Traffic Classification Using Visual Motifs: An Empirical Evaluation

Wilson Lian[1] Fabian Monrose[1] John McHugh[1,2]

[1]University of North Carolina at Chapel Hill

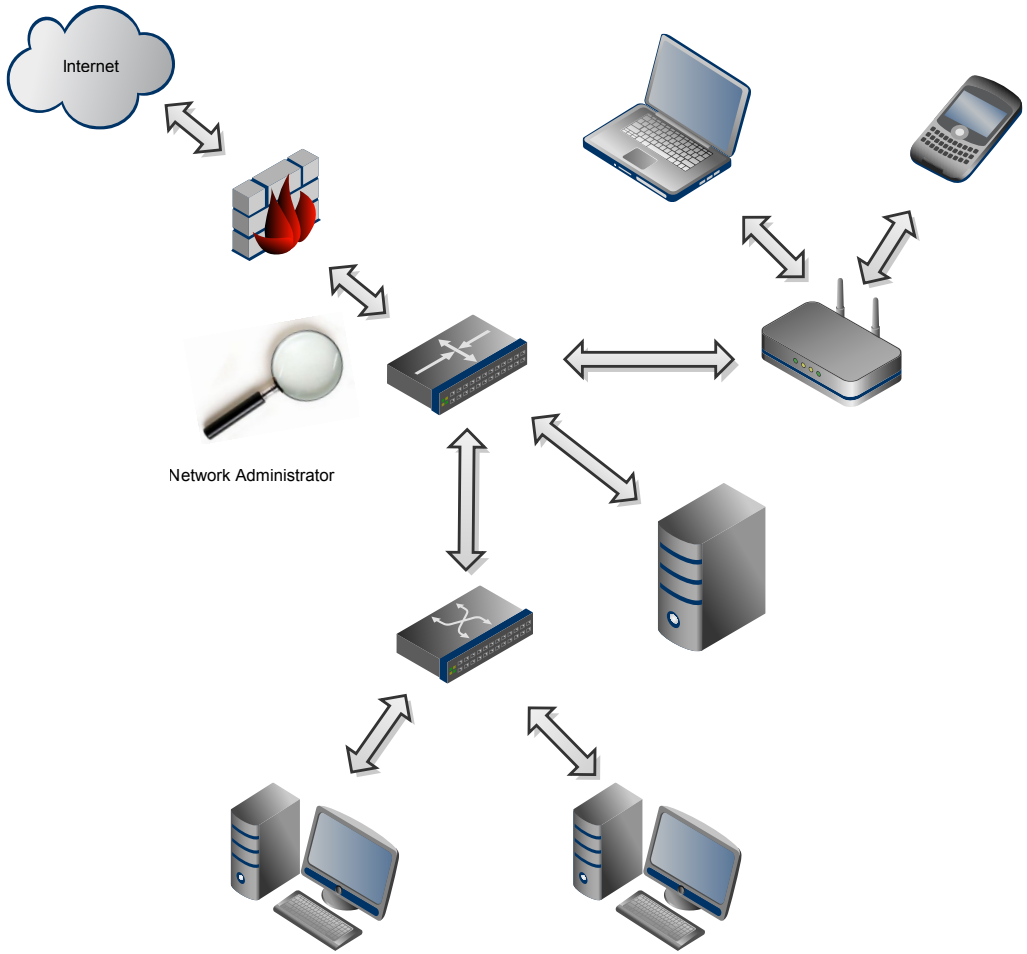
[2]RedJack, LLC

VizSec 2010

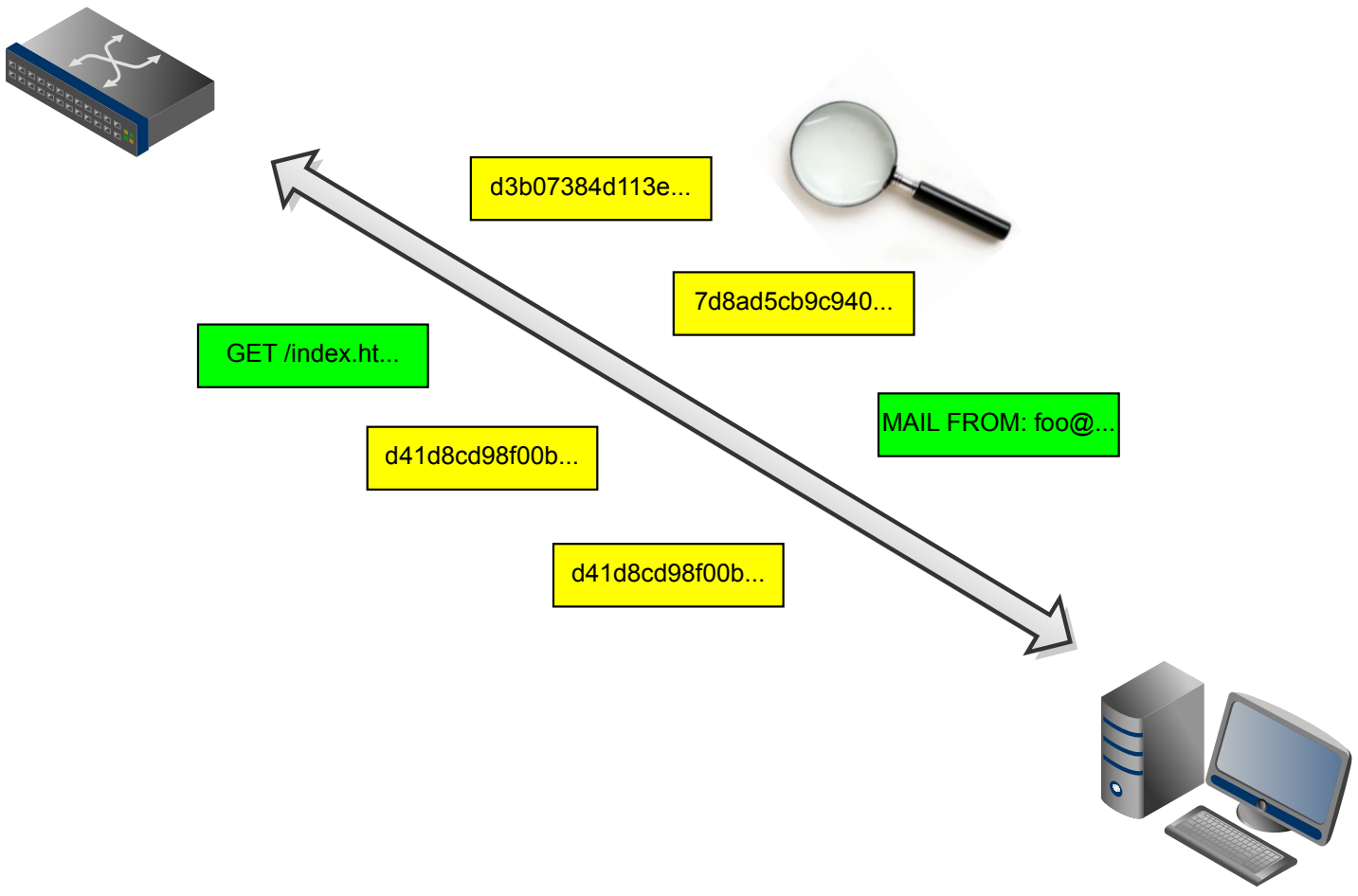
Overview

- Background
- Visual Motifs
- Traffic Classification
- Empirical Evaluation

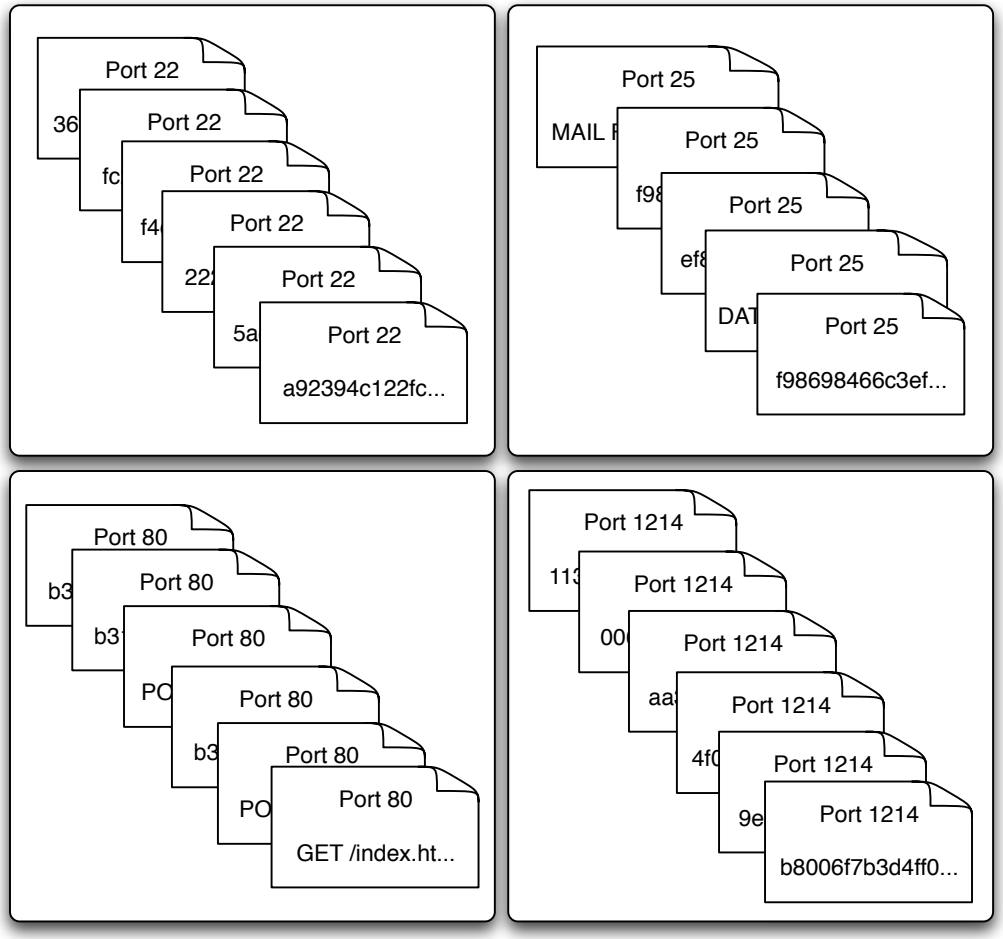
Motivation



Motivation


d3b07384d113e...
7d8ad5cb9c940...
GET /index.ht...
MAIL FROM: foo@...
d41d8cd98f00b...
d41d8cd98f00b...

