Distilling the Internet's Application Mix from Packet-Sampled Traffic

Passive and Active Measurement Conference 2015
New York City

Philipp Richter
TU Berlin

Nikolaos Chatzis
TU Berlin

Georgios Smaragdakis
TU Berlin / MIT

Anja Feldmann
TU Berlin

Walter Willinger
NIKSUN, Inc.

prichter@inet.tu-berlin.de

The Internet's Application Mix

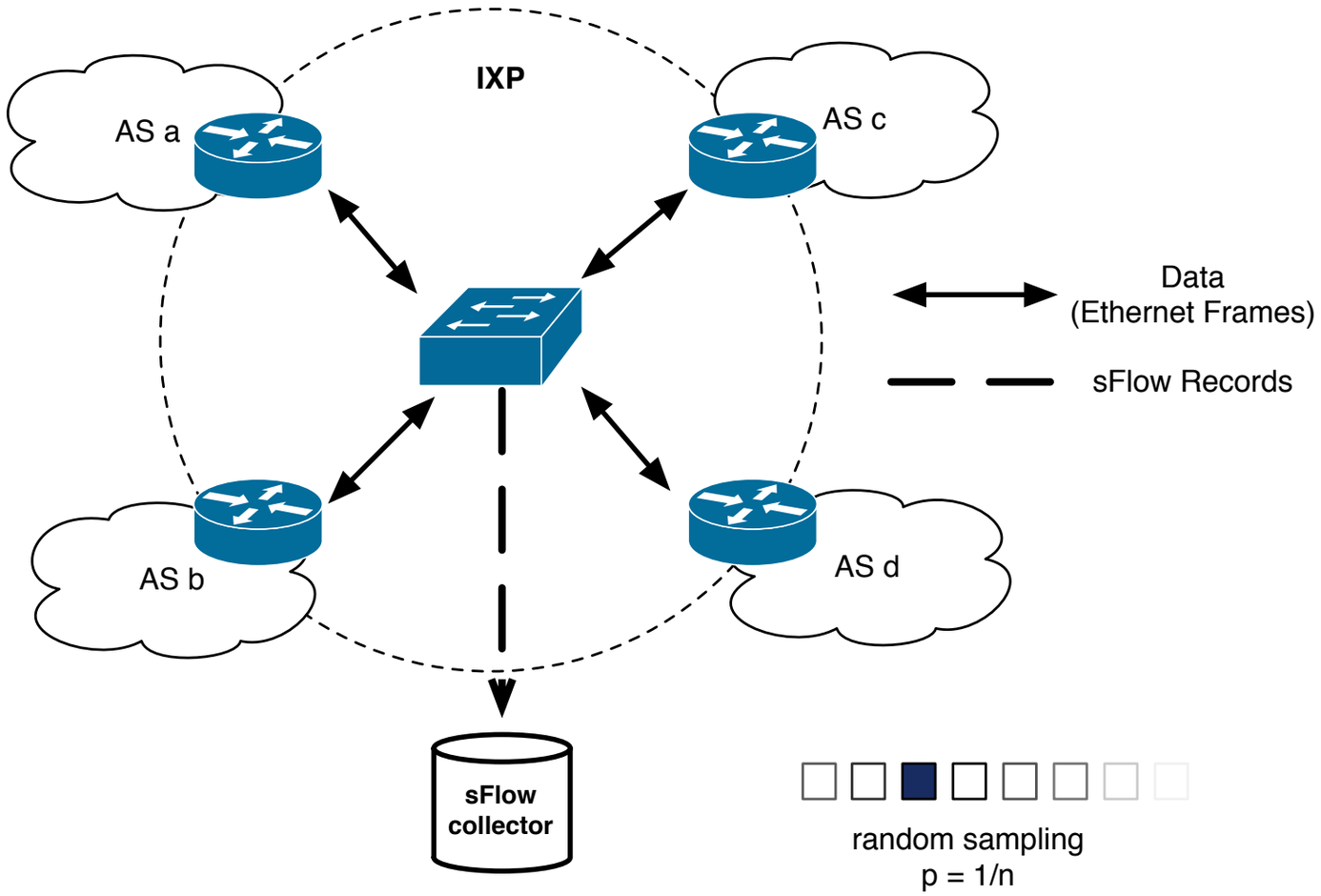
- *"What is the application mix in today's Internet"?*
  - Optimizing network performance
  - Provisioning network resources
  - Identifying new trends in Internet usage

- Numerous academic and commercial studies
- Typically focus on a single or a few locations
- We study the application mix seen on tens of thousands of peering links at a Large European IXP

Agenda

- Dataset & Challenges
- Related Work & Applicability
- Classification Approach
- Results

A Large European IXP



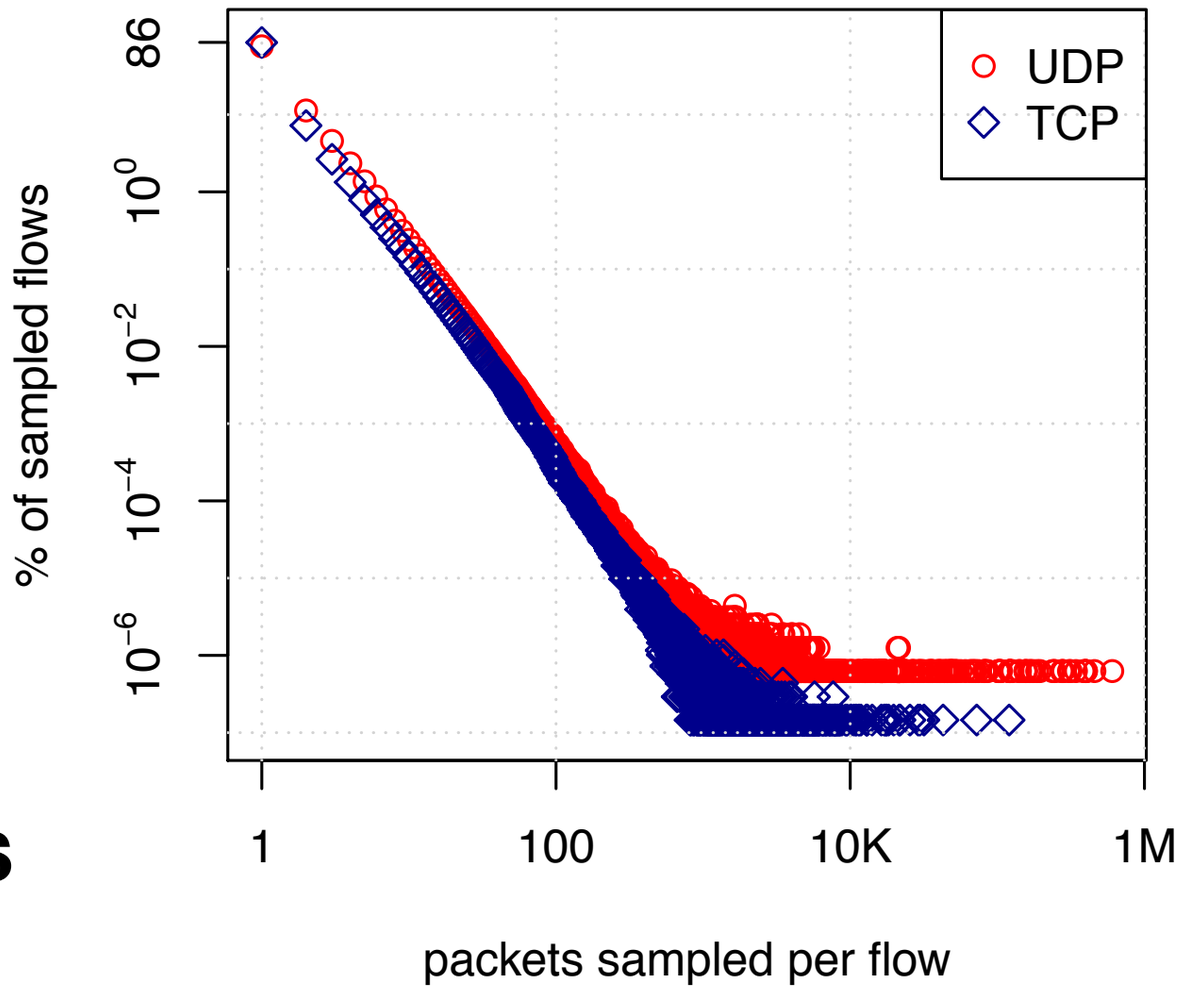
- Daily traffic (2013): ~14 PB
- sFlow export
- Random Sampling 1/16K Packets
- Snaplen 128 Bytes
- Weekly Snapshosts dating back to 2011

most recent snapshot (2013-09, 496 networks, 1 week)

packets sampled	bytes sampled	IPv4 / IPv6	TCP / UDP
9.3B	5.9TB	99.37% / 0.63%	83.7% / 16.3%

