Network Metrology: Traffic Matrix, Application Identification

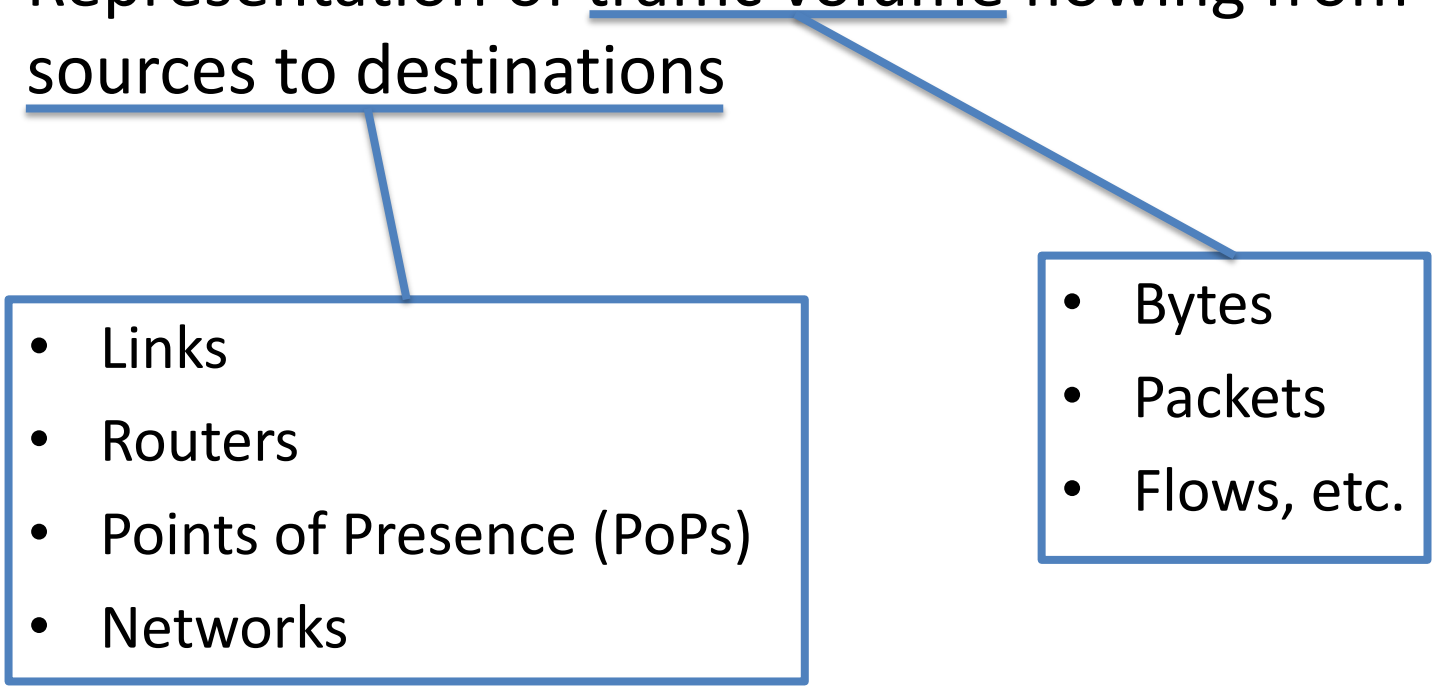
This class

- Traffic Matrix
  - Origin/Destination
  - Ingress/Egress
- Identification of application in a flow
  - Port-, content-, behavior-based methods

TRAFFIC MATRIX

Definition

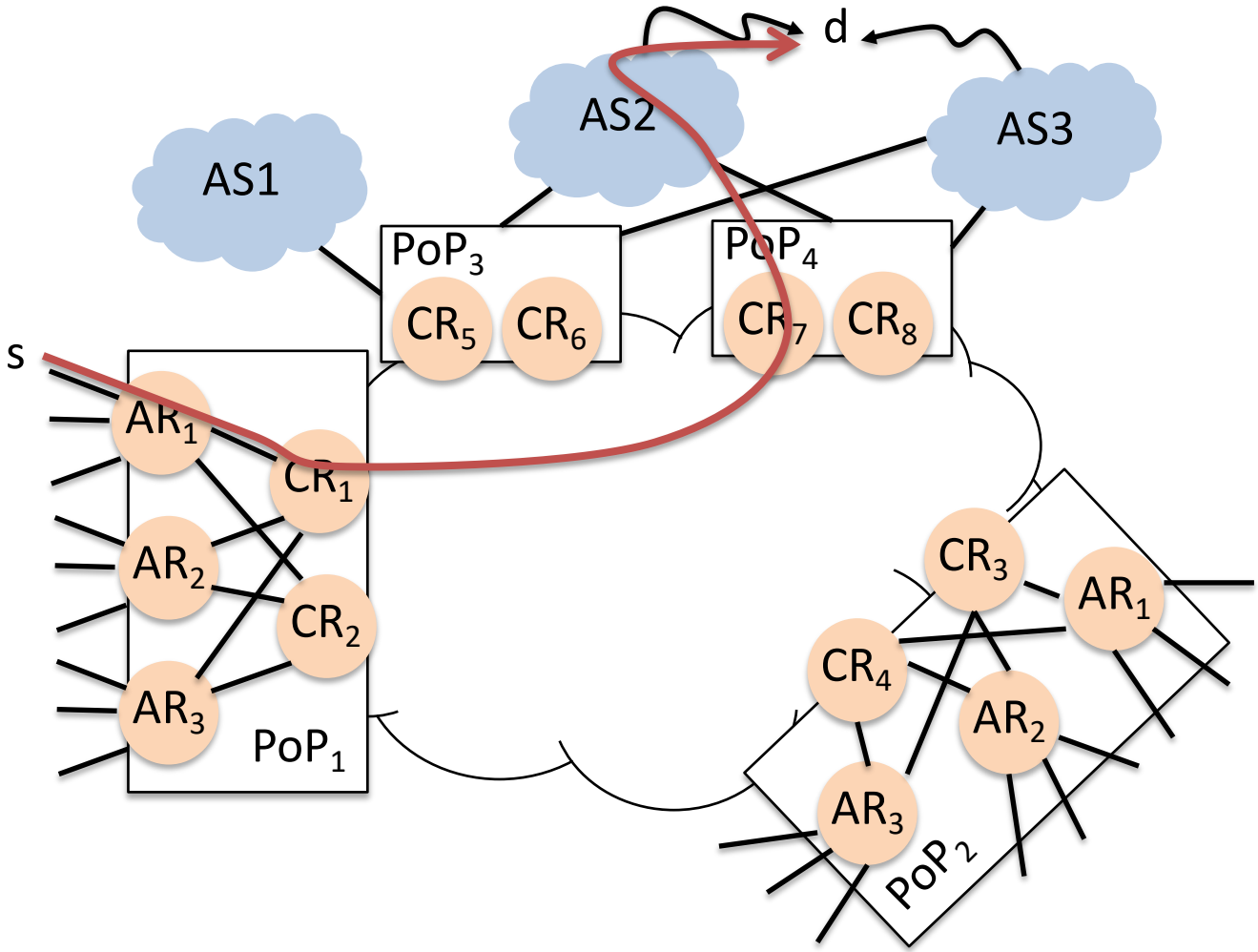
- Traffic matrix
  - Representation of traffic volume flowing from sources to destinations



Usage

- Capacity planning
- Traffic engineering (IGP and BGP)
- Billing
- Peering analysis
- Anomaly detection
- Design of new protocols

