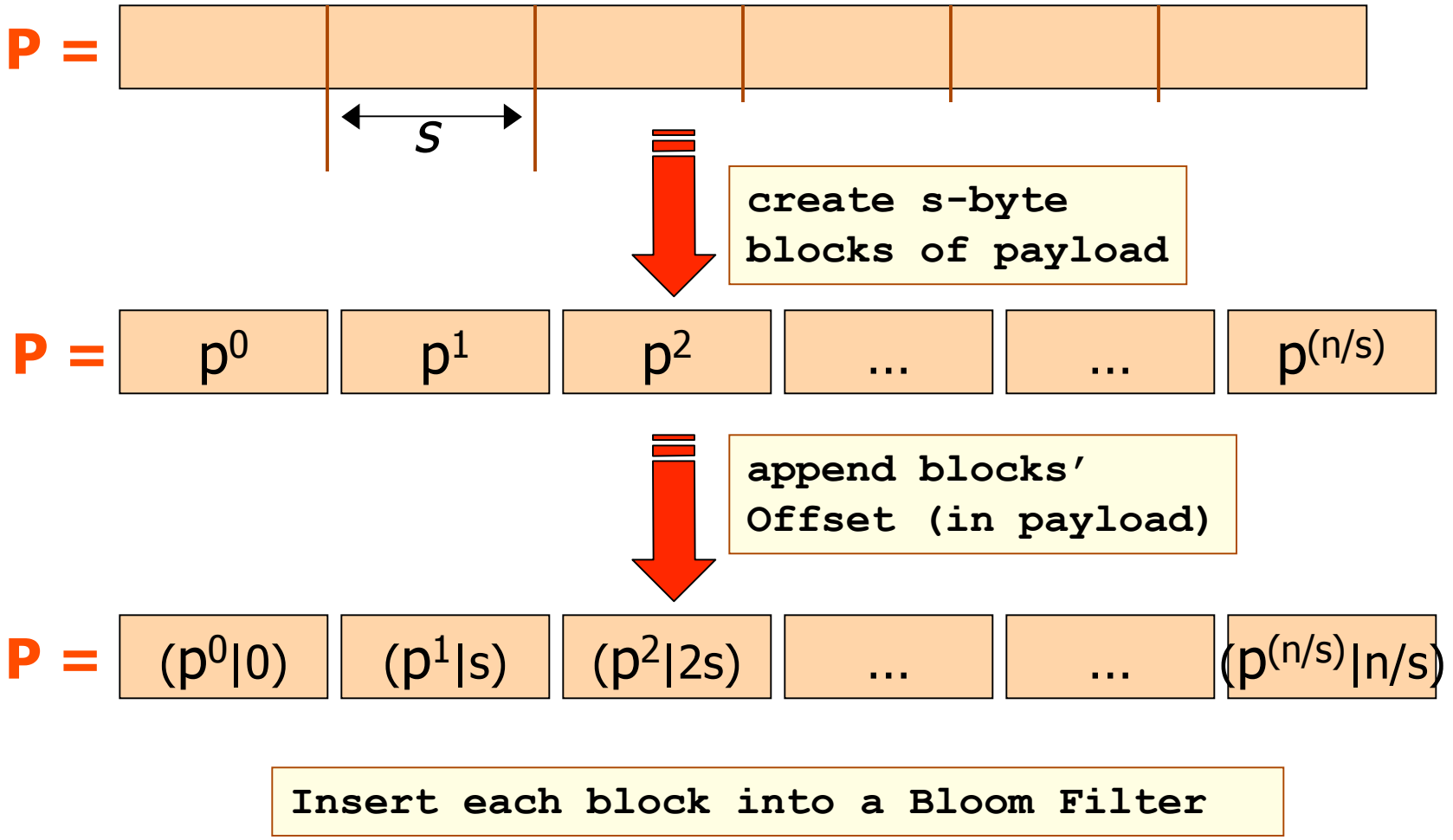
Polytechnic UNIVERSITY

Packet Digests & Bloom Filters

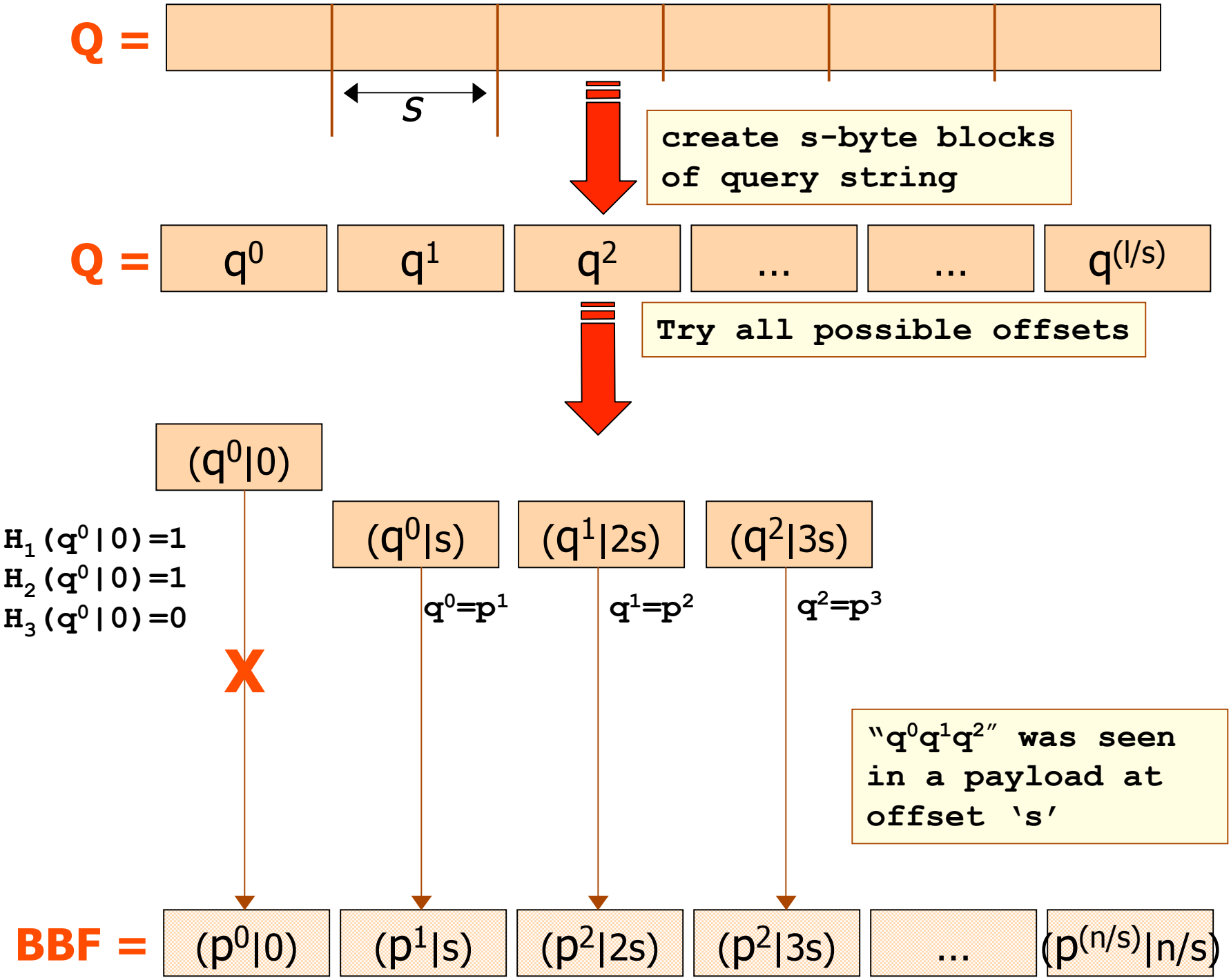
- Snoeren et. al. used it successfully in SPIE for single packet traceback ("**Hash-Based IP Traceback**")
- Space Efficient:
  - 16-bits per packet (m/n=16) and 8 hashes (k=8)
    false positive (FP) = 5.74×10^{-4}
  - No false negatives!
- However, we don't have packets.
  - We only have some excerpt of payload
  - Don't know where the excerpt was aligned in the packet
- Extend Bloom Filters to support excerpt/substring matching

Polytechnic UNIVERSITY

Block-based Bloom Filter



Polytechnic UNIVERSITY


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

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"

