Towards Reliable Traffic Classification Using Visual Motifs

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

[1]University of North Carolina at Chapel Hill

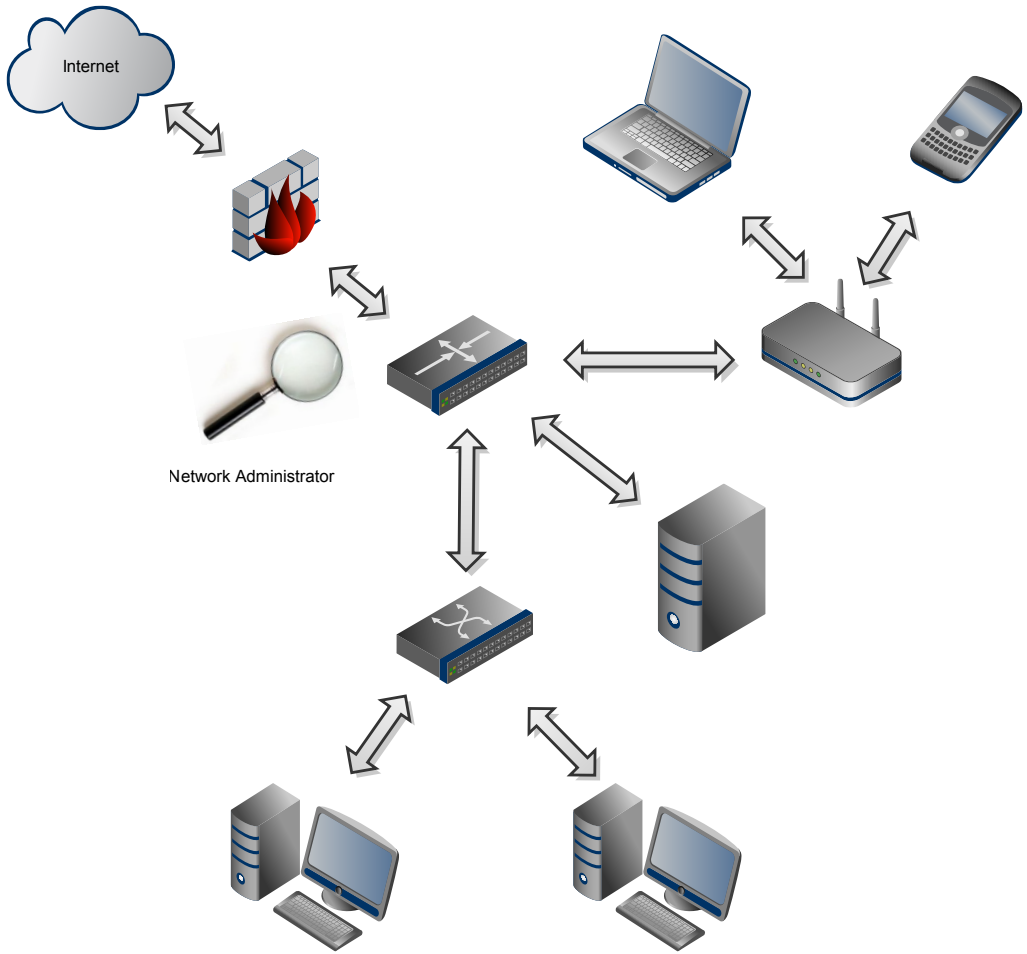
[2]RedJack, LLC

FloCon 2010

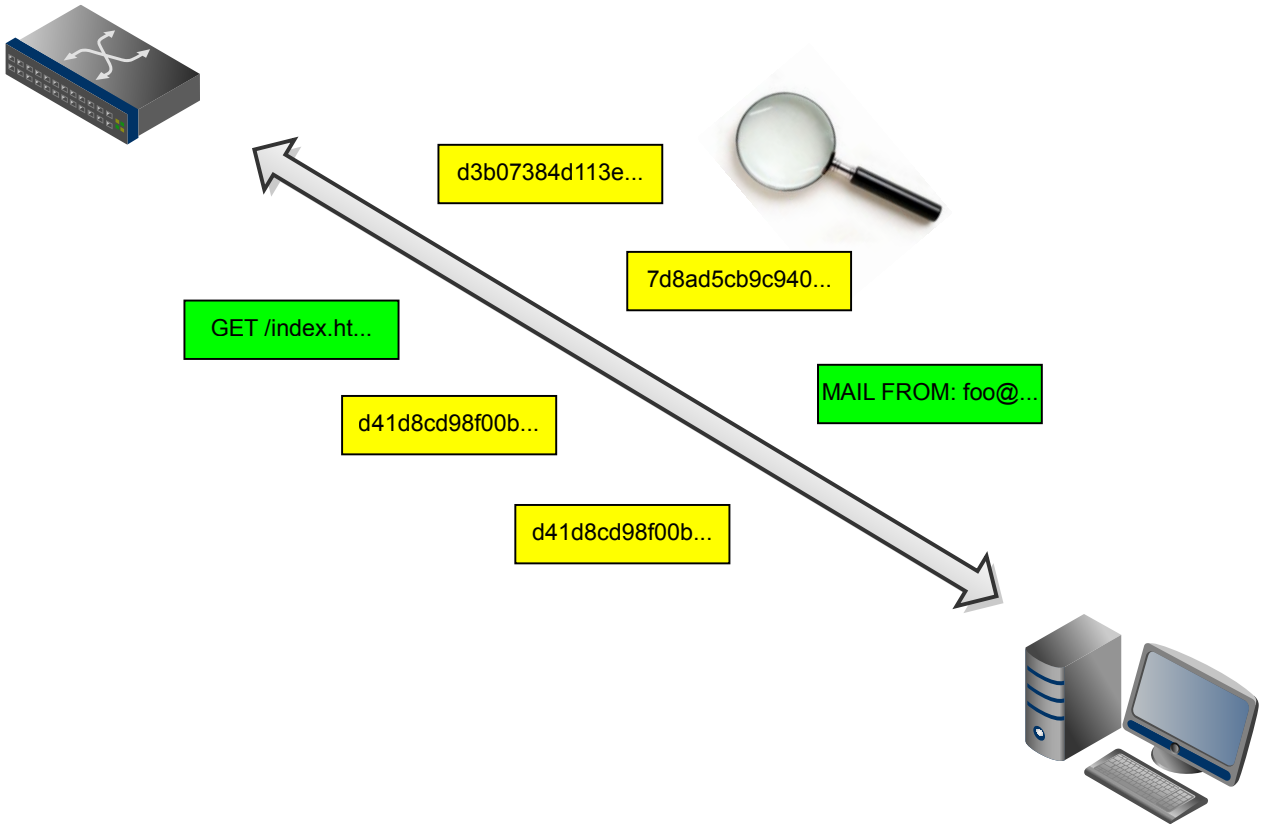
Overview

- Background
- Visual Motifs
- Traffic Classification
- Evaluation

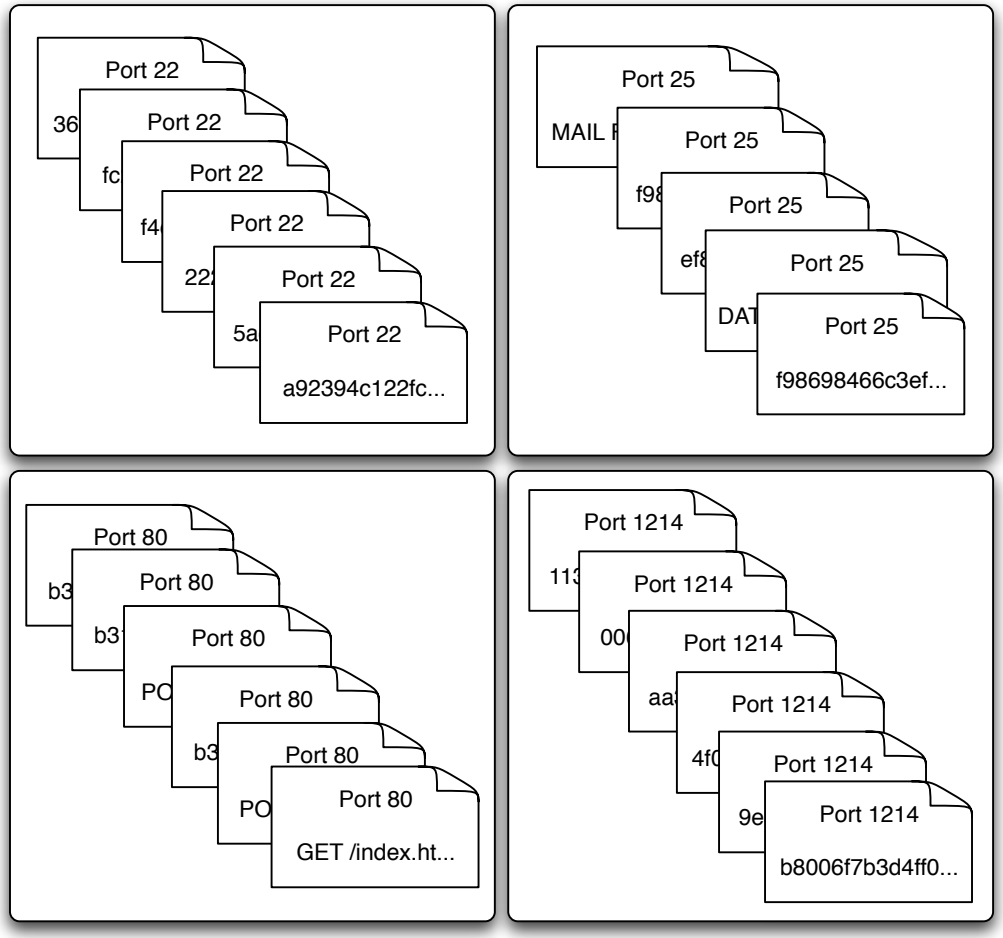
Motivation



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
  - Length preservation
- Negligible packet loss & retransmission

Related Work

- *Scatter (and other) Plots for Visualizing User Profiling Data and Network Traffic*, Goldring 2004.
- *Using Visual Motifs to Classify Encrypted Traffic*, Wright *et al.* 2006
- *Intelligent Classification and Visualization of Network Scans* Muelder *et al.* 2008.
- *FloVis: A Network Security Visualization Framework*, Taylor 2009.

