“Tracking the Evolution of Web Traffic: 1995-2003

Felix Hernandez-Campos, Kevin Jeffay, F. Donelson Smith

IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)

Orlando, FL, October 2003

Outline

- Introduction
- Related Work
- Data Sets Collected at UNC.
- Analysis of UNC Data Sets
- Comparison with Mah, Barford and Crovella Studies
- Sampling Issues
- Conclusions

Introduction

- Web traffic has been the dominant traffic type on the Internet since mid-1990s.
- The Web (implying HTTP and HTML) is the ***de facto*** user-interface for many distributed applications.
- Goal:: To discover and document the evolving nature and structure of Web traffic.

Introduction

What the authors did:

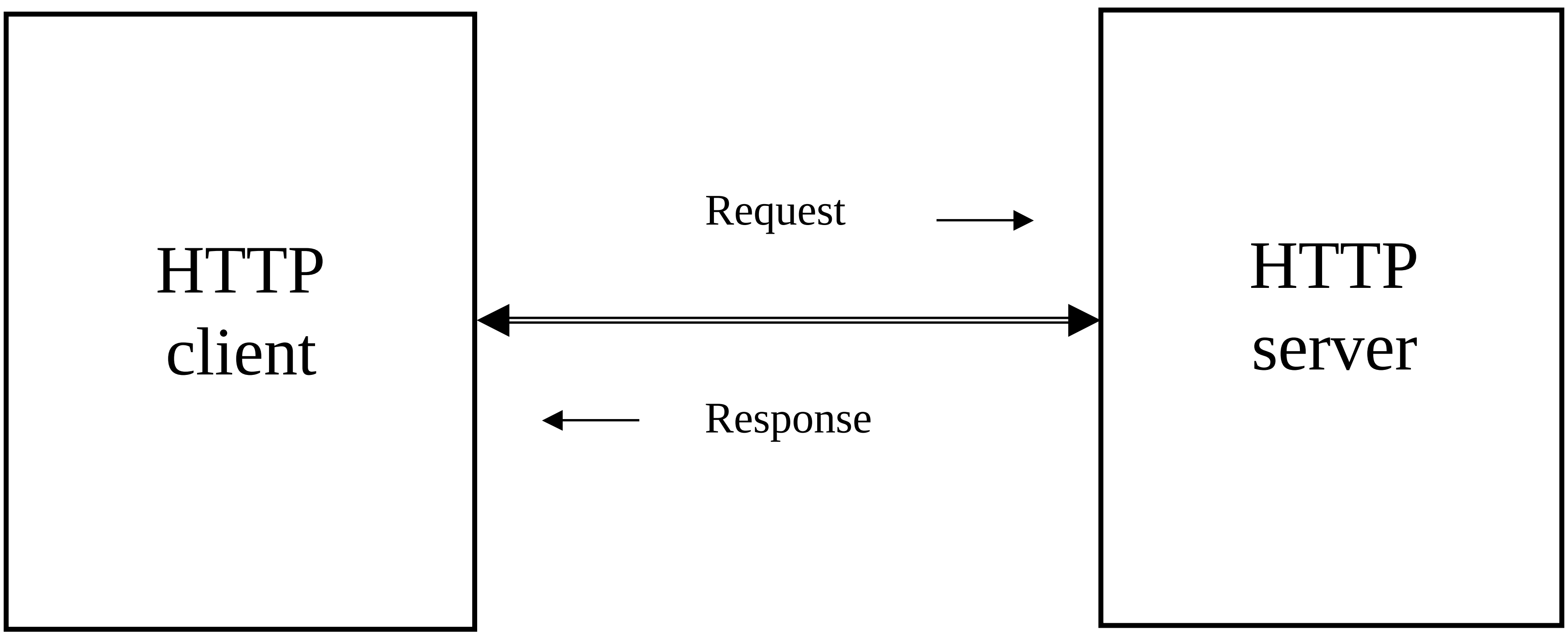
- Analyzed 1 terabyte of TCP/IP header traces collected in 1999, 2001 and 2003 at UNC at Chapel Hill.
- Compared results to similar measurements made from 1995 to 1998.