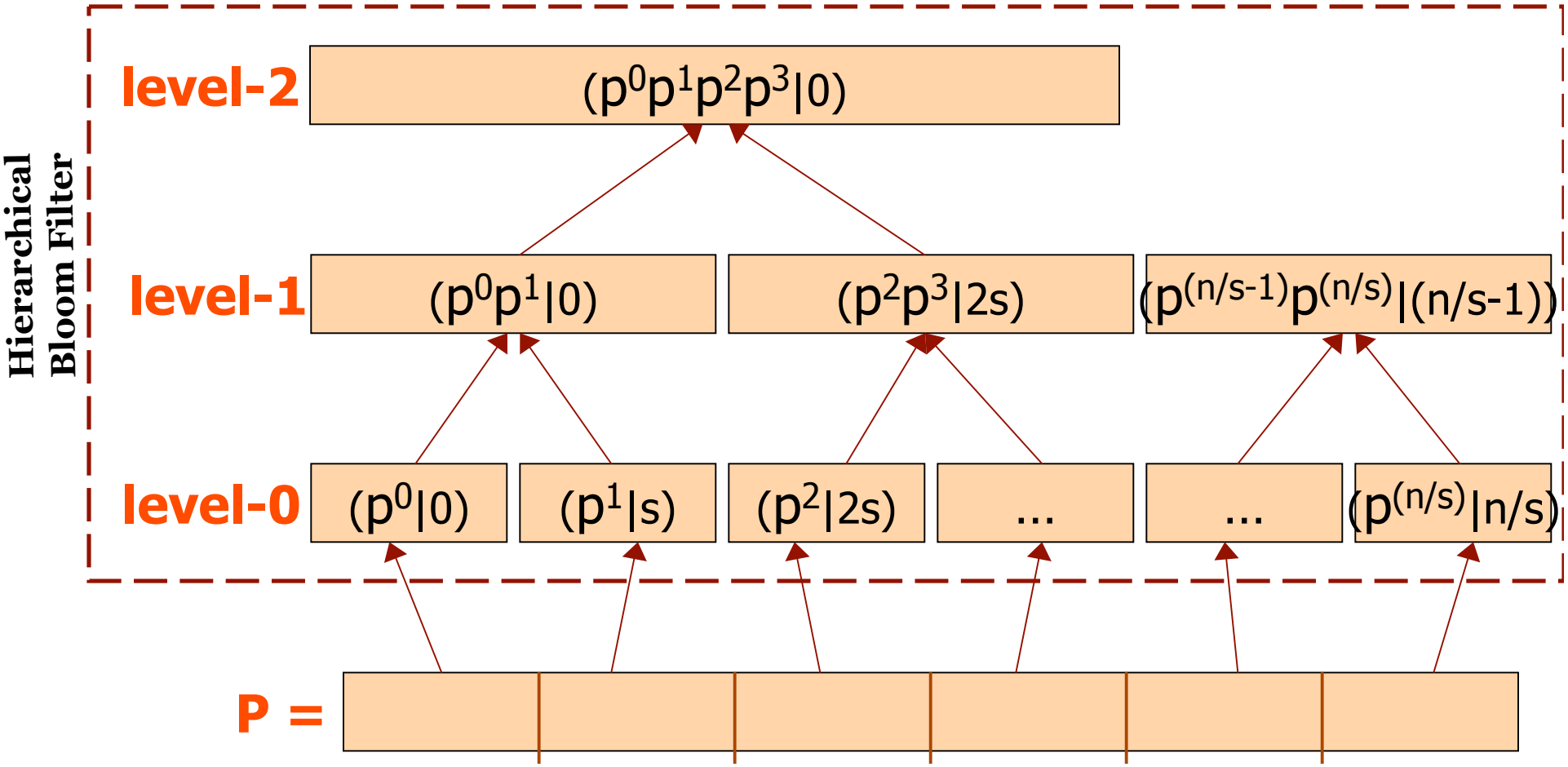
(A|0)	(B|s)	(R|2s)	(A|3s)	(C|4s)	(A|5s)
(C|0)	(D|s)	(A|2s)	(B|3s)	(R|4s)	(A|5s)

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

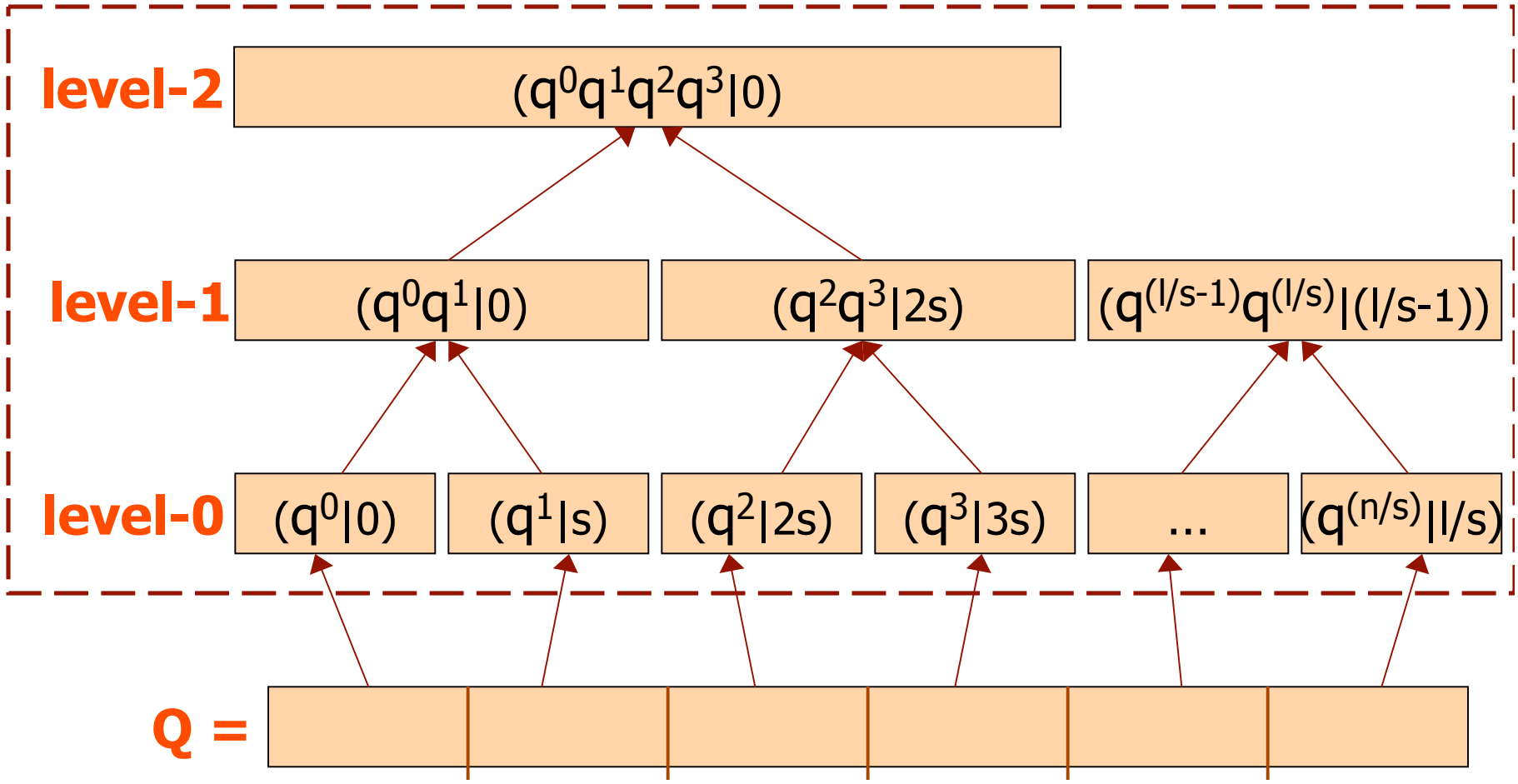
Because BBF cannot distinguish between P1 and P2

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.