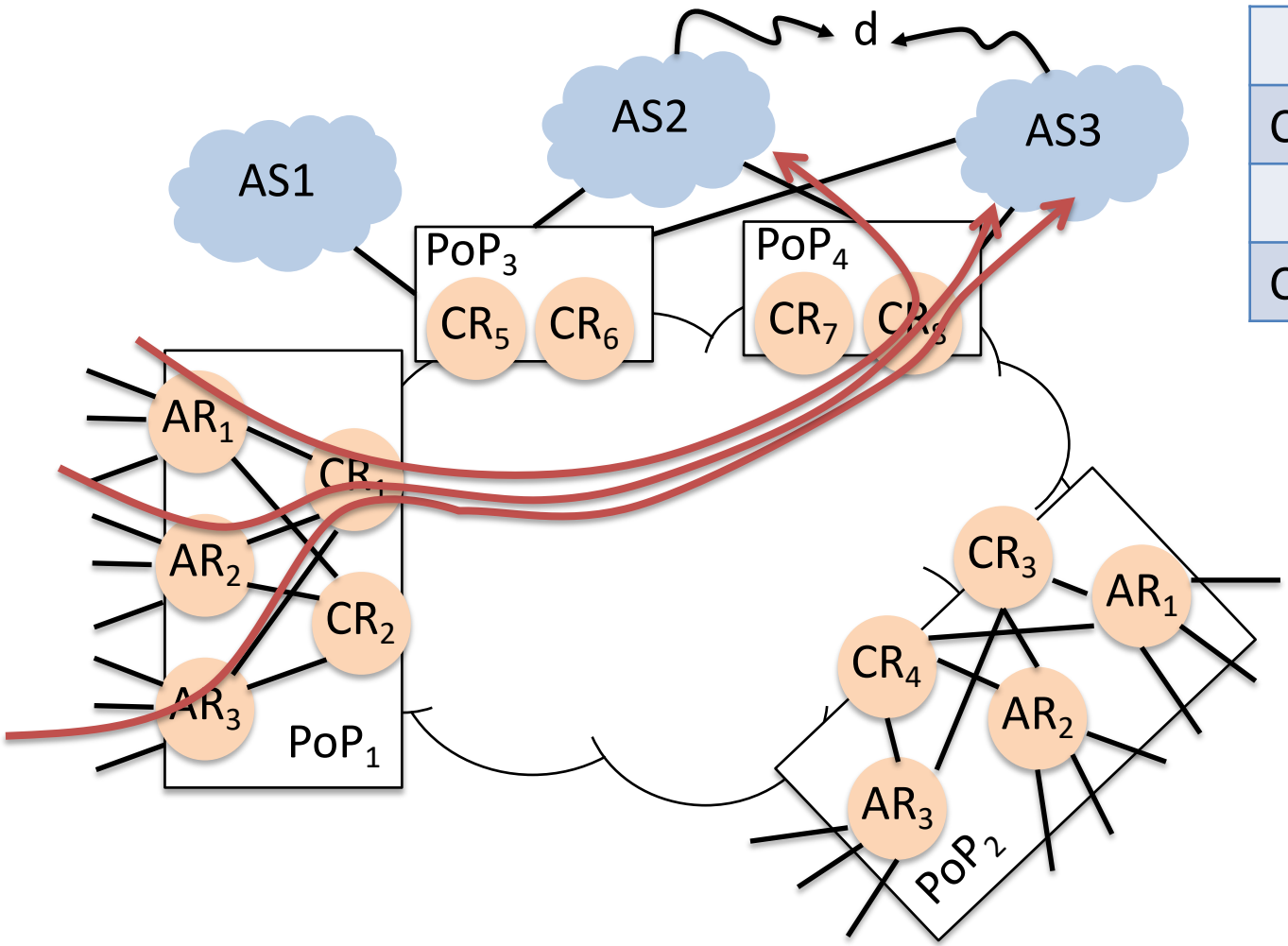
Origin to destination (OD) matrix



Challenges

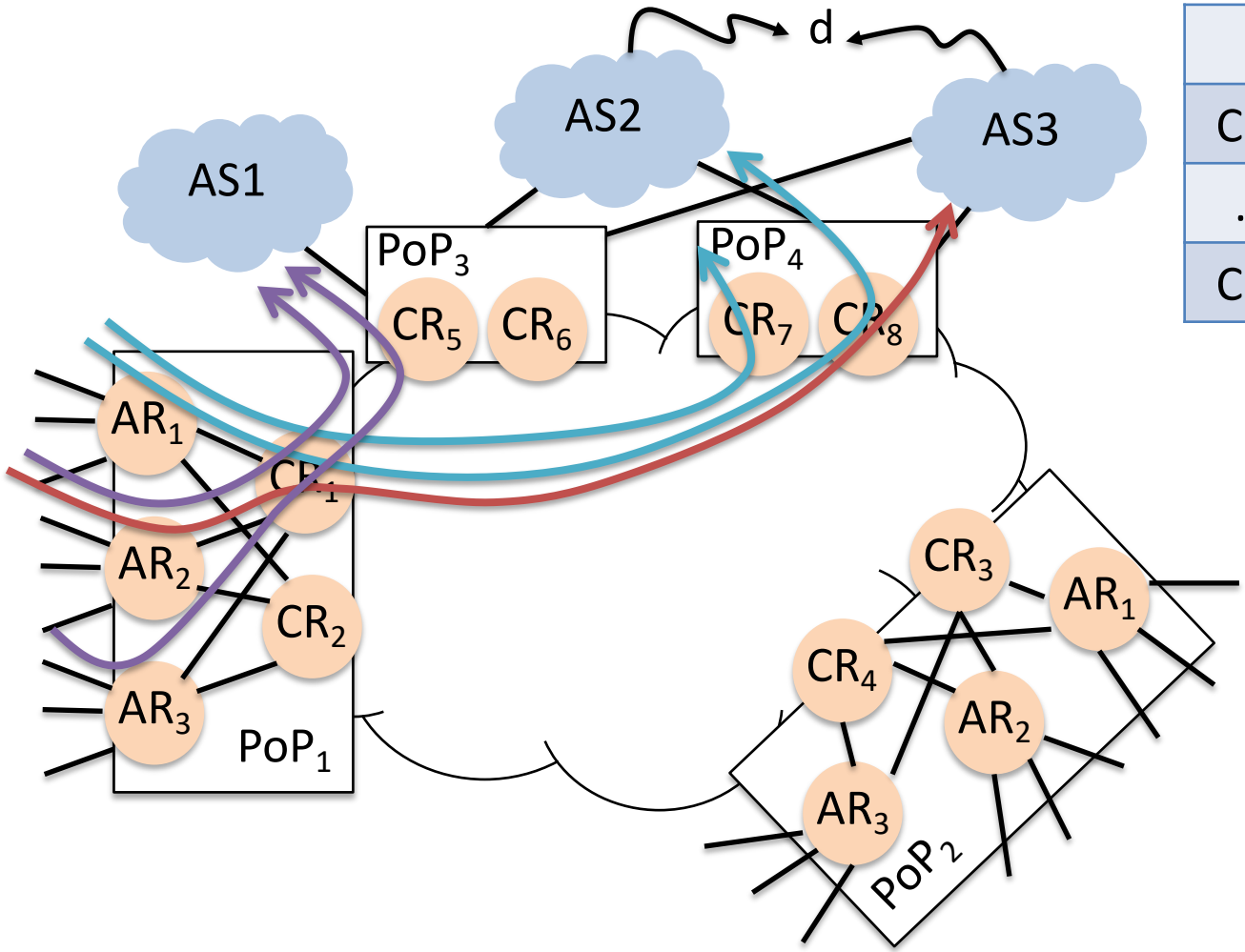
- Large matrices are hard to work with
  - E.g.: 2^{32} IPv4 addresses (worse for IPv6)
  - Computational/storage problems
  - Very sparse matrix
- Aggregation into blocks of IP addresses:
  - IP prefix
- No single entity sees the Internet OD matrix
  - Cannot be fully measured
  - Instead, operators work with ingress to egress (IE) matrix

