Mind The Gap Between HTTP and HTTPS in Mobile Networks

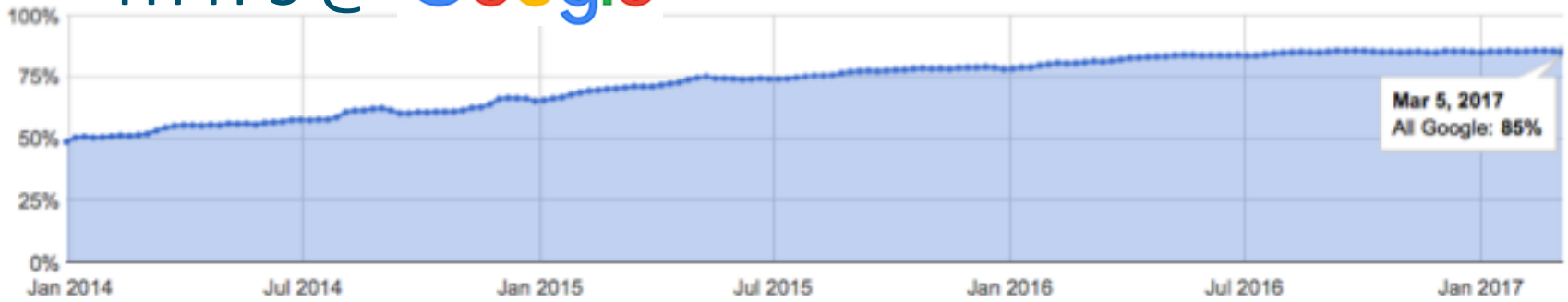
Alessandro Finamore
Matteo Varvello
Kostantina Papagiannaki

Passive and Active Measurement
03.31.2017
Sydney, Australia

01 Why to study HTTPS?

HTTPS is big in the wild

HTTPS @ Google

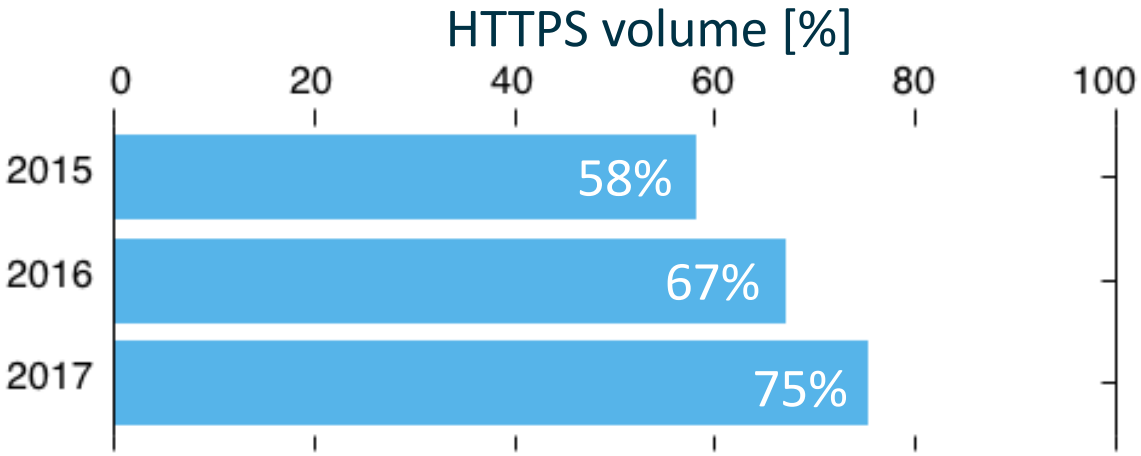


https://www.google.com/transparencyreport/https/

HTTPS @ Telefónica

(1 country only)

At least a +10% y-o-y



Research studies about HTTPS

- Many research studies on mobile network traffic only on HTTP
  - Traffic classification
  - Mobility and users behavior
  - Privacy
  - Etc.

- Why such limited interested towards HTTPS?
  - Difficult (but not impossible) to collect/access dataset at-scale
  - Collecting HTTPS logs is perceived as a “waste of storage”
    - Little information to profile users accessed services
    - Little information to extract performance metrics