Hierarchical Bloom Filter

- Now we have a data structure that:
  - Allows us to do substring matching
  - Avoids "*offset collisions*"
  - Improves accuracy over standard Bloom Filter and BBF

$$FP_e^{HBF} << FP_e^{BBF} << FP$$

HBF Performance Summary:

- Using a Bloom Filter with
  - m/n=5, k=2, FP=0.1090
  - Block size of 128-bytes
- Achieves about 100 fold savings in storage
- For a 512-byte query $FP_{HBF} = 2.00 \times 10^{-5}$

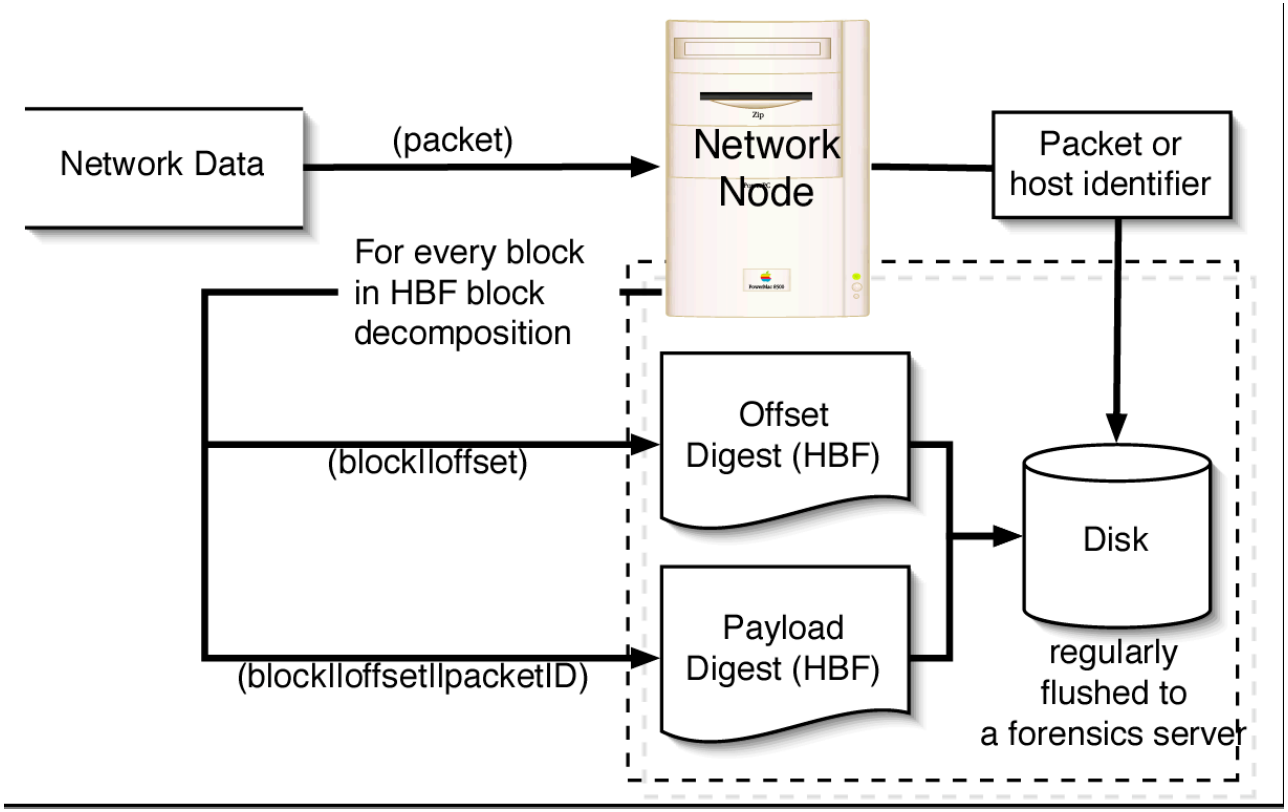
Polytechnic UNIVERSITY

Adapting an HBF for PAS

- 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)