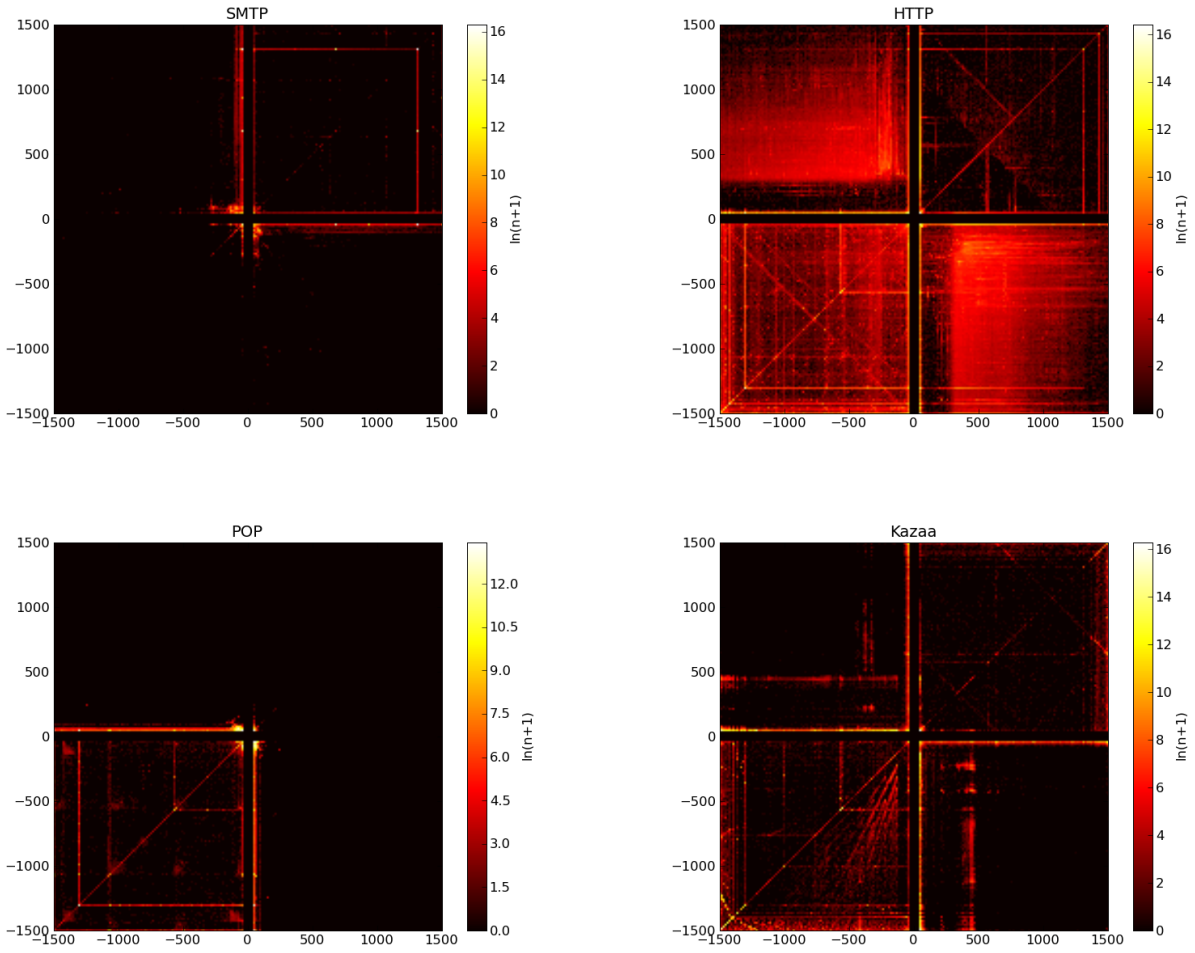
Goals


Port 22
Port 22
Port 22
Port 22
Port 22
Port 22
a92394c122fc...
Port 25
Port 25
Port 25
Port 25
Port 25
f98698466c3ef...
Port 80
Port 80
Port 80
Port 80
Port 80
Port 80
GET /index.ht...
Port 1214
Port 1214
Port 1214
Port 1214
Port 1214
Port 1214
b8006f7b3d4ff0...

Assumptions

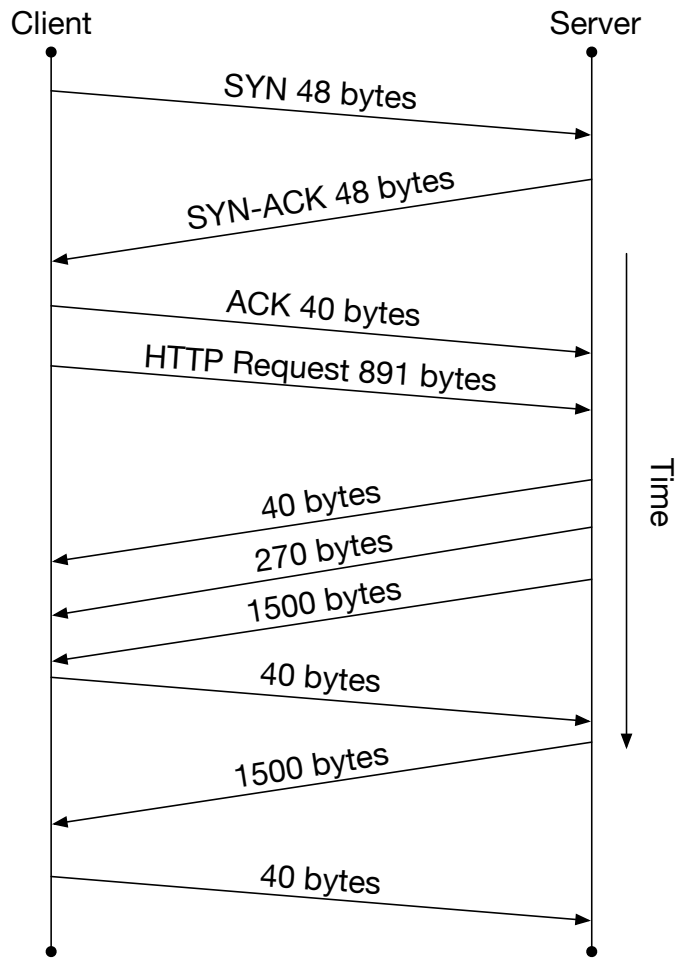
- Reliable transport via TCP
- Stream Cipher
  - No access to payload
  - Payload length preservation
- Encryption at or above Transport Layer