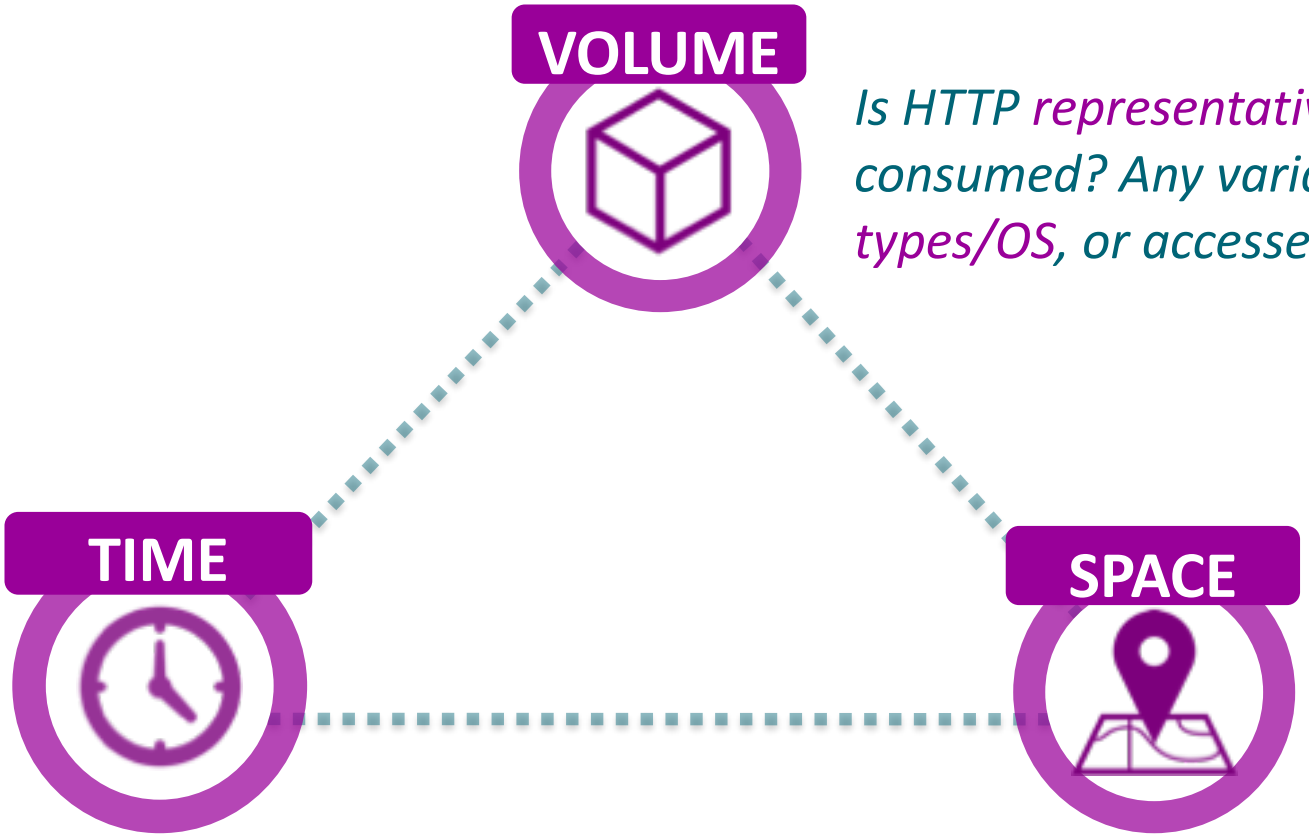
Research questions

An holistic view of what/when/how users access content is key
…so relying only on HTTP introduces “gaps”

Questions

Is HTTP still representative of the overall mobile traffic?
Can we quantify the ”gaps” when monitoring only HTTP?
Is it important to monitor also HTTPS?

Study “gaps” across 3 dimensions



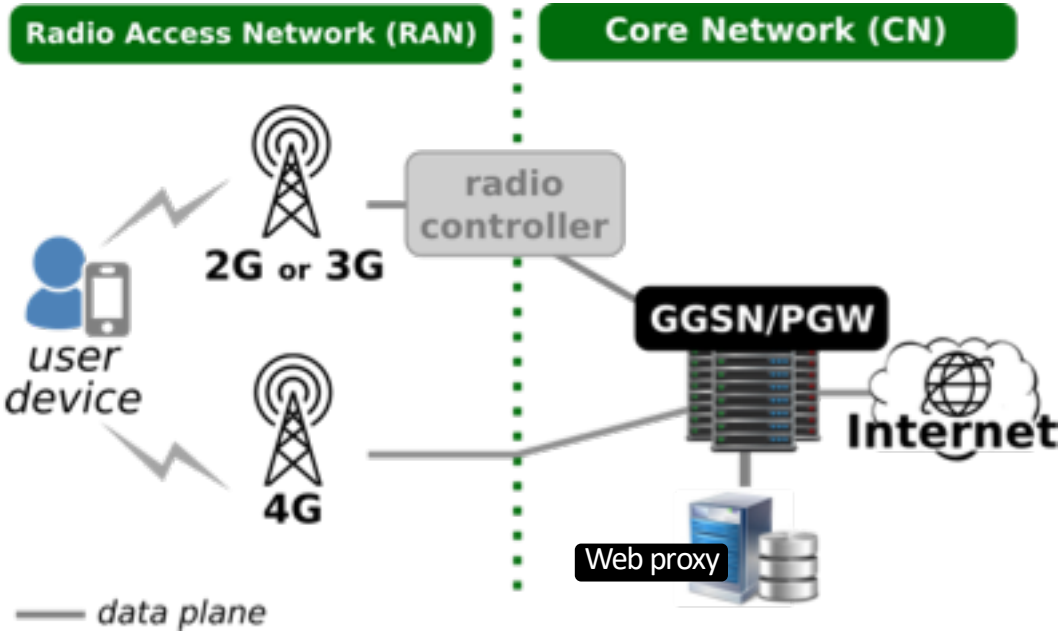
Is HTTP representative of the traffic consumed? Any variation across device types/OS, or accessed services?

Is HTTP presence different at different time of the day?