Polytechnic UNIVERSITY

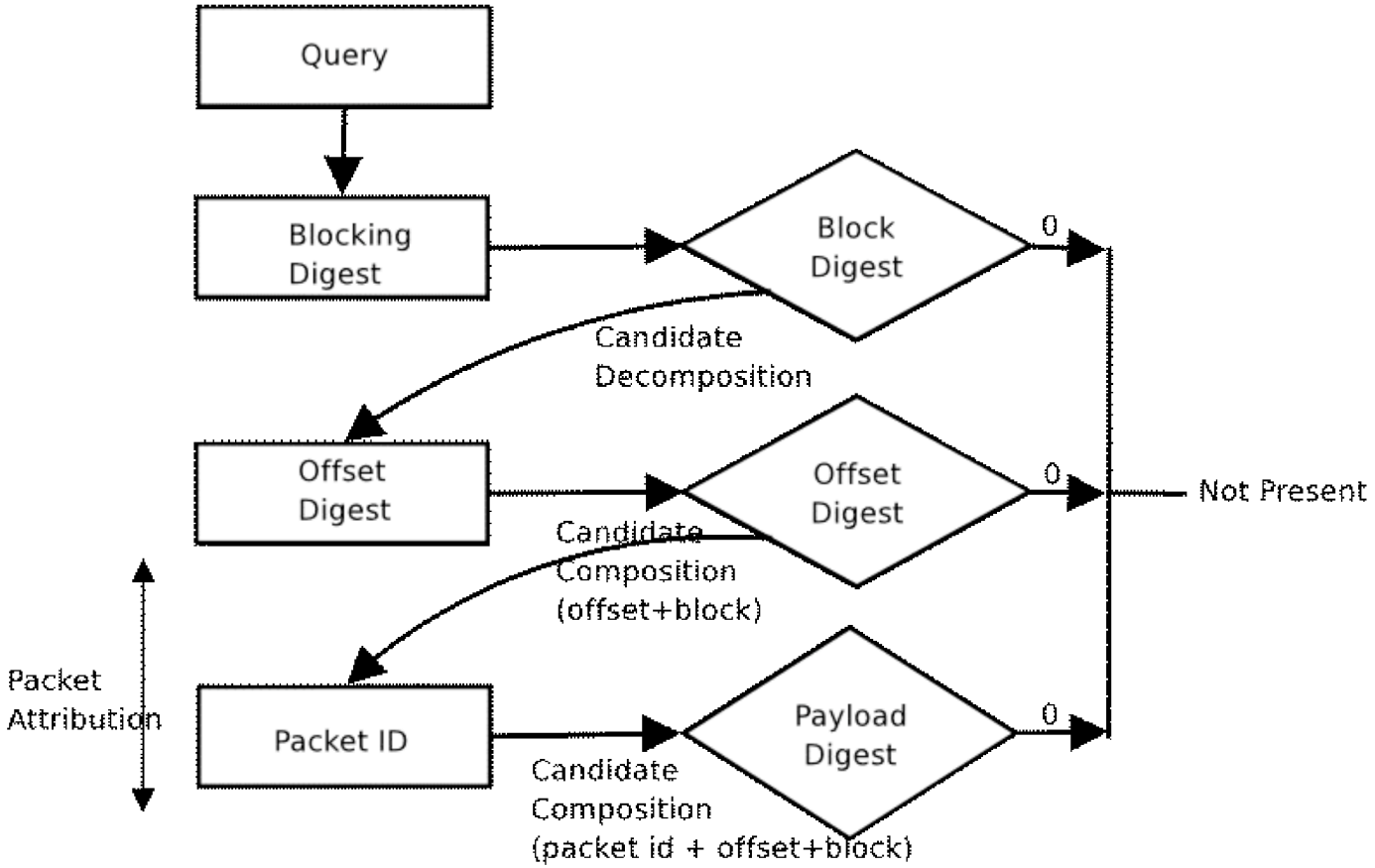
A Payload Attribution System (PAS)



- System design of a payload attribution system
  - Packet id or host identifier is (SourceIP||DestIP)
  - Although host identifier can be obtained from firewalls and routers (Netflow), a list of host ids is maintained by the system

Polytechnic UNIVERSITY

How to run a query?



Polytechnic UNIVERSITY

A Bayesian approach

- Trying to measure:
  - *P* = *conditional* probability of offset collision (i.e., probability that when two equal blocks at the same position in two packets, the next blocks are also equal)
- *Turns out to be too pessimistic for practical use:*

Query Blocks	SMTP	HTTP	FTP
16	0.542259	0.562018	0.168082
32	0.762793	0.547919	0.050411
64	0.410678	0.362524	0.006253
128	0.313077	0.322957	0.003306
256	0.213254	0.282814	0.003717

Table 1: Probabilities of offset collisions by protocol, measured on a network trace of 1M packets.
Note: these are *not* the false positive rates of our filters, see Table 2 for those.

- *FP rate analysis must depend on the query string to be useful!*

Polytechnic UNIVERSITY

FP Rate Analysis

FP_e	=	effective FP rate for payload query (after all the filtering)
FP_o	=	individual FP rate of (content) and (content||offset) BF for BBF or HBF,
FP_p	=	individual FP rate of (content||offset||packetID) BF for BBF or HBF
n_b	=	number of blocks inserted in the BBF,
n_h	=	number of blocks inserted in the HBF ($n_b \leq n_h \leq 2n_b$),
N	=	the number of packetIDs to check for the payload digest.