Timeline Heatmaps

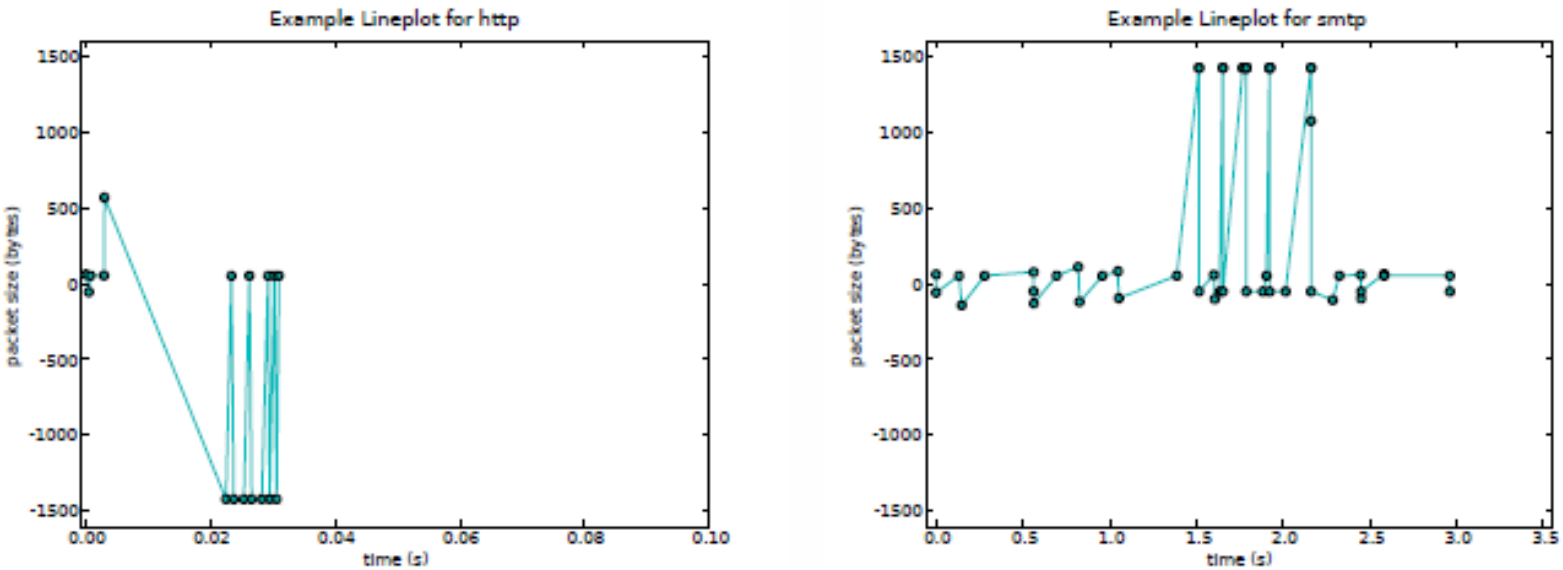


Image credit: Wright *et al.* 2006

Timeline Heatmaps

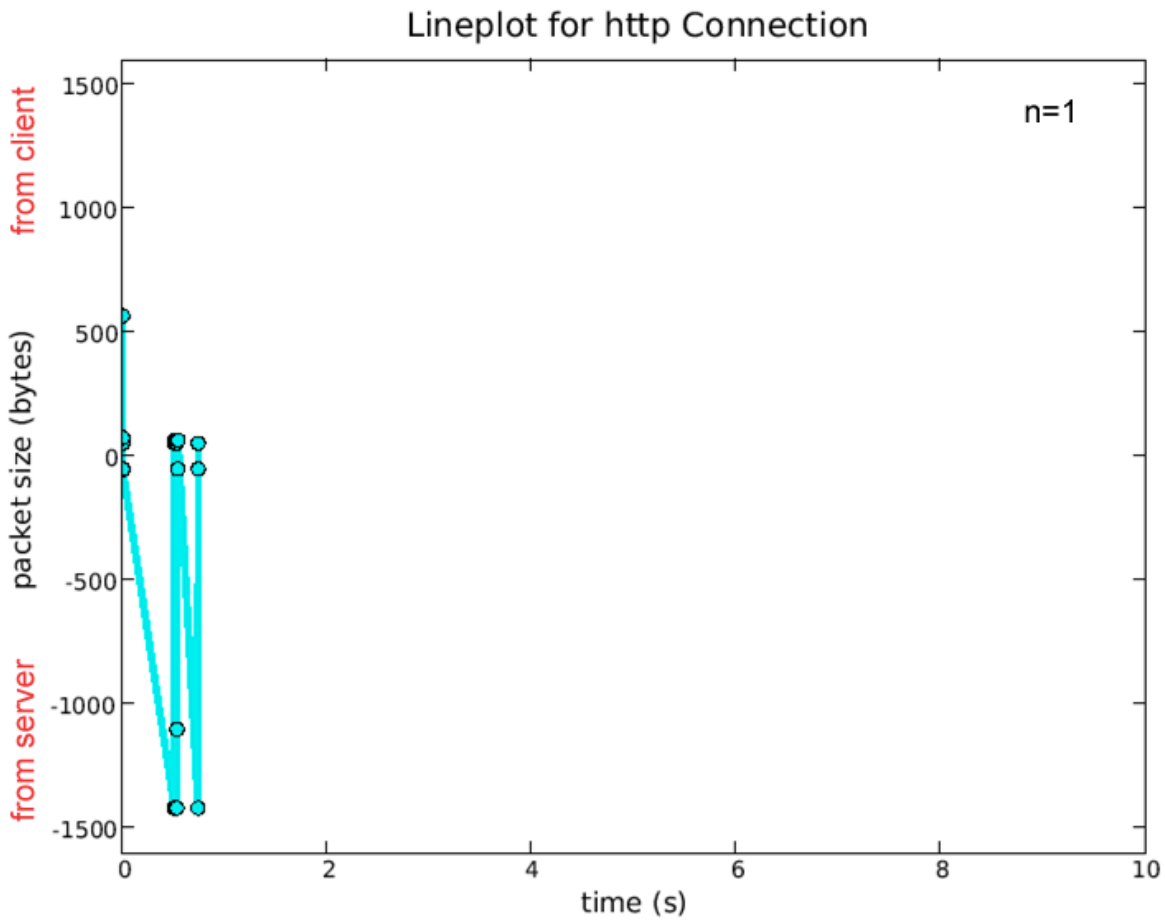


Image credit: Wright *et al.* 2006

Timeline Heatmaps

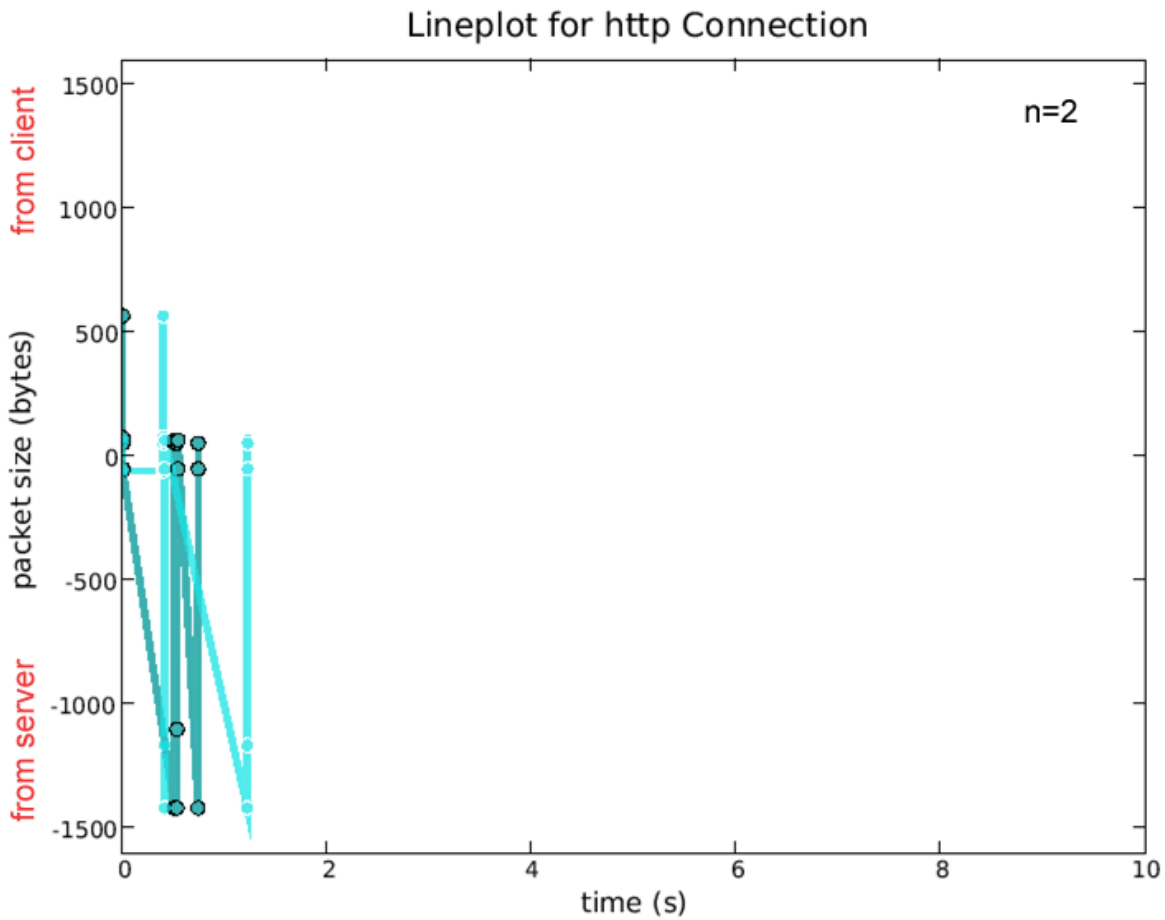


Image credit: Wright *et al.* 2006

Timeline Heatmaps

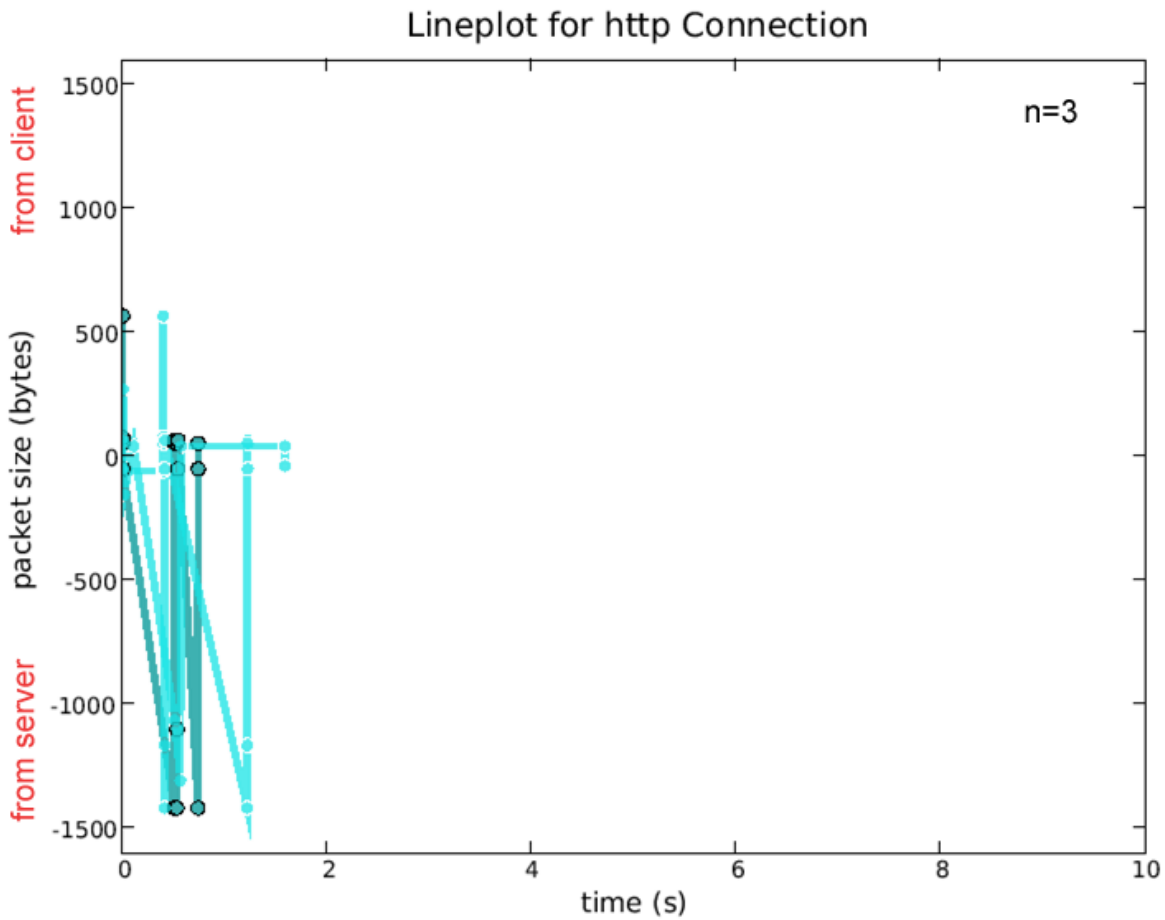


Image credit: Wright *et al.* 2006

Timeline Heatmaps

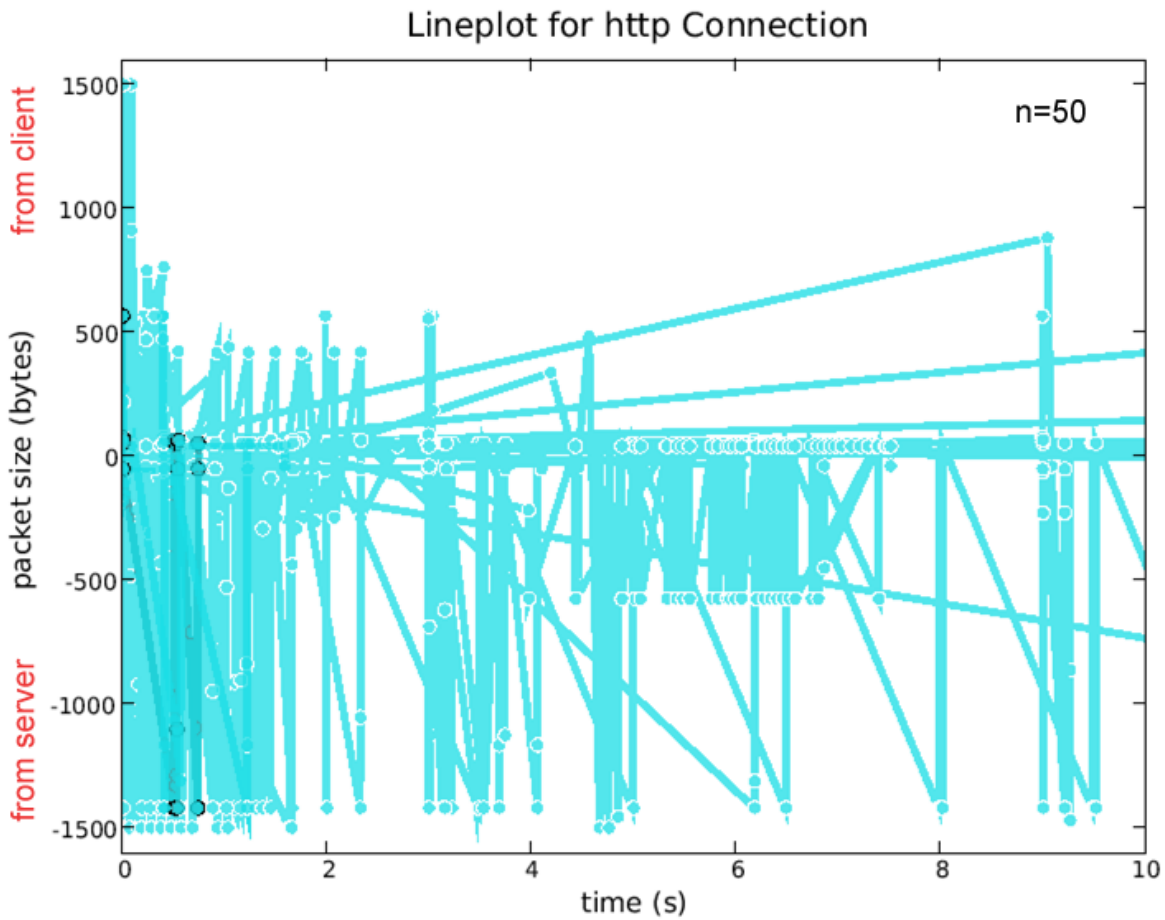


Image credit: Wright *et al.* 2006

Timeline Heatmaps

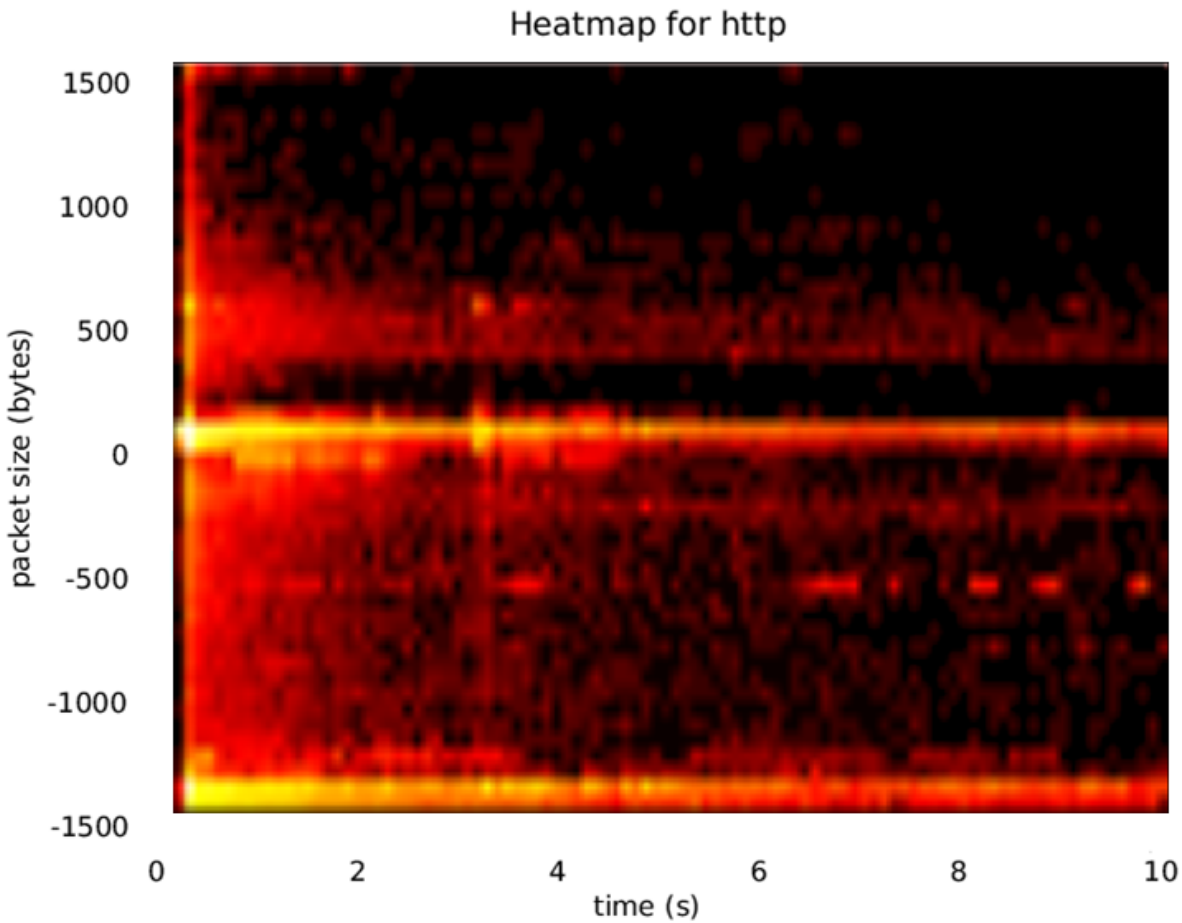
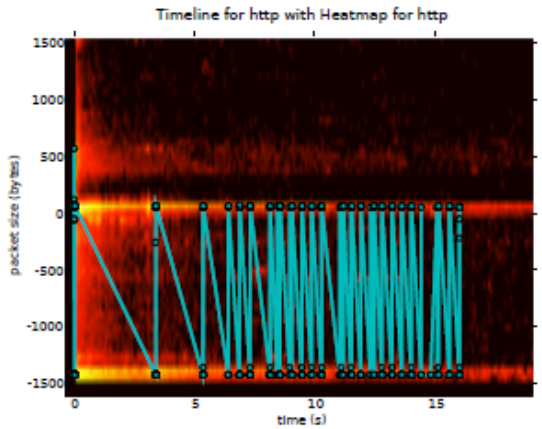
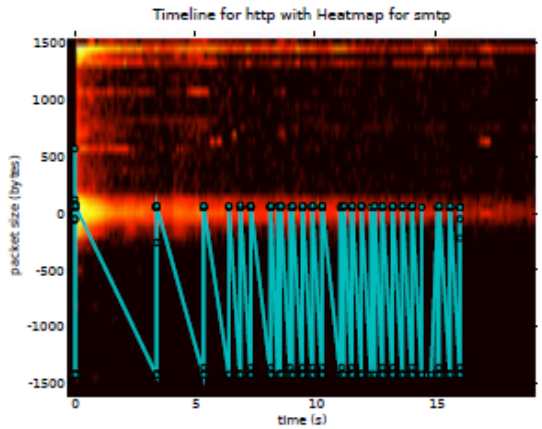