Ingress router to egress router matrix



	CR_1	...	CR_8
CR_1			
...			
CR_8			

Ingress router to neighbor AS matrix



	AS_1	AS_2	AS_3
CR_1			
...			
CR_8			

Example: Internet2

	Destination												
Source	1	2	3	4	5	6	7	8	9	10	11	12	Row sum
1	0.07	0.07	0.43	0.00	0.06	0.12	0.06	0.00	0.05	0.00	0.00	0.25	1.12
2	0.00	4.09	6.42	0.06	7.07	4.42	1.59	0.02	3.24	0.03	0.16	11.09	38.18
3	0.00	4.70	25.48	4.11	13.99	11.53	3.31	87.27	5.22	0.01	0.08	7.70	163.38
4	0.00	1.93	10.25	1.68	5.63	6.11	2.59	0.01	4.11	2.60	0.04	5.92	40.88
5	0.00	4.76	0.25	0.01	24.06	0.04	0.01	0.02	1.24	0.02	0.03	18.05	48.49
6	0.00	2.87	23.73	1.55	13.53	4.78	2.89	0.01	9.45	0.08	0.50	7.64	67.02
7	0.00	0.67	4.79	1.92	3.50	2.24	1.25	0.00	0.93	0.02	0.03	3.31	18.67
8	0.00	4.18	2.58	5.80	26.35	0.17	0.16	1.41	10.88	2.11	3.64	16.67	73.97
9	0.00	8.61	12.34	5.71	18.21	11.05	3.84	0.41	36.36	0.02	0.52	17.31	114.37
10	0.00	0.18	0.04	1.71	1.69	0.00	0.06	5.61	0.96	1.82	8.44	0.36	20.86
11	0.00	3.47	3.28	0.54	8.60	0.13	0.93	3.92	1.77	0.81	0.61	2.32	26.38
12	0.00	18.20	16.04	0.83	34.03	11.18	5.64	0.09	25.57	0.08	0.80	47.02	159.47
Column sum	0.07	53.74	105.61	23.94	156.73	51.76	22.34	98.77	99.77	7.59	14.84	137.65	772.80

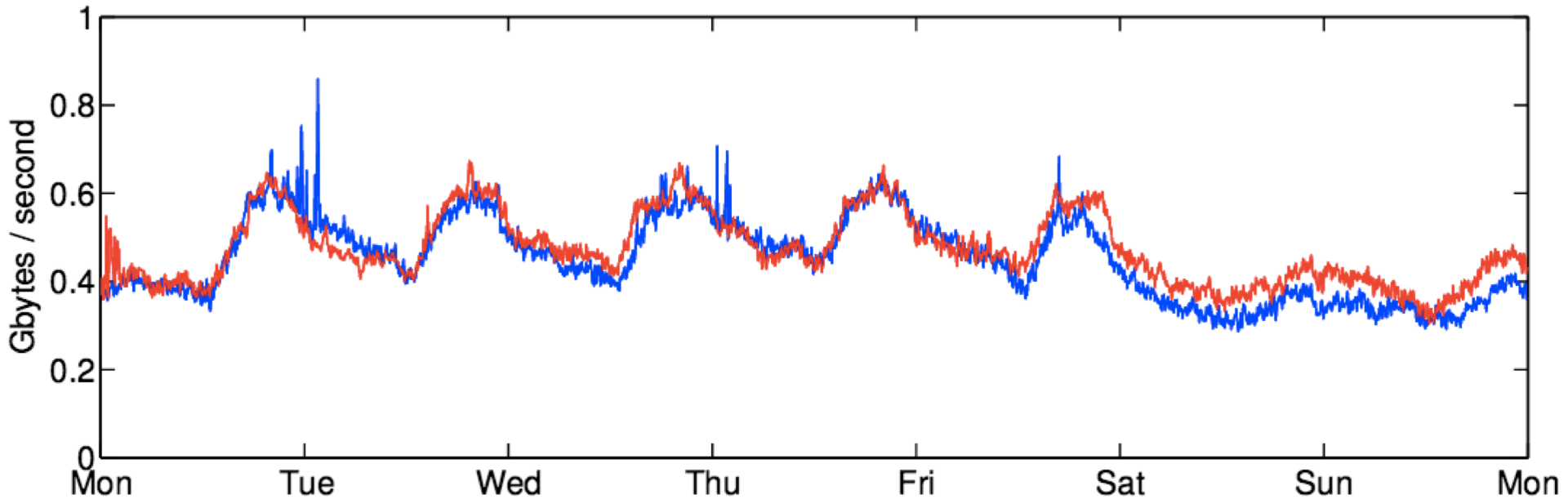
5-min PoP-PoP traffic matrix for Internet2 for April 2004 (in Mbps)

From: Tune, Roughan, “Internet Traffic Matrices: A Primer”

Question: How to bin traffic over time?

- Traffic properties depend on the time scale
  - Highly variable at time scale of seconds
  - Strong periodicity when considering days, weeks
  - Long-term growth over years

Example: Internet2's traffic in two weeks in March 2004



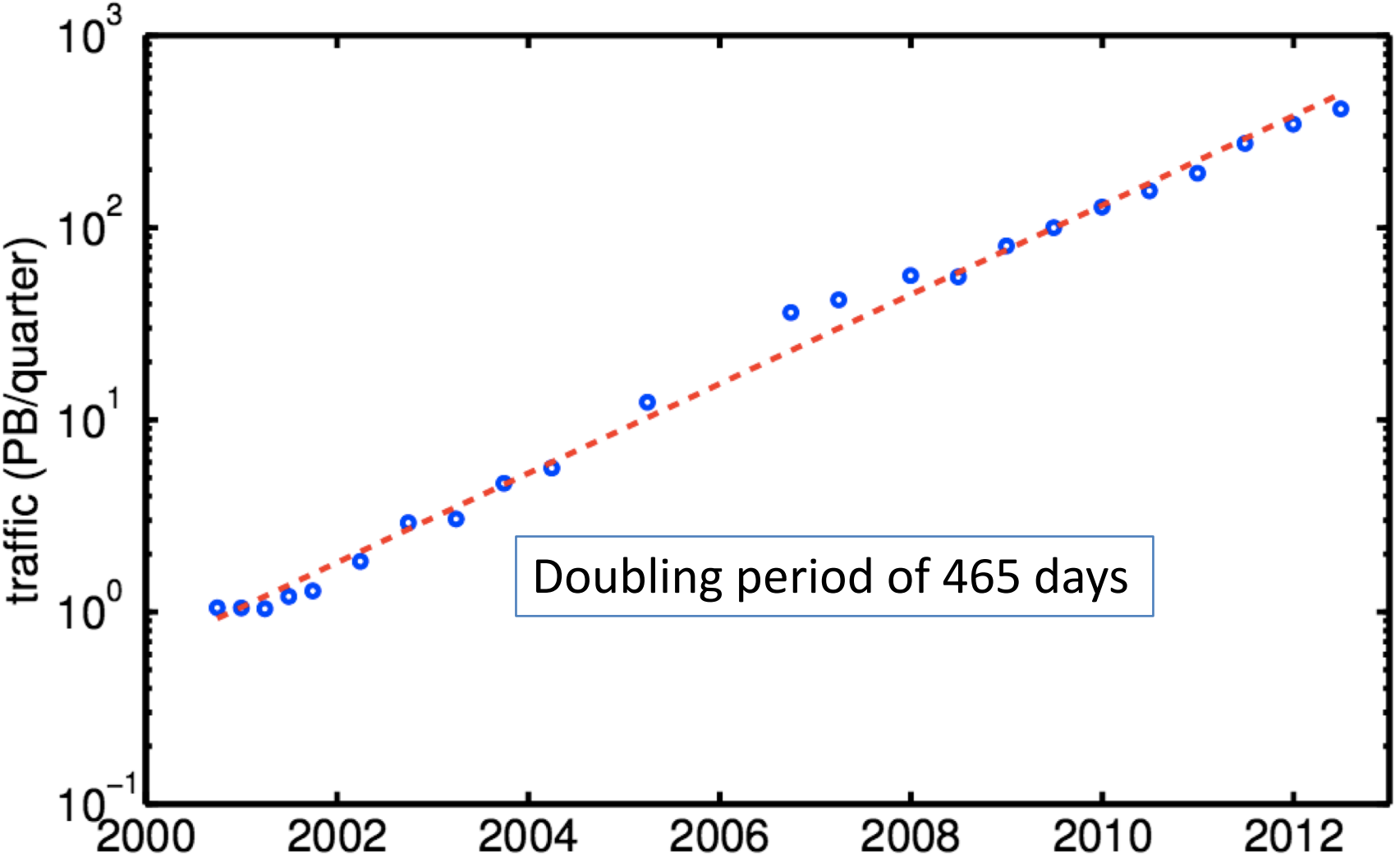
Total traffic in 5-min time bins

From: Tune, Roughan, "Internet Traffic Matrices: A Primer"

Timescales of hours/days

- Signal: Daily variation, weekday vs. weekend
  - Most traffic is generated by humans
  - Time of day effects depend on how a network's population is distributed geographically

Example: Total Australian Internet traffic



From: Tune, Roughan, “Internet Traffic Matrices: A Primer”

Timescales of years

- Exponential growth
  - Increasing number of users connected
  - More devices per user
  - Applications increasing networking demands

How to bin traffic over time?