Polytechnic UNIVERSITY

FP Rate (BBF)

- L = max length of a payload
- u = number of blocks not in the query or at wrong offsets (offset collisions)
- v = number of blocks at correct offsets but wrong packet (packet collisions)

To prevent offset collisions, since we must test N packet IDs for each passing combination of blocks with offsets, we expect to see a false positive after both (content||offset) and (content||offset||packetID) BFs with probability at most

$$FP_e^{(BBF)} = L \cdot N \cdot FP_b^u \cdot FP_p^{u+v}.$$

Polytechnic UNIVERSITY

FP Rate (HBF)

- L = max length of a payload
- u = number of blocks not in the query or at wrong offsets (offset collisions)
- v = number of blocks at correct offsets but wrong packet (packet collisions)
- u' and v' induced by the hierarchy, and
- u'' = offset collisions in the hierarchy

To prevent offset collisions, since we must test N packet IDs for each passing combination of blocks with offsets, we expect to see a false positive after both (content||offset) and (content||offset||packetID) BFs with probability at most

$$FP_e^{(HBF)} = L \cdot N \cdot (FP'_o)^{u+u'+u''} (FP'_p)^{u+u'+u''+v+v'}.$$

As a conclusion, with equal memory requirements, HBF is never worse than BBF, and sometimes much better.

Note: $u+u' \geq 2u-1$

$u'' \approx v$

$FP' \approx \sqrt{FP}$

Polytechnic UNIVERSITY

Actual FP Rates

	Basic False Positive Rates (FP_o)							
Blocks	0.3930	0.2370	0.1550	0.1090	0.0804	0.0618	0.0489	0.0397
1	1.000000	0.999885	0.996099	0.976179	0.933179	0.870477	0.798657	0.728207
2	0.063758	0.064569	0.048981	0.036060	0.026212	0.021024	0.015881	0.012538
3	0.012081	0.002620	0.000744	0.000275	0.000172	0.000046	0.000023	–
4	0.000820	0.000230	0.000060	0.000020	–	–	–	–
> 4	–	–	–	–	–	–	–	–

Table 2: Measured effective false positive rate (FP_e) of HBF as a function of both the basic false positive rate (FP_o) and the length of the query (in blocks; 1block=32 bytes). Note that for $blocks > 4$, we encountered no false positives, hence the measured FP_e is equal to 0 (indicated by –).

Query Blocks	2	3	4	5
BBF	0.049621	0.035129	0.000560	0.000088
HBF	0.016457	0.000720	0.000110	0.0

Table 3: Performance comparison of a BBF and an HBF with the same memory footprint. (Query strings of size > 5 resulted in 0 measured false positives for both BBF and HBF, hence are not listed.)