Dataset Characteristics: Sampling

- Typically one packet per sampled flow
- Can be any packet (e.g., just a TCP ACK)
- One packet = unidirectional visibility
- **a “random set” of packets**

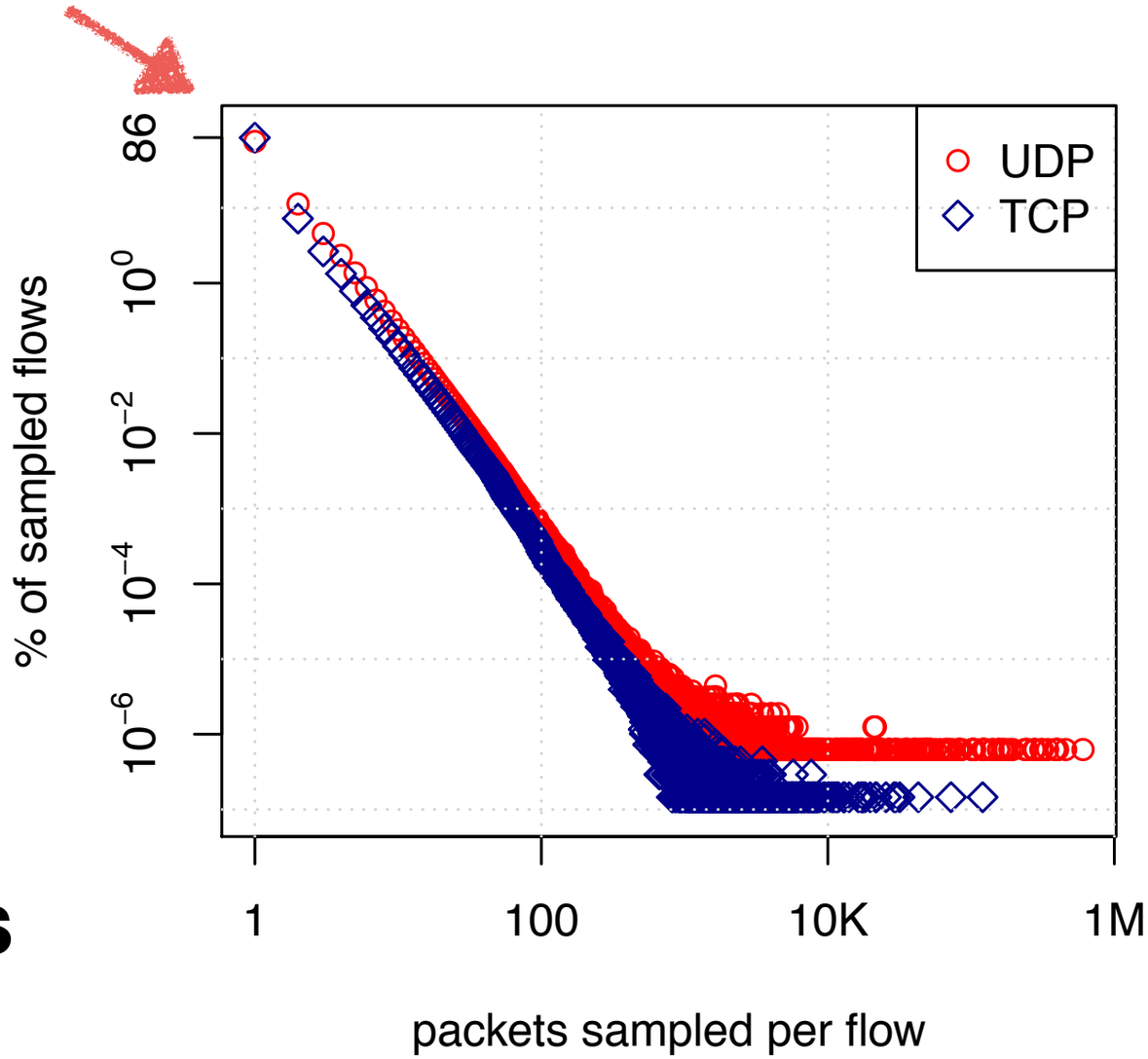


5-tuple aggregation: packets seen per sampled flow

Dataset Characteristics: Sampling

86% of sampled TCP flows: one packet

- Typically one packet per sampled flow
- Can be any packet (e.g., just a TCP ACK)
- One packet = unidirectional visibility
- **a “random set” of packets**



5-tuple aggregation: packets seen per sampled flow

Classification Approaches

(I) Payload-based Approach

- Match application signatures (i.e., handshakes)
- Produces accurate results
- *Challenge: Most sampled packets are "in the middle" of a flow and contain only binary data.*

(II) Port-based Approach

- Match port-numbers to well-known applications
- Problems: Applications hiding behind well-known ports, applications using random ports (P2P)
- *Applicable as-is*

taxonomy based on Kim et al.

Classification Approaches (cont.)

(III) Flow feature-based Approach

- Match per-flow properties (i.e., #packets, #avg. packet size etc.)
- *Not applicable, no per-flow statistics available*

(IV) Host behavior-based Approach

- Social interaction between hosts (e.g., BLINC)
- Network-wide interaction of hosts (e.g., TDGs)
- *Partially applicable*

taxonomy based on Kim et al.

Classification Approaches

(III) Flow feature-based Approach

- Match per-flow properties (i.e., #packets, #avg. packet size etc.)
- *Not applicable, no per-flow statistics available*

(IV) Host behavior-based Approach

- Social interaction between hosts (e.g., BLINC)
- Network-wide interaction of hosts (e.g., TDGs)
- *Partially applicable*