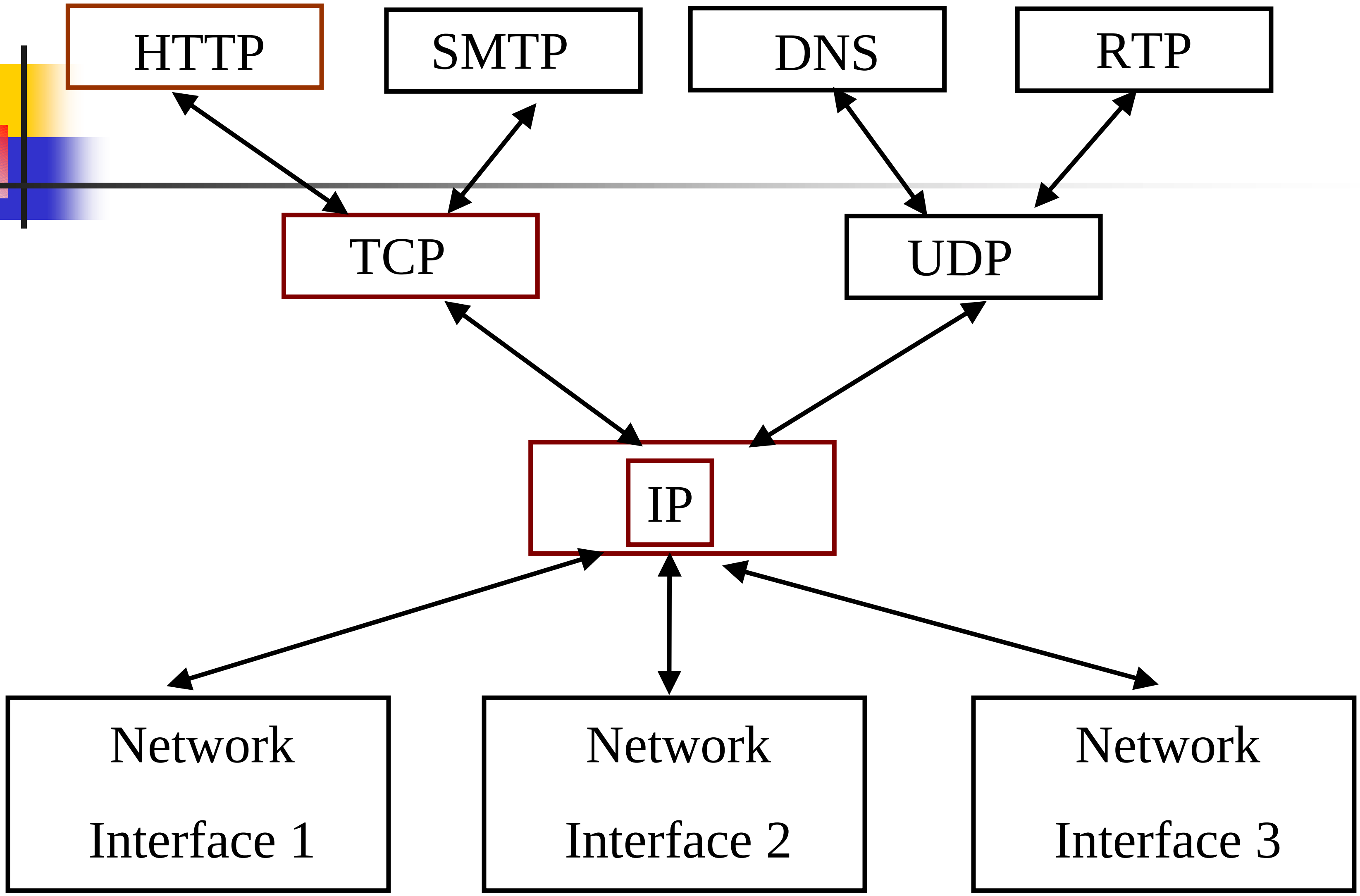
Introduction

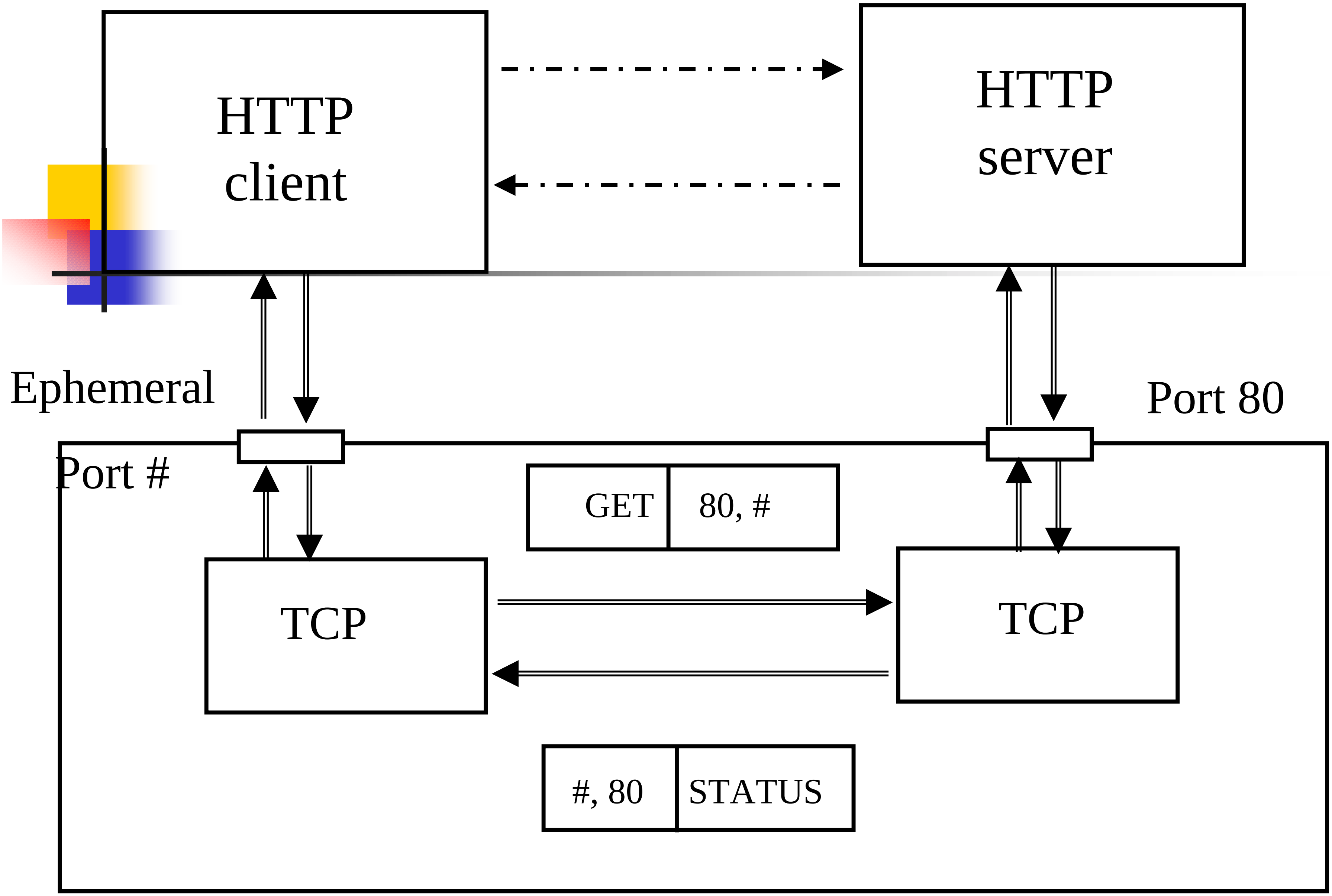
Contributions of this research:

- Empirical data for traffic generating models of Web traffic.
- Characterization of TCP usage including the effects of HTTP 1.1
- Characterization of Web usage that includes "new influences" such as banner ads, server load balancing and content distribution.

HTTP client/server interaction




HTTP
SMTP
DNS
RTP
TCP
UDP
IP
Network Interface 1
Network Interface 2
Network Interface 3


HTTP client
HTTP server
Ephemeral
Port #
Port 80
GET
80, #
TCP
TCP
#, 80
STATUS

Related Work

- Bruce Mah [10] captured 1.7 million TCP traces from UC Berkeley grad student population in 1995.
- Barford and Crovella, et al, [2,4,7] collected in aggregate around 1 million references to Web objects from undergrad CS students at BU in 1995 and 1998.
- Considering the evolution of the Web, this data is old and before the deployment of HTTP 1.1

Data Collected

- 1.6 billion TCP segments generated by a user population of 35,000 users and the transfer of almost 200 million Web objects.
- Analyzed *unidirectional* traces sent from Web servers to client browsers.
- Used TCP sequence and ACK numbers to determine request and response sizes.