Image credit: Wright *et al.* 2006

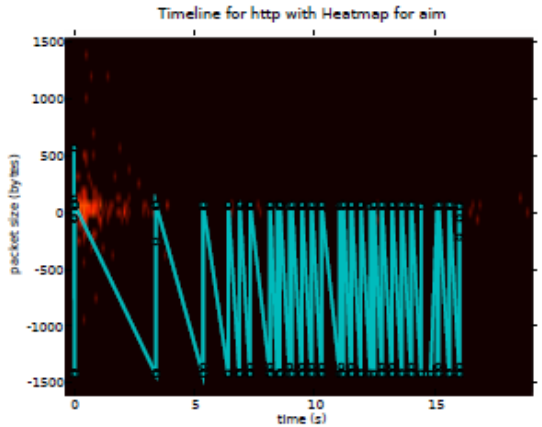
Timeline Heatmaps



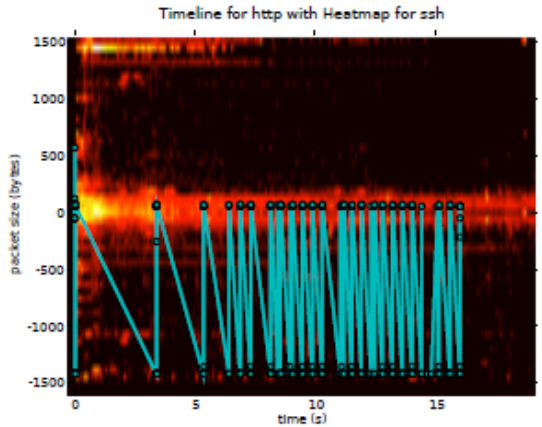
(a) HTTP



(b) SMTP



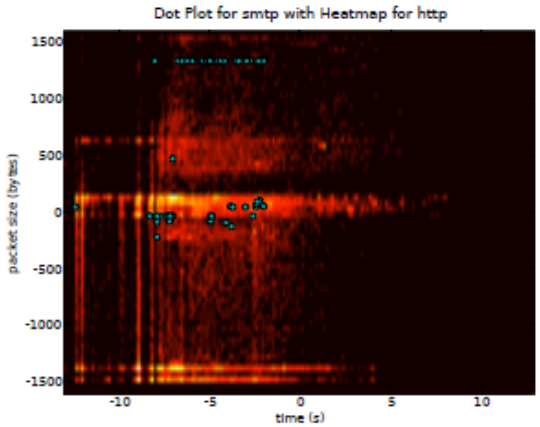
(c) AIM



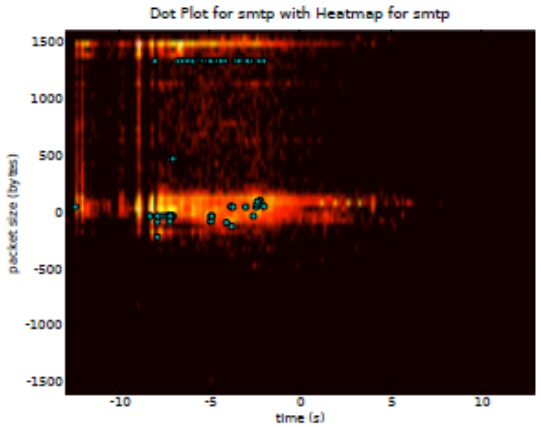
(d) SSH

Image credit: Wright *et al.* 2006

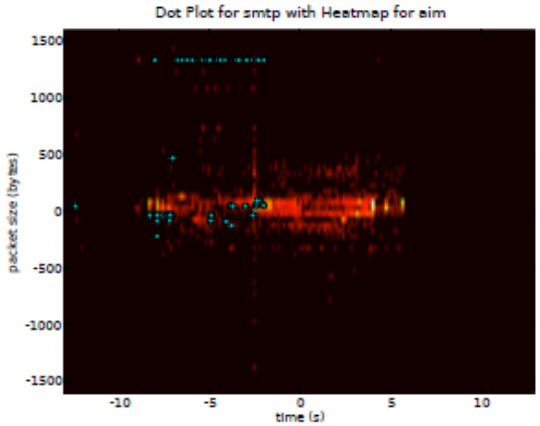
Unigram Heatmaps



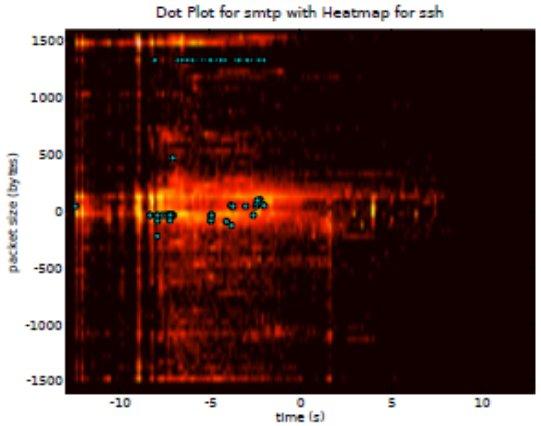
(a) HTTP



(b) SMTP



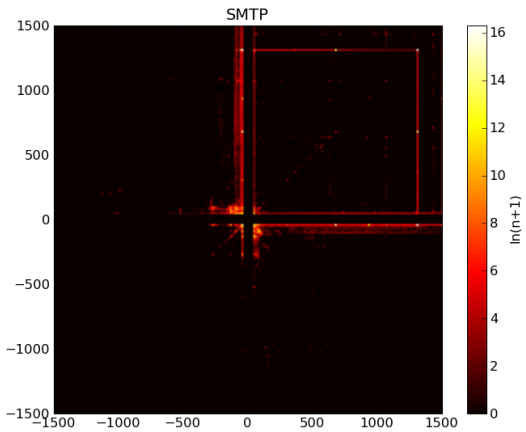
(c) AIM

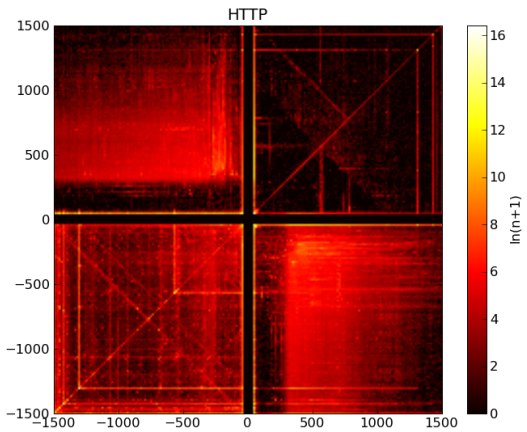


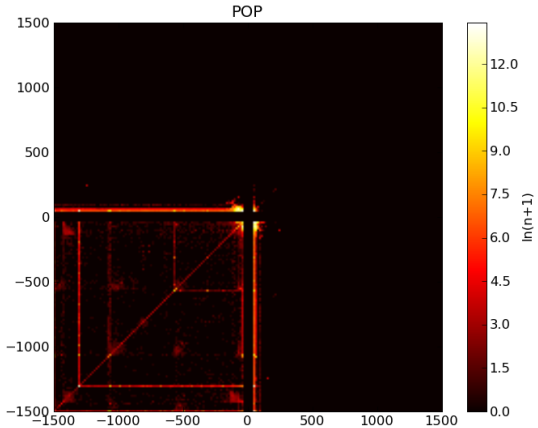
(d) SSH

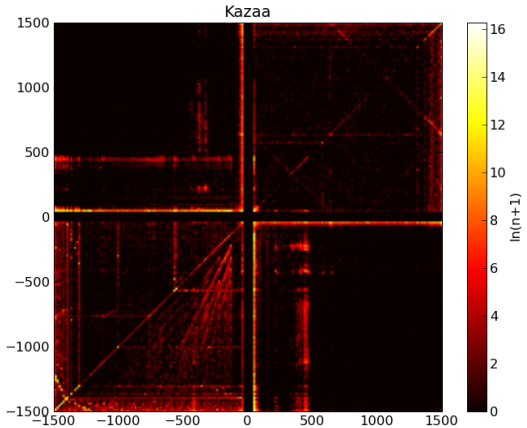