Bigram Heatmaps

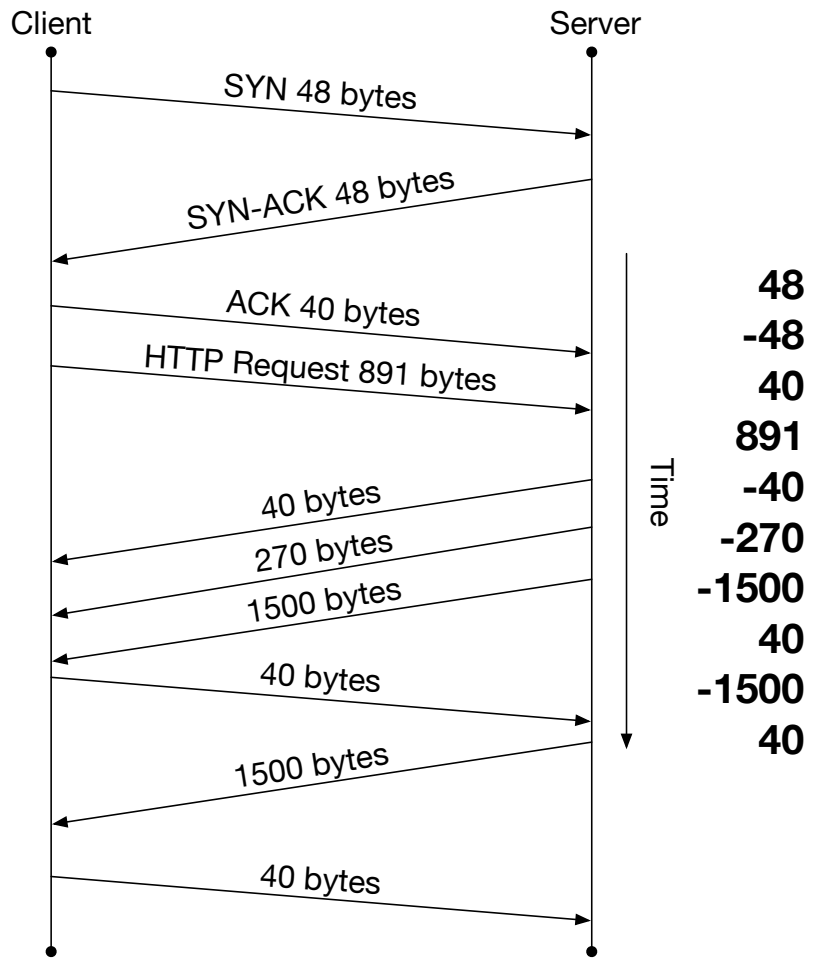


C. Wright, F. Monrose, and G.M. Masson. Using visual motifs to classify encrypted traffic. 2006

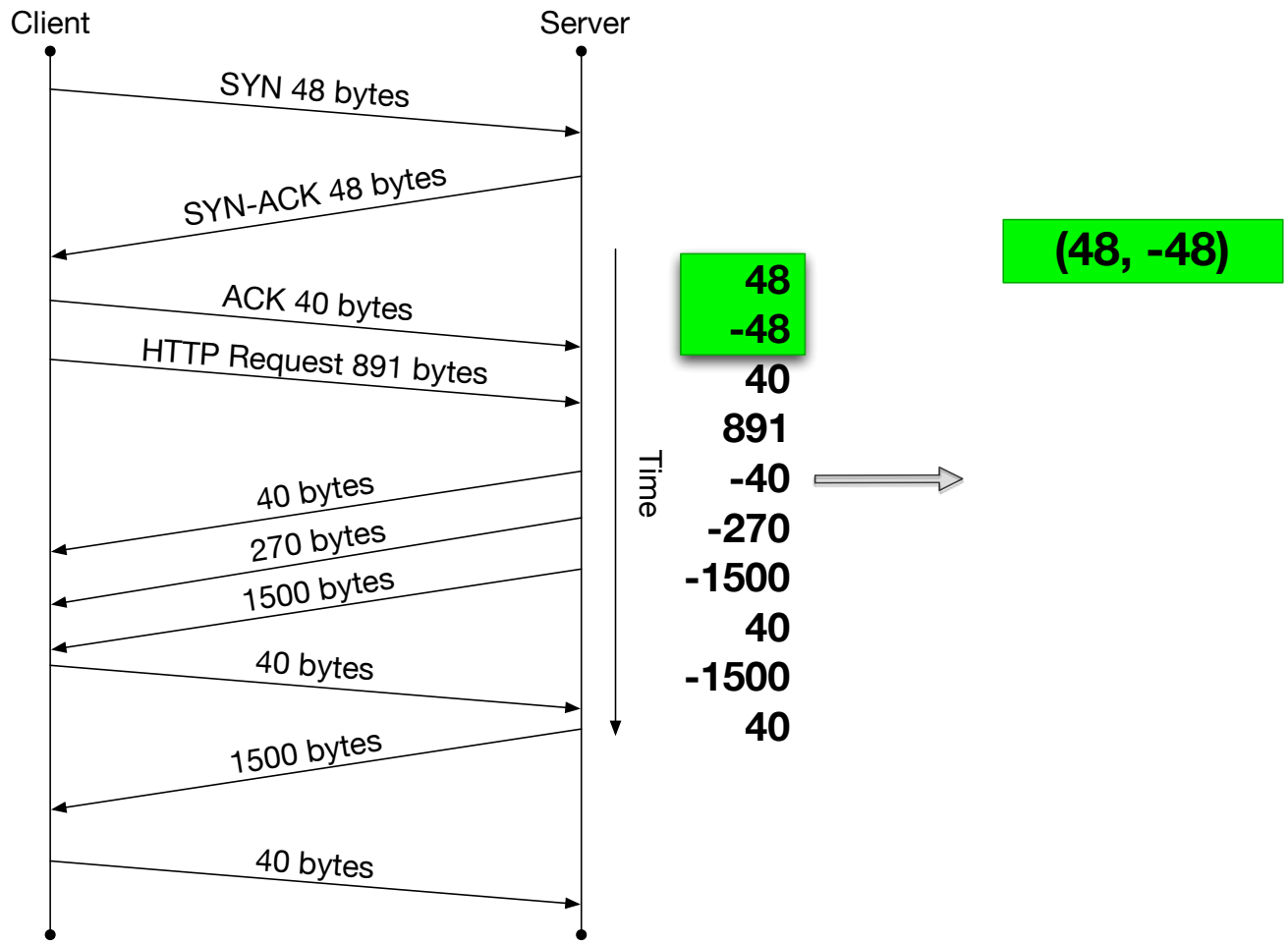
Heatmap Construction



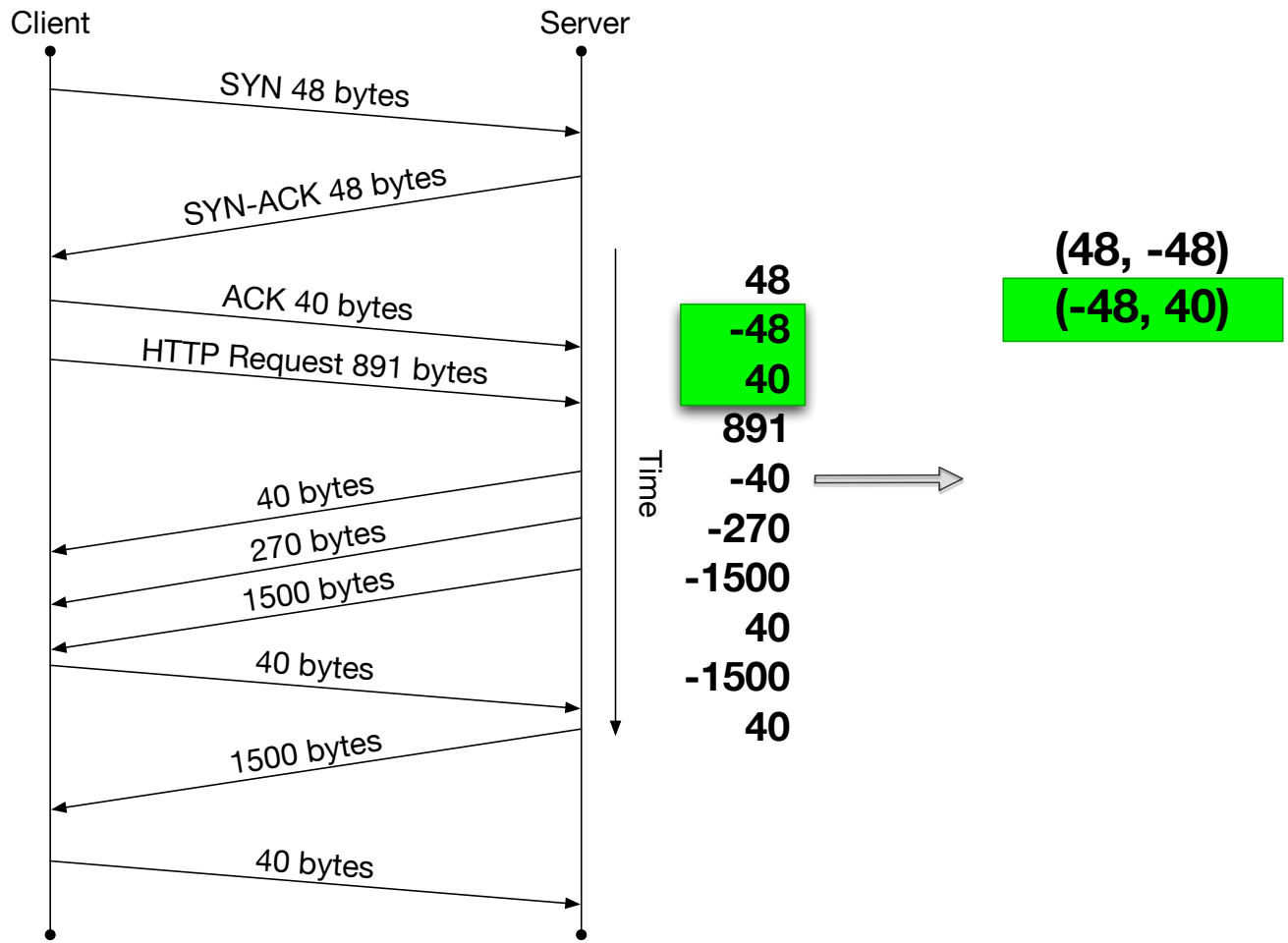
Heatmap Construction



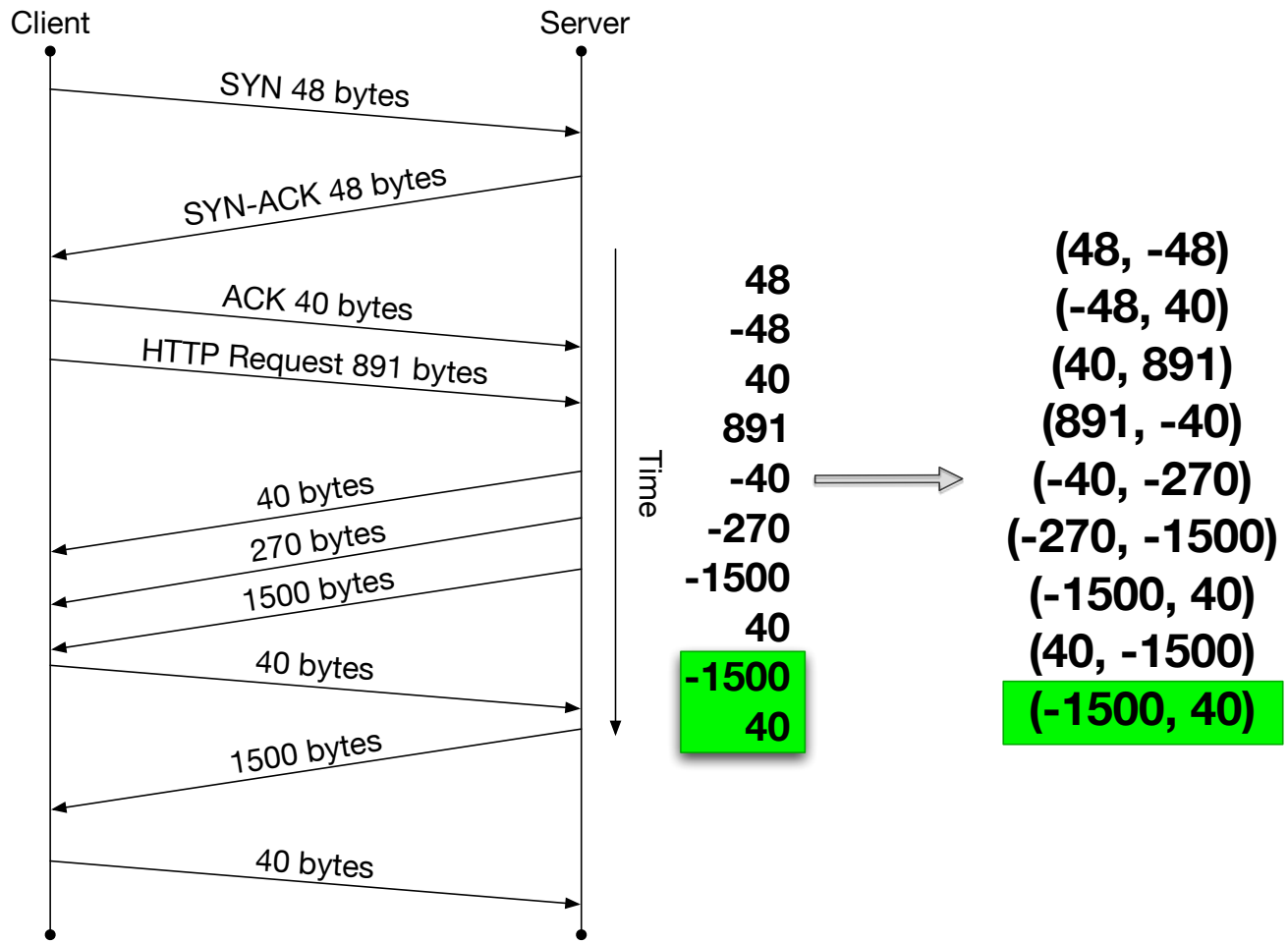
Heatmap Construction



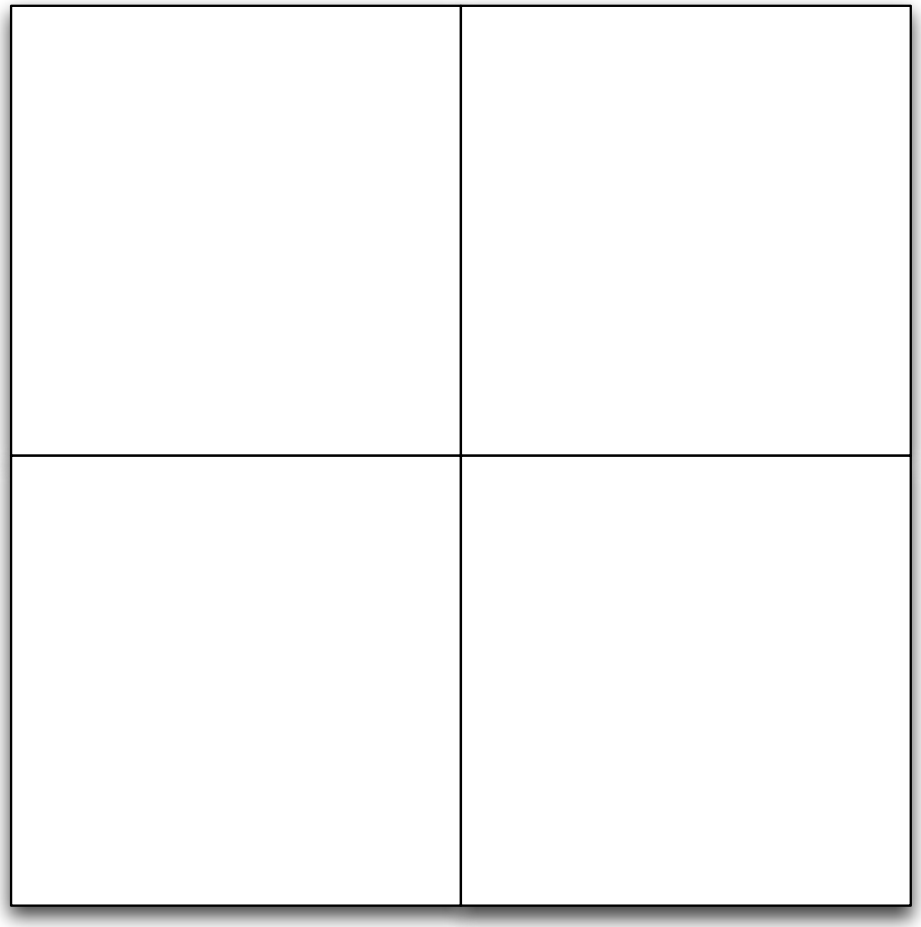