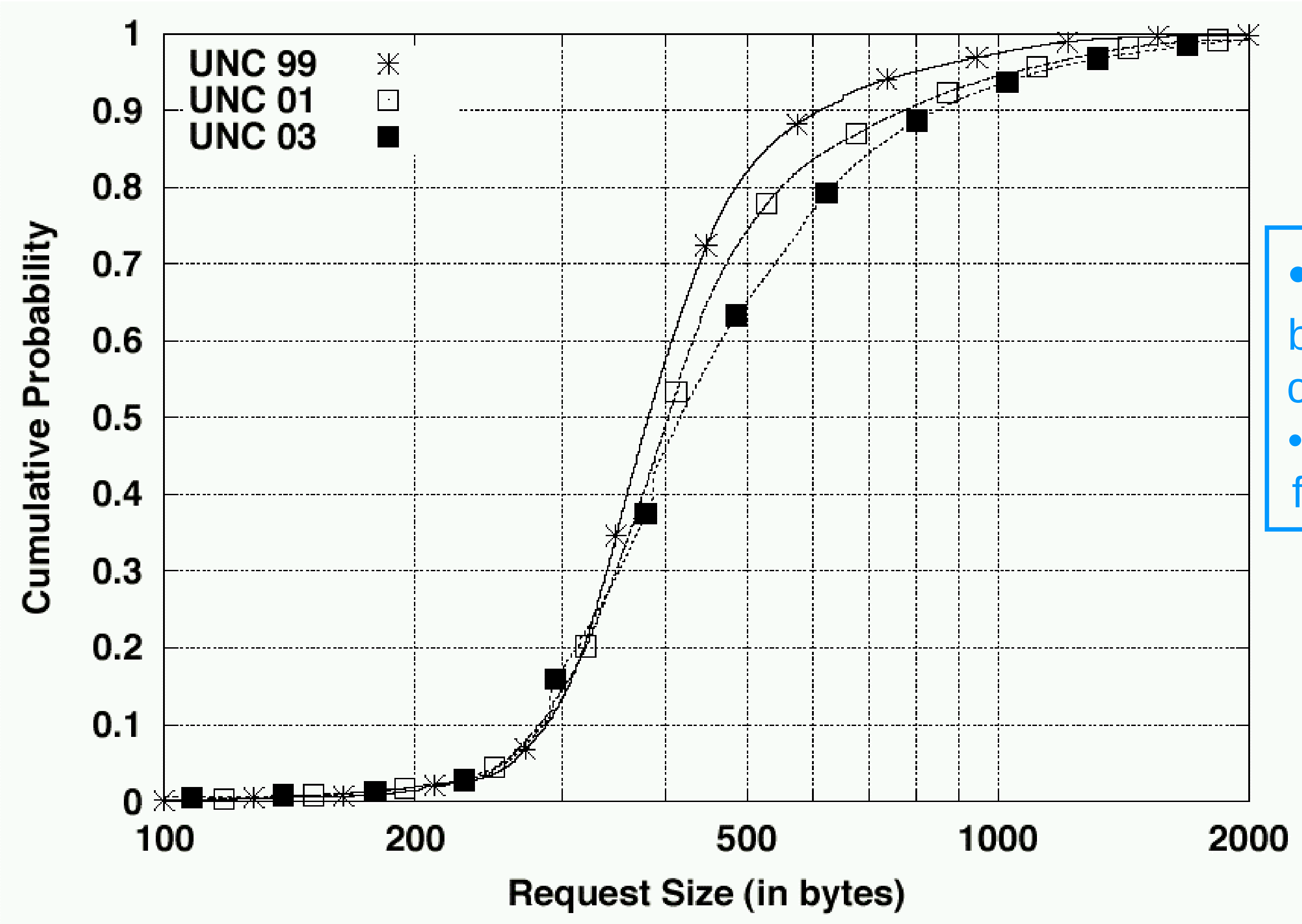
Data Sets

- **[UNC 99]** Fall 1999 (6 one-hour samples, over 7 consecutive days)
- **[UNC 01]** Spring 2001 (3 four-hour samples, 7 consecutive days)
- **[UNC 03]** Spring 2003 (8 one-hour traces over 7 consecutive days)
- Network:
  - 1999: OC-3 (155 Mbps) ATM link
  - 2001 and 2003: OC-48 (2.4 Gbps) Cisco DPT technology; However traffic monitor placed on Gigabit Ethernet link (1 Gbps).

Analysis of UNC Data Sets

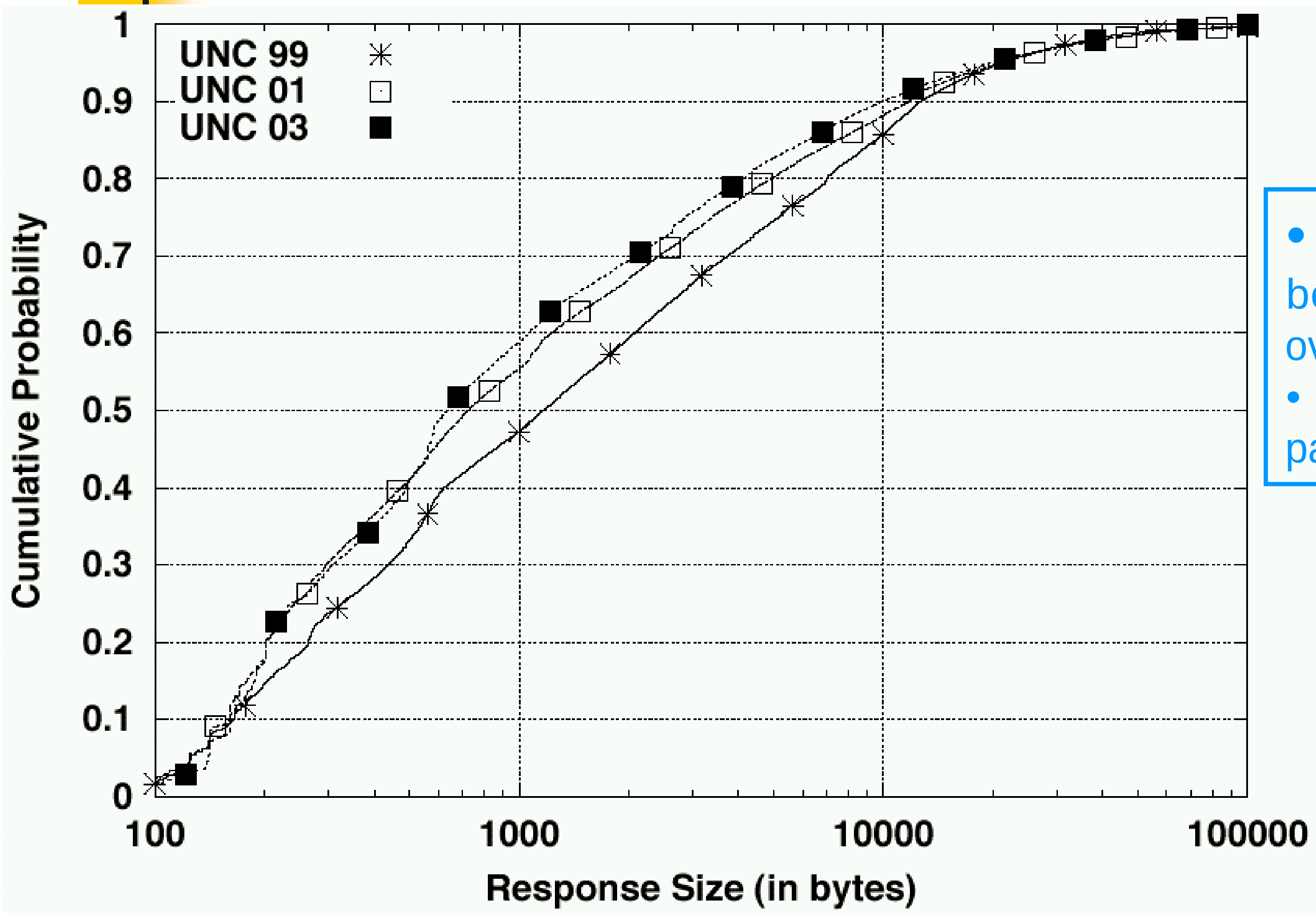
- TCP Request and Response Data Sizes
- User and Web Content Characterizations
  - Distribution of number of objects per page.
  - Distinction between primary and non-primary servers with respect to number of objects requested and size of response objects.