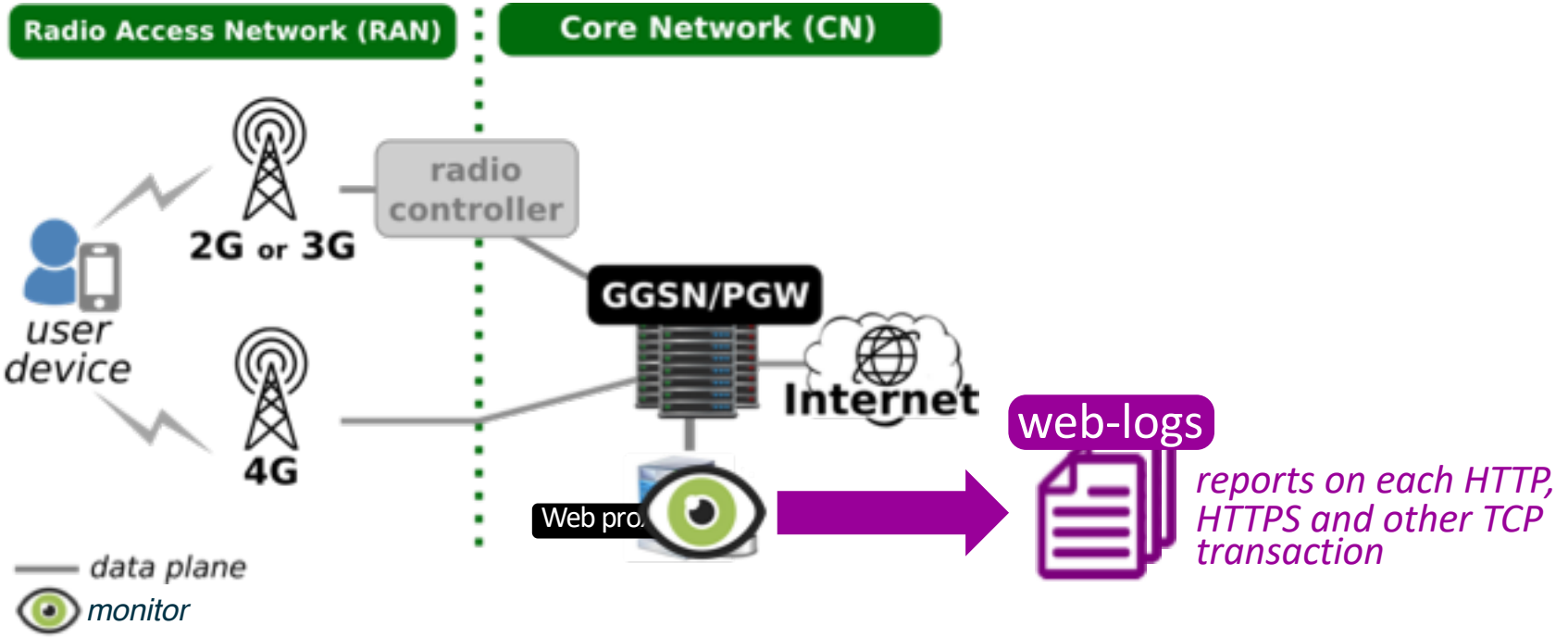
Is HTTP alone sufficient to identify where users consume content?

02 Dataset

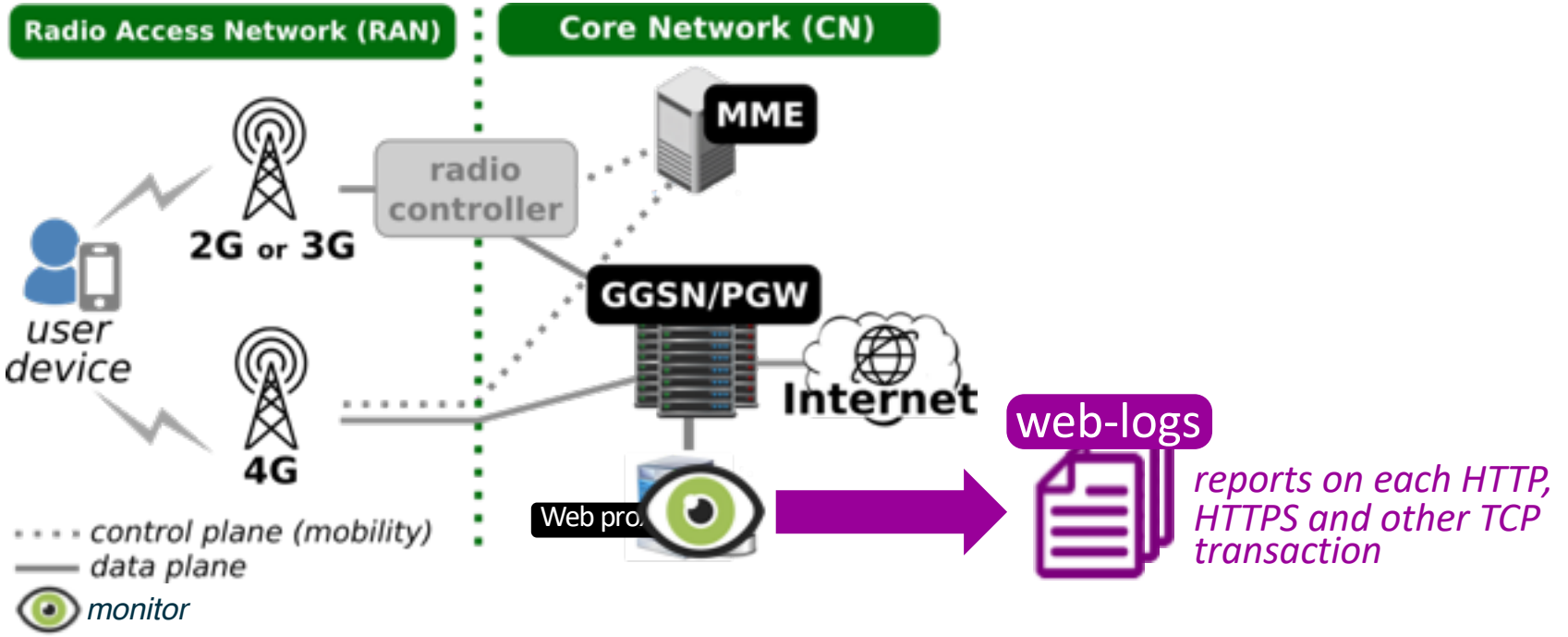
Where to collect data?