Heatmap Construction



Heatmap Construction



Heatmap Construction

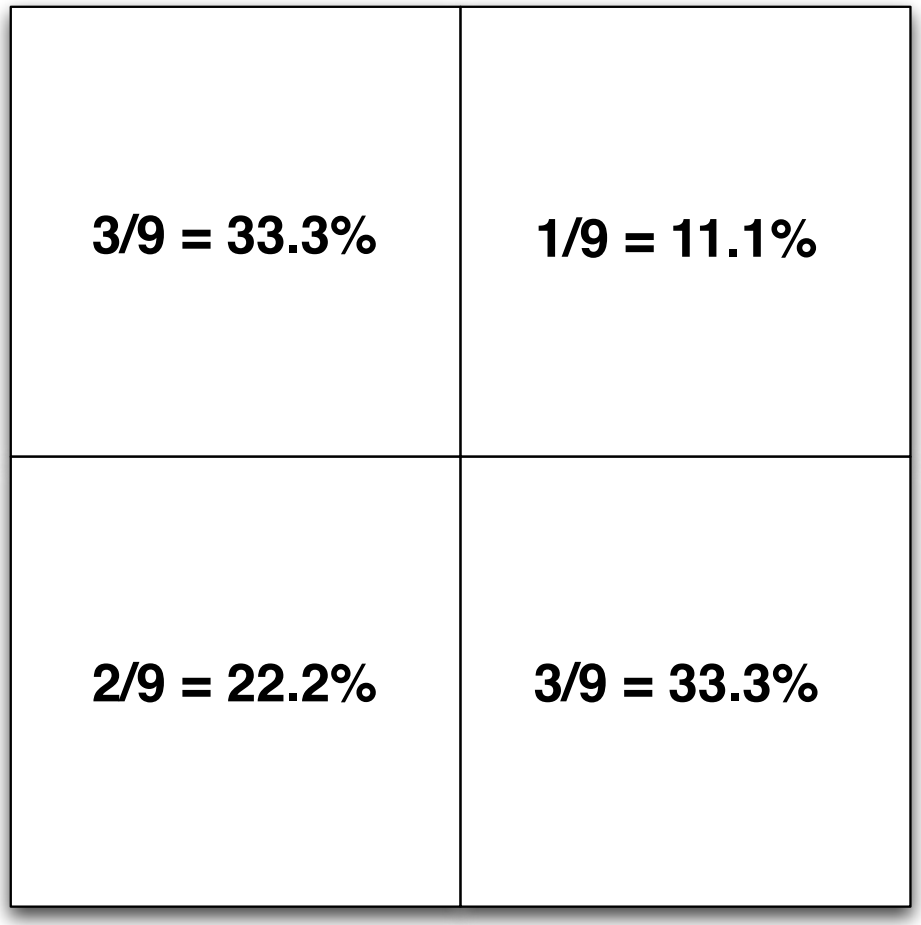


(48, -48)
(-48, 40)
(40, 891)
(891, -40)
(-40, -270)
(-270, -1500)
(-1500, 40)
(40, -1500)
(-1500, 40)

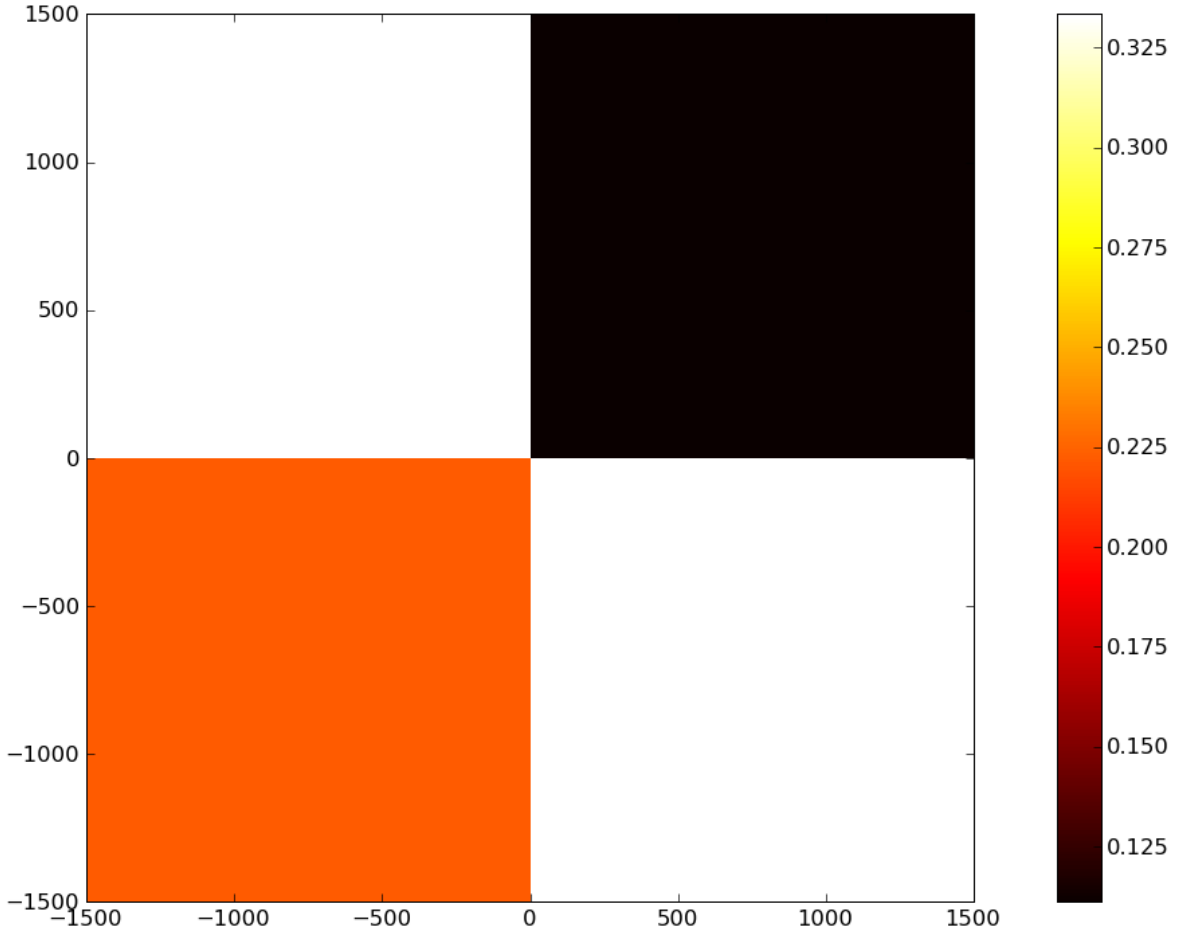
Heatmap Construction

(-48, 40) **(-1500, 40)** **(-1500, 40)**	**(40, 891)**
(-40, -270) **(-270, -1500)**	**(48, -48)** **(891, -40)** **(40, -1500)**