Image credit: Wright *et al.* 2006

Bigram Heatmaps

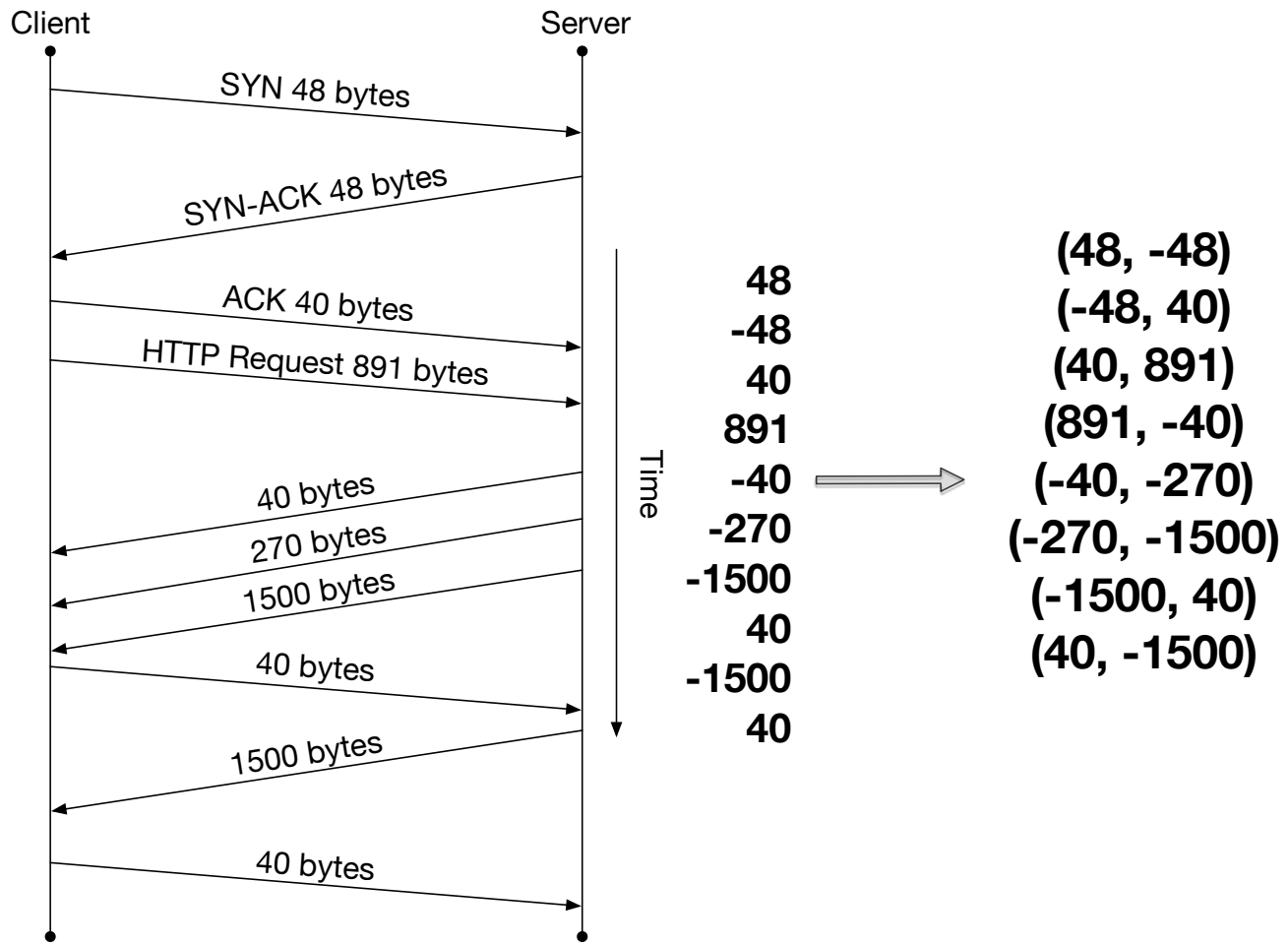








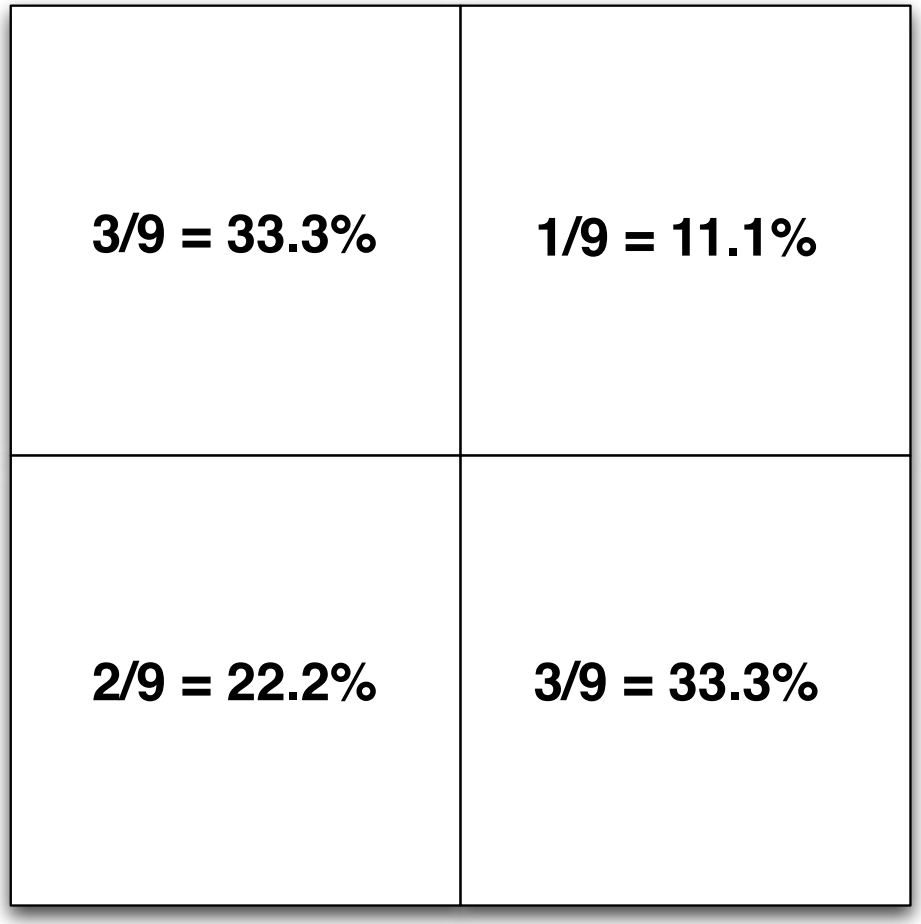
Heatmap Construction



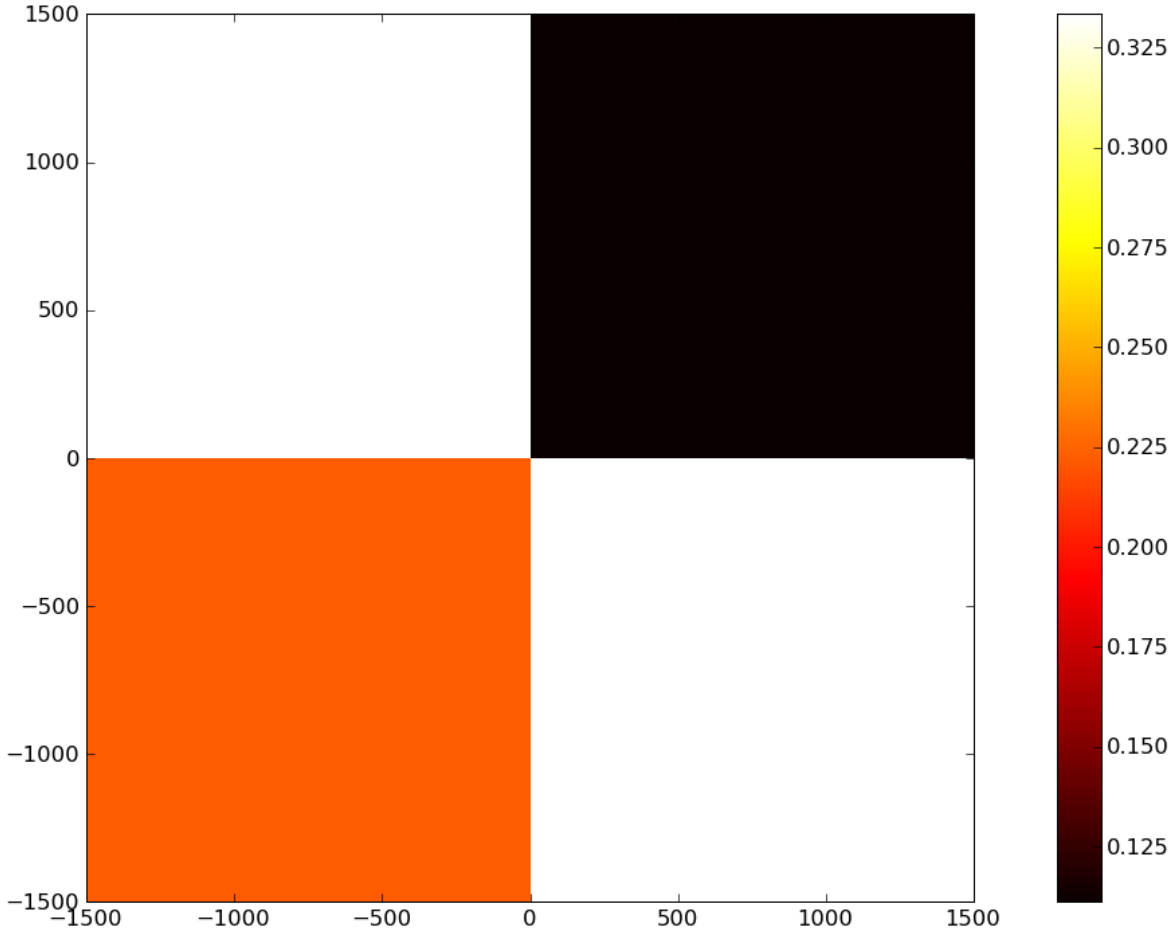
Heatmap Construction

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

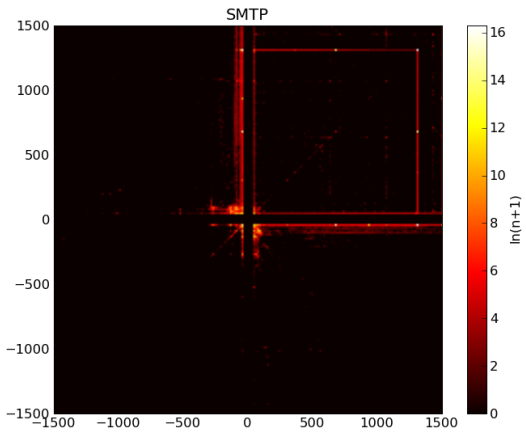
Heatmap Construction

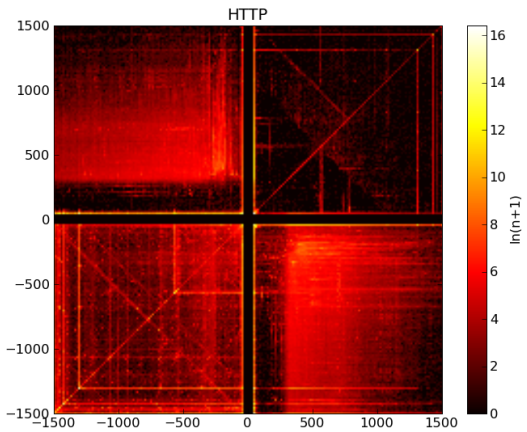


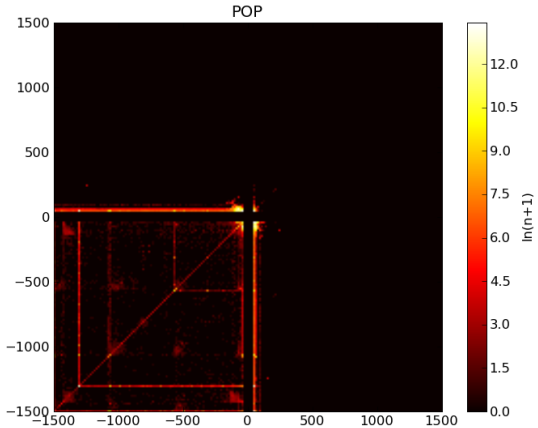
Heatmap Construction

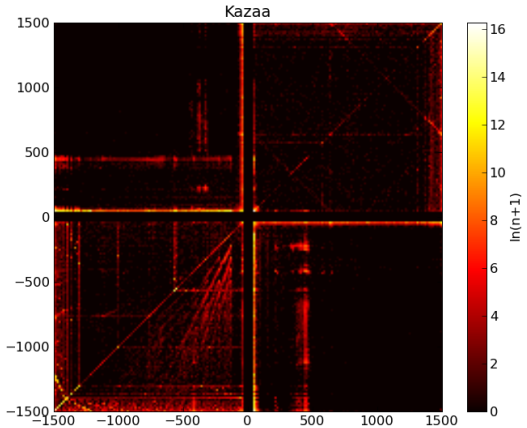


Bigram Heatmaps





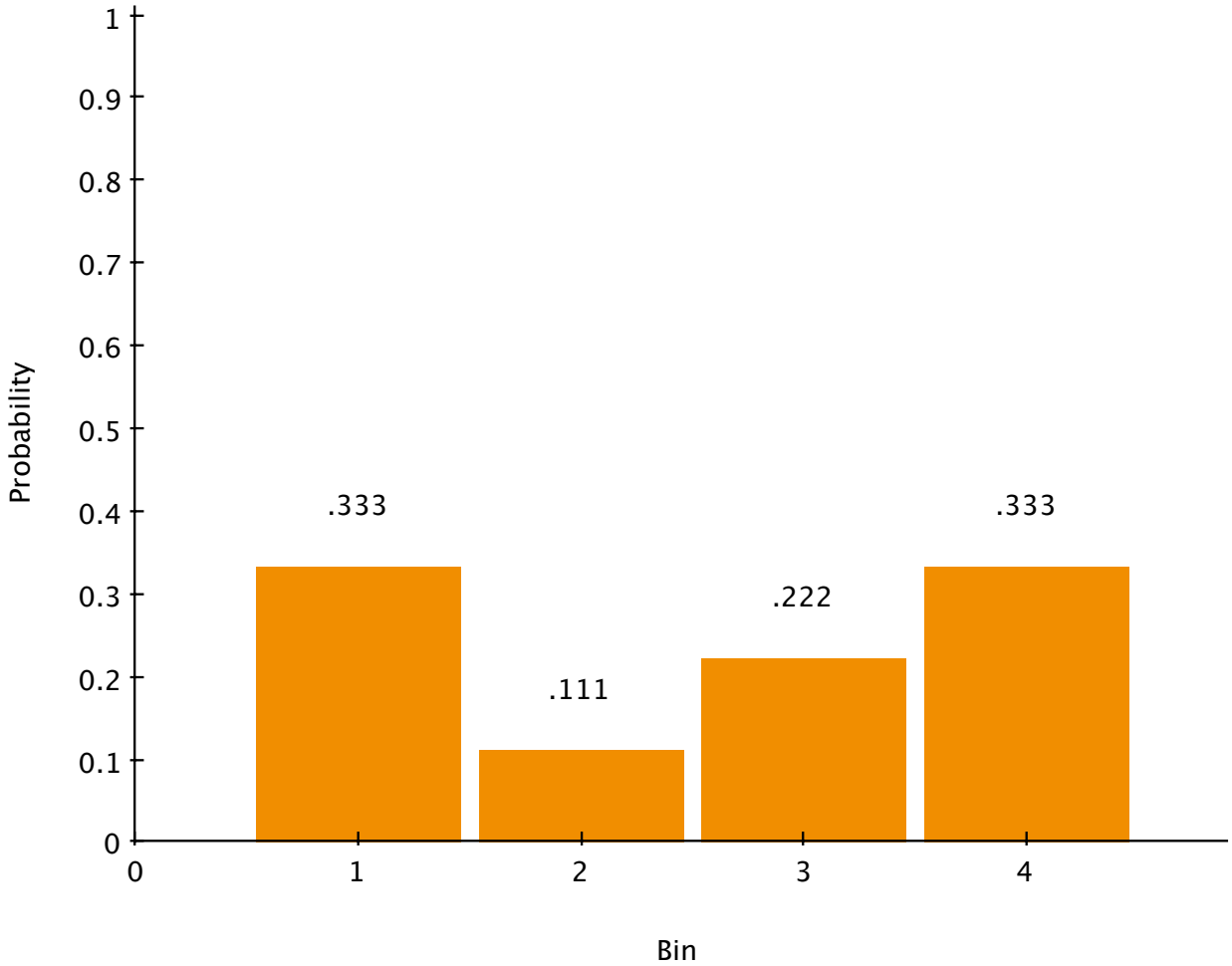