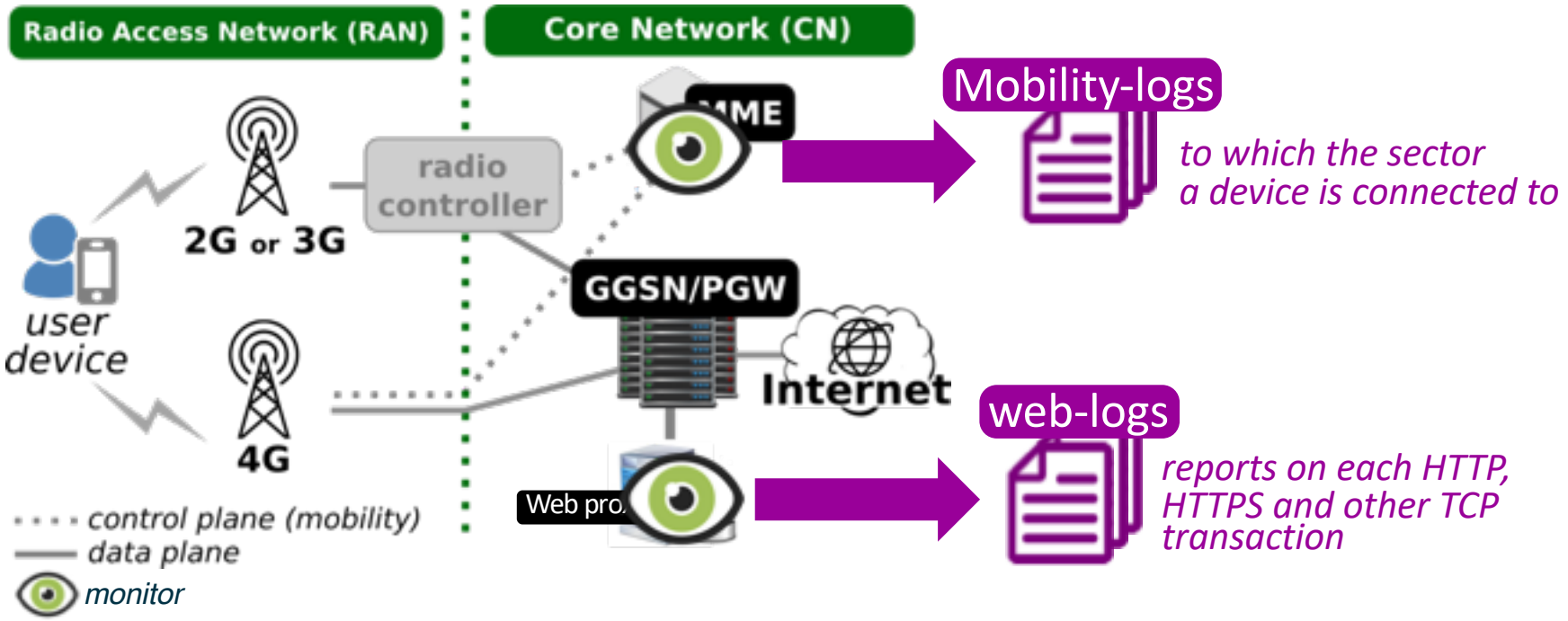
Where to collect data?



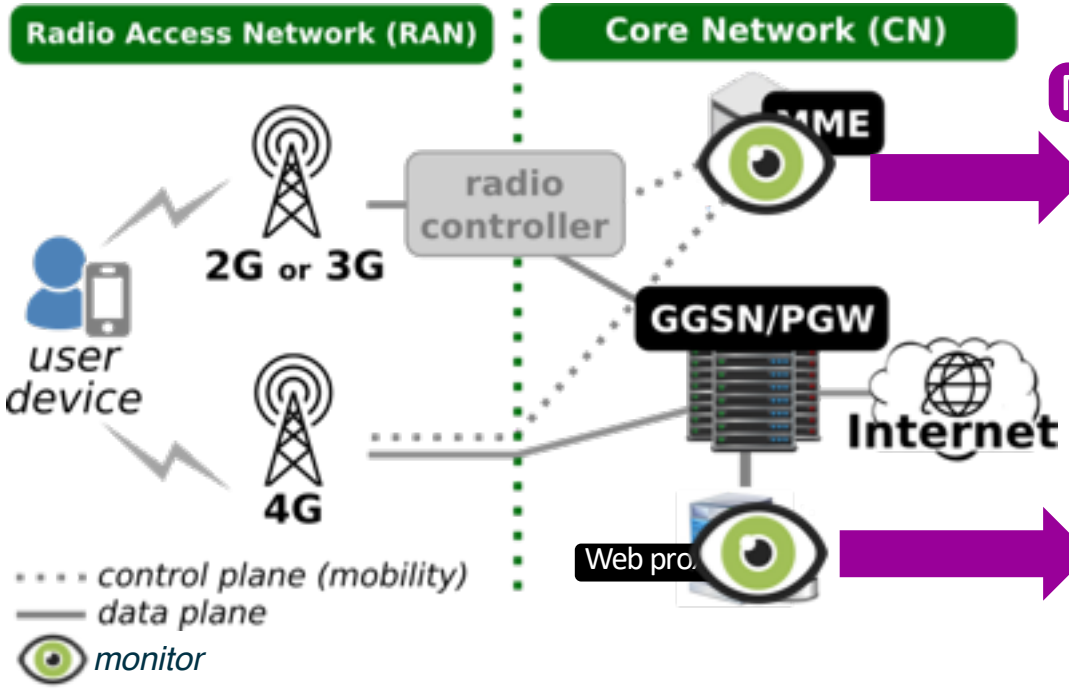
Where to collect data?



Where to collect data?



Where to collect data?