- Best binning depends on task
  - Examples
    - Anomaly detection: minutes (maybe even seconds)
    - Capacity planning: busy hour
- In practice, binning depends on:
  - Measurement tools/apparatus available
  - Storage facilities available.

Question: How to measure the traffic matrix of a network?

- Packet capture
  - Gives the most detailed view of traffic
  - But, expensive and high collection overhead
- Flow capture
  - Enough to build traffic matrix
  - Lower collection overhead (in particular with sampling)
- Interface counts
  - Cannot directly measure traffic matrix, must estimate
  - Lowest overhead, widely available

Measuring an IE traffic matrix with flow capture

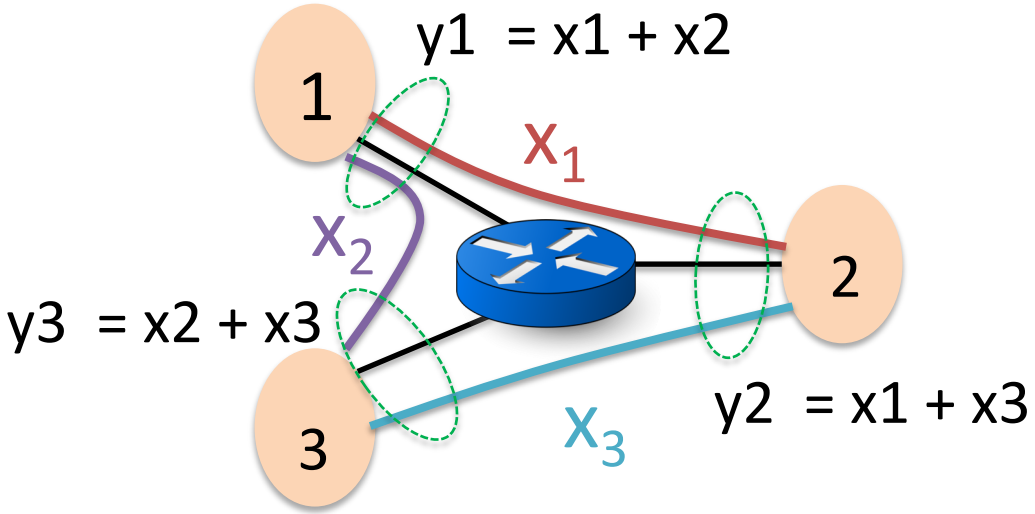
- Basic method
  - Deploy flow capture in border routers
  - Ingress: router(s) capturing flows
  - Egress: routing lookup to identify next hop(s)

- Considerations
  - Sampling leads to errors/noise in data
    - Should estimate sampling errors for each setting
  - Same flow may be sampled in multiple points
    - Should measure only at ingress links

Estimating an IE traffic matrix from interface counts

- Basic method
  - Monitor traffic volumes per interface (e.g., with SNMP)
  - Monitor topology (e.g., from routing messages)
  - Estimate traffic matrix using system of equations
- Assumptions
  - Routing remains stable
  - Traffic matrix is stationary (statistical properties remain stable during measurement interval)

Estimating an IE traffic matrix from interface counts



Unidirectional traffic in a small network

	1	2	3
1	?	?	?
2	?	?	?
3	?	?	?

Desired Traffic Matrix!

Solution: Stack the equations! 3 equations with 3 variables!

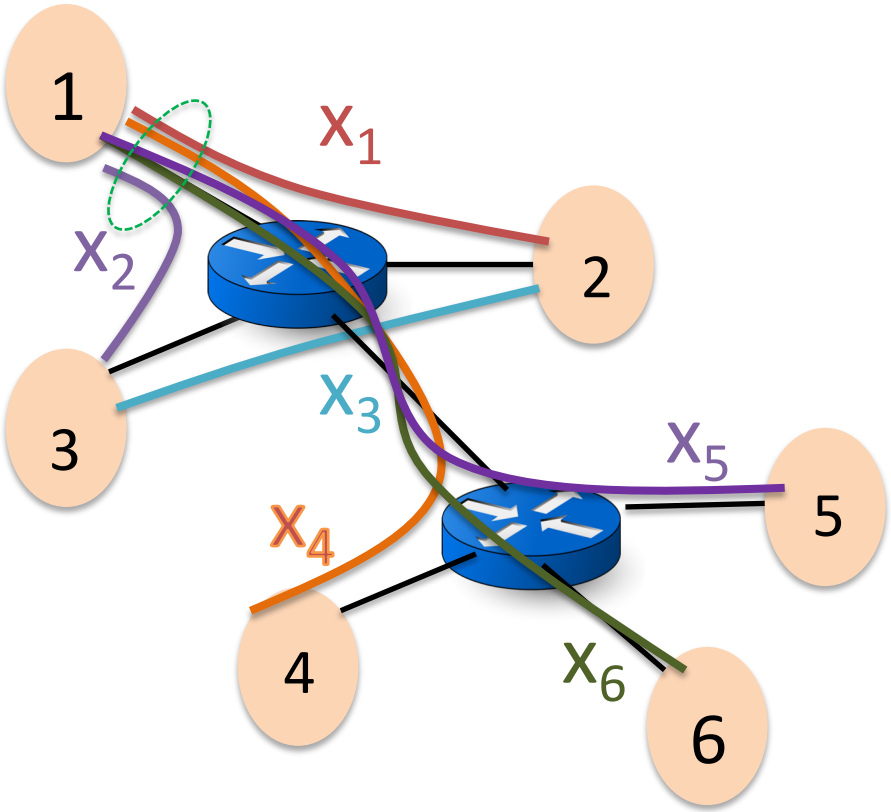
y1 = x1 + x2 → x2 = y1 – x1

Y2 = x1 + x3 → x3 = y2 – x1

Y3 = x2 + x3 → y3 = y1 + y2 -2x1 → X1 = (y1 + y2 –y3)/2

Estimating Traffic Matrix is hard in practice

- Number of links much smaller than total IE pairs
  - Problem is highly under-constrained