We combine several approaches

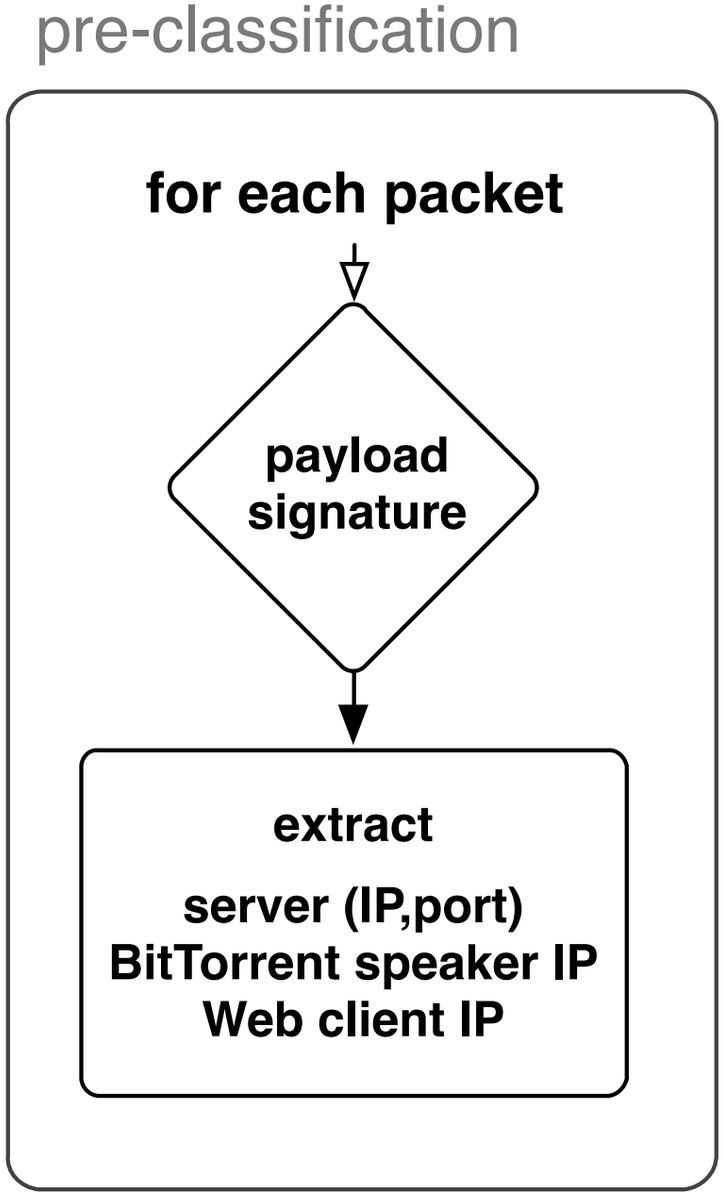
taxonomy based on Kim et al.

Our Classification Approach

1. Pre-Classification Phase
   derive state which will be leveraged later

2. Classification Phase
   actual classification of packets

Pre-Classification Phase



state - server-side:

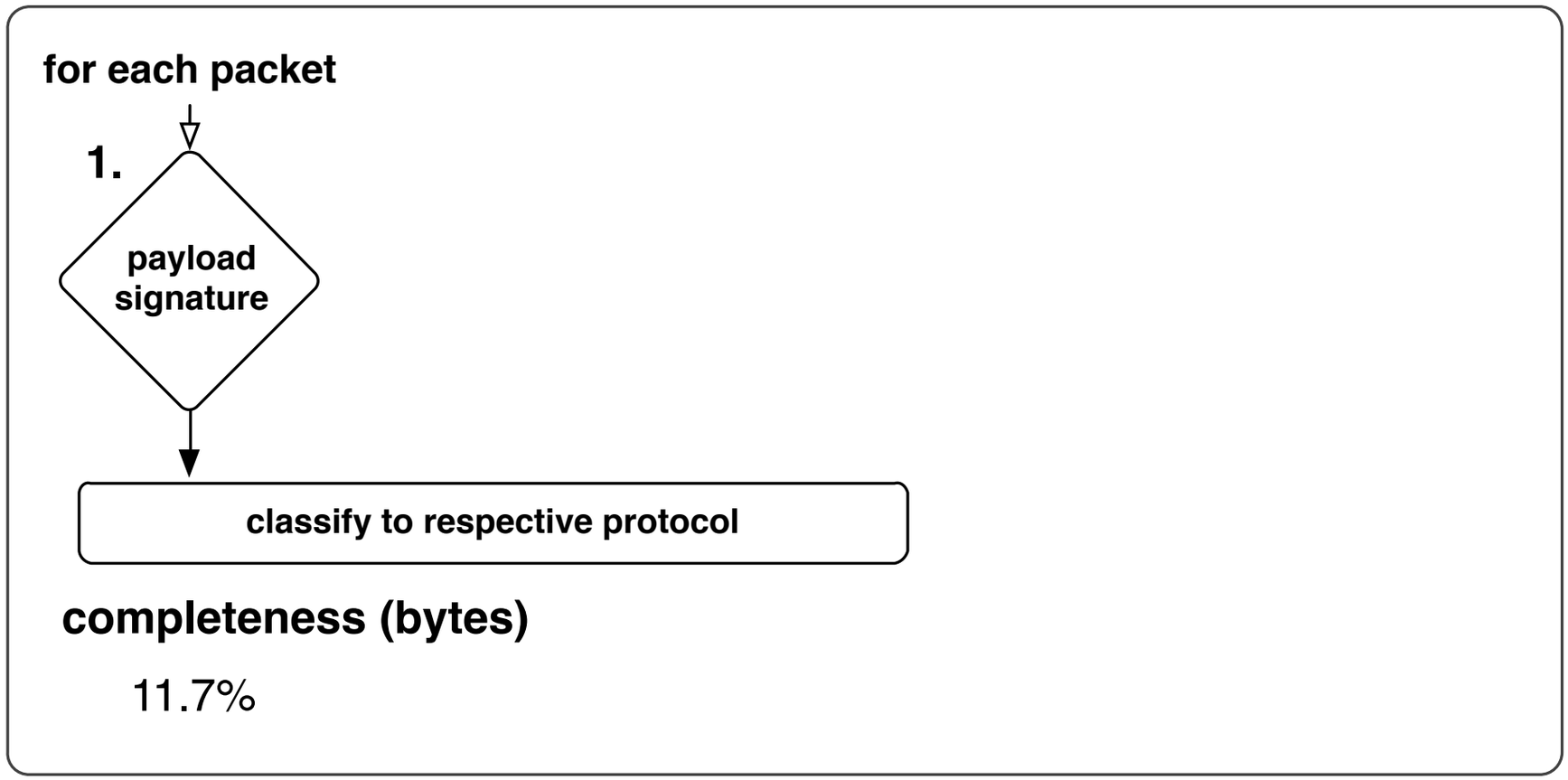
- Server-Endpoint (IP,port)
  e.g., ~2.7M HTTP endpoints
- For SSL-based applications:
  SSL signature on well-known port
  e.g., ~210K HTTPS endpoints (IP,port)

state - client-side:

- BitTorrent peer IPs (~38.9M)
- Web Client IPs (~37.7M)

We extract Connection Endpoints (*"state"*)