Mobility-logs

to which the sector a device is connected to

web-logs

reports on each HTTP, HTTPS and other TCP transaction

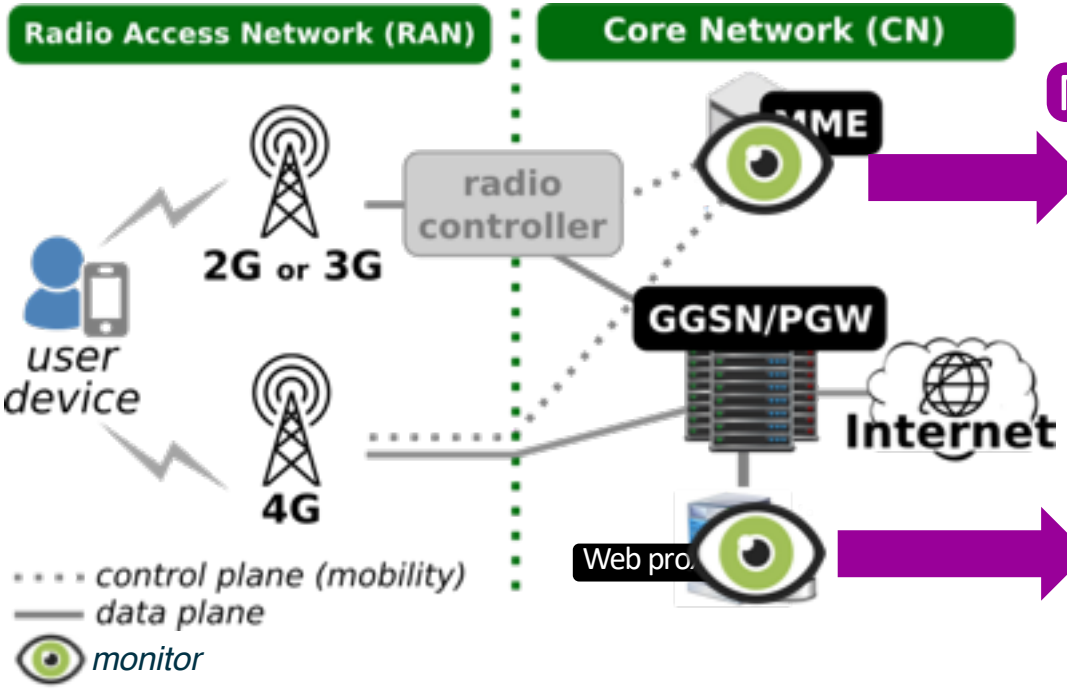
TAC DB

Device type classification Based on IMEI TAC

GeoIP2 IPS

The organization "owning" an IP address

Where to collect data?



Mobility-logs

to which the sector a device is connected to

web-logs

reports on each HTTP, HTTPS and other TCP transaction

TAC DB

Device type classification Based on IMEI TAC

GeoIP2 IPS

The organization "owning" an IP address

By "blending" different data sources,

we obtain a fine grained view

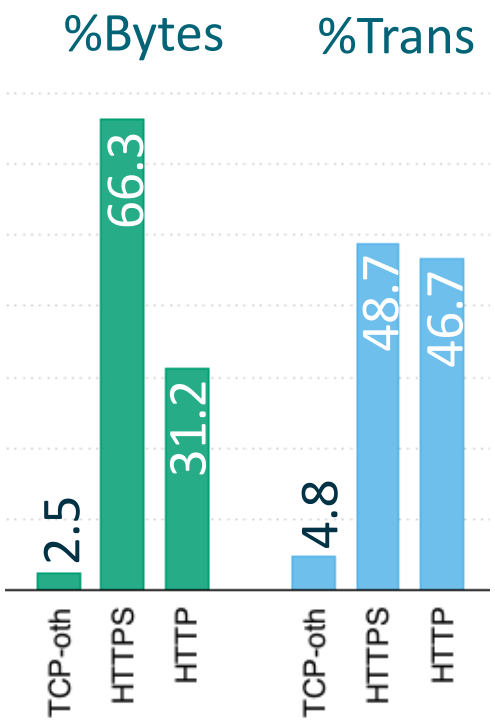
- Spatial and temporal
- Device and OS
- Accessed services

Dataset

- One single day of traffic (in Oct'16) for all users (>10M)
  - 50B transactions = ~5TB (compressed)
- Full view of HTTP, HTTPS and the remaining TCP traffic (TCP-oth)

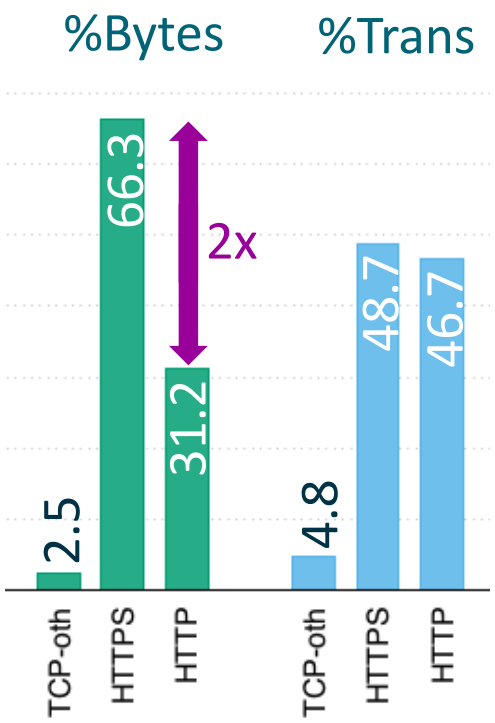
Dataset

- One single day of traffic (in Oct'16) for all users (>10M)
  - 50B transactions = ~5TB (compressed)
- Full view of HTTP, HTTPS and the remaining TCP traffic (TCP-oth)