Figure 1: Request Sizes



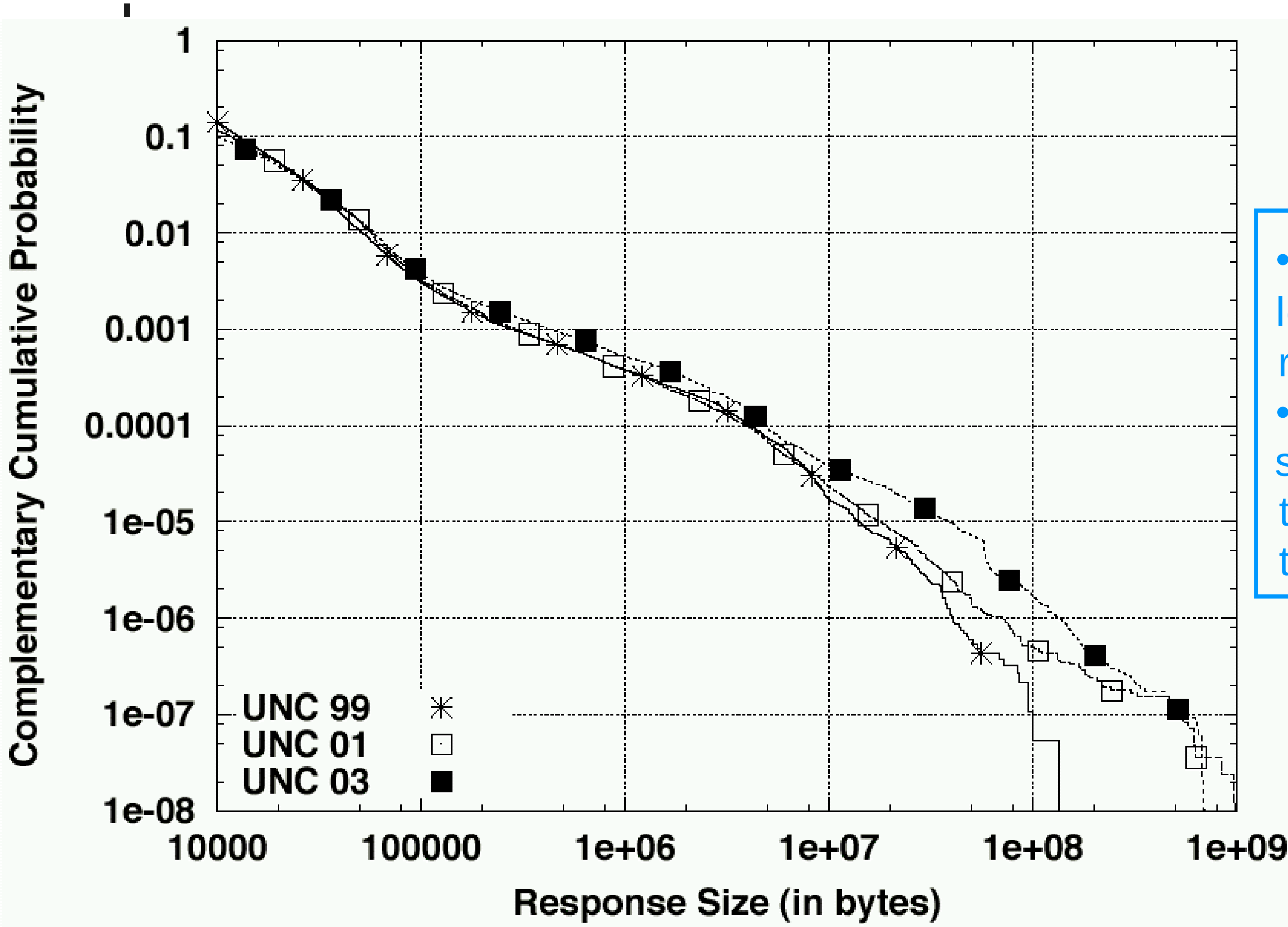
- Requests becoming larger over time.
- But, still typically fit in one packet.

Figure 3: Response Sizes



- Responses becoming smaller over time.
- Median fits in one packet.

Figure 4: Response Size Tail



- CCDF shows long-tailed responses.
- There is a slight increase in the tail over time.