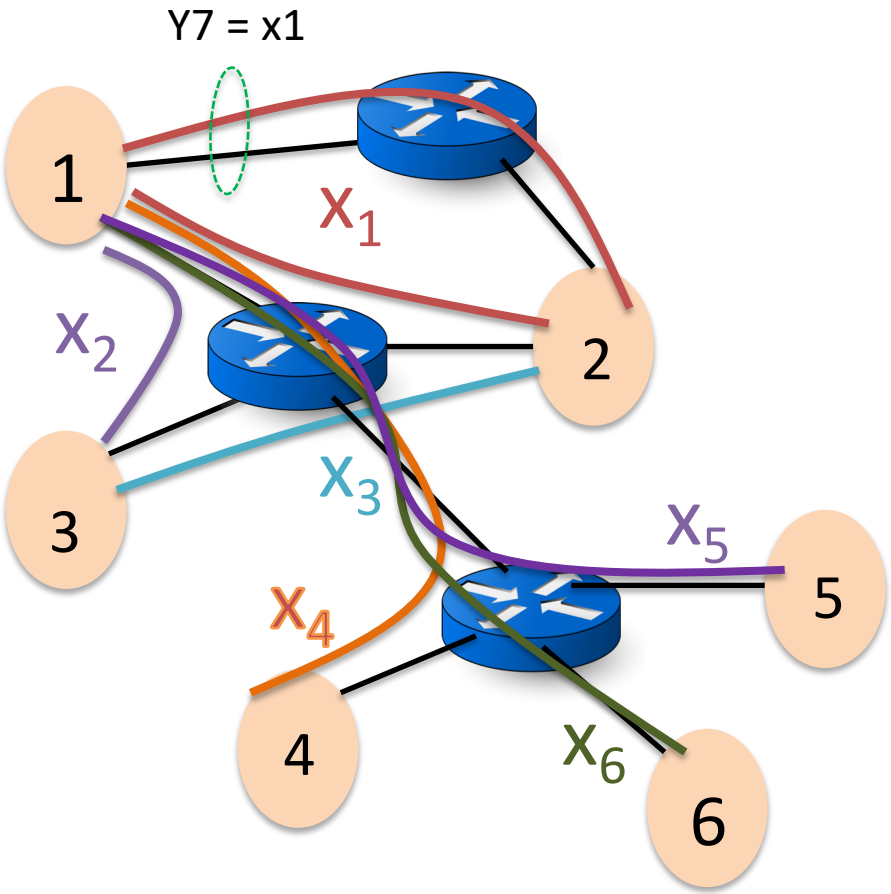


y1 = x1 + x2 + x4 + x5 + x6
Y2 = x1 + x3
Y3 = x2 + x3
Y4 = …
Y5 = …
Y6 = …

15 variables and 6 equations!

Need More Constraints!

- Change IGP weights to get new constraints

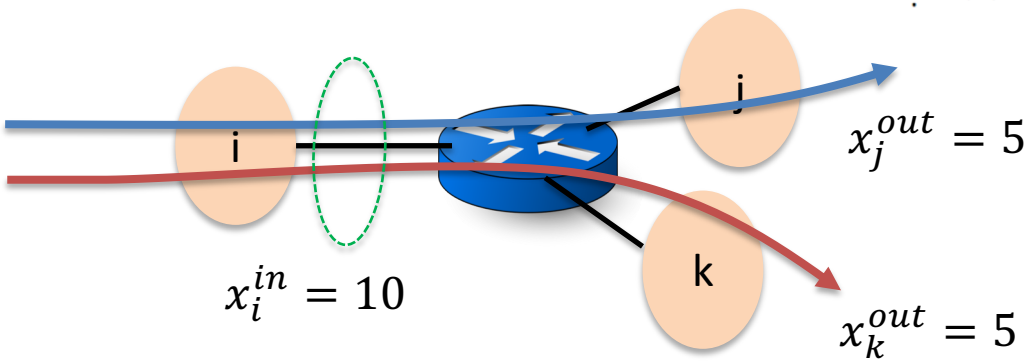


y1 = x2 + x4 + x5 + x6
Y2 = x3
Y3 = x2 + x3
Y4 = …
Y5 = …
Y6 = …
Y7= x1

Estimating traffic matrix is hard in practice

- Need more constraints
  - Flow capture at a single point gives one row of the matrix
  - Gravity model
    - Traffic from ingress *i* to egress *j* is proportional to total traffic entering in *i* and existing in *j*
    - Assumes independence between *i* and *j*
    - A number generalizations of the gravity model exist

$$T(n_i, n_j) = T \frac{T^{\text{in}}(n_i)}{\sum_k T^{\text{in}}(n_k)} \frac{T^{\text{out}}(n_j)}{\sum_k T^{\text{out}}(n_k)} = T p^{\text{in}}(n_i) p^{\text{out}}(n_j) \quad (1)$$



APPLICATION IDENTIFICATION

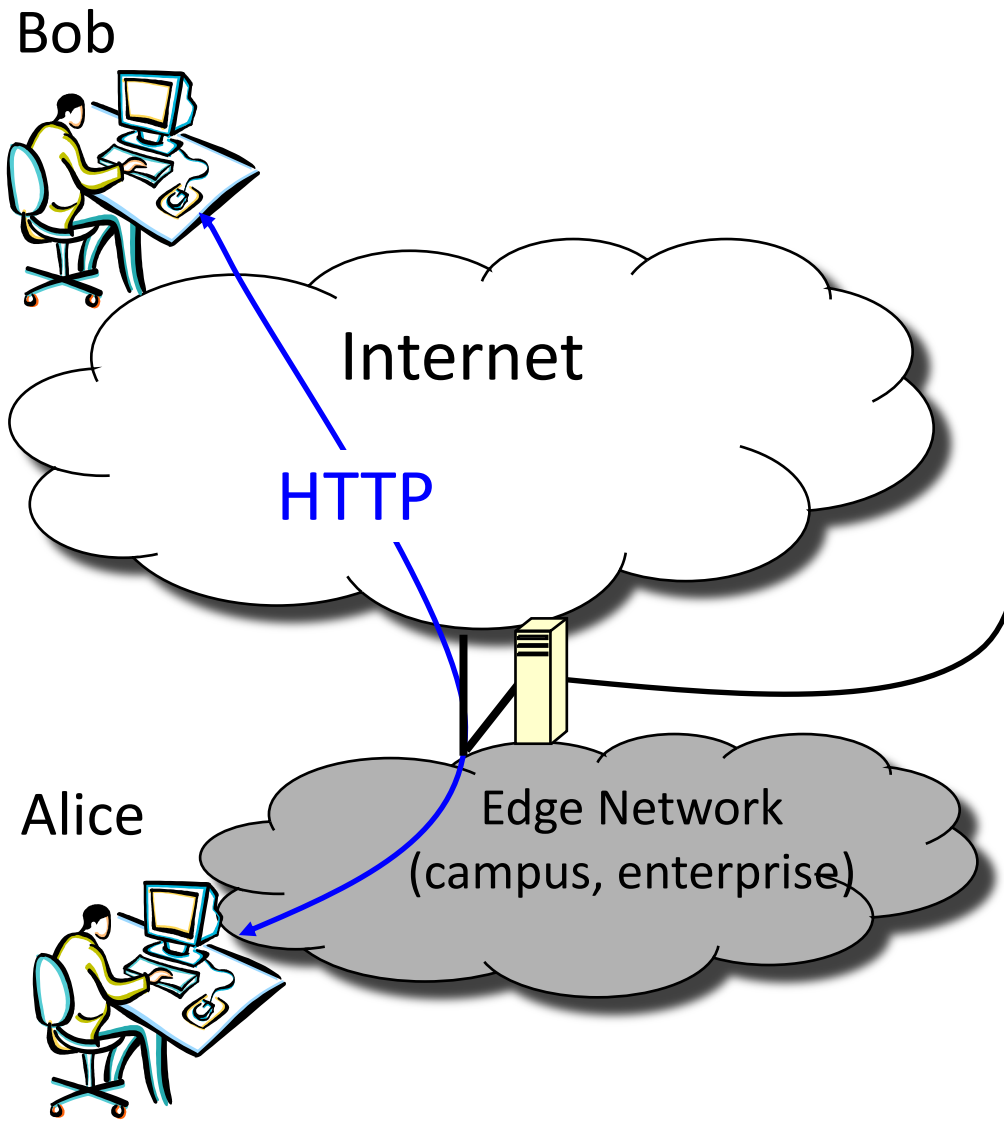
Why identify applications?