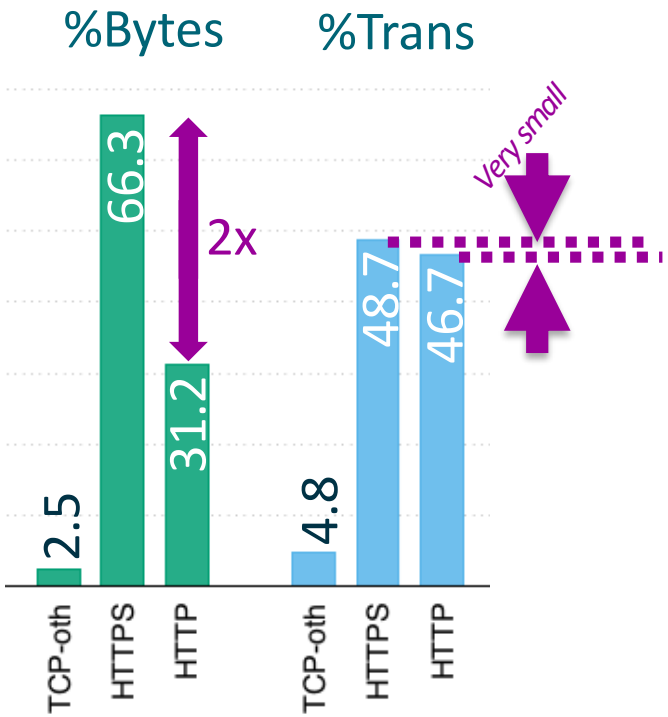
Dataset

- One single day of traffic (in Oct'16) for all users (>10M)
  - 50B transactions = ~5TB (compressed)
- Full view of HTTP, HTTPS and the remaining TCP traffic (TCP-oth)



Dataset

- One single day of traffic (in Oct'16) for all users (>10M)
  - 50B transactions = ~5TB (compressed)
- Full view of HTTP, HTTPS and the remaining TCP traffic (TCP-oth)

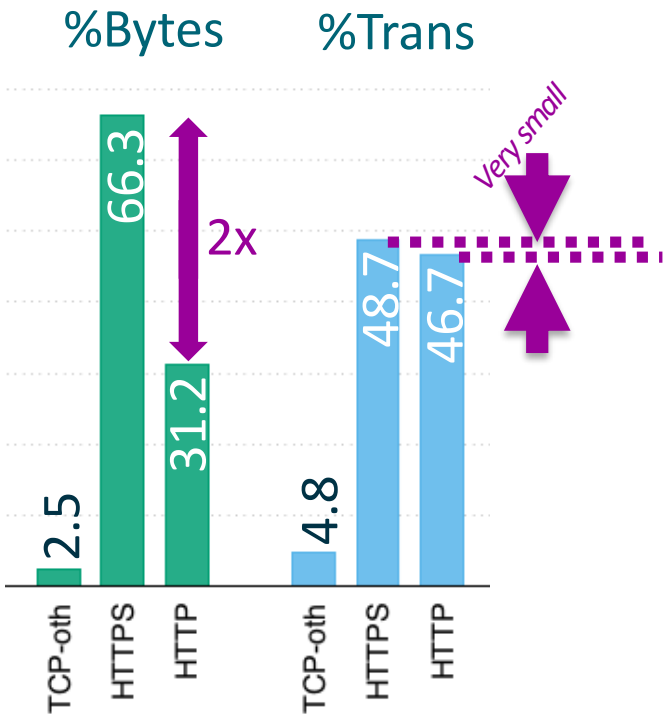


2 observations

1. No HTTPS logs = 2x HTTP logs retention window

Dataset

- One single day of traffic (in Oct'16) for all users (>10M)
  - 50B transactions = ~5TB (compressed)
- Full view of HTTP, HTTPS and the remaining TCP traffic (TCP-oth)



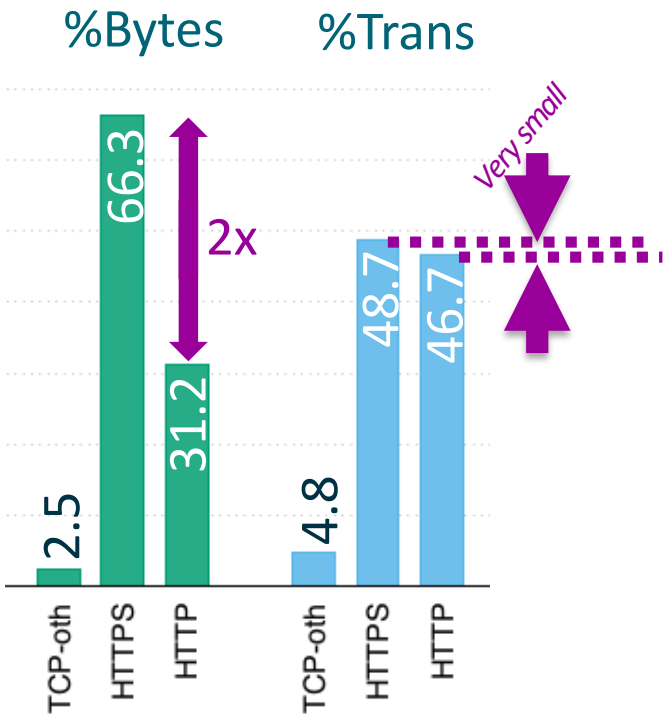
MIND THE GAP

2 observations

1. No HTTPS logs = 2x HTTP logs retention window
2. Different traffic classes = different granularity
   - if HTTP ➔ info on individual requests
   - otherwise ➔ info on whole connection

Dataset

- One single day of traffic (in Oct'16) for all users (>10M)
  - 50B transactions = ~5TB (compressed)
- Full view of HTTP, HTTPS and the remaining TCP traffic (TCP-oth)