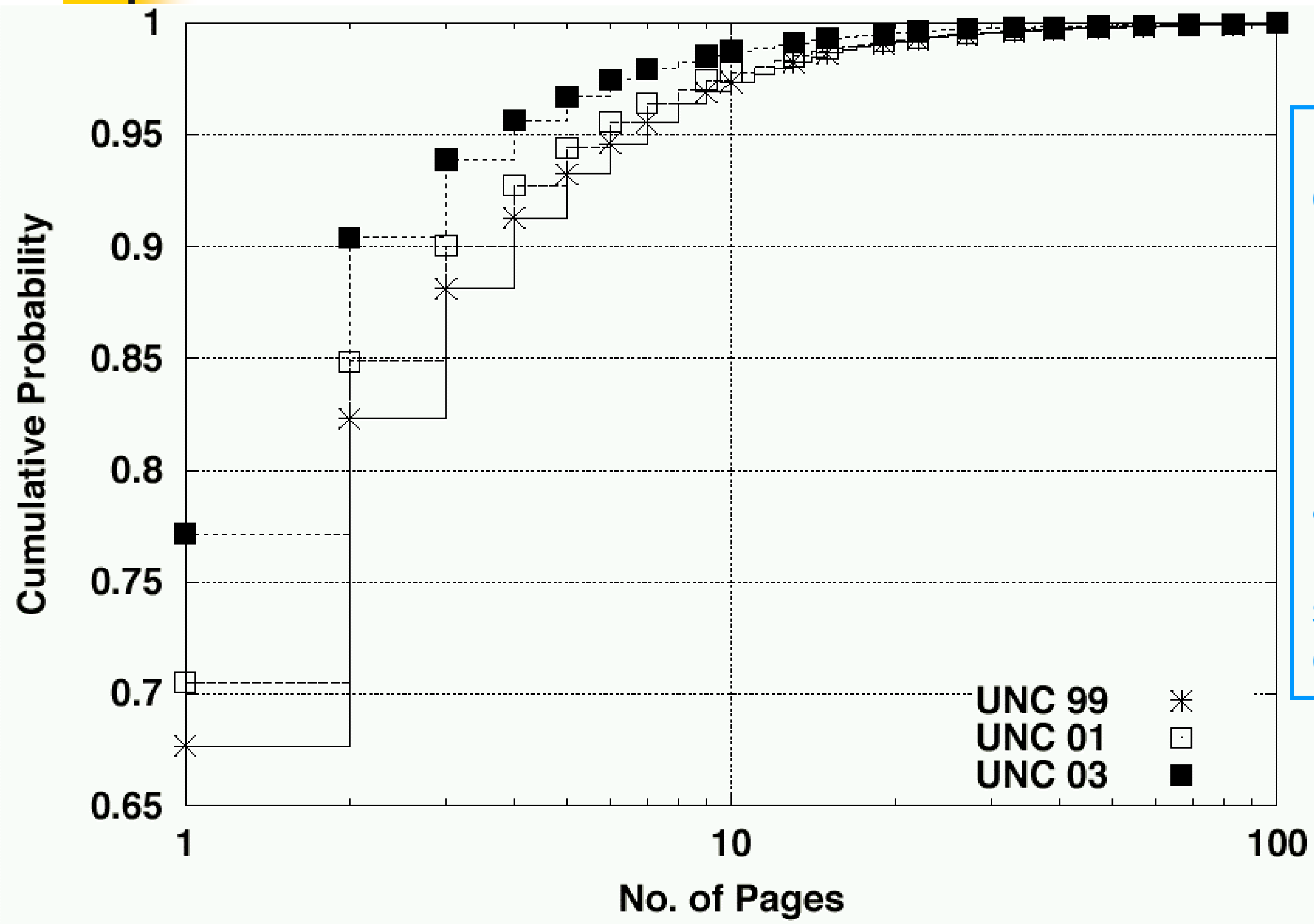
User and Browser Characteristics

- Without HTTP headers, authors “infer” HTTP behavior from TCP connections.
- Aggregate by unique client IP address and then time-sorted all flows between clients and servers.
- Assume each IP address is one user (fewer NATs on campus).
- Used previous researcher’s heuristic approach to estimate the first request is “page”.

User and Browser Characteristics

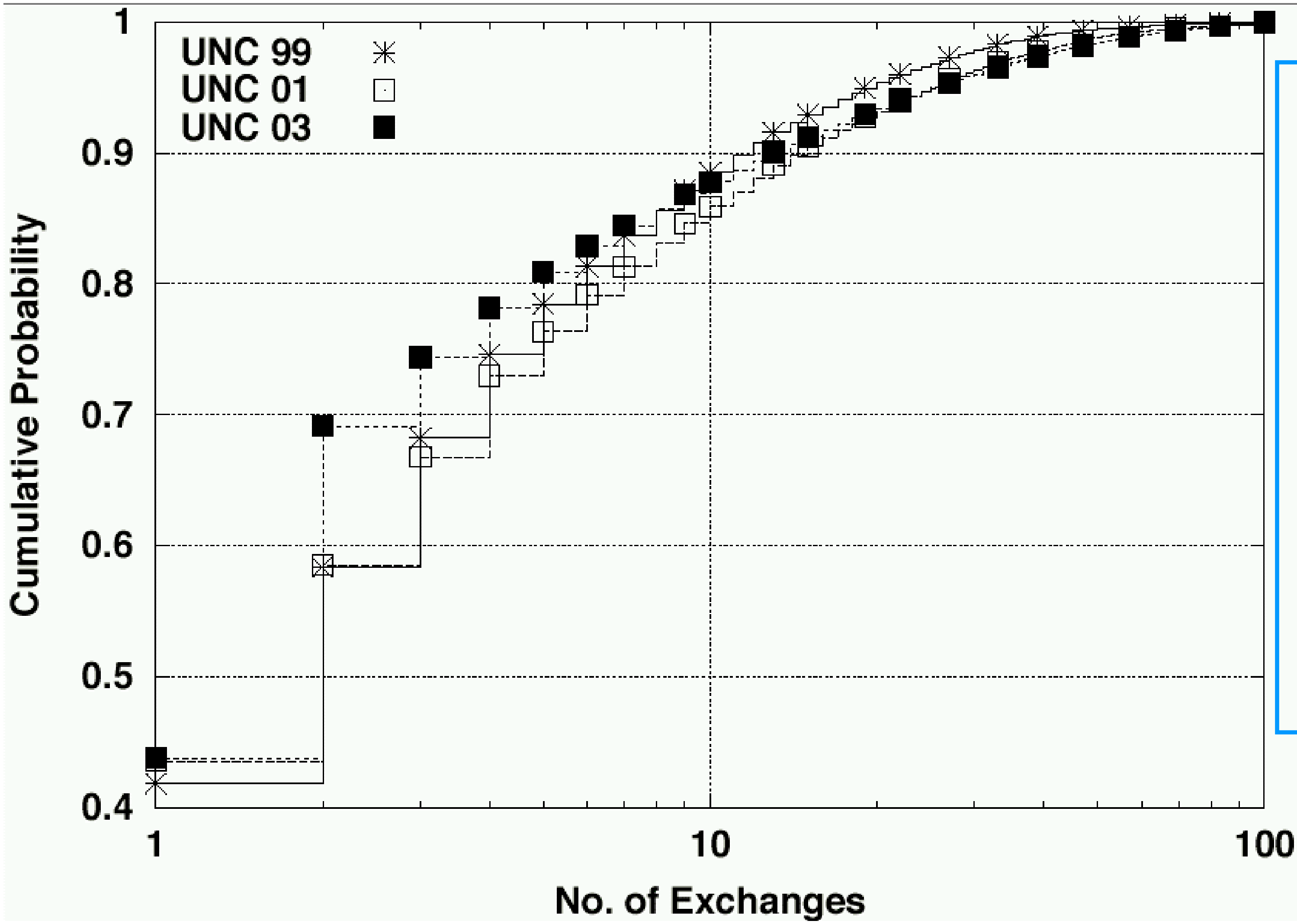
- An “object” is synonymous with a server response. Note – this includes error reports.
- A threshold of 1 second is used to distinguish “idle time” (or “think time”).
- Note – all Web traffic observed does not include objects from the local browser cache.