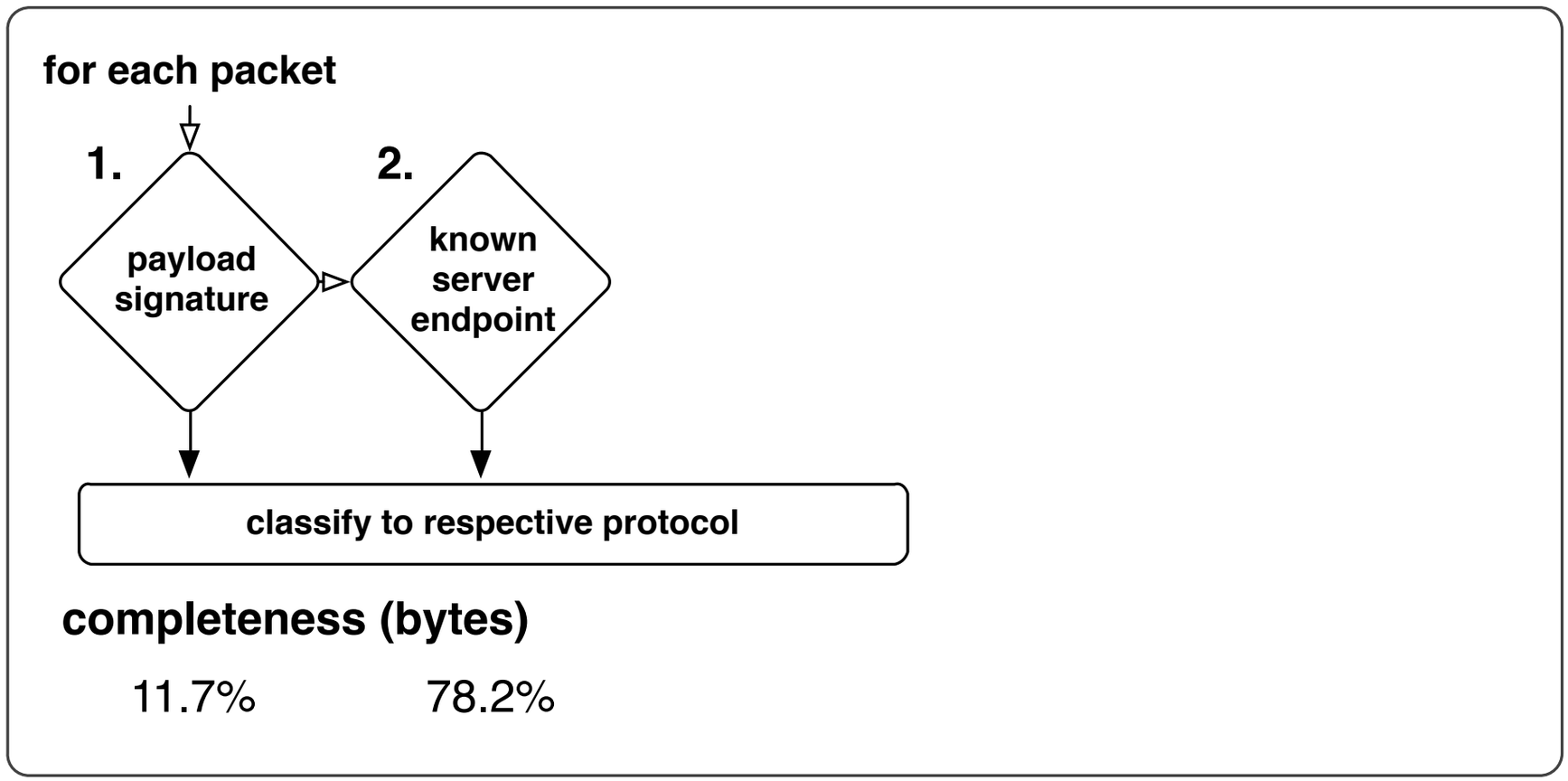
(1) Payload Signatures

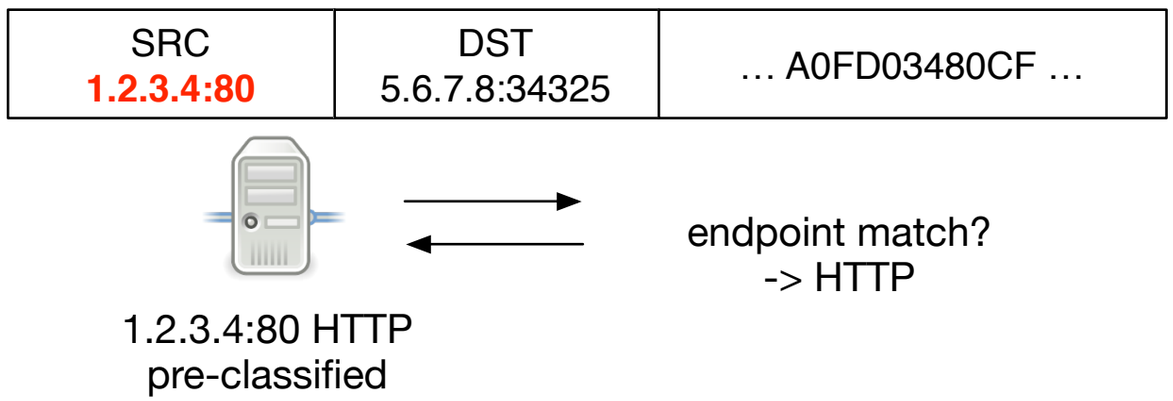


SRC 1.2.3.4:80	DST 5.6.7.8:34325	… HTTP/1.1 200 OK …

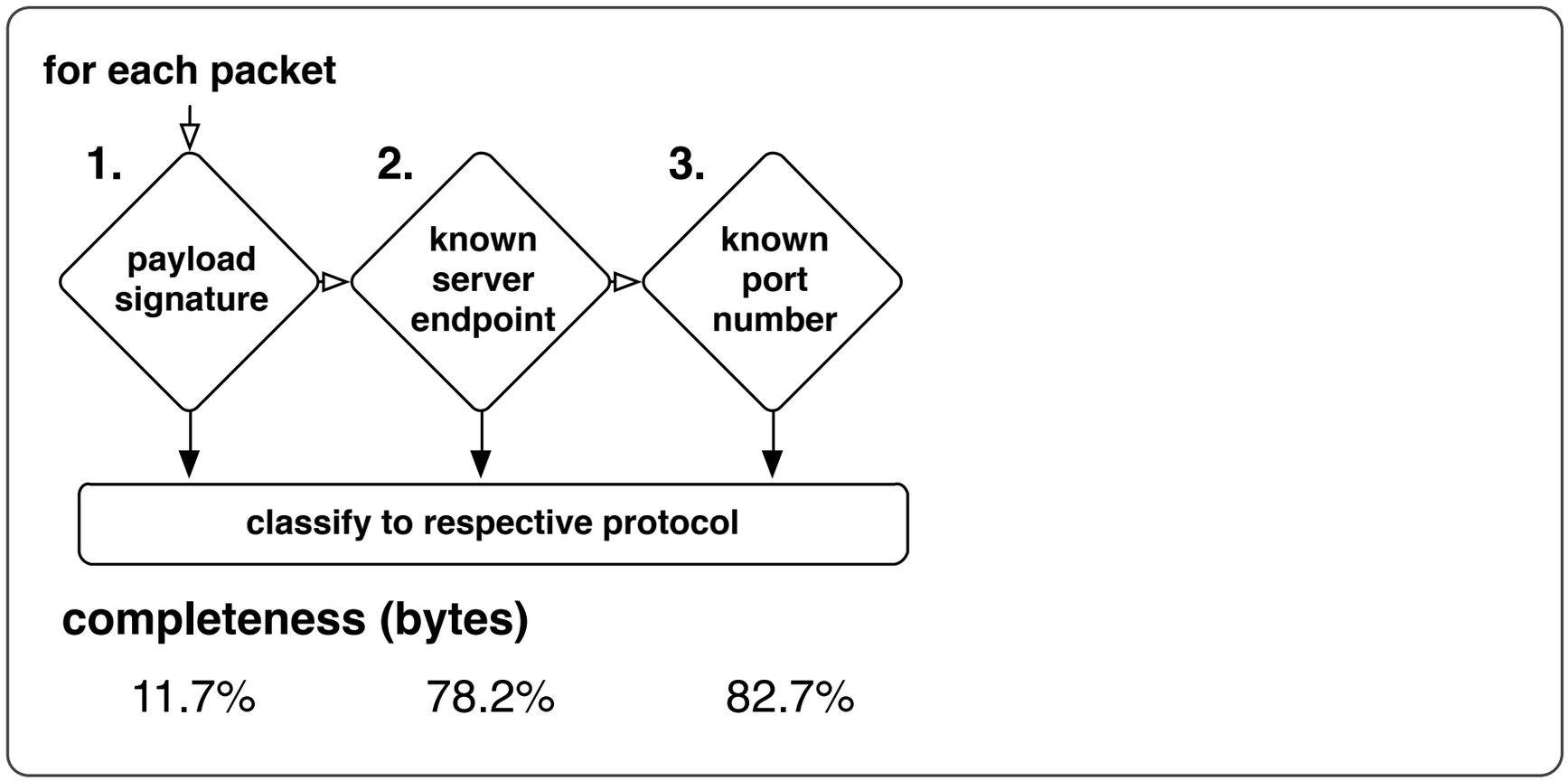
signature match?
-> HTTP

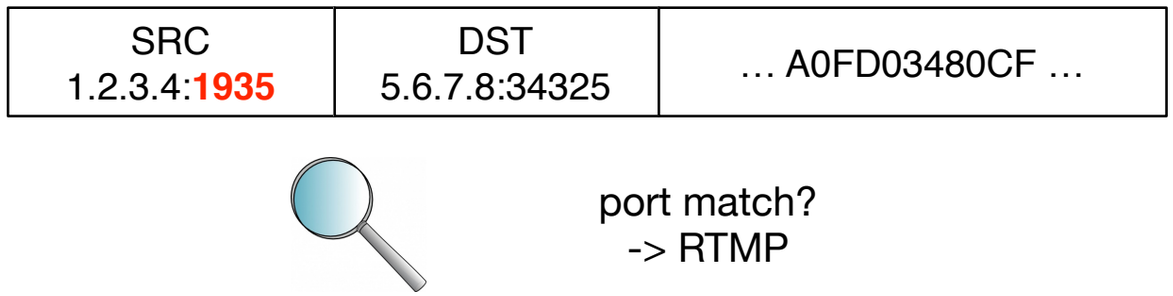