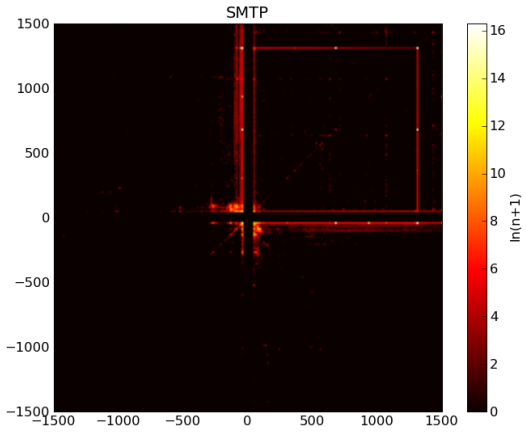
Heatmap Construction

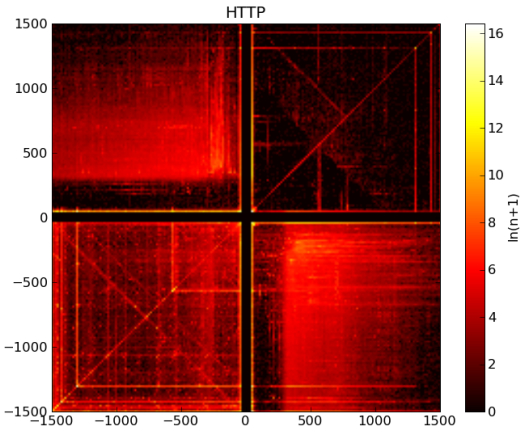


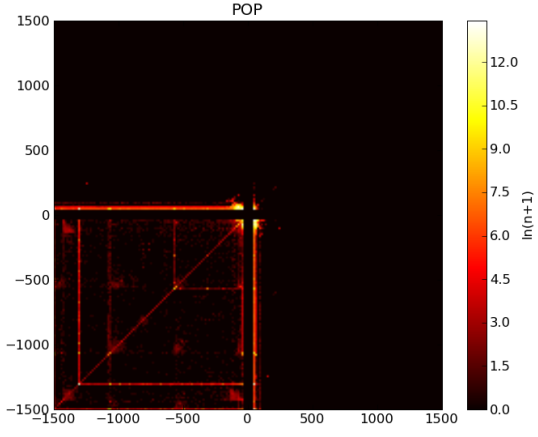
Heatmap Construction

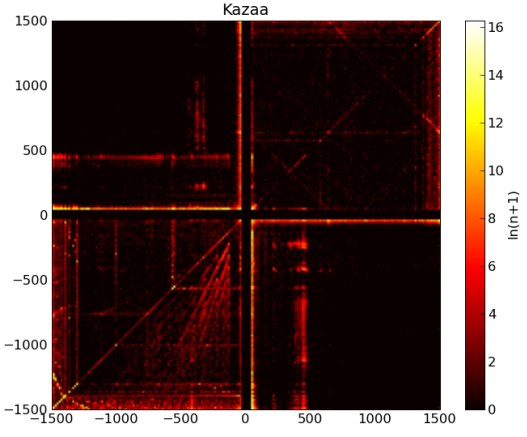


Bigram Heatmaps





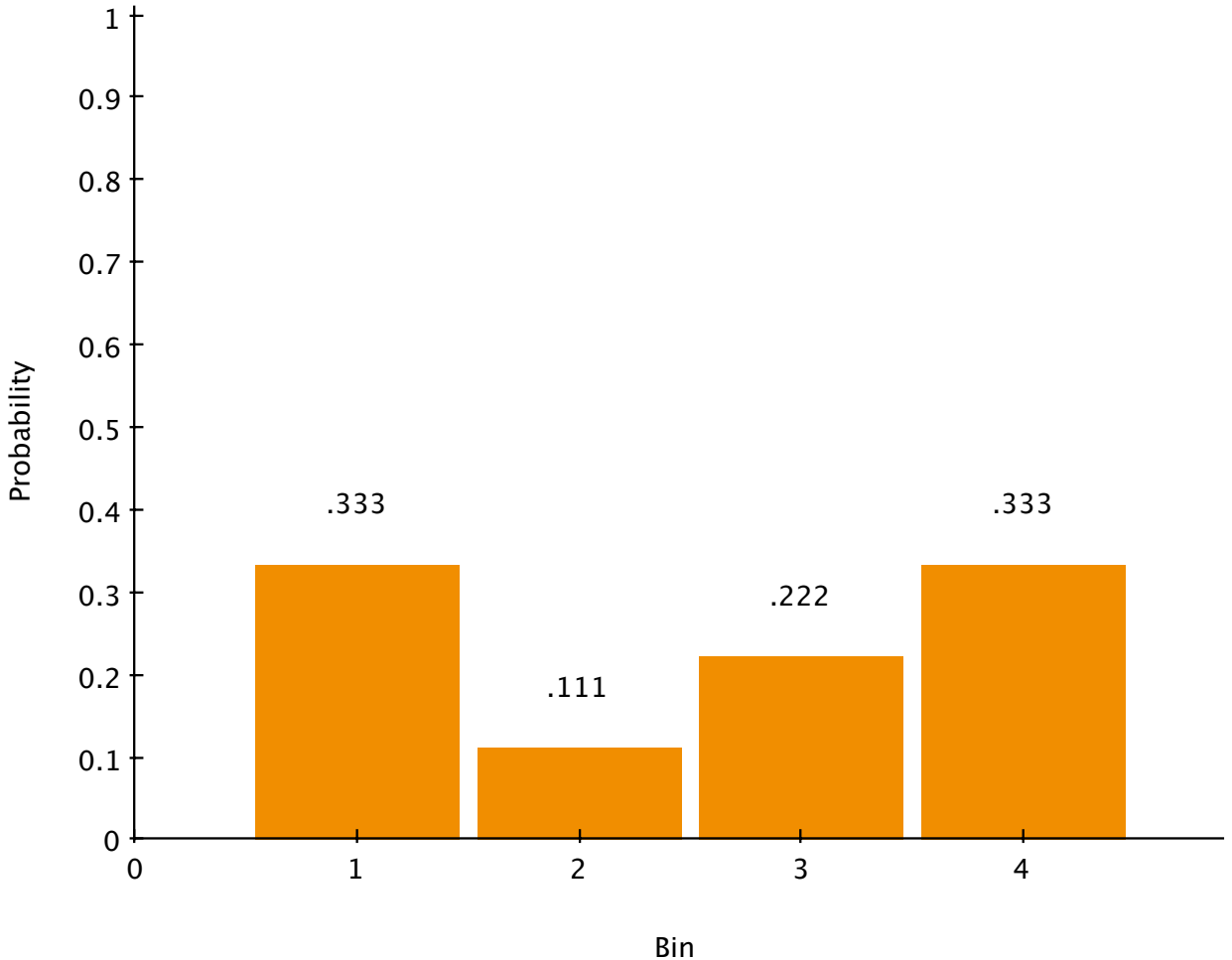




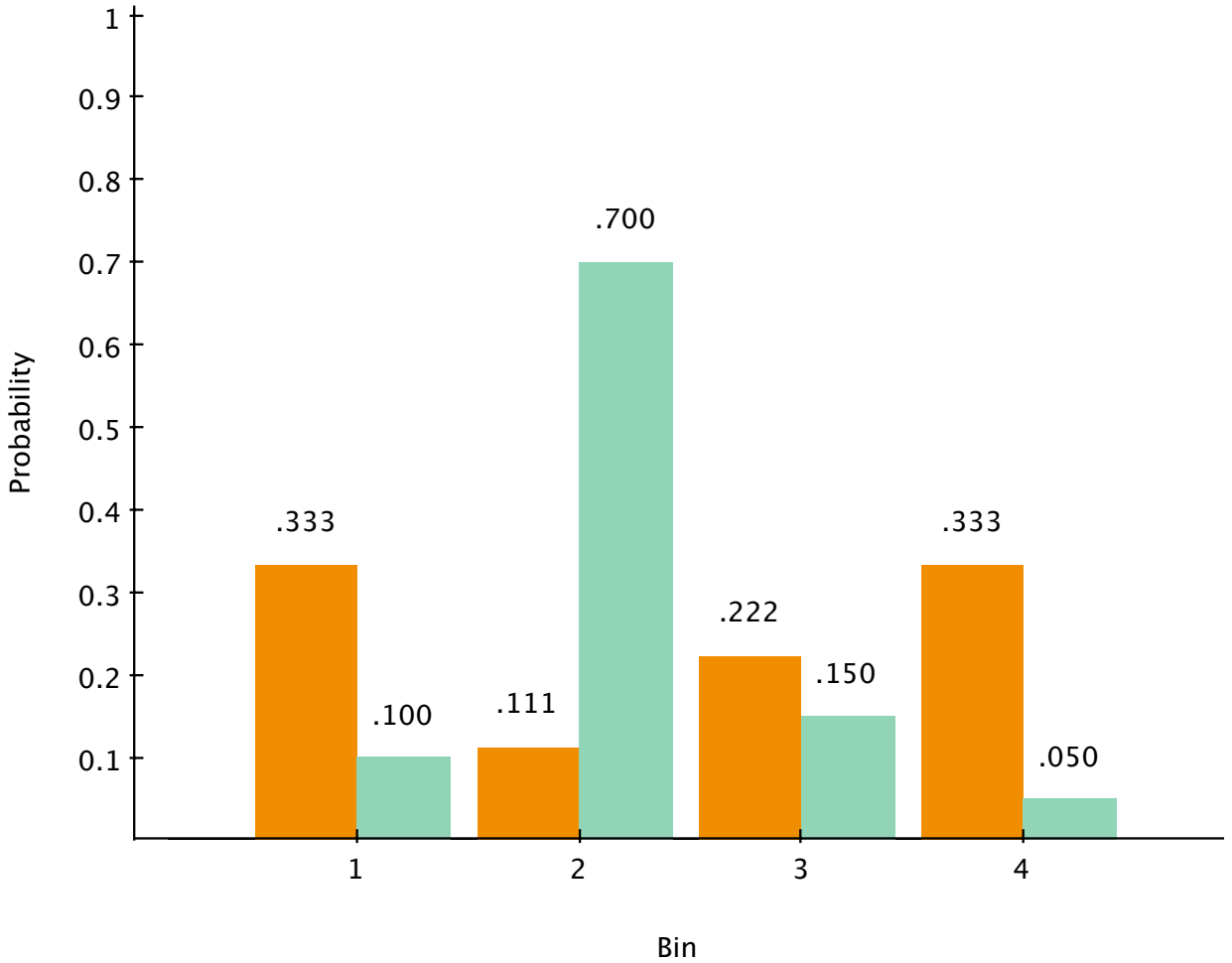
Modeling Protocol Behavior

3/9 = 33.3% **1**	**1/9 = 11.1%** **2**
2/9 = 22.2% **3**	**3/9 = 33%** **4**

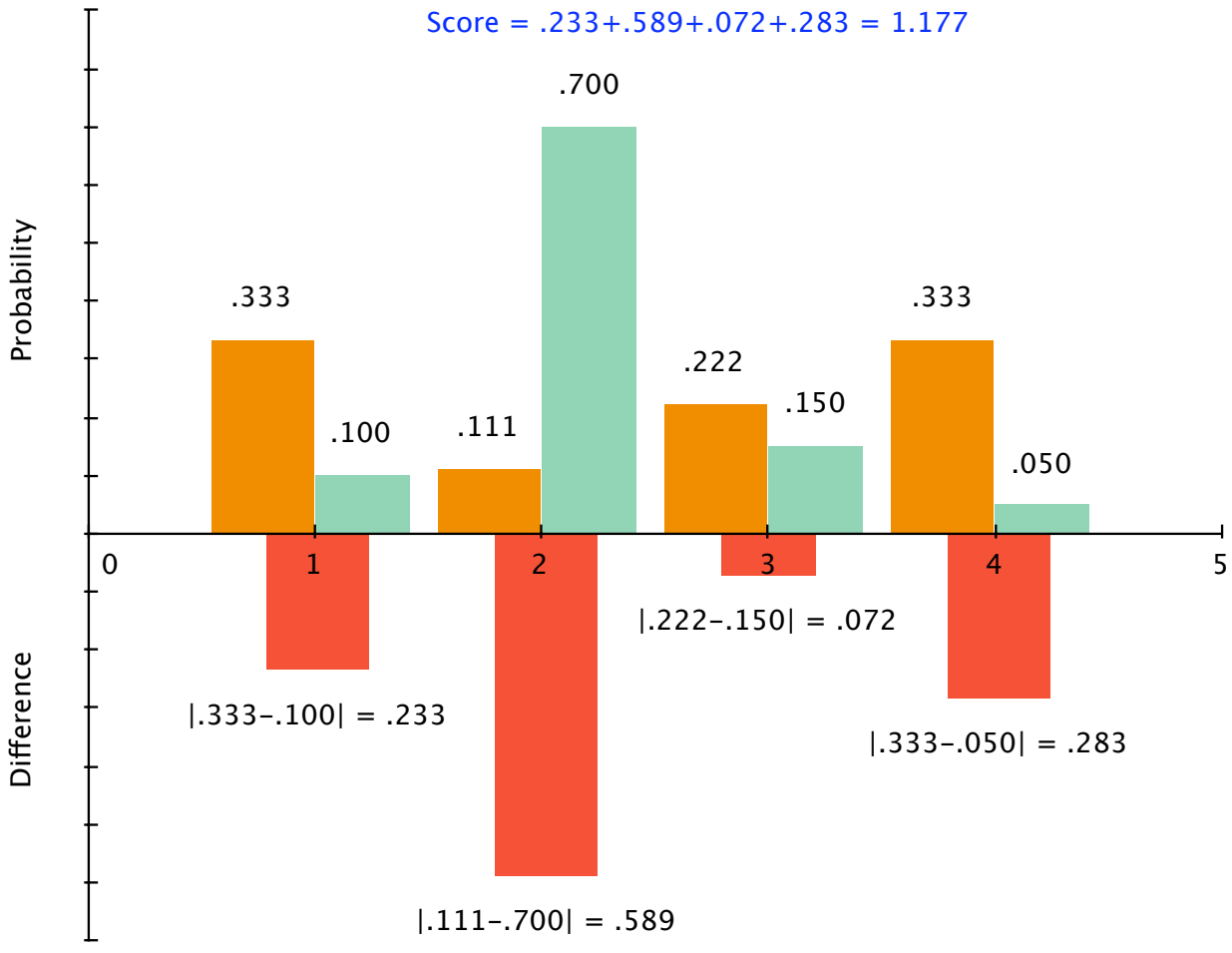
Modeling Protocol Behavior



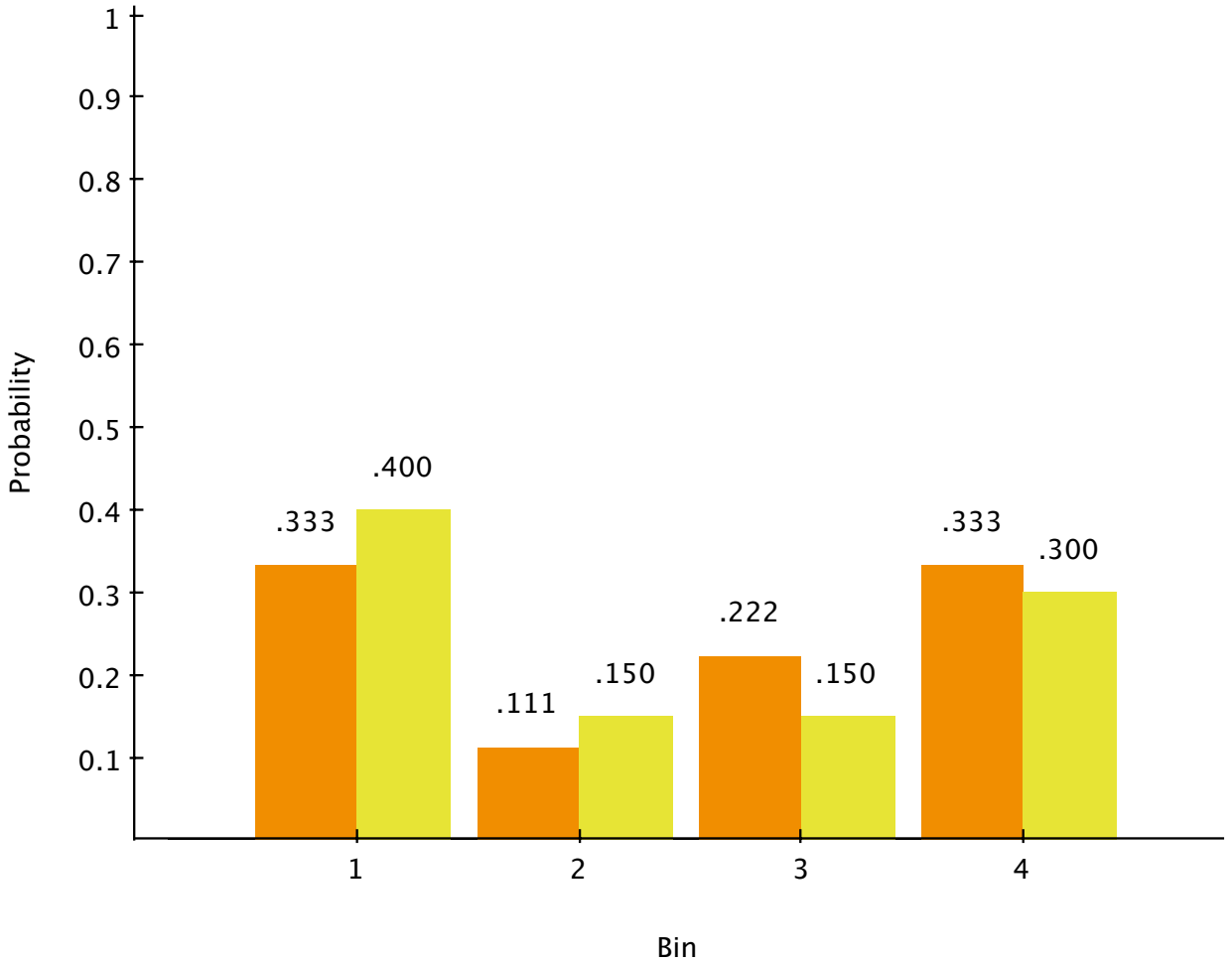
Comparing Protocol Models



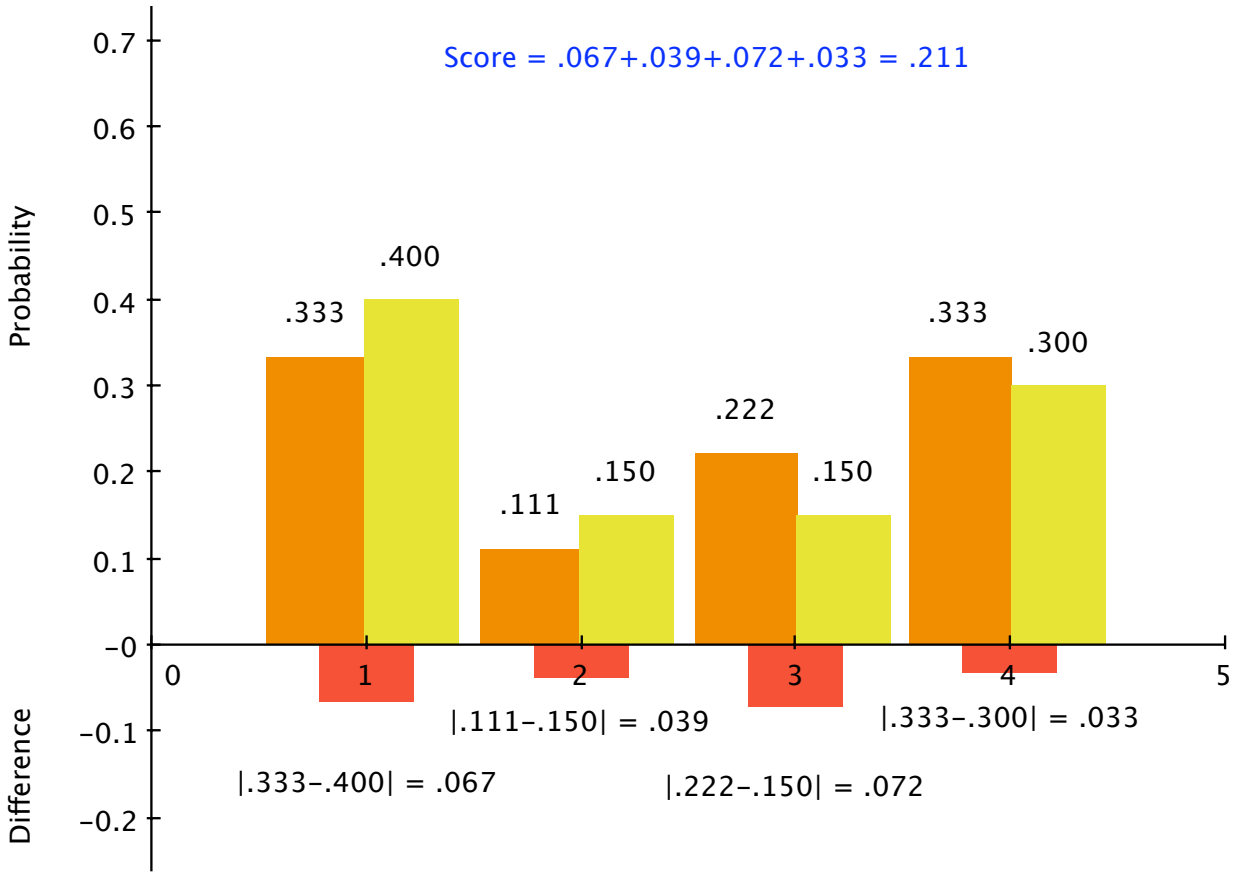
Comparing Protocol Models



Comparing Protocol Models



Comparing Protocol Models



Classifying Samples

1. Create training models for desired protocols
2. Build testing model for network trace to be classified
3. Find training model with lowest L_1 distance

$$L_1(\mathcal{A}, \mathcal{B}) = \sum_{i=1}^{n} \left| \frac{\mathcal{A}_{\phi_i}}{\mathcal{A}_\tau} - \frac{\mathcal{B}_{\phi_i}}{\mathcal{B}_\tau} \right|$$

Classification Confidence Threshold

- Goal: Eliminate close calls
- Require 1st place candidate to lead 2nd place by certain amount to make decision
- Based on standard deviation of L_1 distances

Evaluation

- Parameter Selection