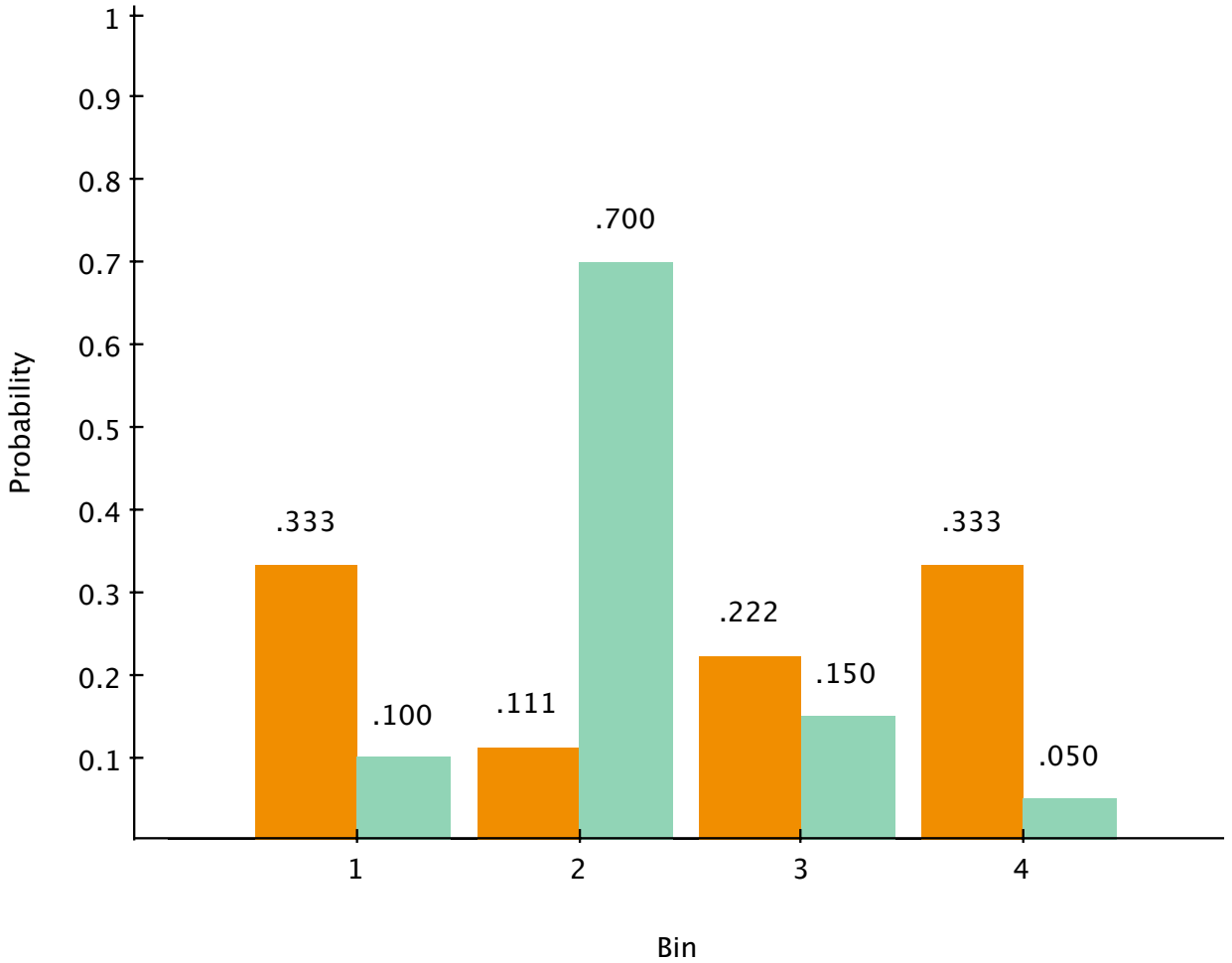
Modeling Protocol Behavior

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

Modeling Protocol Behavior



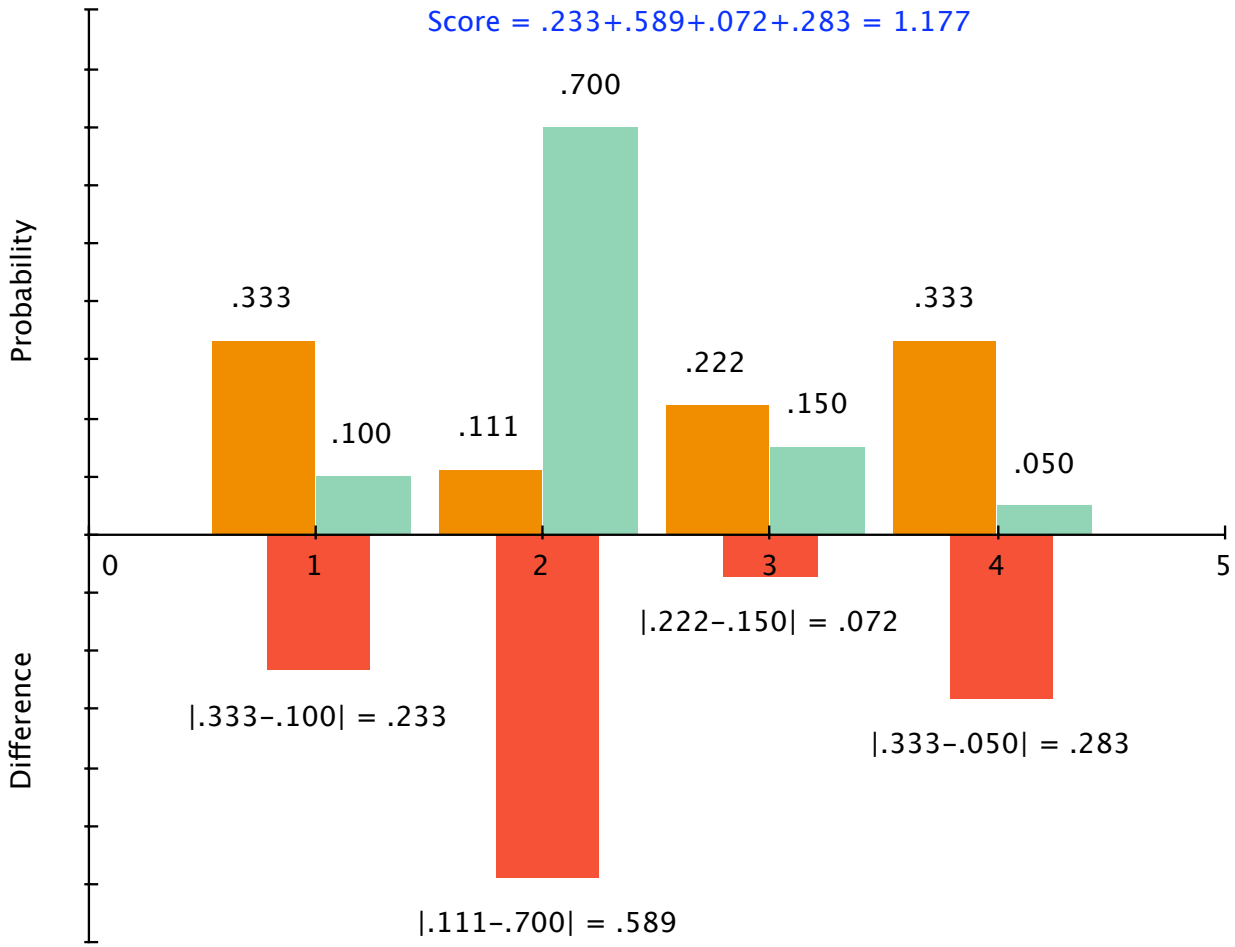
Comparing Protocol Models



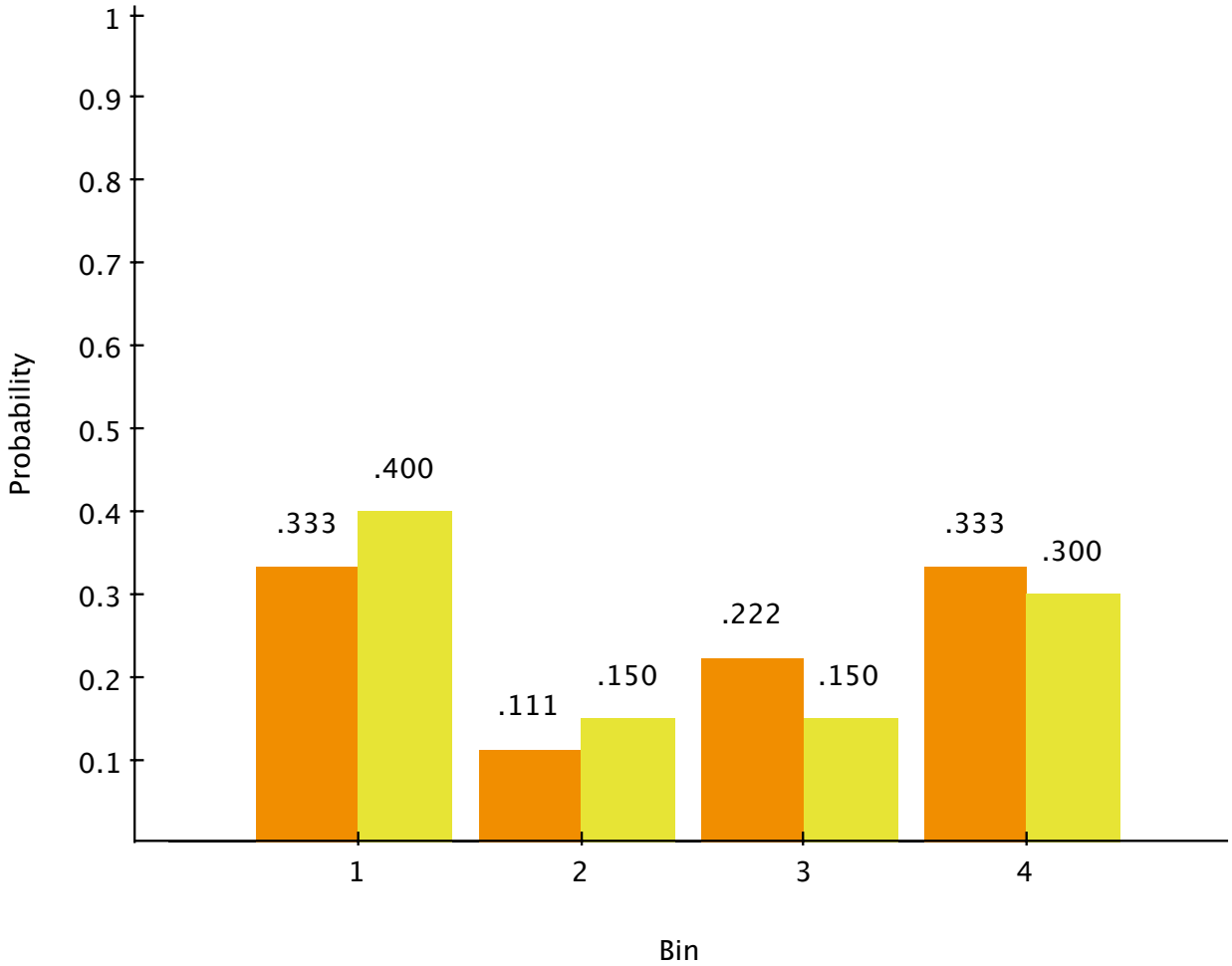
Comparing Protocol Models

$$
\begin{aligned}
A_{total} &= \sum_{k=1}^{n} A_k \\
B_{total} &= \sum_{k=1}^{n} B_k \\
Score_{A \leftrightarrow B} &= \sum_{i=1}^{n} \left| \frac{A_i}{A_{total}} - \frac{B_i}{B_{total}} \right| \\
&= \frac{1}{A_{total} \cdot B_{total}} \sum_{i=1}^{n} |A_i \cdot B_{total} - B_i \cdot A_{total}|
\end{aligned}
$$