Figure 7: Number of Consecutive Requests to Same Server



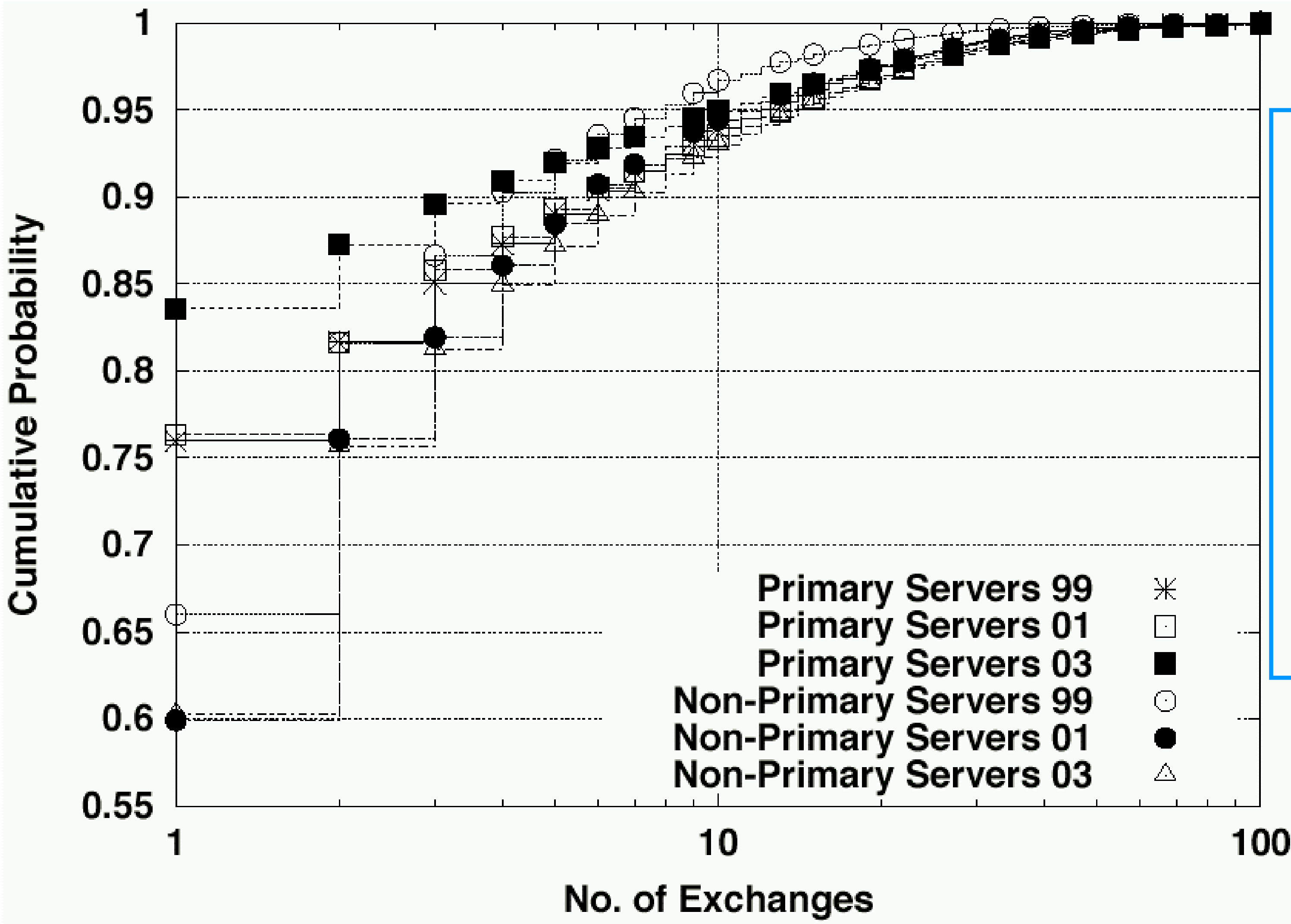
- Most requests go to one page per server.
- This trend is increasing over time.
- Results are attributed to load balancing in server farms and CDNs.

Figure 9: Number of Objects per Page



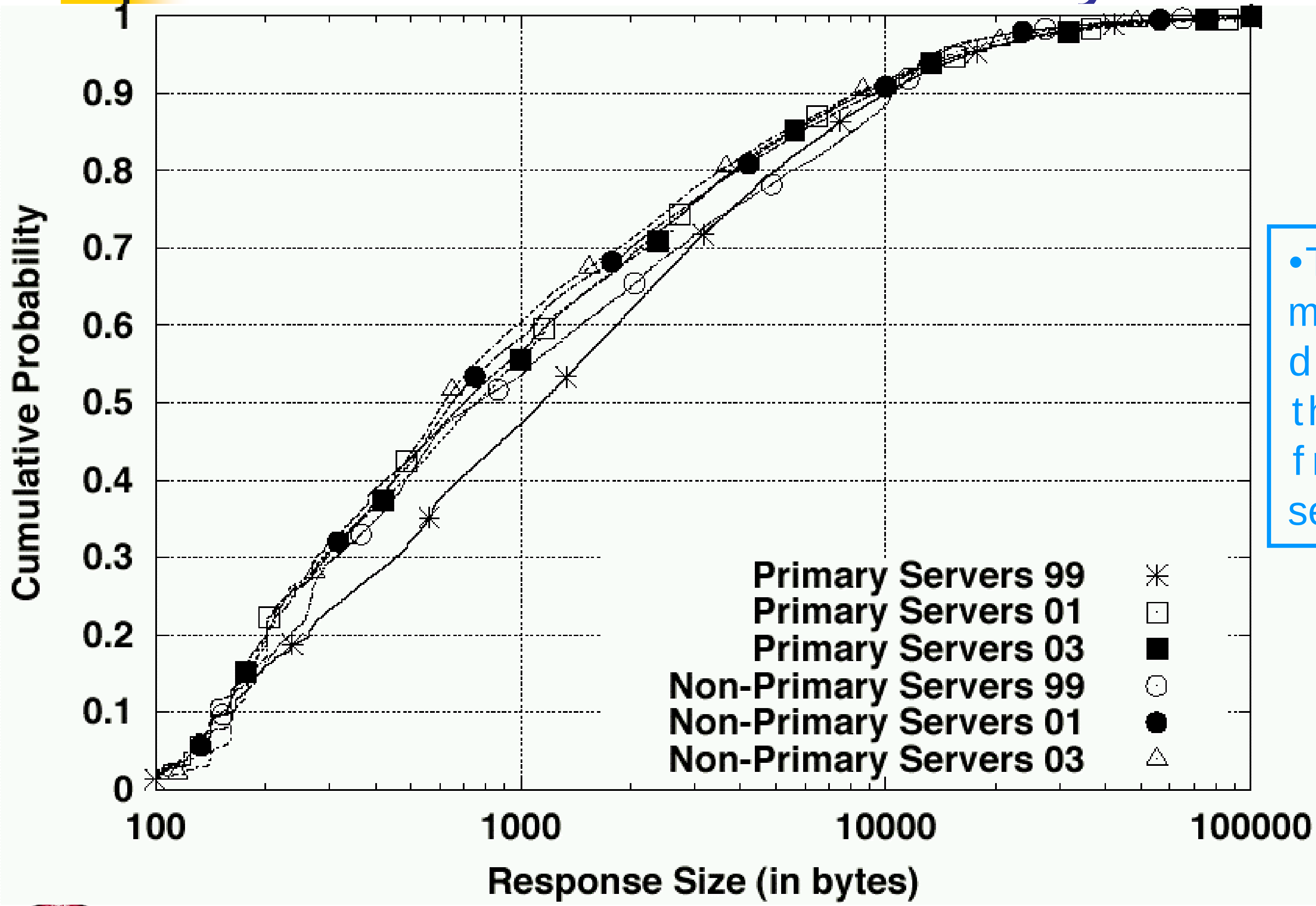
- 40% are simple pages with no imbedded objects.
- Some pages are quite complex with 100 objects
- Both trends increasing
- Data “fuzzy” due to browser caches.

Figure 10: Primary vs Secondary Servers



- Trend of only one object from primary server increases.
- Trend of more than one object from non-primary server increases.
- Note – graph is poor!!

Figure 13: Response Sizes Primary vs. Non-Primary



- There are only minor differences in the object sizes from different servers.