  - Bin size
  - Confidence threshold
  - Training set size
- Precision

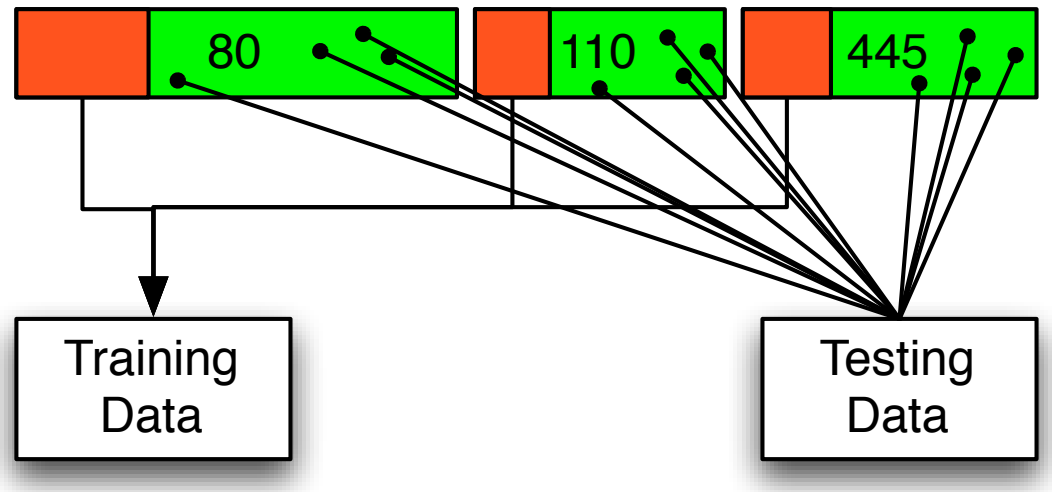
$$\frac{true\ positives}{true\ positives + false\ positives}$$

- Recall

$$\frac{true\ positives}{true\ positives + false\ negatives}$$

Evaluation

- Trial :=
  1. Randomly sample some percentage of available data for each port and train classifier
  2. Randomly sample some number ($\lambda = 45,000$) of the remaining data points for each port and create testing samples
  3. Classify testing samples
- 50 Trials



dartmouth Dataset

- Weekdays: January 19, 2004 – February 6, 2004
- Top 10 ports by number of sessions observed
  - 25 (SMTP), 80 (HTTP), 88 (Kerberos), 110 (POP3), 135 (DCE), 139 (NetBIOS), 443 (HTTPS), 445 (MDS), 902 (VMware), 1214 (Kazaa)
- No payload data
- Used for parameter selection

Total Packets	1.3 Billion
Traffic Volume	707 GB
Observed Ports	64,214
Sessions	5.2 Million

darpa Data

- *1999 DARPA Intrusion Detection Evaluation*
- Weeks 1, 3, 4, 5
- Ports 21 (`FTP`), 23 (`Telnet`), 25 (`SMTP`), 79 (`finger`), 80 (`HTTP`), 110 (`POP3`)
- Additional cross-validation using parameter values determined from `dartmouth`.

Total Packets	60 Million
Traffic Volume	12 GB
Observed Ports	10,274
Sessions	1.6 Million