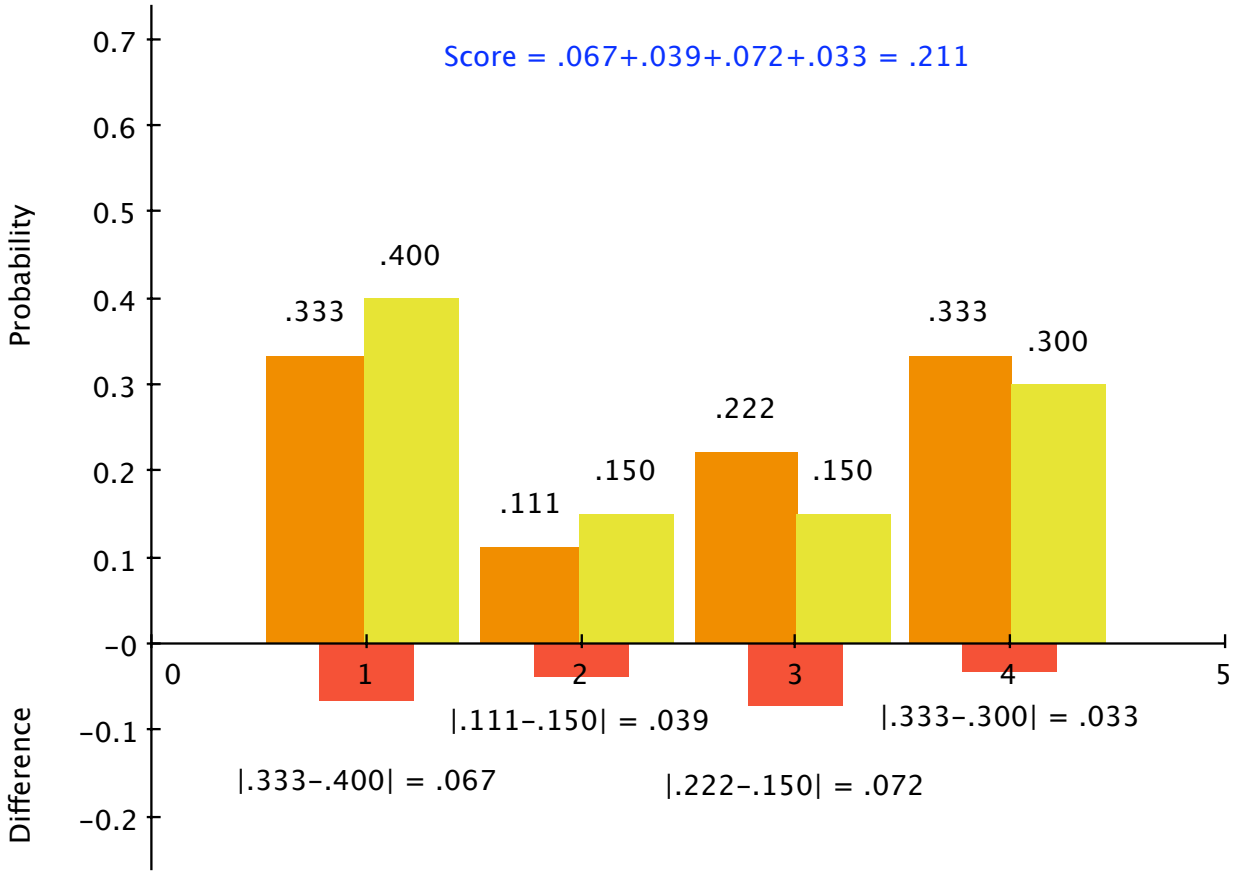
Comparing Protocol Models



Comparing Protocol Models



Comparing Protocol Models



Classifying Samples: Easy as 1-2-3

1. Create training models for desired protocols
2. Build distribution for sample network trace
3. Find training model with lowest difference score

$$
\begin{aligned}
Score_{A \leftrightarrow B} &= \sum_{i=1}^{n} \left| \frac{A_i}{A_{total}} - \frac{B_i}{B_{total}} \right| \\
&= \frac{1}{A_{total} \cdot B_{total}} \sum_{i=1}^{n} |A_i \cdot B_{total} - B_i \cdot A_{total}|
\end{aligned}
$$

Evaluation

- How much traffic must be collected for:
  - Training
  - Testing
- Precision?

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

- Recall?

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

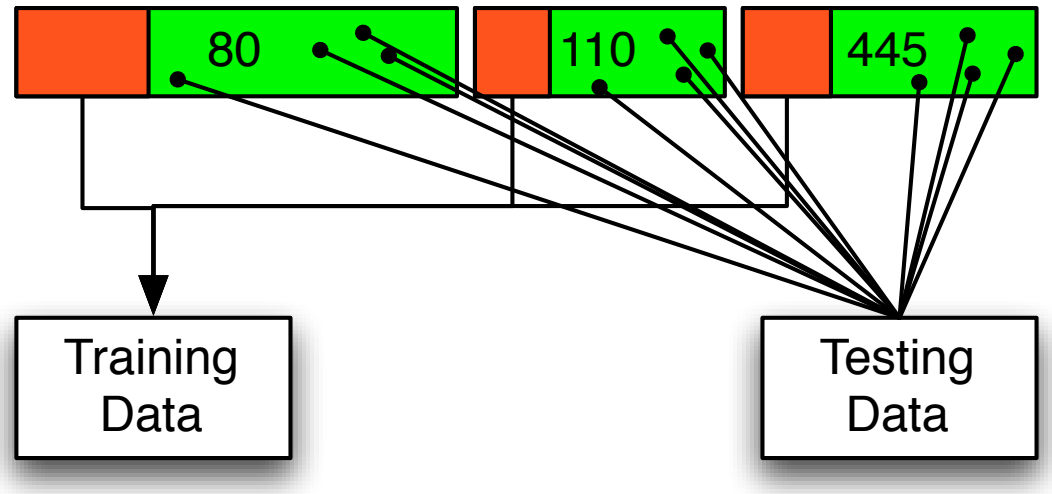
Data

- CRAWDAD Dataset
- Weekdays: January 19, 2004 – February 6, 2004
- Ports with sufficient traffic
  - $\geq$ 1M packets
  - 0.3% of ports $\rightarrow$ 95.21% of packets
- Keep top 10 ports by number of sessions observed
- No ground truth

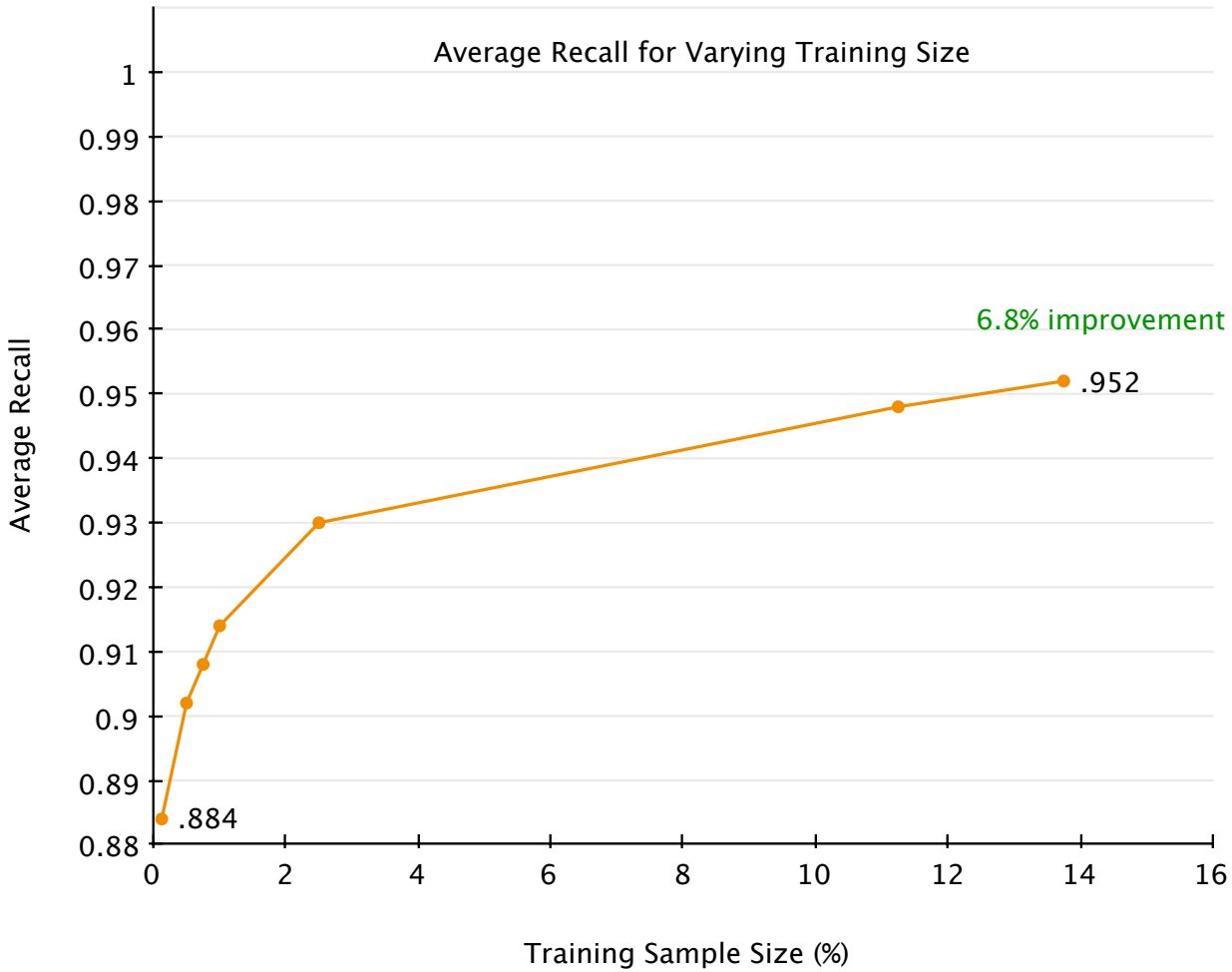
Total Packets	1.3 Billion
Traffic Volume	707 GB
Observed Ports	64,214
Sessions	5.2 Million
Port 80 Sessions	1.7 Million

Methodology

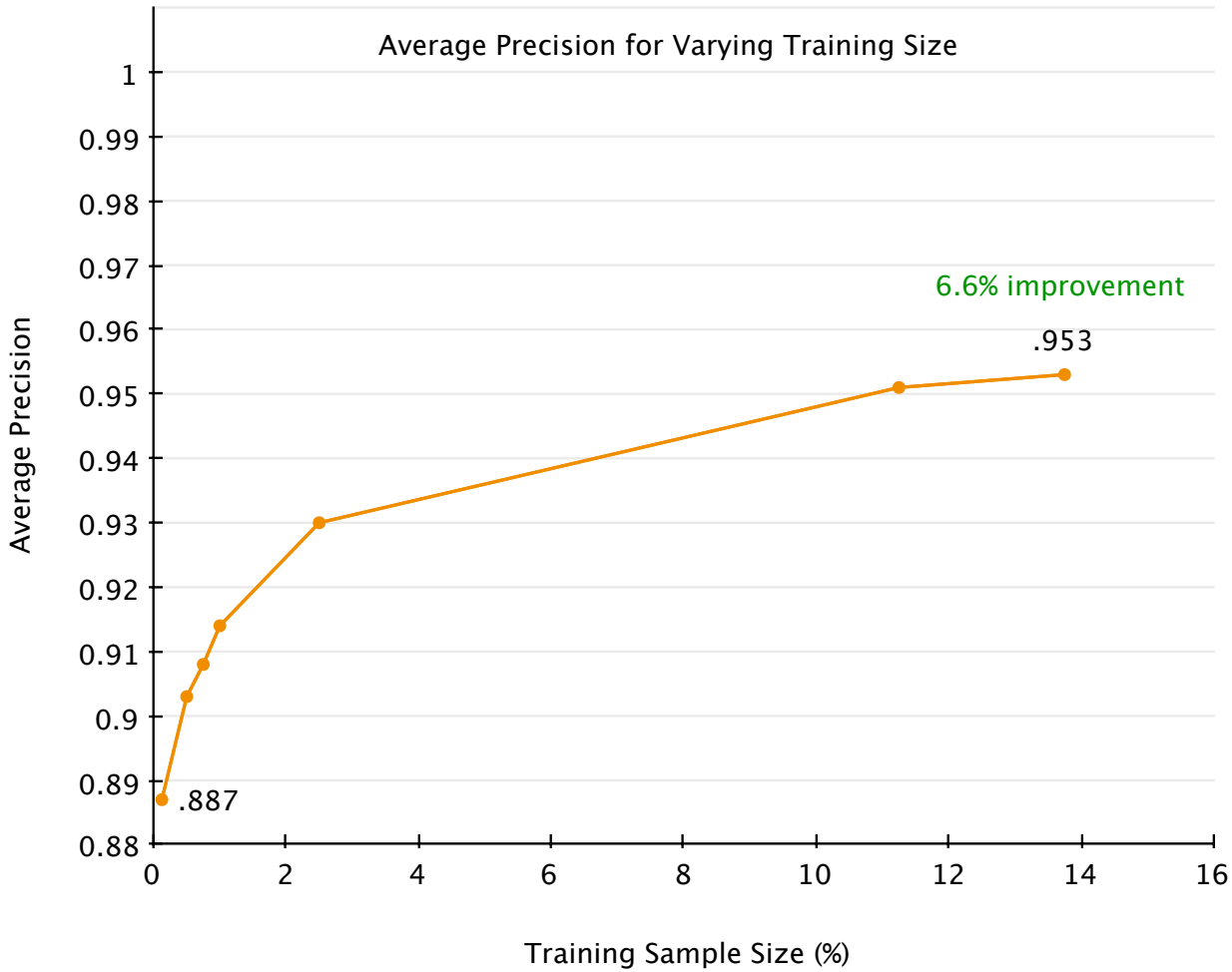
- Trial :=
  1. Randomly sample some percentage of available data for each port and train classifier
  2. Randomly sample some number of the remaining data points for each port and create testing samples
  3. Classify testing samples
- 50 Trials