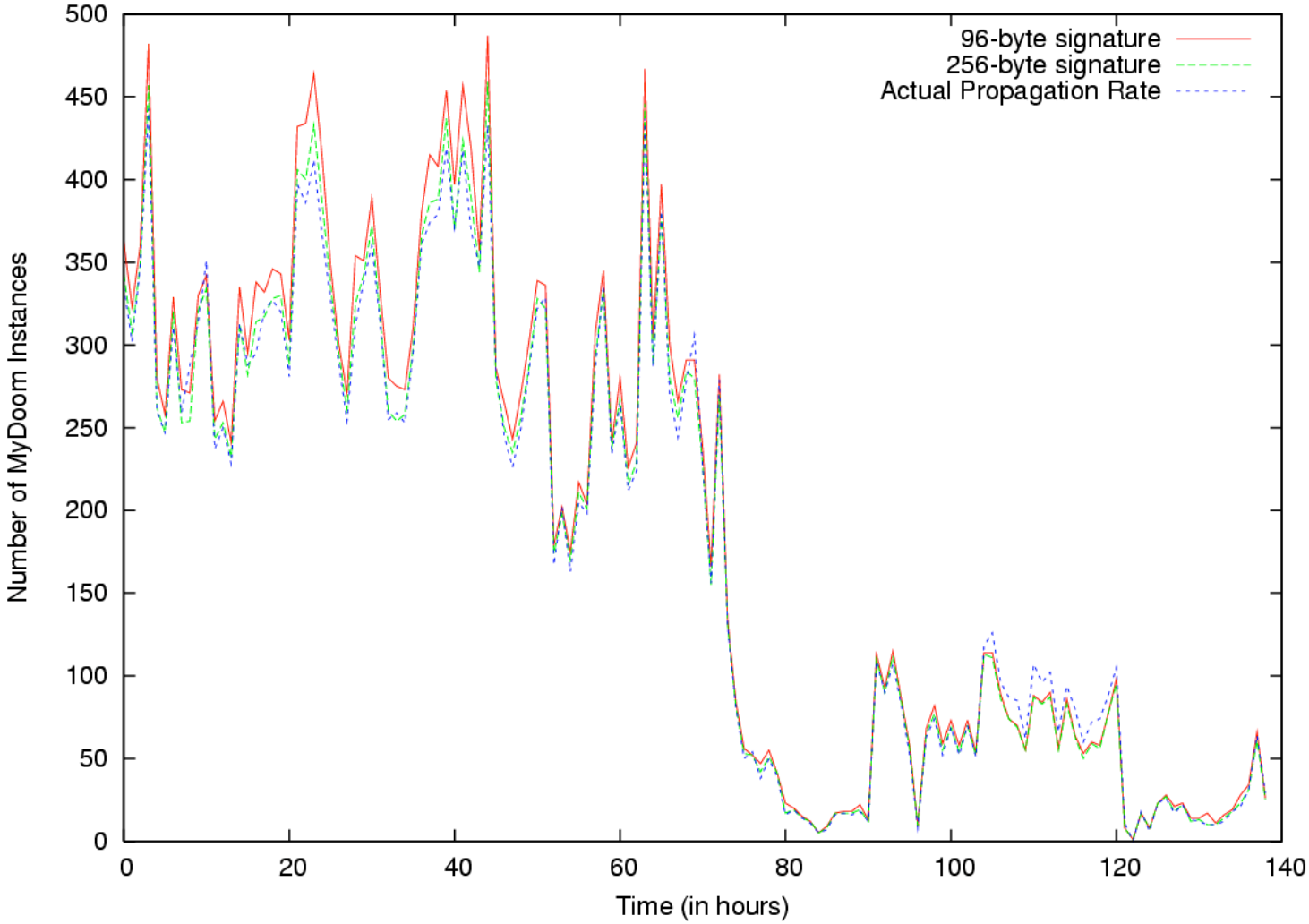
Po

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

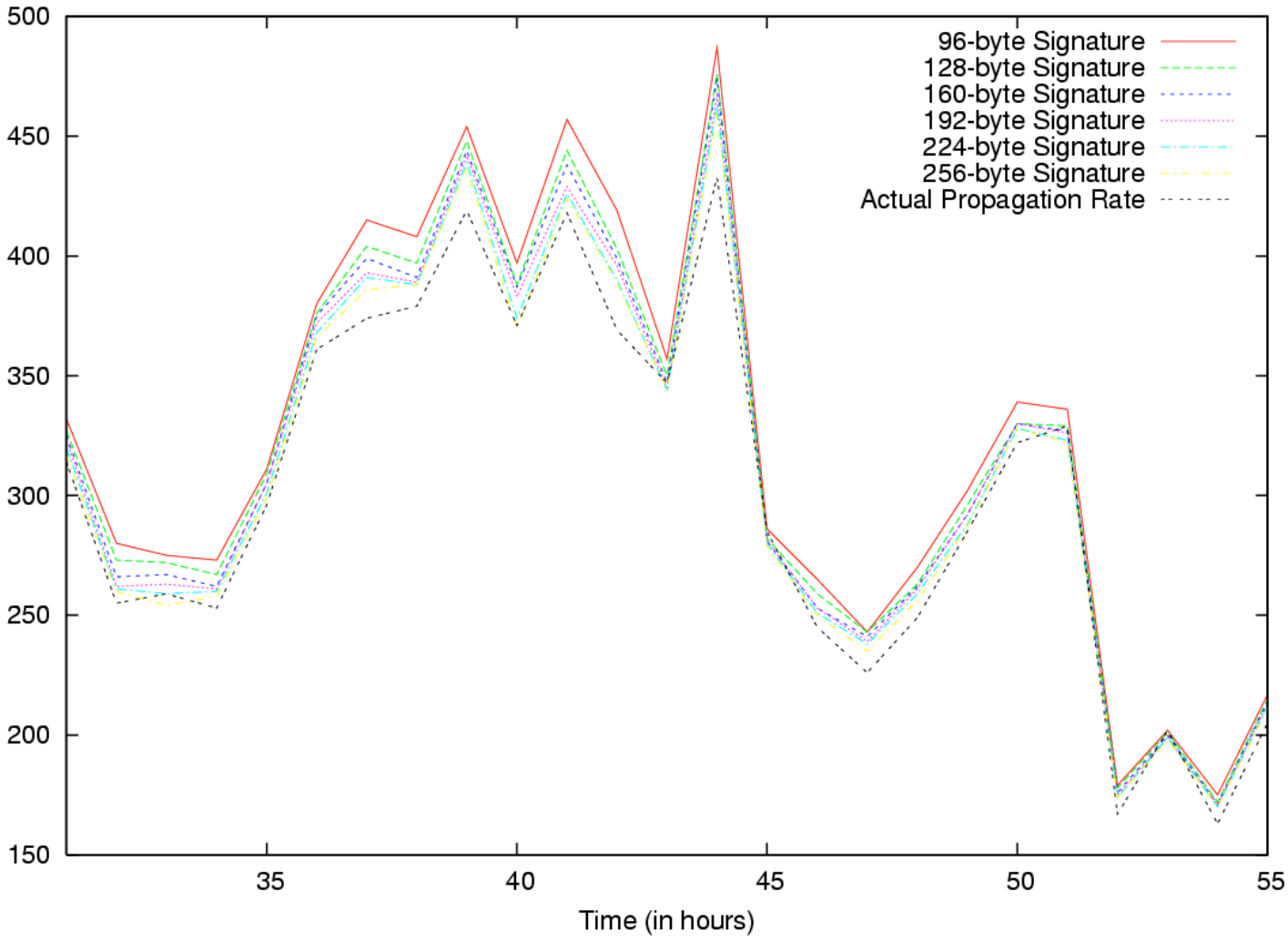
Polytechnic UNIVERSITY

MyDoom's Weekly Progress...