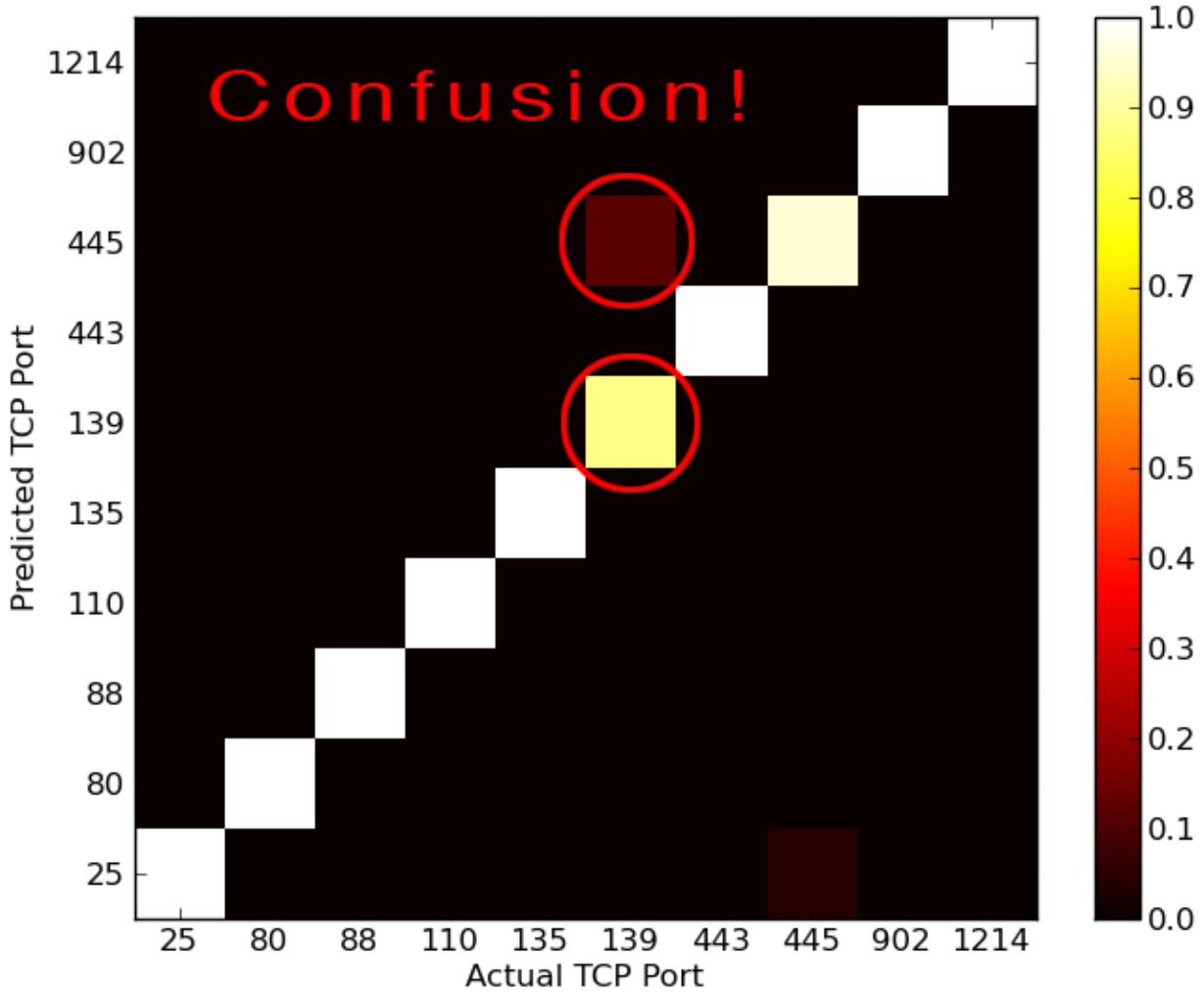
dartmouth Results

50 Trials

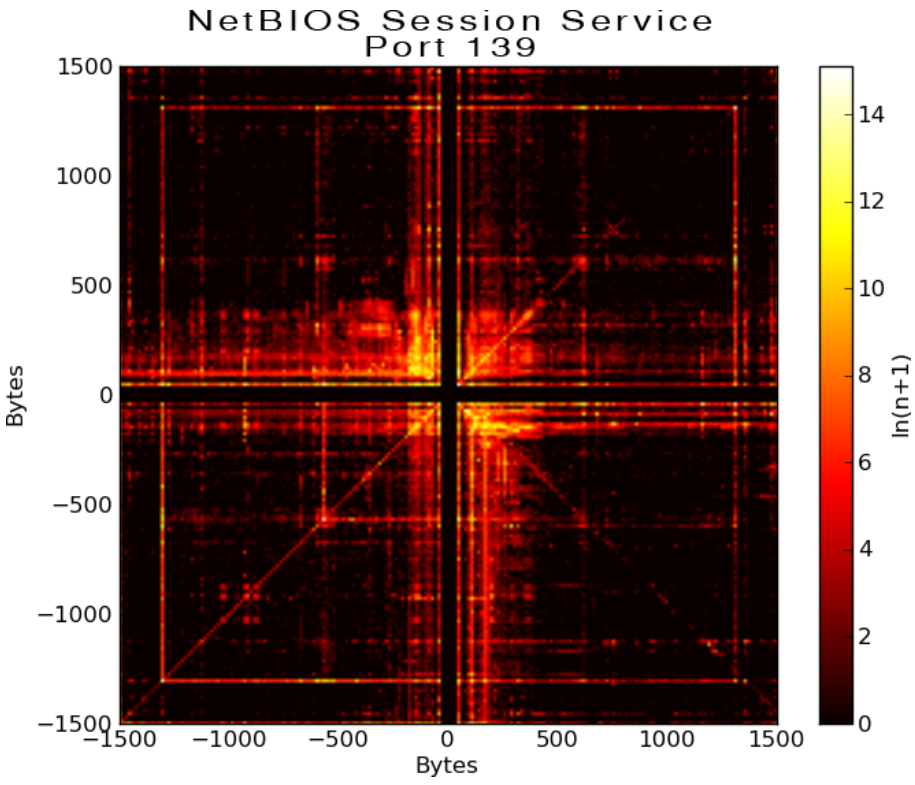
15% Training Set Size

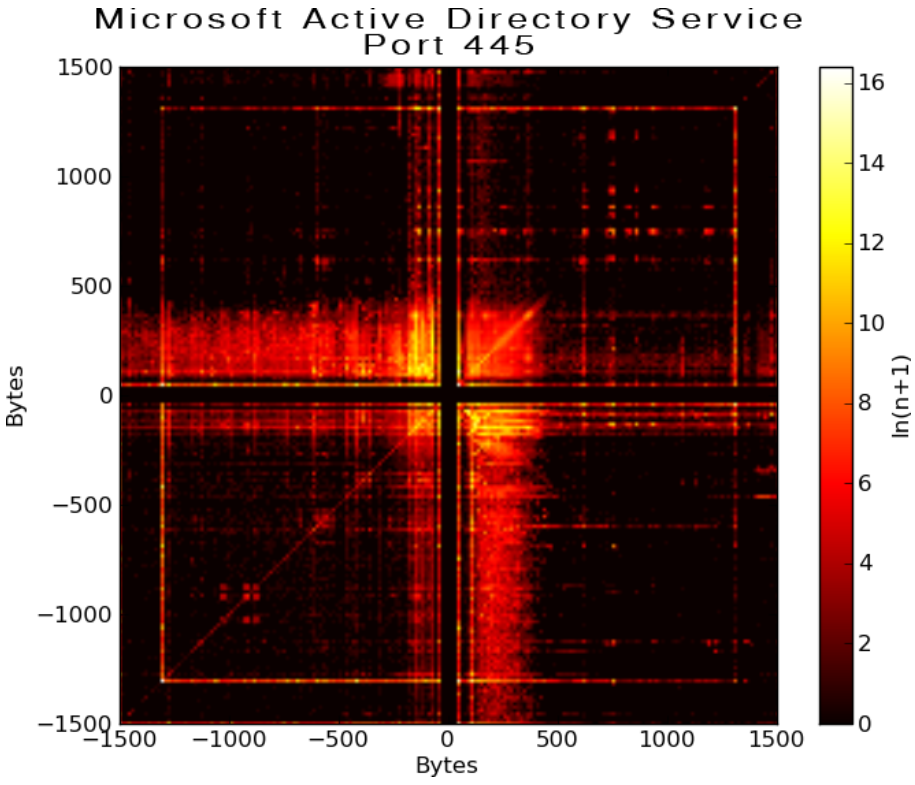
45,000 Data Points Testing Set Size

$k = 0.75$ Confidence Threshold



dartmouth Results





dartmouth Results

- Server Message Block (SMB) protocol
  - Used for Windows resource sharing
  - Windows NT: via NetBIOS (NBT) on TCP 139
  - Windows 2000: directly on TCP 445
- Clients with NetBIOS enabled try connections on TCP 139 and TCP 445