MIND THE GAP

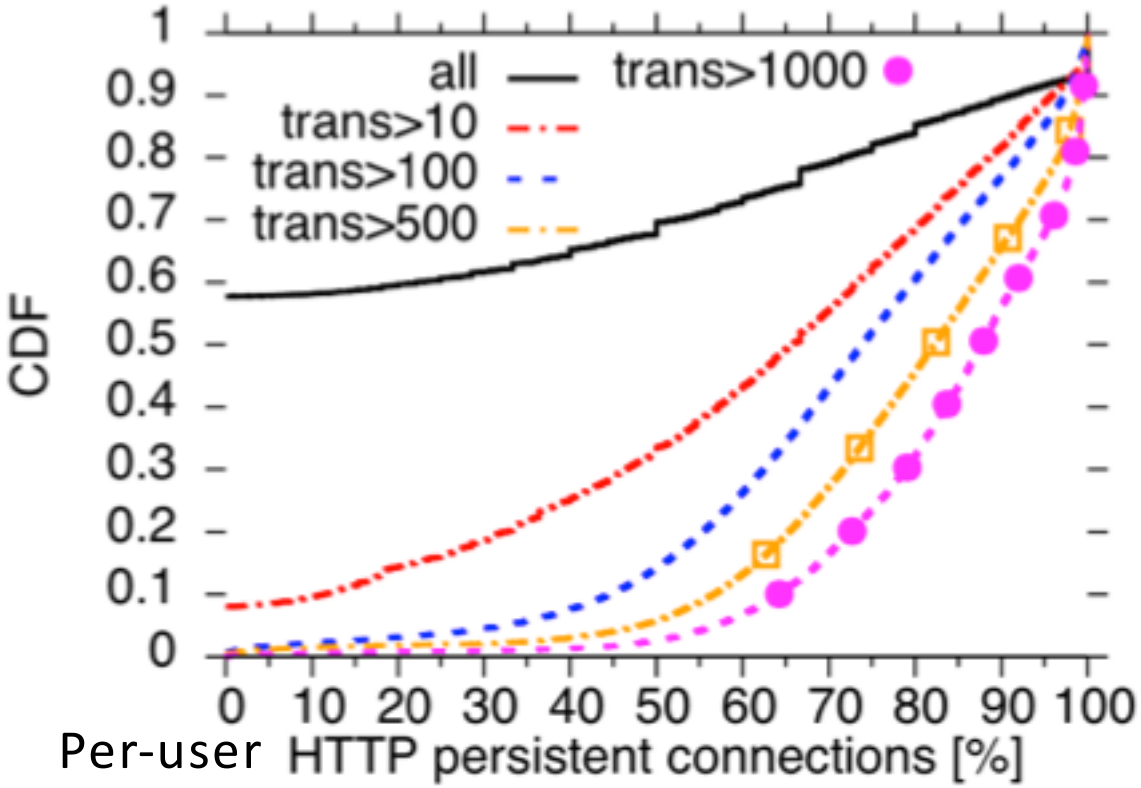
2 observations

1. No HTTPS logs = 2x HTTP logs retention window
2. Different traffic classes = different granularity
   - if HTTP ➔ info on individual requests
   - otherwise ➔ info on whole connection

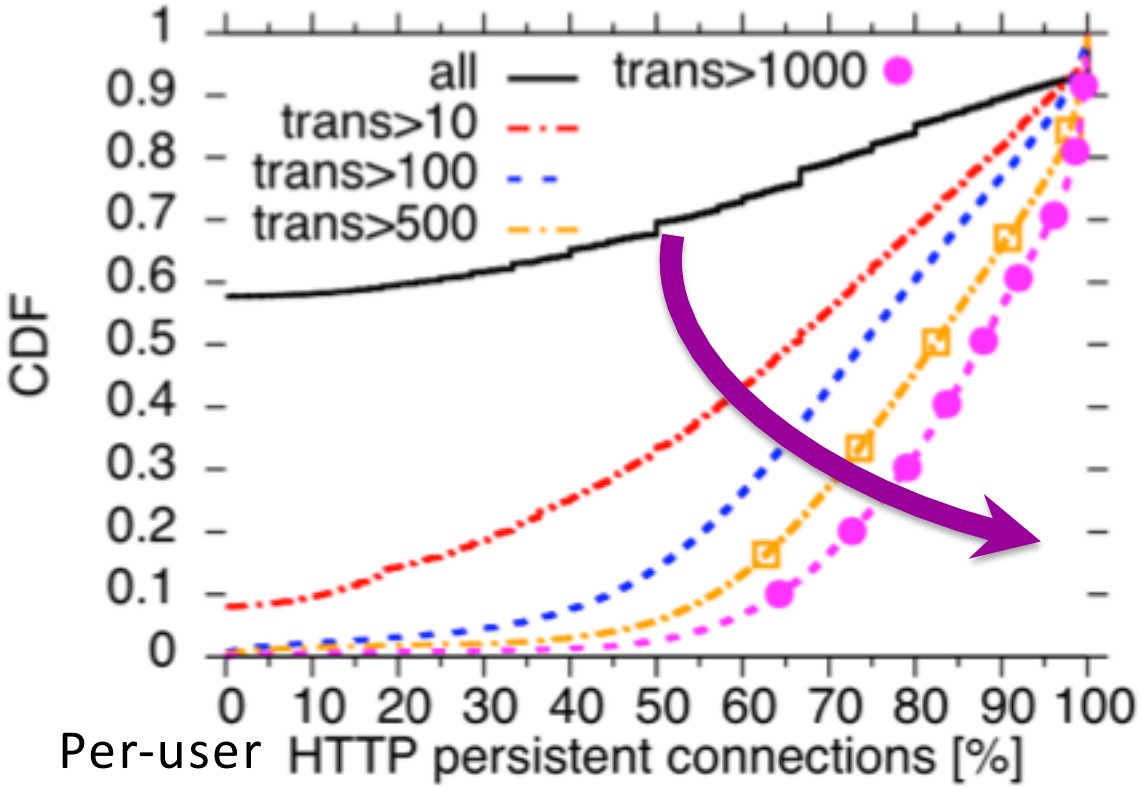
…but, is this relevant, and why?