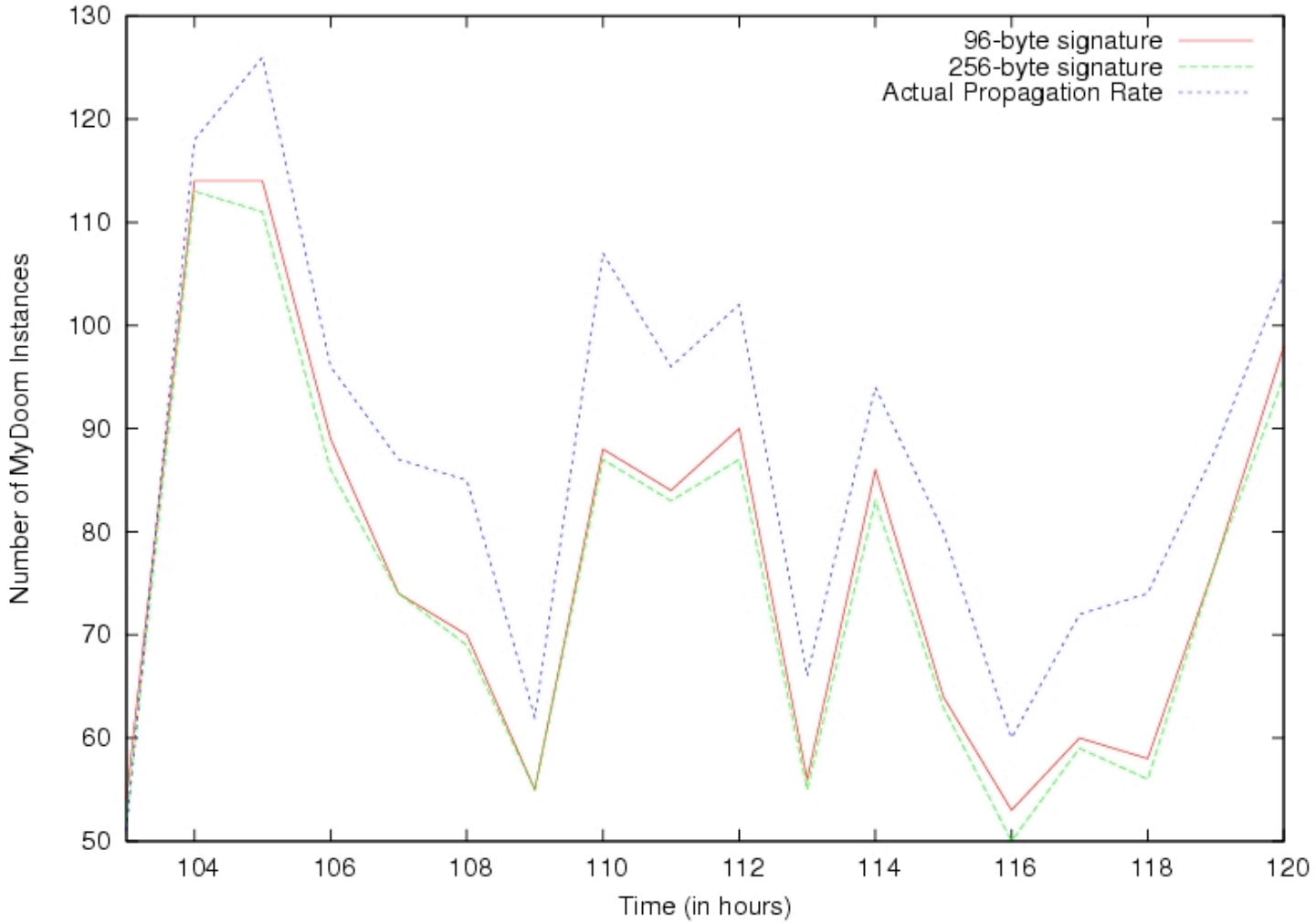
Polytechnic UNIVERSITY

MyDoom's Daily Progress



Polytechnic UNIVERSITY

HBF Cannot Account for Duplicates



Polytechnic UNIVERSITY

False Positives

- Actual number of virus instances seen is 25328
- Number of incorrect attributions for various lengths of the signature used to query the system

Length	96	128	160	192	224	256
Incorrect	1375	932	695	500	293	33

Polytechnic UNIVERSITY

Attacks on PAS

- **Streaming transformations:**
  - Encryption, compression
- **Malicious Transformation:**
  - Create packets of length `(blocksize – 1)`
- **Stuffing:**
  - Stuff every other block with application dependent escape characters
  - For smaller blocks we can try to guess for larger blocks it is not possible!
- **Resource Exhaustion:**
  - Flood the network with random bits of data
- **Exploiting Collisions:**
  - Hash collisions
    - Very unlikely for strong hash functions
    - We use a random seed for every HBF so it makes it more difficult
  - Packet collisions
    - A possibility in Block-based Bloom Filters but not in HBFs

Polytechnic UNIVERSITY

Conclusion & Future Work

- Summary:
  - A data structure for digesting payloads
    - Supports queries on excerpts
    - Reduces storage requirements
    - Provides some privacy guarantees
  - Payload Attribution System:
    - Capable of attributing excerpts of payload to source/destination
    - In alpha-tests in our campus network

- Future work:
  - Value-based blocking using Rabin fingerprints
  - Enhancing storage of host ids
  - Hardware implementations

Polytechnic UNIVERSITY


Questions