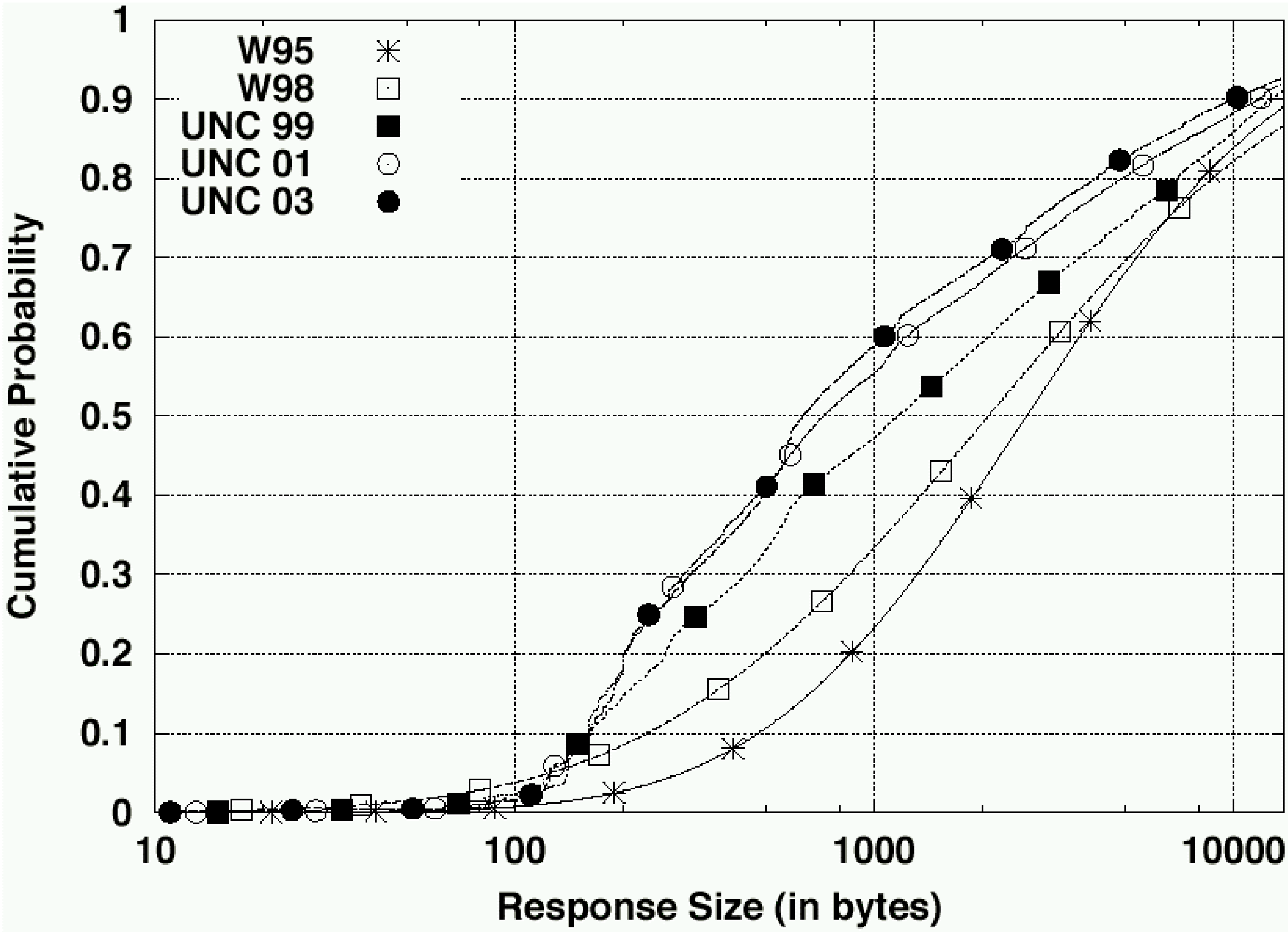
Limitations of Methodology

- TCP analysis solid (inferences about the number of packets and flows are reasonable.)
- HTTP analysis less certain due to:
  - Pipelined exchanges
  - User/browser interactions (Stop and Reload)
  - Browser and proxy caches
  - TCP processing dealing with loss, duplication and re-ordering of packets in the network.

Comparison with Mah, Barford and Crovella, et al. Studies

- Distribution of response sizes has evolved over time.
- Data fits Barford's lognormal-Pareto models of response times.
- Change in distribution of objects per page reflect increased complexity in Web page layout.

Figure 15: SURGE (BU) vs UNC



- A clear reflection of the evolution of Web objects
- Figure 17 with Mah data is very similar.

Table 1: Summary Data

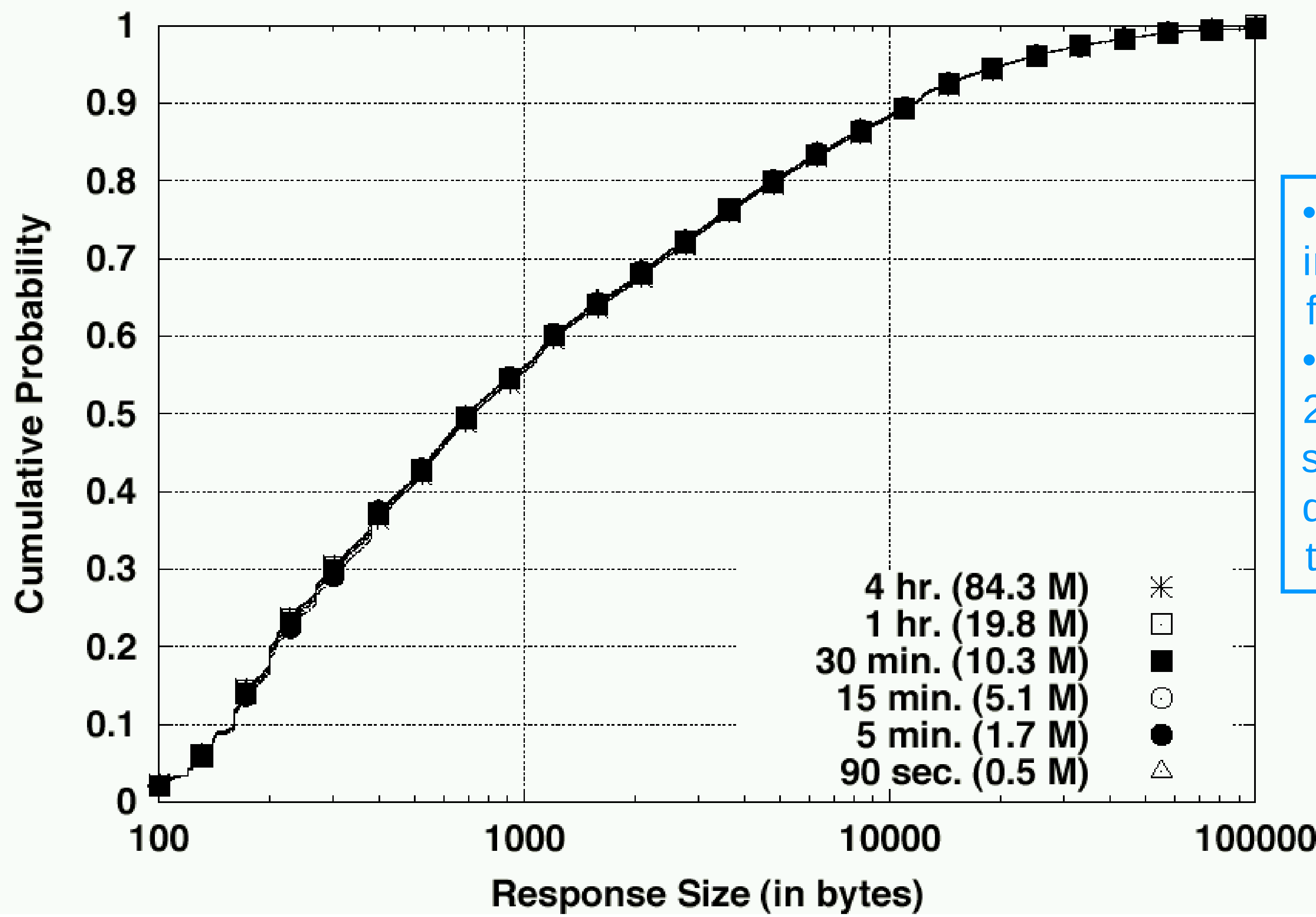
Data Set	Sample Size (Number of responses)	Min Response Size	Max Response Size	Mean Response Size	Median Response Size
W95	269,811	3	20,135,435	14,826	2,245
W98	66,988	1	4,092,928	7,247	2,416
Mah 95	5,300	62	8,146,796	10,664	2,035
UNC99	18,526,201	1	135,294,044	6,734	1,164
UNC01	84,343,238	1	984,871,070	6,397	722
UNC03	96,836,703	1	718,067,386	7,296	632