- Traffic engineering
- Billing
- Network planning
- Security

Application Identification Methods

- Port-based Identification
- Content-based Identification
- Behavior-based identification

Port-based identification



IP src: A	IP dst: B
Port src: X	Port dst: 80

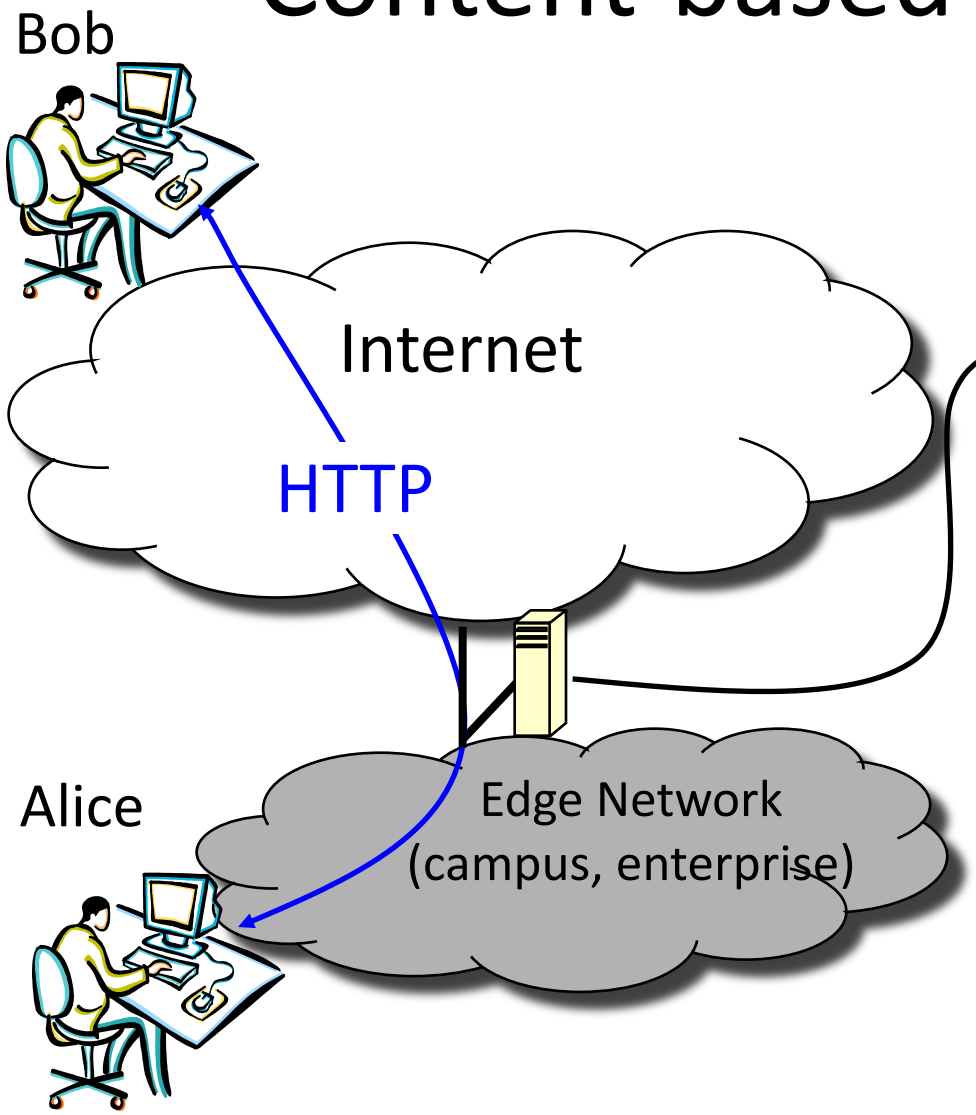
IANA Mapping

25	SMTP
80	HTTP
110	POP3

Pros and cons of port-based identification

- Simple and fast
- Used in firewalls
- Becoming more and more inaccurate
  - Non-standard ports
  - Masquerade traffic

Content-based identification



IP src: A	IP dst: B
Port src: X	Port dst: 80
GET / HTTP/1.1 Host: www.bob.com…	

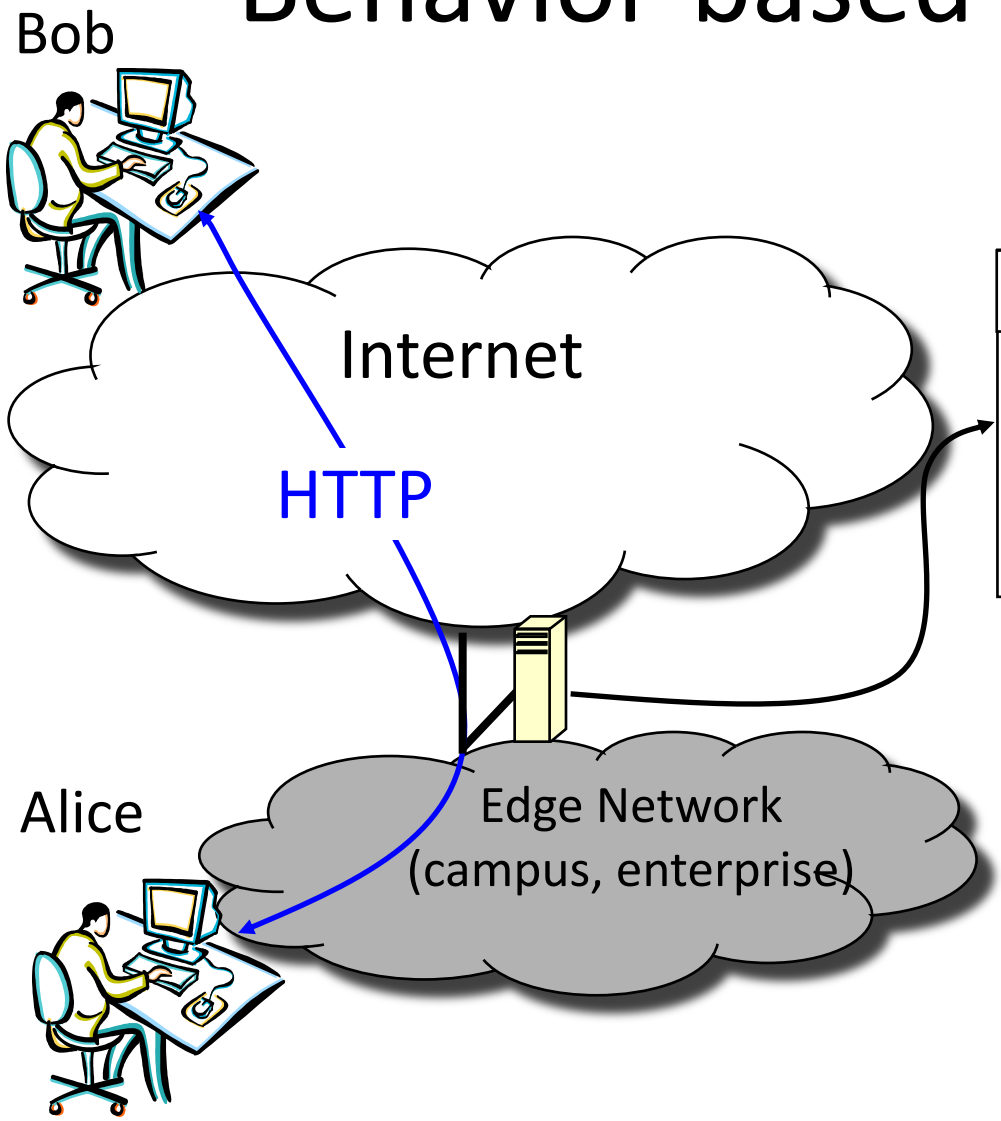
List of signatures

220 * SMTP *	SMTP
GET * HTTP/1.?	HTTP

Pros and cons of content-based identification

- Widely used
  - Intrusion detection systems, e.g., SNORT, BRO
  - Commercial classifiers, e.g., Sandvine
- Very accurate, but…
  - Privacy issues
  - Finding signatures is hard
  - Computationally intensive
  - Harder with encryption

Behavior-based identification



Connection A-B
Packets: 25 Duration: 400ms Volume: 18918 Bytes

Application model

20 < Packets < 40 Duration < 500ms Volume < 50kB	HTTP

Pros and cons of behavior-based identification

- Very rich literature
  - Identification of classes of applications
  - Identification of protocols
  - Identification of applications on a host
- Not very practical yet
  - Not as accurate as content analysis
  - Need to re-train models for new applications
  - Often provide coarse-grained application classes

How is application identification done in practice?