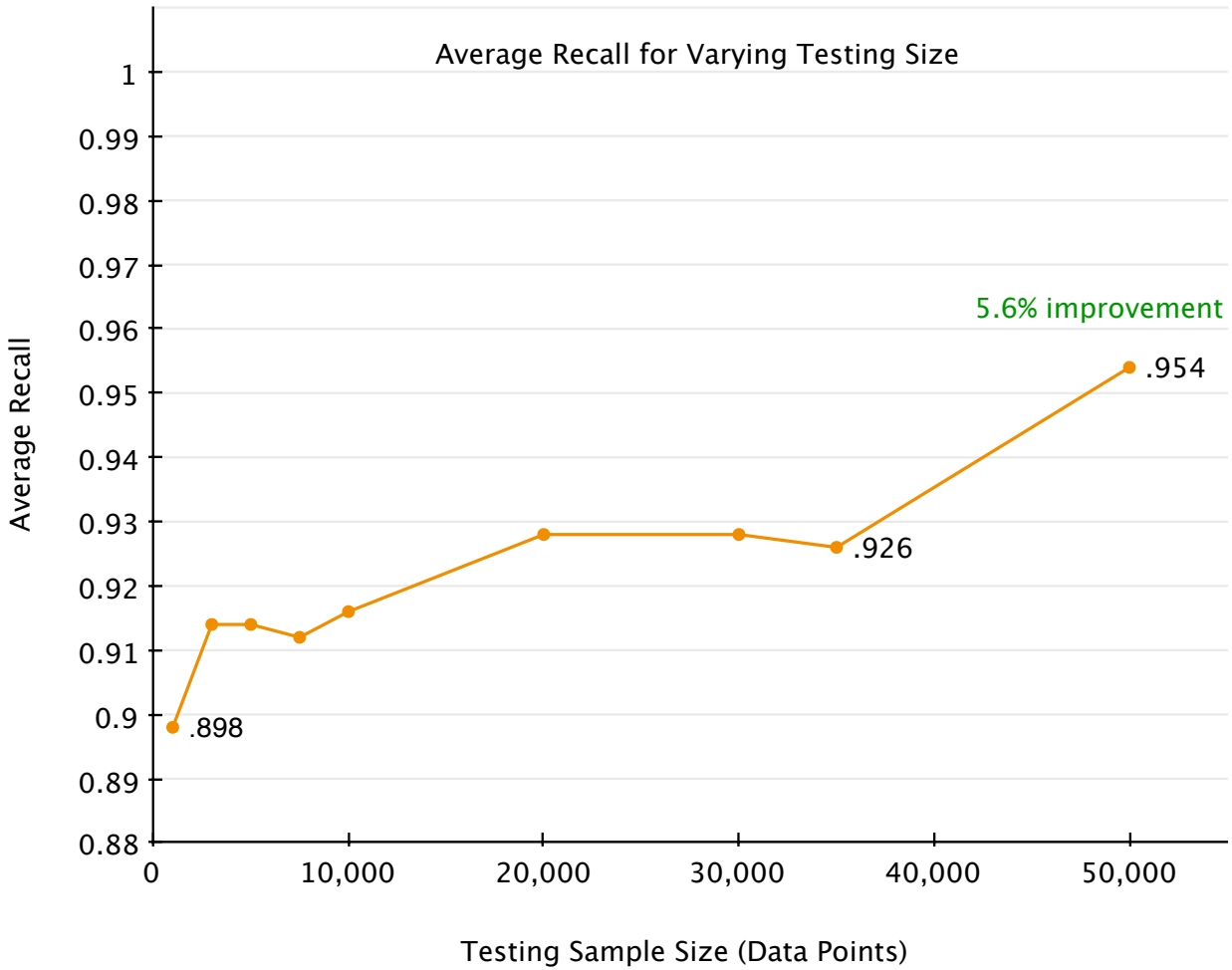
Training Size Selection



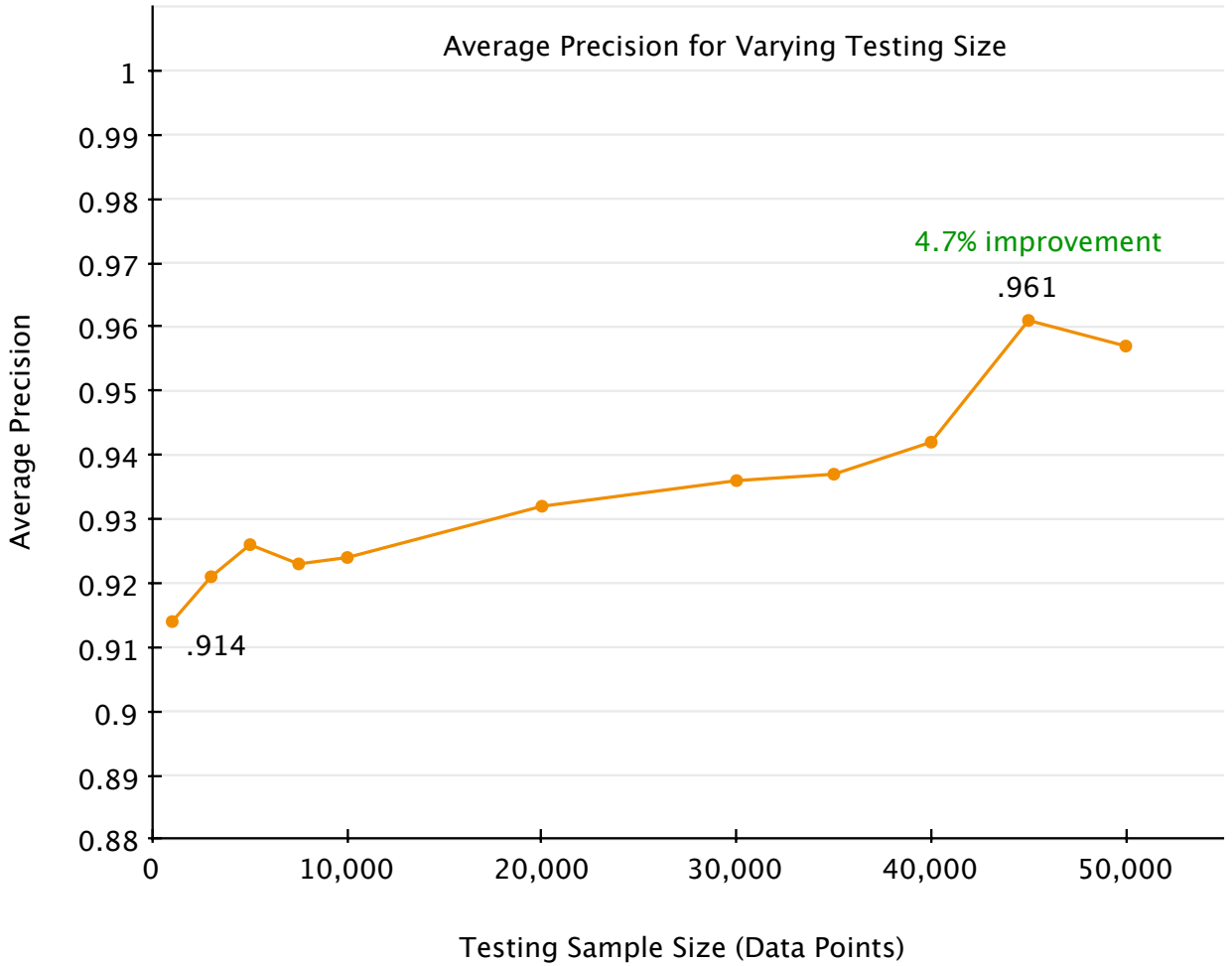
Training Size Selection



Testing Size Selection



Testing Size Selection



Results

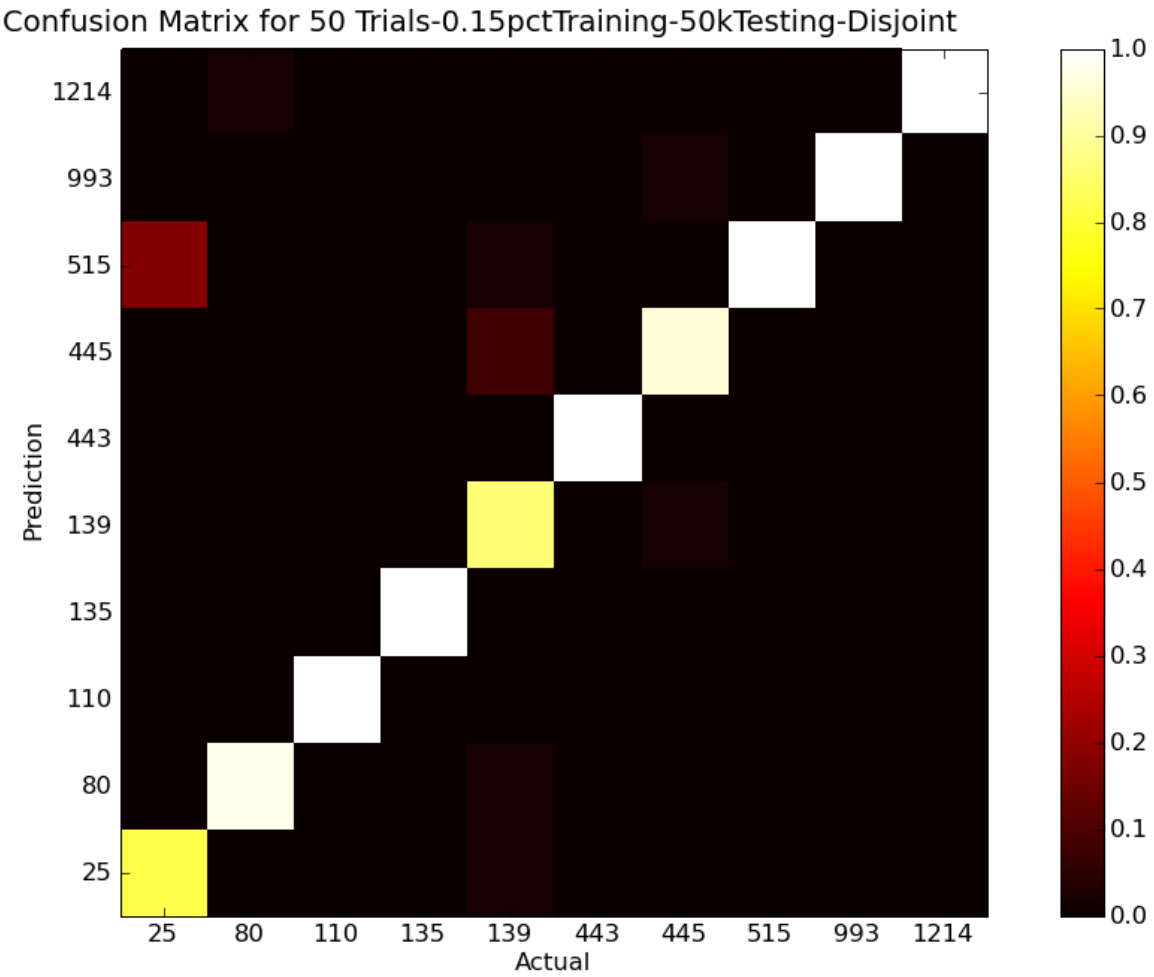
50 Trials

15% Training Set Size

50,000 Data Points Testing Set Size

96.5% Precision

96.0% Recall



Classification Confidence Threshold

- Goal: Eliminate close calls
- Require 1st place candidate to lead 2nd place by certain amount to make decision
- Standard deviation of scores

Methodology v2.0

- Randomly sample some percentage of available data for each port and train classifier
- Randomly sample some number of the remaining data points for each port and create testing samples
- Attempt to classify testing samples
  - If all testing samples reach threshold, done.
  - If any testing sample fails, rebuild testing samples and try again.