(2) Server Endpoint Matching



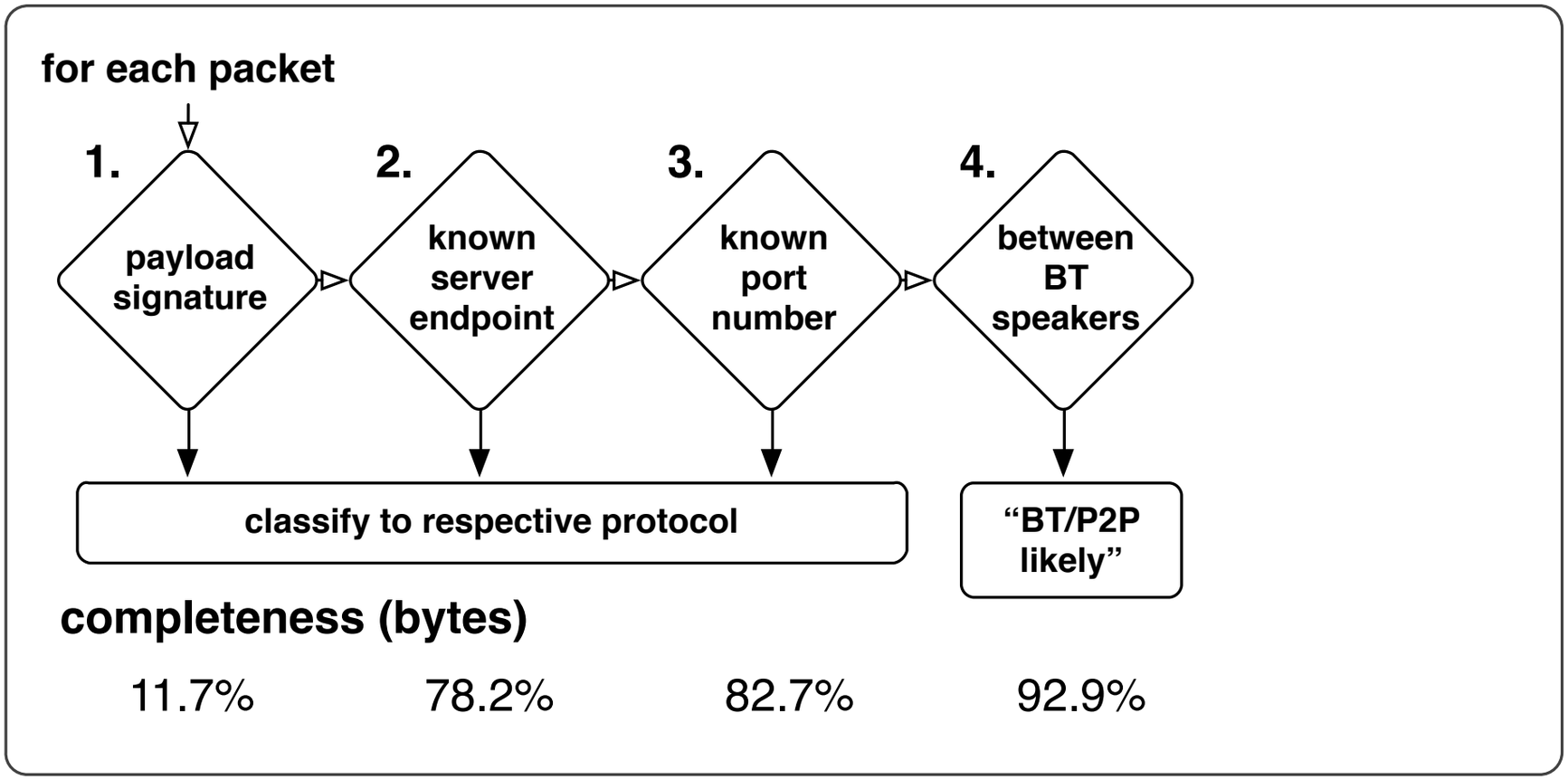


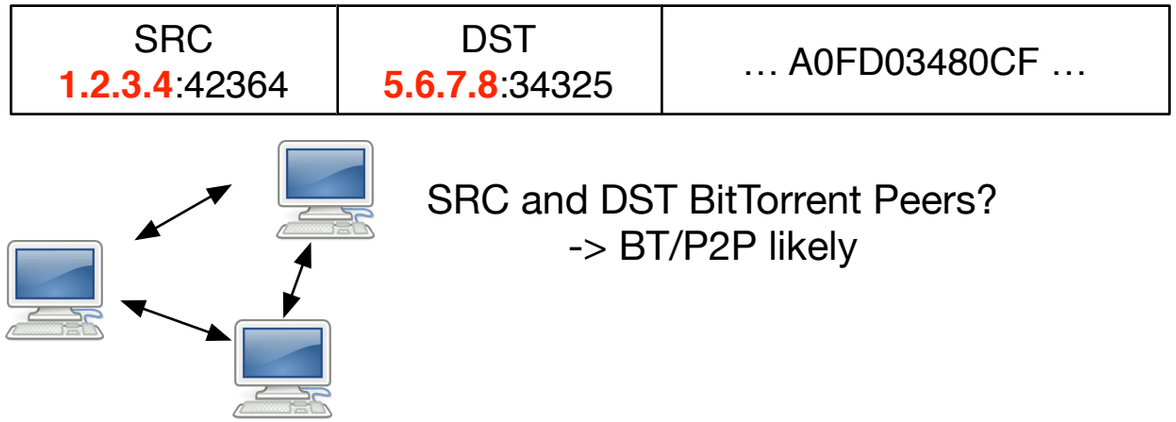
(3) Fallback: Port-Based Classification



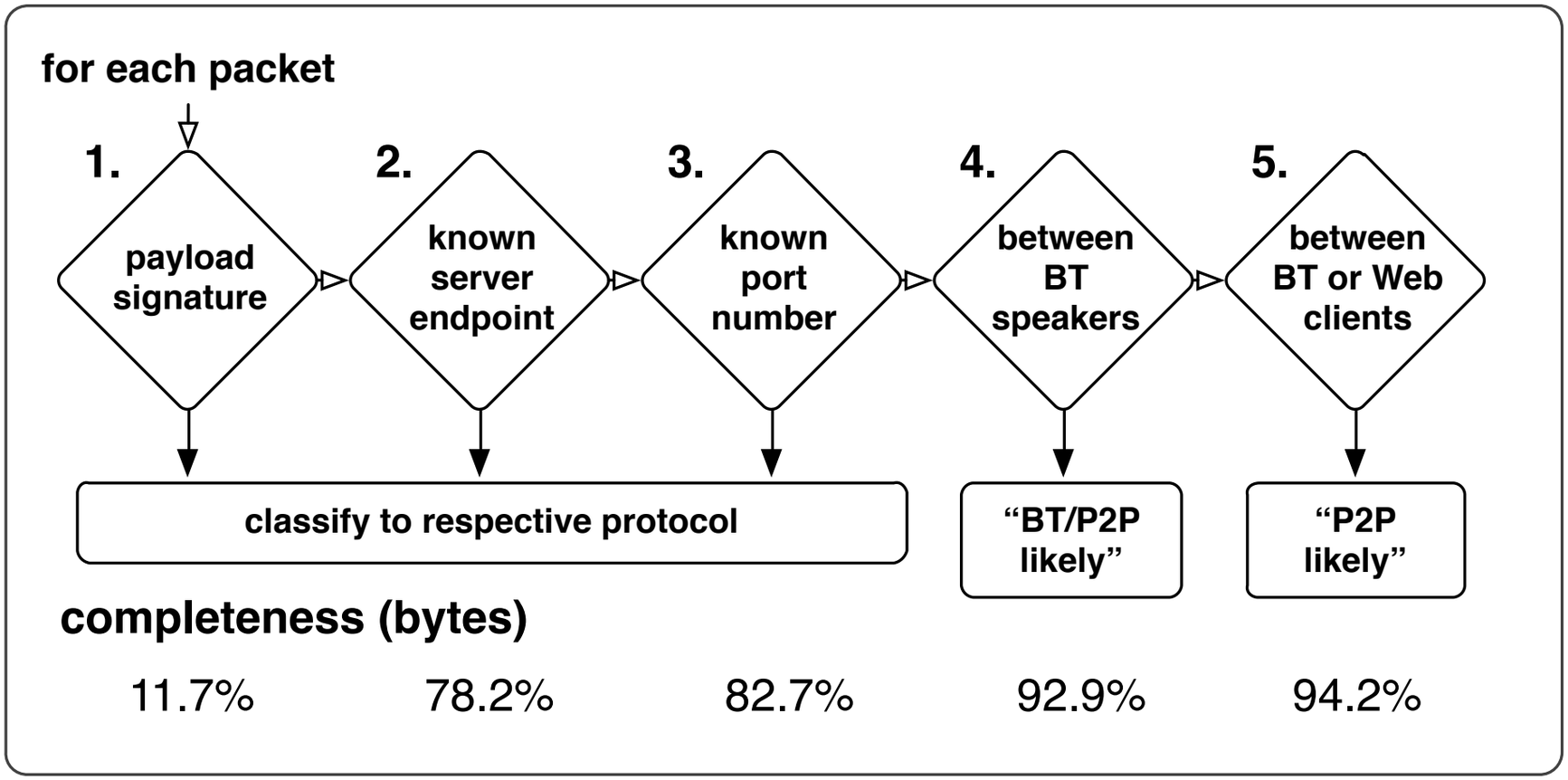


(4) Catching the TCP BitTorrent Traffic