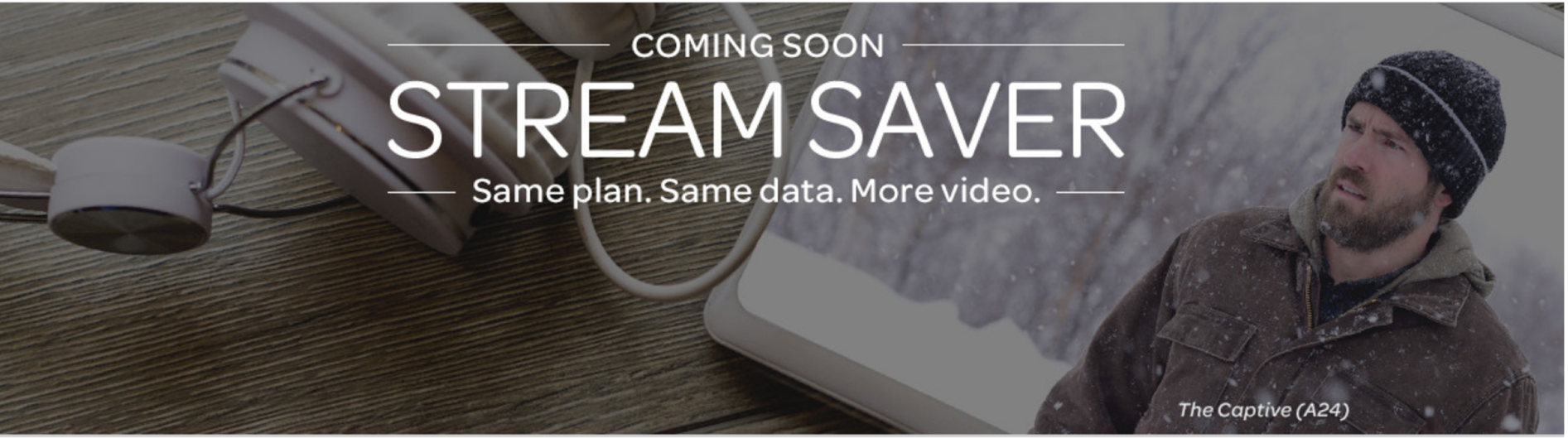
Classifiers Unclassified
An Efficient Approach to Revealing IP Traffic Classification Rules

Fangfan Li, Arash Molavi Kakhki, David Choffnes,
Phillipa Gill, Alan Mislove

Northeastern

UMASS AMHERST

Traffic Differentiation



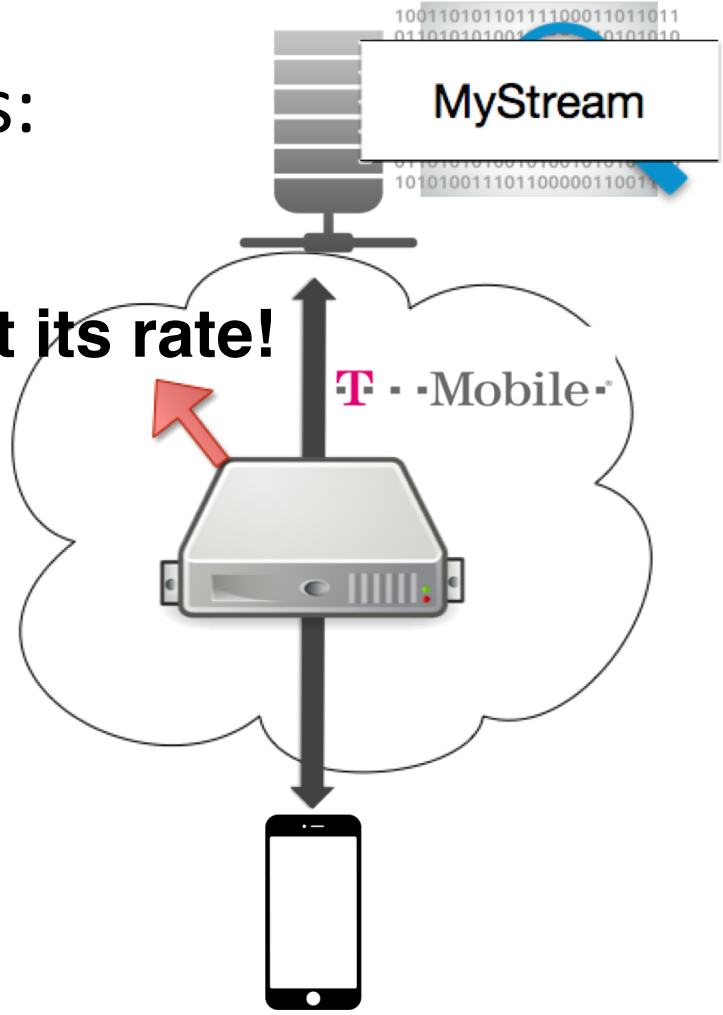
Do more with your data

Stream more video on the go with your same data package at a quality similar to DVD (about 480p). Extend your data so you can surf more, play harder, and enjoy more of what you love on your smartphone or tablet.* Best of all, you won't have to pay more for Stream Saver.

*Our most popular plans with data will include the Stream Saver feature which allows you to save data on content it recognizes as video by streaming higher definition video at Standard Definition quality on compatible devices (unless the video provider has opted out). AT&T will activate the feature for you. Check your account online to see if the feature is active. Once active, you can turn it off or back on at any time

Differentiation through Classifiers

- Previous work [IMC 15] shows:
  - Deep packet inspection
  - Uses text “matching rules”
  - **Precise rules unknown**
  - **Classification might be wrong**



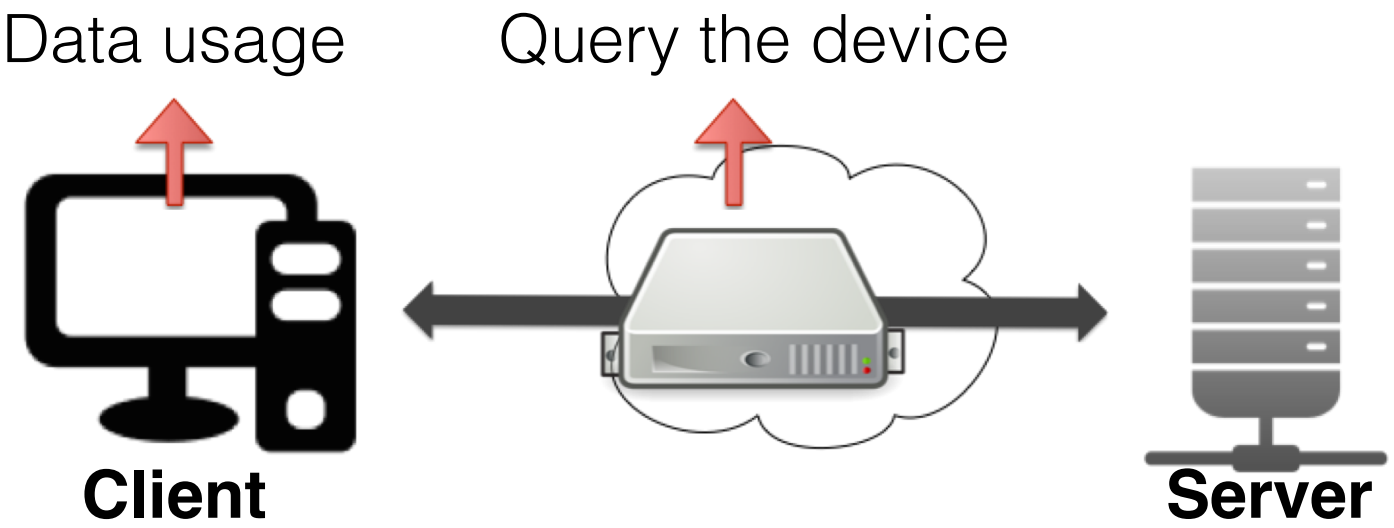
Roadmap

- Goal: Understand classifiers deployed in today's networks
- Methodology to identify the matching rules
- Validation in testbed with carrier-grade device
- Case study classifiers "in the wild"