Reference: http://www.ntsecurity.nu/papers/port445/

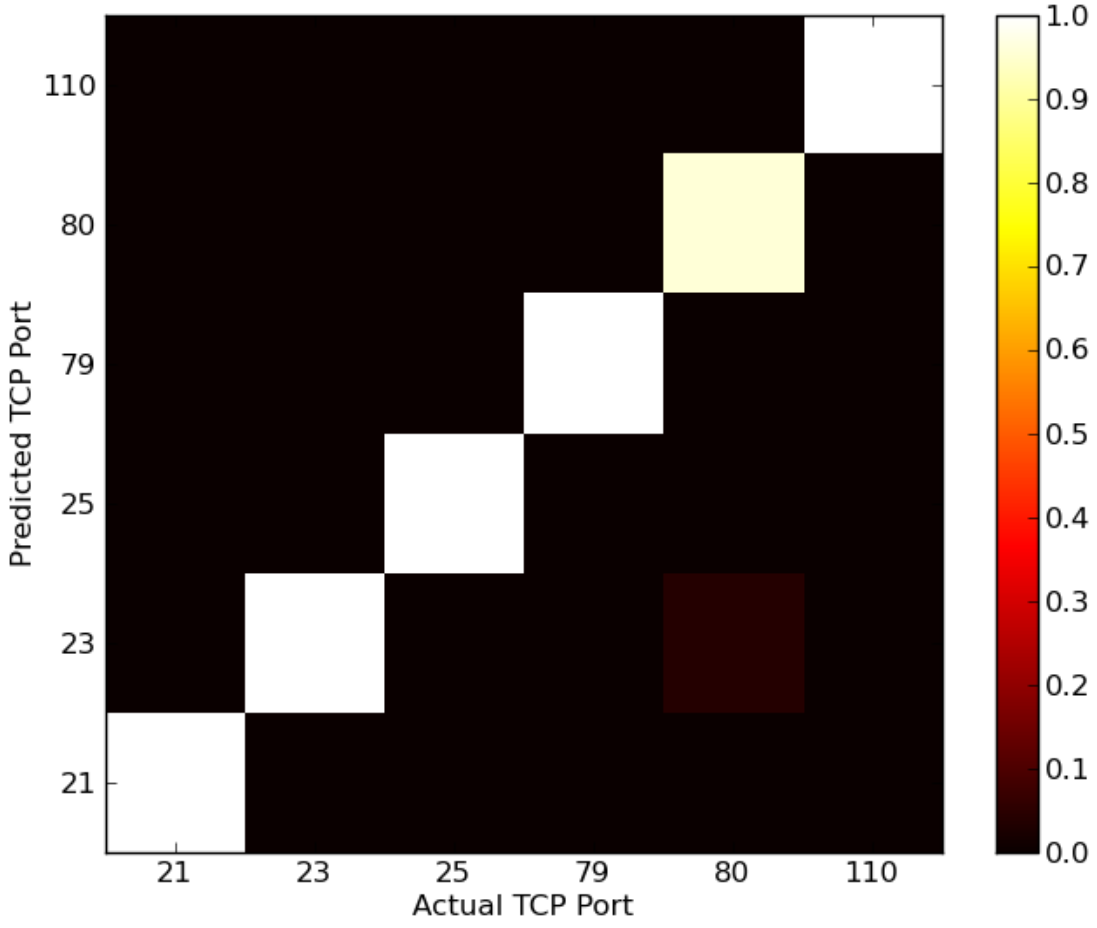
darpa Results

50 Trials

15% Training Set Size

45,000 Data Points Testing Set Size

$k = 0.75$ Confidence Threshold



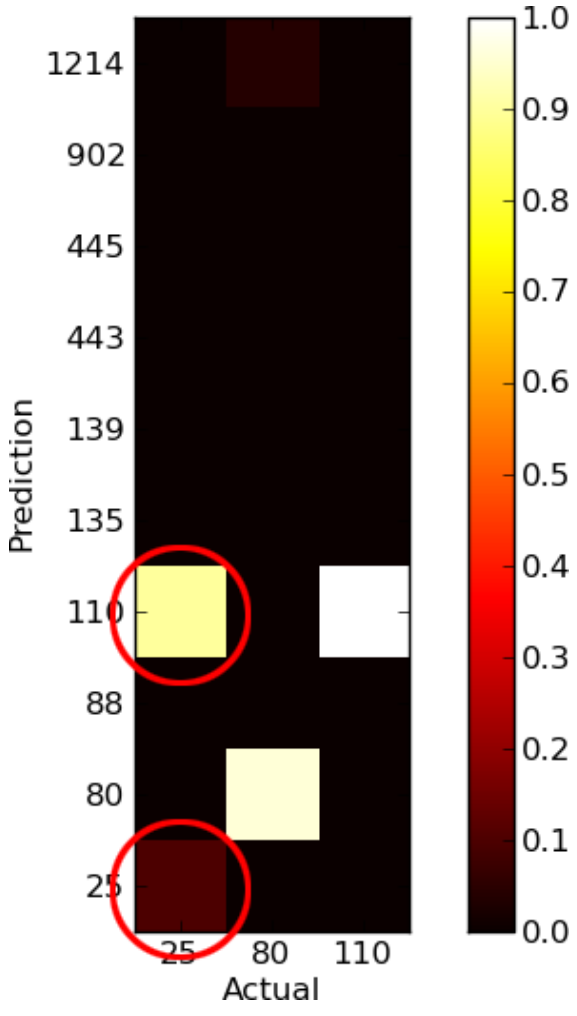
Inter-dataset Analysis

50 Trials

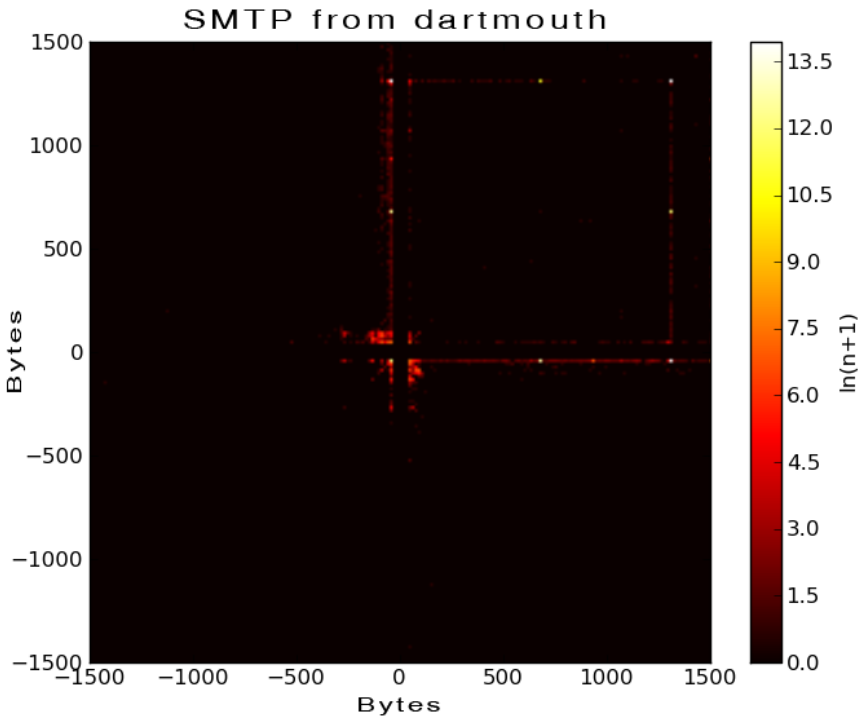
15% Training Set Size (`dartmouth`)

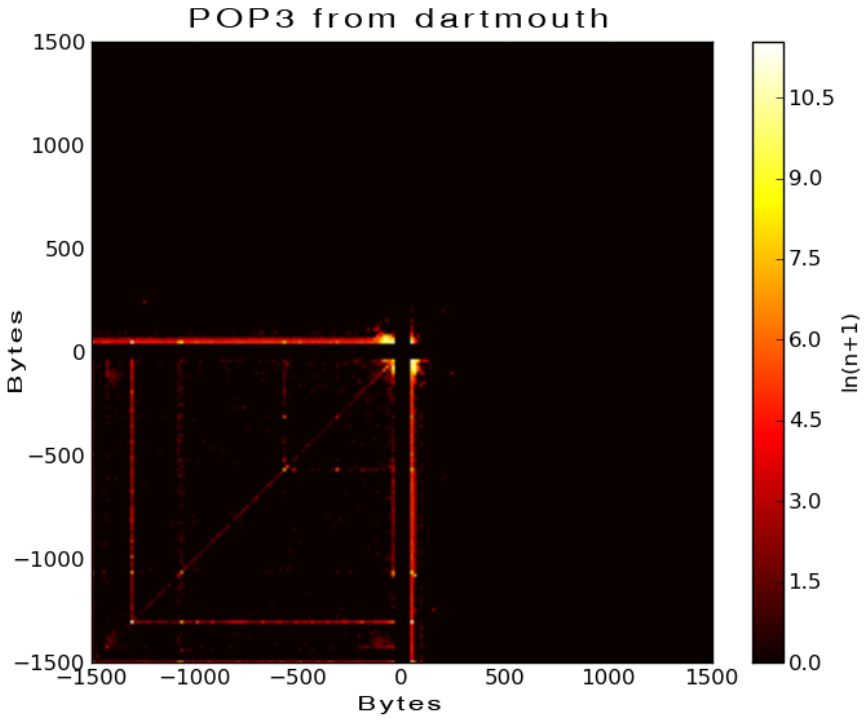
45,000 Data Points Testing Set Size (`darpa`)

$k = 0.25$ Confidence Threshold

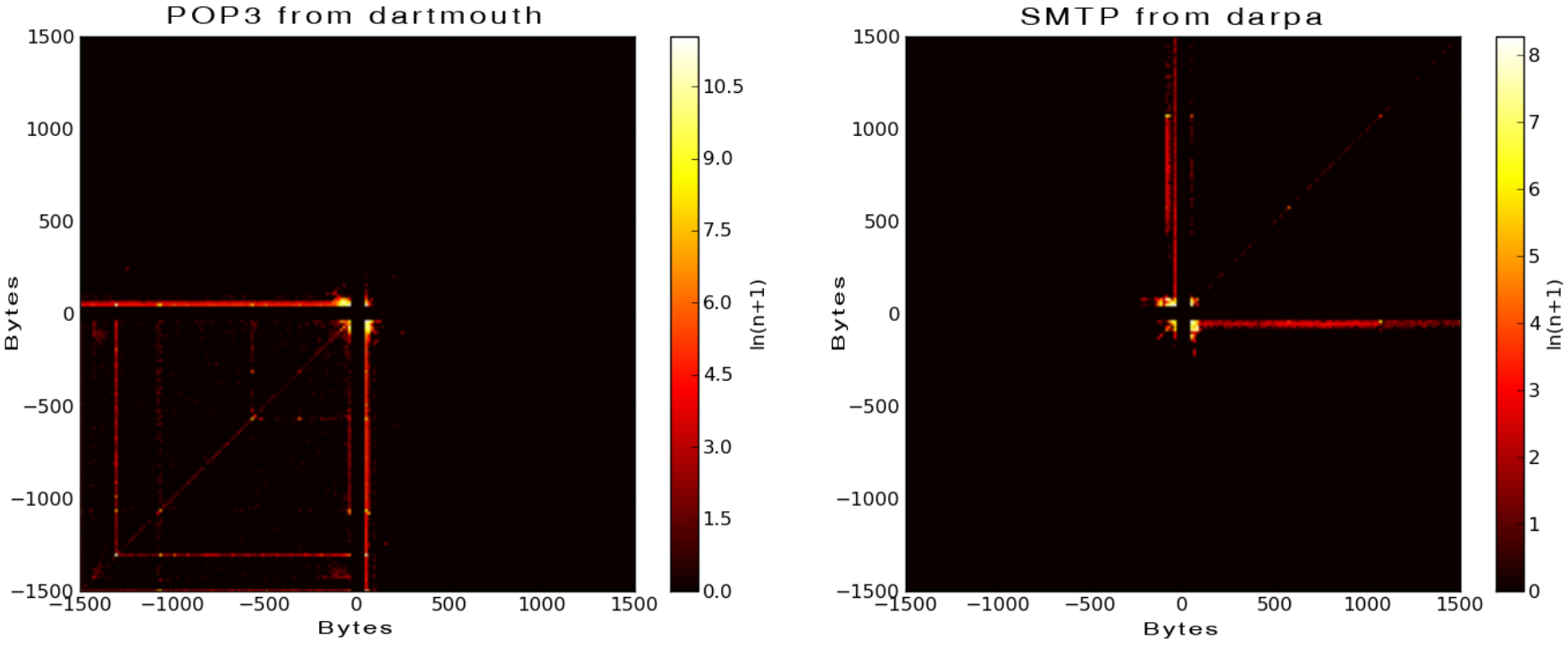


Inter-dataset Analysis



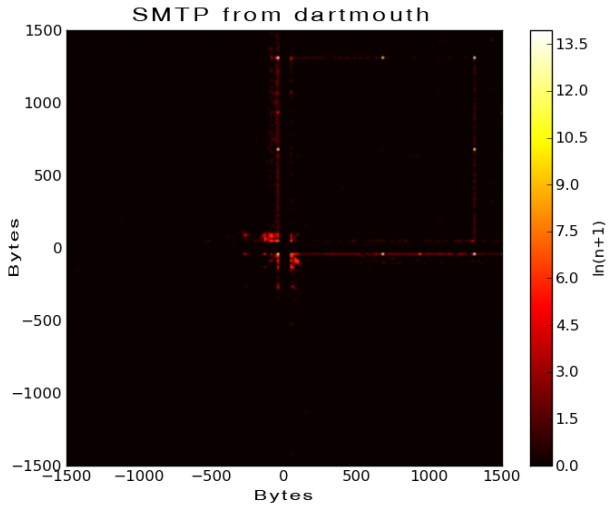


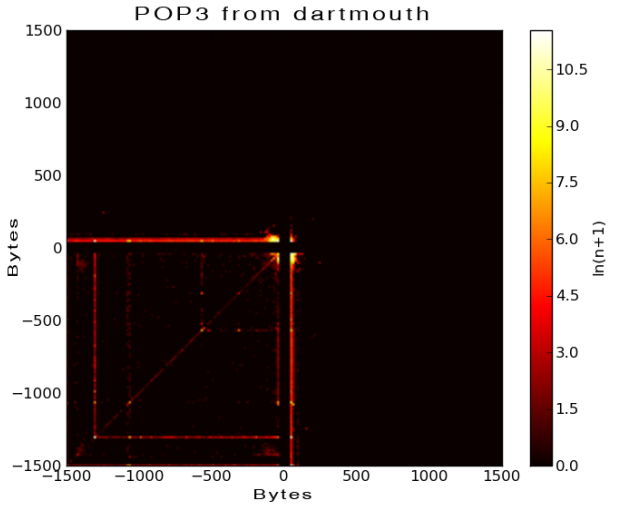
Inter-dataset Analysis

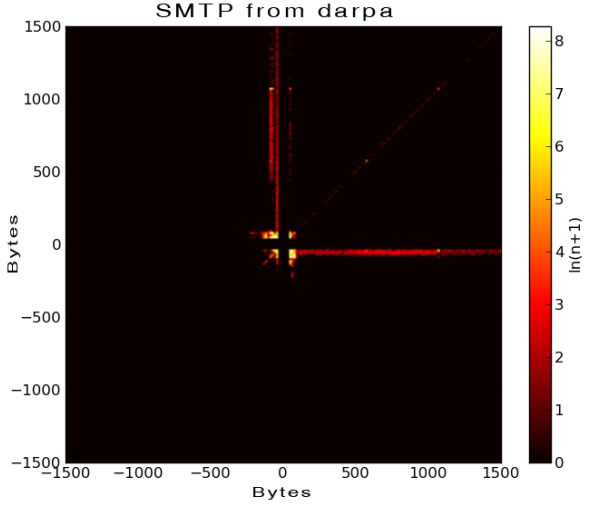


Mailbomb attack?

Inter-dataset Analysis







Limitations

- Must be trained on network to be tested
- Reliance on session de-multiplexing
- TCP only
- Deliberate attempts to disguise traffic (Folga et al., Wright et al. 2009)

Future Work

- On-line classification
- Protocol subcategorization (e.g., SSH interactive vs. SCP file transfer)

Conclusion

- Modeling protocol behavior using only packet size, direction, and order
- Resistant to encryption
- High precision and recall
- Quick and reliable traffic inspection
- Useful for anomaly and misuse detection

Questions?

Thanks for listening.
Q & A

wwlian@gmail.com