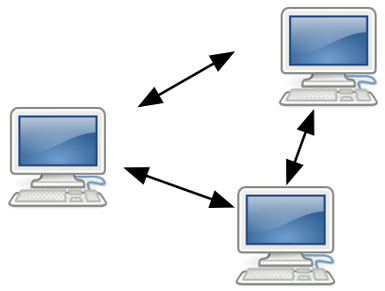




(5) Tie Breaker: Other P2P likely

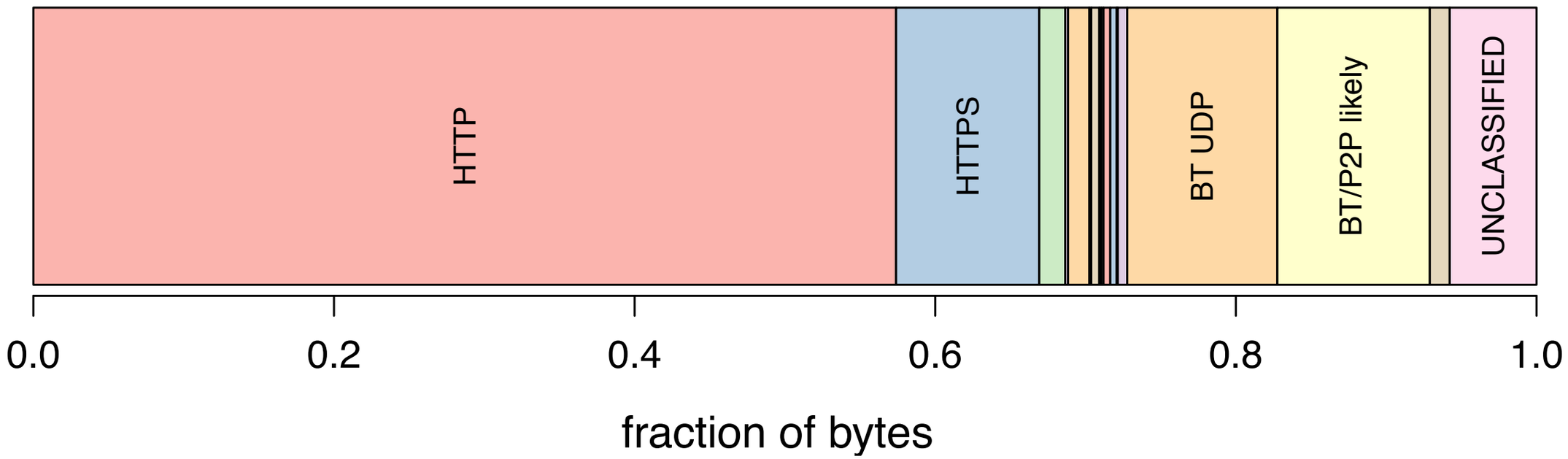


SRC	DST	… A0FD03480CF …
1.2.3.4:42364	5.6.7.8:34325	



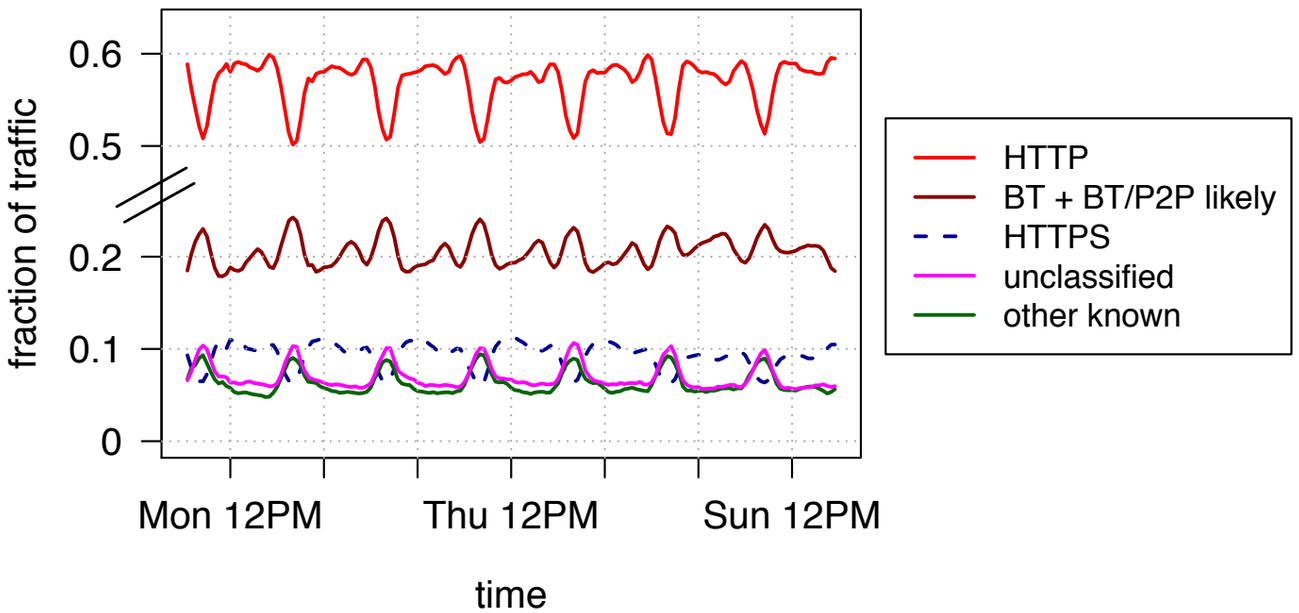
SRC and DST BitTorrent Peers
OR Web Clients?
-> P2P likely