- Notice decreasing trend in median response sizes.
- Caveat - larger sizes in some experiments are partially due to larger samples.

Sampling Issues

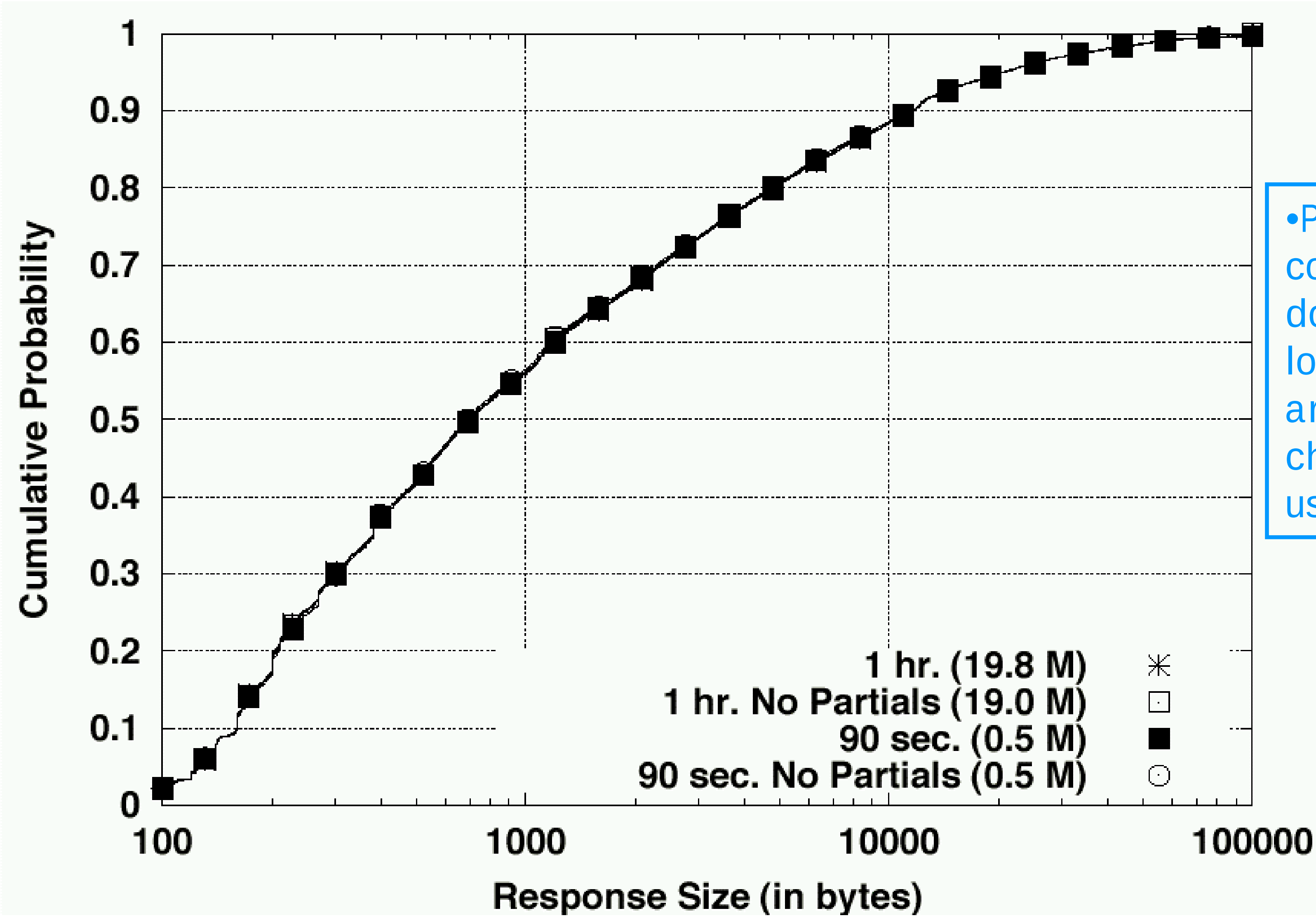
- Number and duration of trace intervals bring up important analysis issues.
  - 1 hour of only 68 byte TCP headers consumes 30 Gigabytes of storage at UNC.
  - 90-second trace only requires 200MB for each of inbound and outbound traces.
  - Processing takes *hours.*
  - Capturing can slow down routers.
- Questions
  - Do lengths of traces affect the distribution shape?
  - Do incomplete TCP connections affect the distribution shapes?

Figure 23: Response Sizes for Sub-Samples



- 90-second intervals work fine.
- Although Fig 24 shows slight difference in the tail.

Figure 25: Complete and Partial Connections



- Partial connections do not hurt as long as you are NOT characterizing user activity.

Conclusions

- Captured and analyzed Web traffic for 35,000 UNC people, three data sets from 3 years
- General Results:
  - HTTP request sizes are increasing.
  - HTTP response sizes are decreasing.
  - Largest HTTP responses are increasing.
  - Web pages complexity is increasing (more objects per page).

Future Work

- Effects of persistent connections and pipelining?
- What about other (non-port 80) traffic over HTTP?
  - About ½ of all TCP traffic "other"
- Are all objects Web objects?
  - As opposed to re-direction requests, error messages
  - This may help understand Web structure.