Methodology

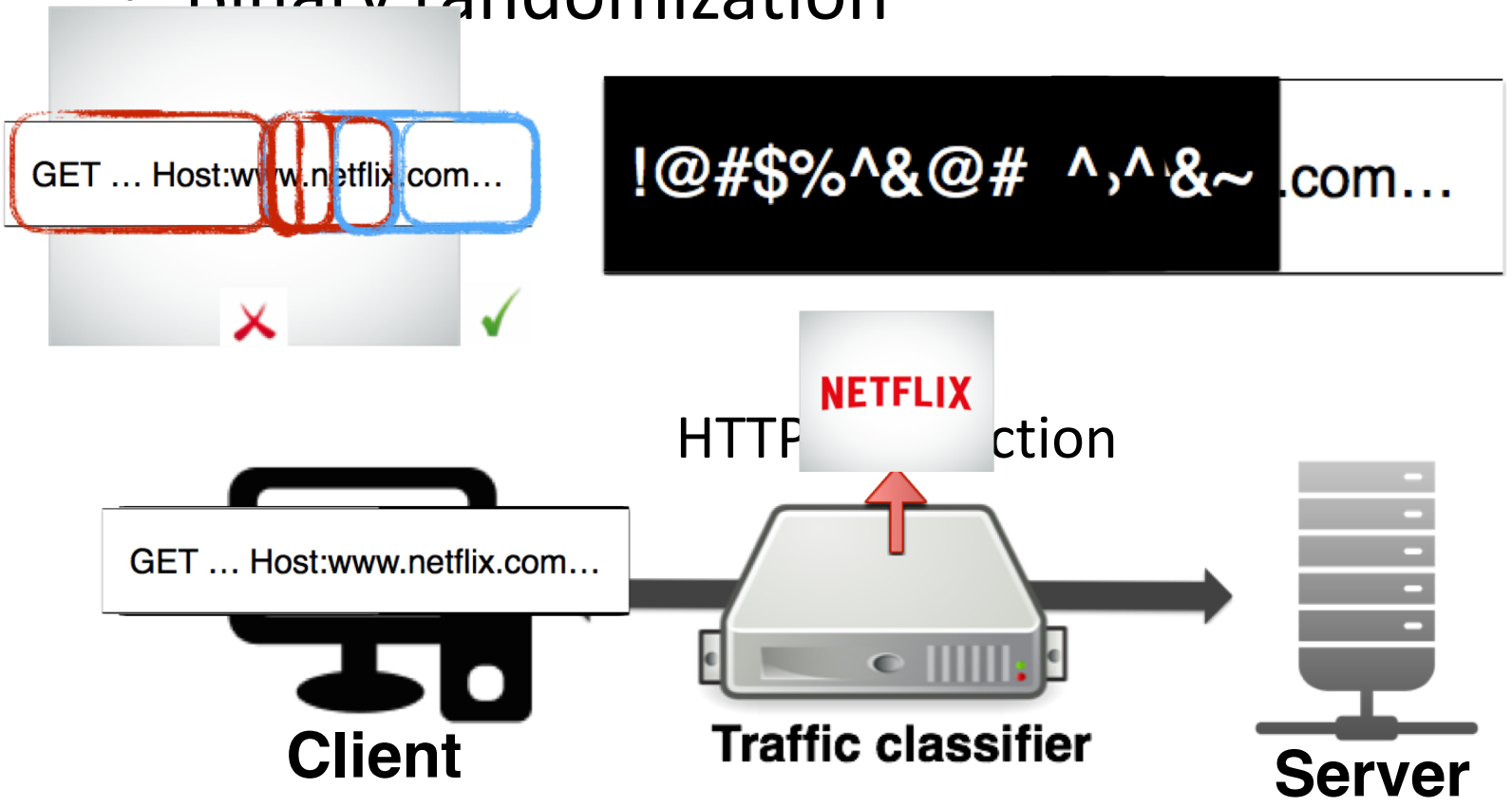
1. Network traffic
2. Feedback signal

NETFLIX Record&Replay [IMC 15]



Identifying Matching Rules

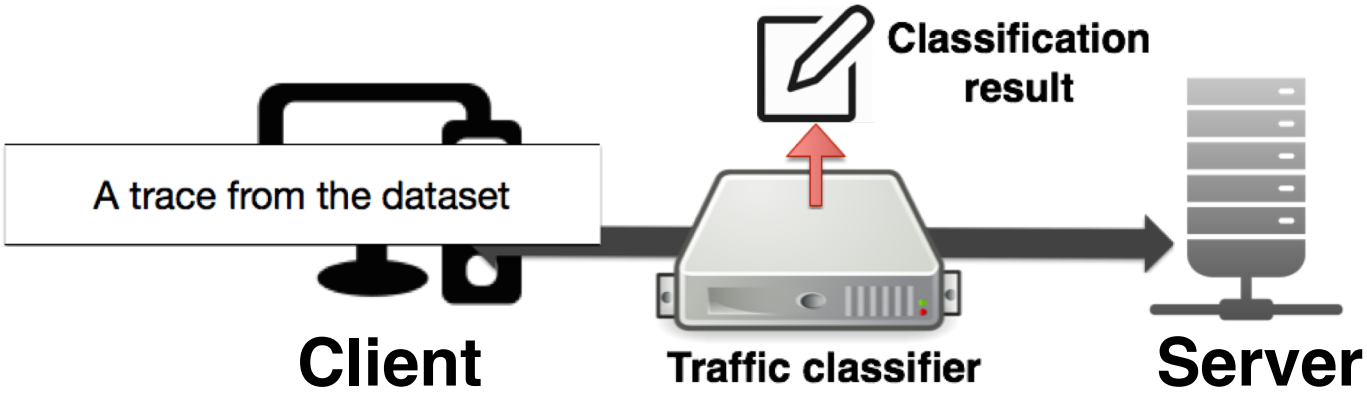
- Binary randomization



Locating Matching Fields

Dataset: Traces from ReCon [Mobisys 16]

- 1 million HTTP/S requests from 300 users
- Where are the matching fields?
  - HTTP: URI, Host, User-Agent and Content-Type
  - HTTPS: Text strings in the TLS handshake



Example Matching Rules

Header	Example Value	Application
URI	site.js{…}-**nbcsports**-com	NBC Sports
Host	Host: www**.spotify.**com	Spotify
User-Agent	User-Agent: **Pandora** 5.0{…}	Pandora
Content-Type	Content-Type: video/**quicktime**	QuickTime

Summary

- Traffic matrix
  - Measured from flow capture
  - Estimated from interface counts
- Application identification
  - Ports more and more inaccurate
  - Content: mostly used in practice, but becoming harder with encryption
  - Behavior: alternative for encrypted traffic, but not very practical yet

References

- M. Crovella, B. Krishnamurthy, “Internet Measurement: Infrastructure, traffic & applications” Section 6.4.1
- P. Tune, M. Roughan, “Internet Traffic Matrices: A Primer”, in H. Haddadi, O. Bonaventure (Eds.), Recent Advances in Networking, (2013)
  - http://sigcomm.org/education/ebook/SIGCOMMeBook2013v1_chapter3.pdf
- A. Callado, C. Kamienski, G. Szabó, B. Péter-Gerö, J. Kelner, S. Fernandes, “A Survey on Internet Traffic Identification”, IEEE Communications Surveys & Tutorials, Vol. 11, N. 3, 2009.
  - https://www.researchgate.net/profile/Arthur_Callado/publication/220250167_A_Survey_on_Internet_Traffic_Identification/links/0c96053ab69c5b6c71000000.pdf
- A.K. Marnerides, A. Schaeffer-Filho, A. Mauthe, “Traffic anomaly diagnosis in Internet backbone networks: A survey”, Computer Networks 73 (2014)
- F. Li, A.M. Kakhki, D. Choffnes, P. Gill, A. Mislove, “Classifiers Unclassified: An efficient approach for revealing IP traffic classification rules”, IMC 2016.