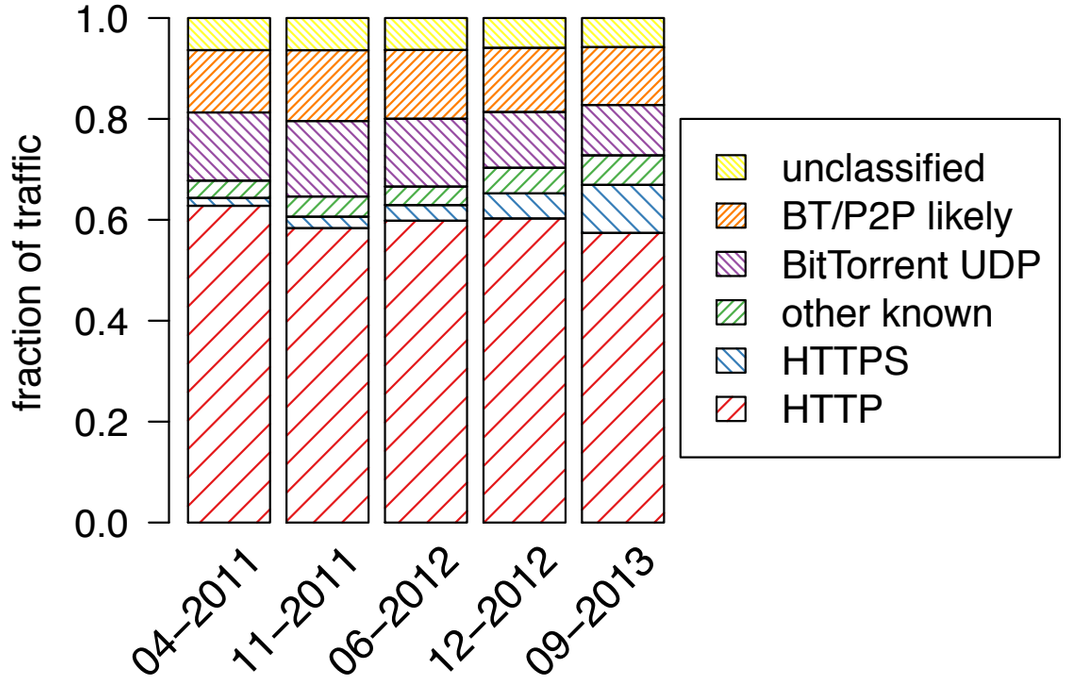
The Application Mix: Aggregate



- HTTP(S) dominates ~67%
- other applications (e.g., RTMP, mail, news) ~6%
- BitTorrent/BT/P2P likely ~22%
- unclassified ~5%

The Application Mix: Over Time





- Diurnal patterns, e.g., P2P dominates in off-hours
- Historical view shows increasing dominance of HTTP(S) and significant HTTPS increase in 2013.

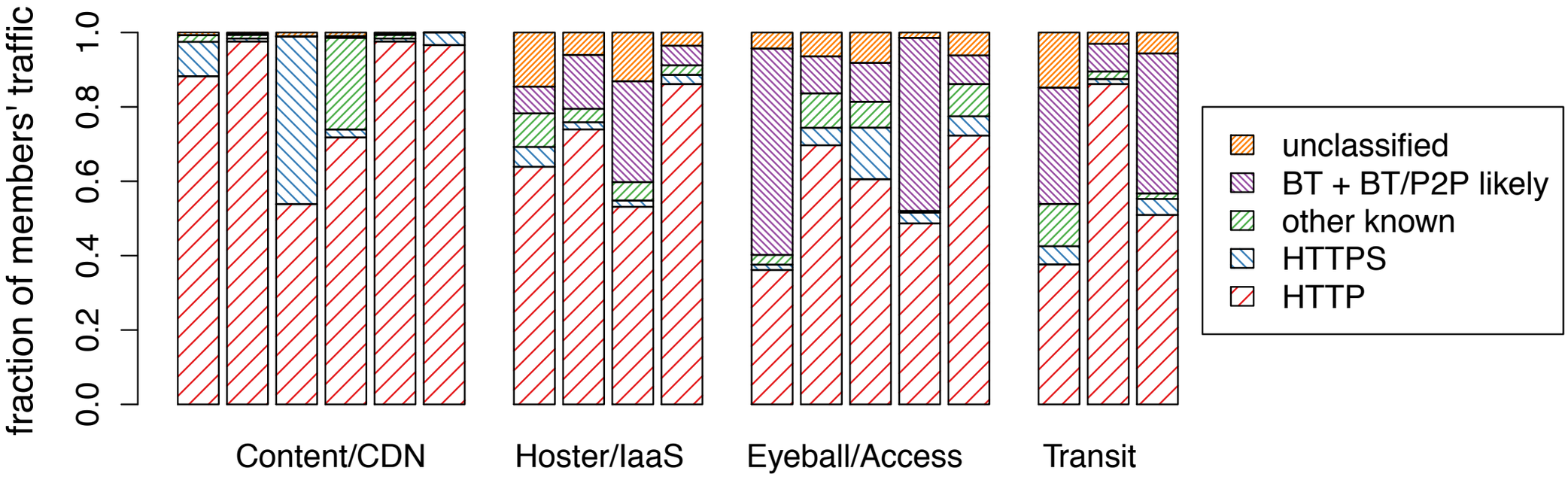
The Application Mix: Same Everywhere?

Study	Vantage Point(s)	Method	HTTP(S)	other	BT/P2P	unclassified
Labovitz, 2010	5 large ISPs	payload	52.1	24	18.3	5.5
Labovitz, 2010	110 Networks	port	52	10	1	37
Maier, 2009	Access ISP	payload	57.6	23.5	13.5	10
Gerber, 2011	Backbone ISP	payload	60	28	12	N/A
Czyz, 2014	260 Networks	port	69.2	4	<7	20
Sandvine, 2014	VA, North America	payload	~70	N/A	6	N/A
Sandvine, 2014	VA, Europe	payload	~65	N/A	15	N/A
Sandvine, 2014	VA, Asia-Pacific	payload	~60	N/A	30	N/A
Sandvine, 2014	VA, Latin America	payload	~65	N/A	9.4	N/A
this study	European IXP	various	66.9	5.9	21.4	5.8

around 55-70% HTTP(S), another 10%-20% Peer-to-Peer across different vantage points.

Is that what I can expect on any link I measure?
Should I design my network for these applications?