Classification Confidence

50 Trials

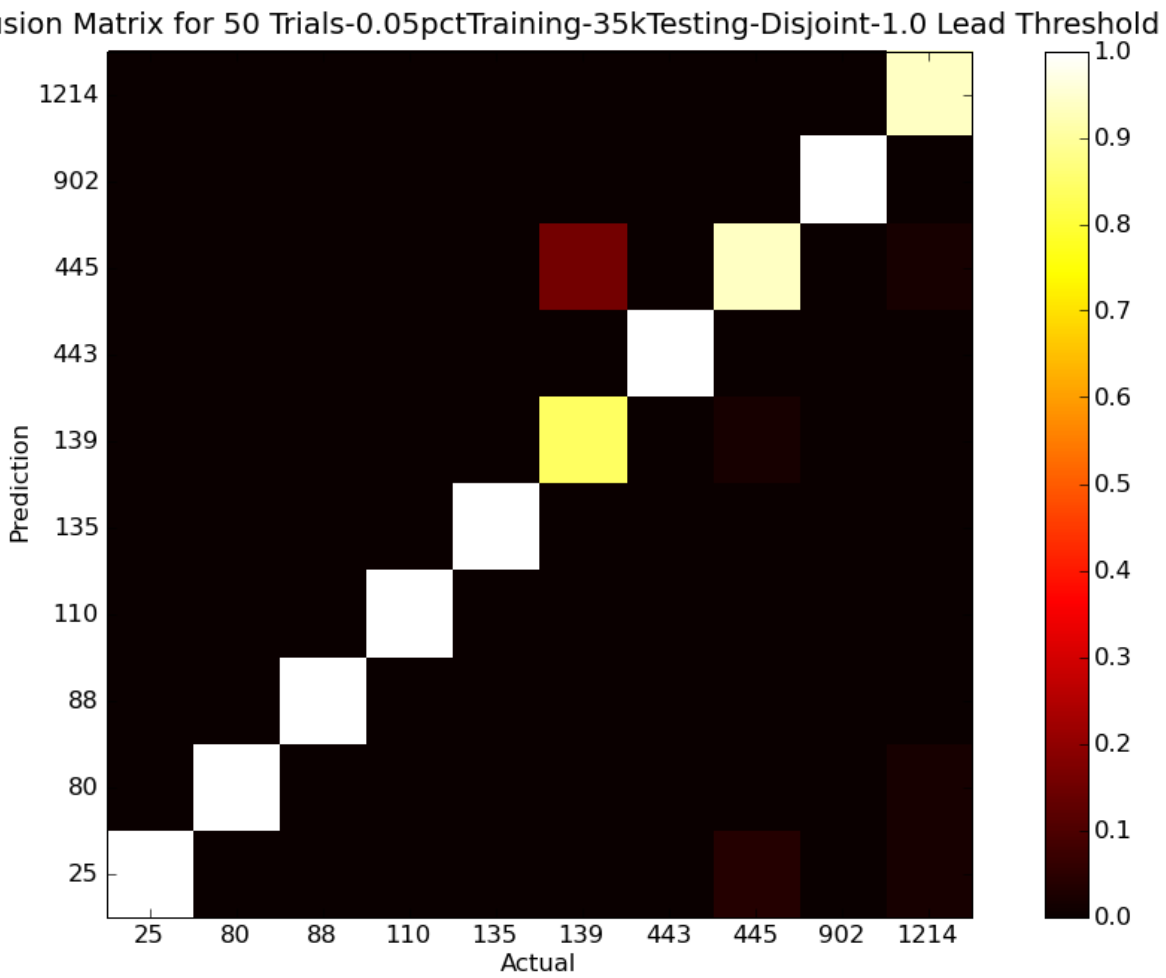
5% Training Set Size

35,000 Data Points Testing Set Size

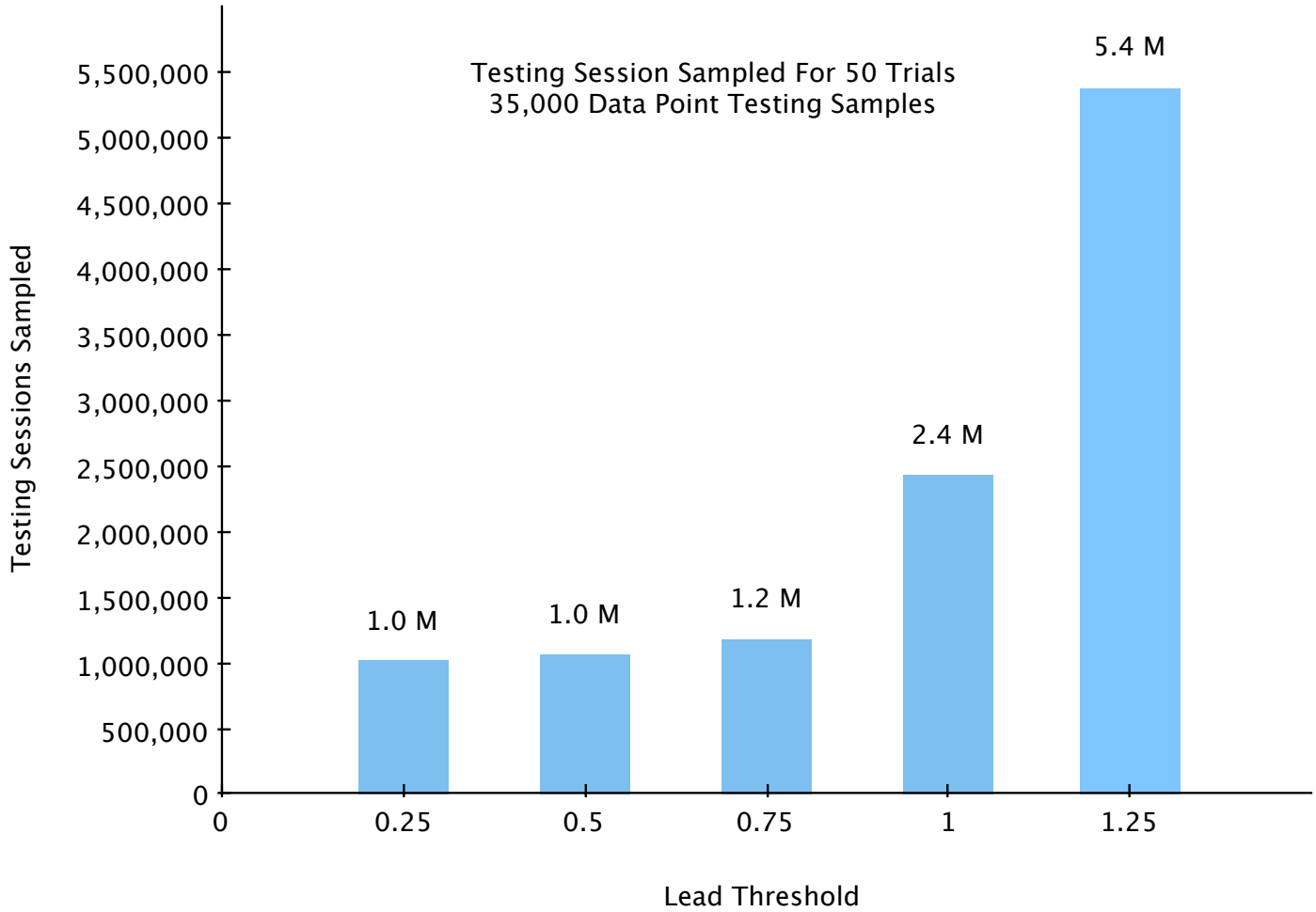
1.0 Lead Threshold

96.9% Precision

96.6% Recall



Classification Confidence



Ground Truth Testing

MIT Lincoln Labs DARPA Data
50 trials, 5% training sample size, 35,000 data point testing sample size, 1.25 lead threshold

- Precision: 98.3%
- Recall: 98.0%

Results

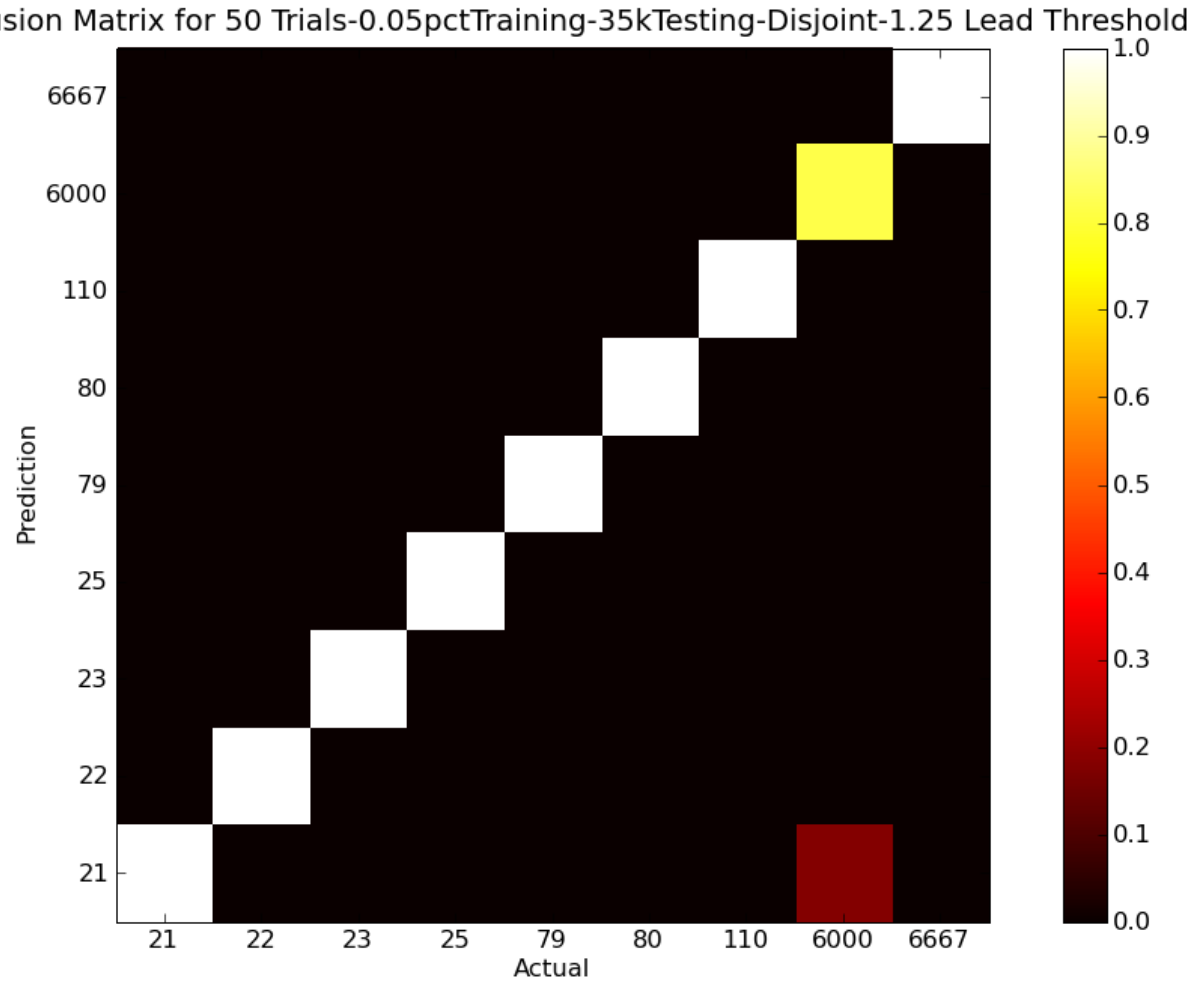
50 Trials

5% Training Set Size

35,000 Data Points Testing Size

1.25 Lead Threshold

Ground Truth

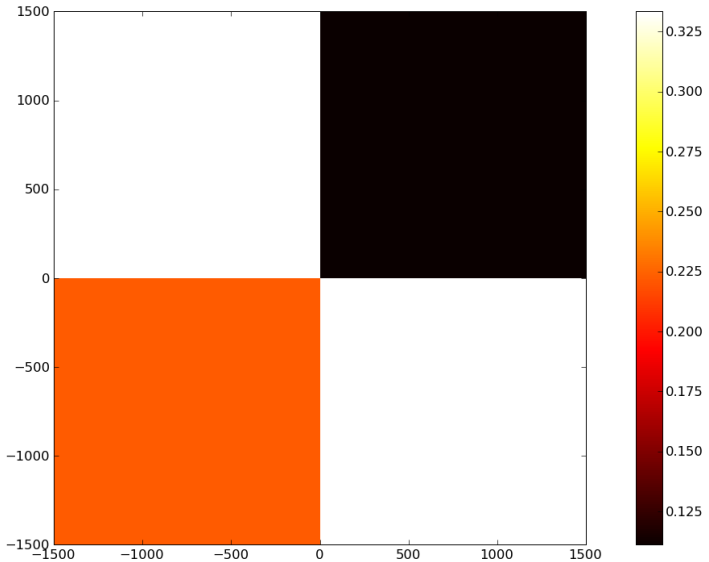


Evasion

One might attempt to thwart our technique by padding all packets to MTU.

Reduces problem to 4-quadrant problem.

Can still make decisions based on relative prevalence of each quadrant.



Current/Future Work

- Packet loss/re-transmission may cause unpredictable results
- On-line classification
- Training and testing from separate datasets
- UDP
- Subcategorization

Conclusion

- Modeling protocol behavior using only packet size, direction, and order
- Resistant to encryption and padding
- Average precision and recall $> 97\%$
- Quick and reliable traffic inspection
- Useful for pre-screening traffic for deeper analysis

Questions?

Thanks for listening.
Q & A

wwlian@gmail.com