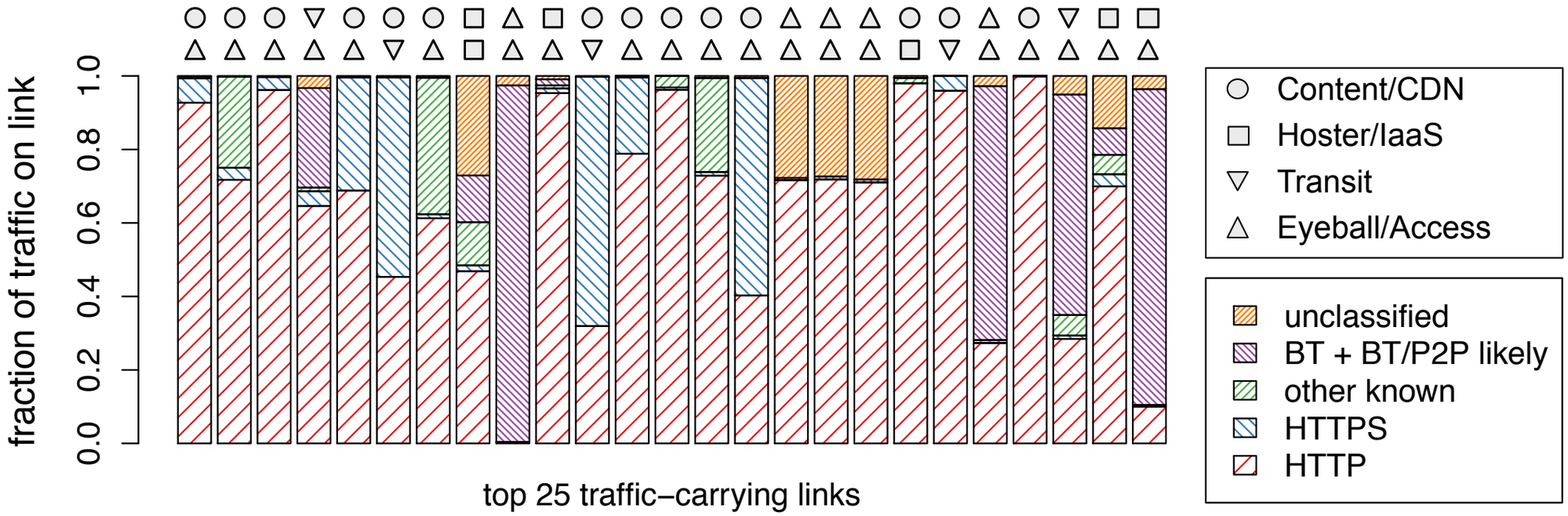
The Application Mix: Per Network Type



- Content/CDN almost 100% HTTP
- HTTPS increase driven by only a few networks
- P2P not only between Eyeballs! Hoster/IaaS too!

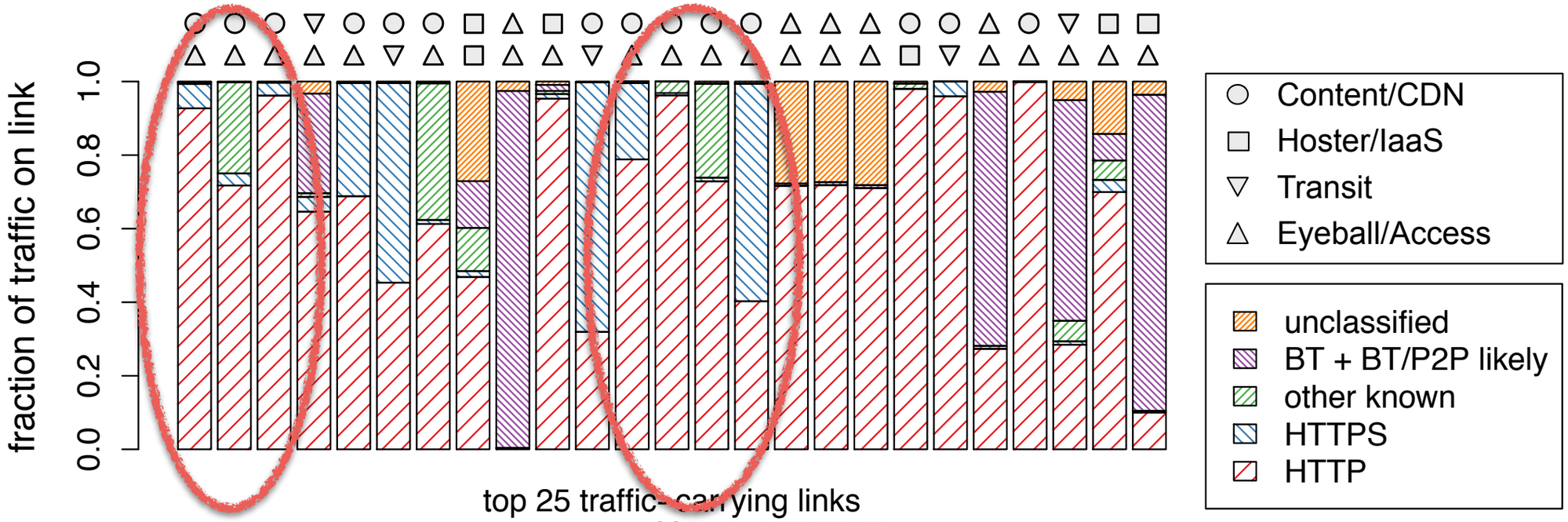
Dissecting per network shows a different appmix!

The Application Mix: Per Link



- Aggregate mix by no means representative of single link
- Many links just have one dominant protocol
- The business type of the ASes gives hints on app mix

The Application Mix: Per Link



content <> eyeball: HTTP

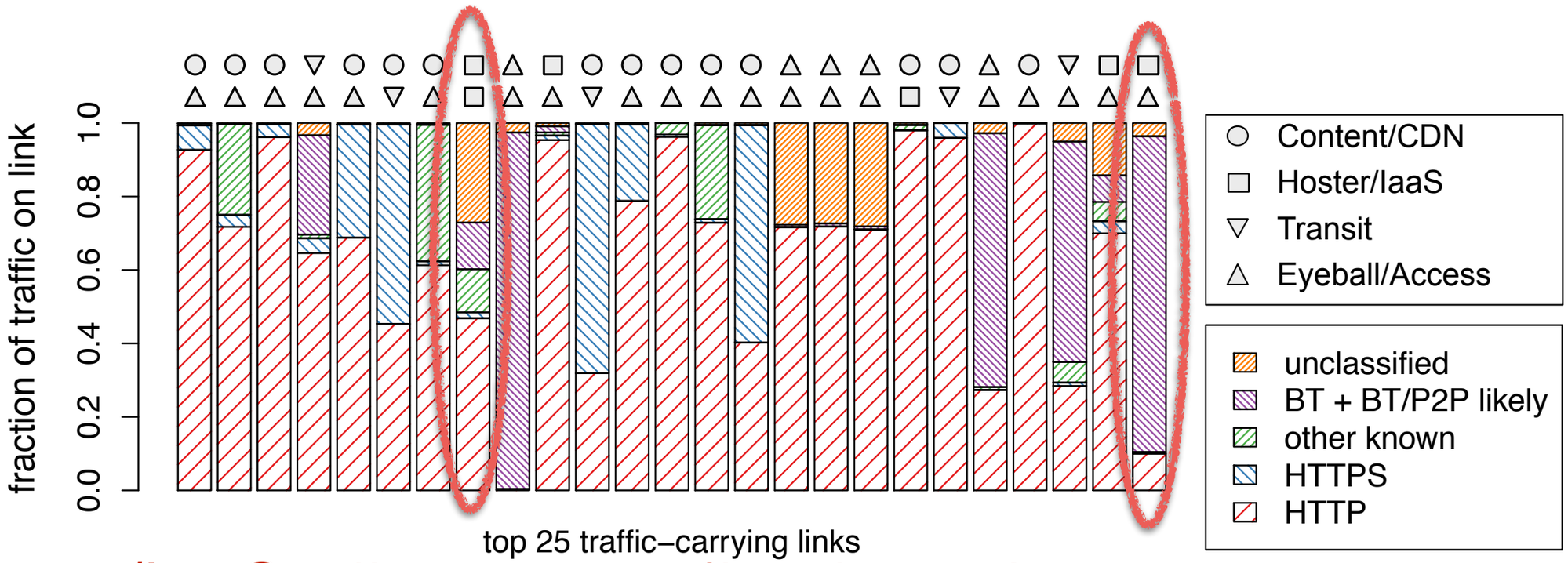
- Aggregate mix by no means representative of single link
- Many links just have one dominant protocol
- The business type of the ASes gives hints on app mix

The Application Mix: Per Link

eyeball <> eyeball: P2P

- Aggregate mix by no means representative of single link
- Many links just have one dominant protocol
- The business type of the ASes gives hints on app mix

The Application Mix: Per Link



hoster/IaaS: diverse application mix

- Aggregate mix by no means representative of single link
- Many links just have one dominant protocol
- The business type of the ASes gives hints on app mix

Summary

- By using a *stateful* approach, we can largely overcome the limitations of random packet sampling
- We can classify up to 95% of the bytes exchanged
- Our results:
  - Application mix similar to commonly reported
  - Dissecting per Network Type reveals different appmix
  - Business types of involved networks give hints