HTTP requests are carried over persistent connections

HTTP requests are carried over persistent connections

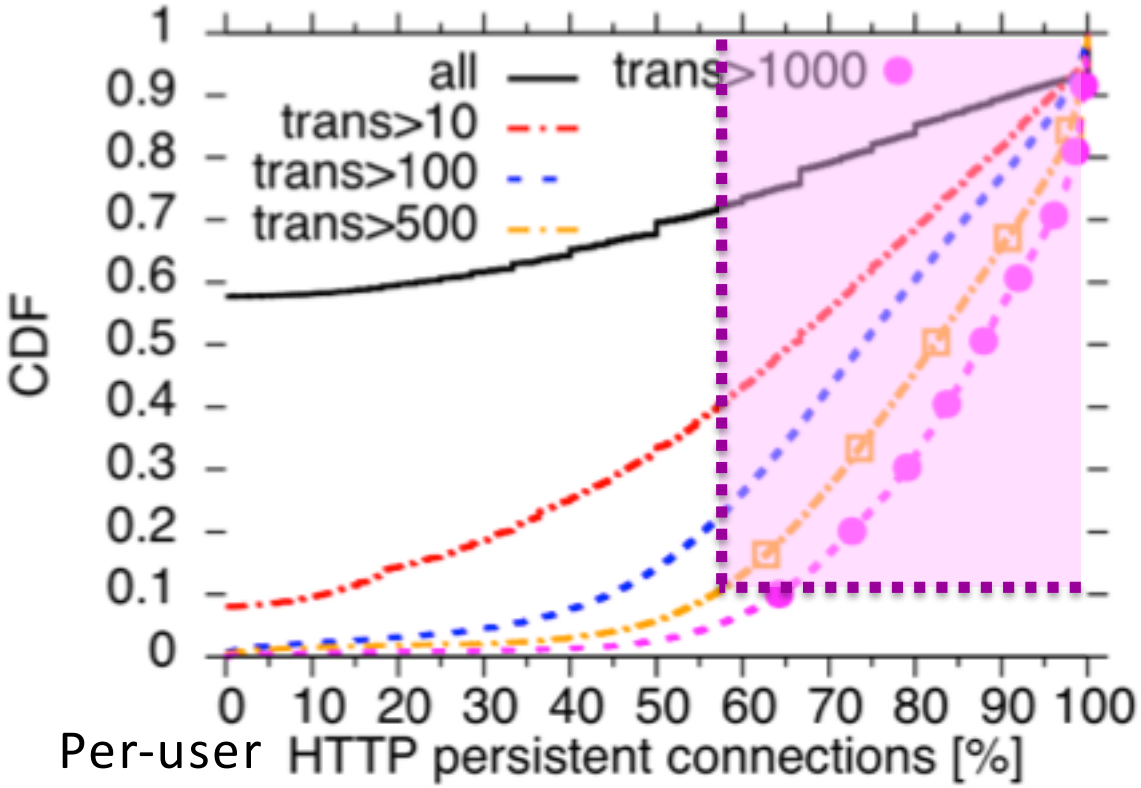


HTTP requests are carried over persistent connections



The higher the user activity, the higher the usage of persistent connections

HTTP requests are carried over persistent connections



The higher the user activity, the higher the usage of persistent connections

About 90% of active users have >55% of HTTP content carried over persistent connections

HTTP requests are carried over persistent connections

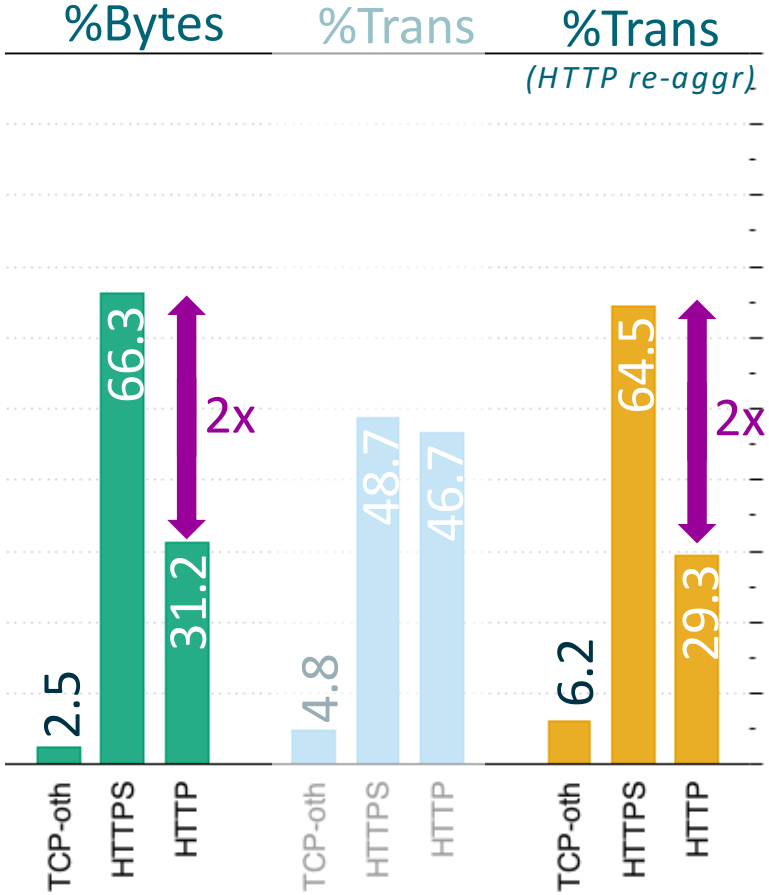
The higher the user activity, the higher the usage of persistent connections

Mind the gap

1. The same effect is expected for HTTPS although we cannot measure it
2. For a fair comparison, we need to aggregate HTTP requests with respect to their associated TCP connection

Dataset (revised)

- One single day of traffic (in Oct'16) for all users (>10M)
  - 50B transactions = ~5TB (compressed)
- Full view of HTTP, HTTPS and the remaining TCP traffic (TCP-oth)



After re-aggregation, number of transactions has 2x as for bytes

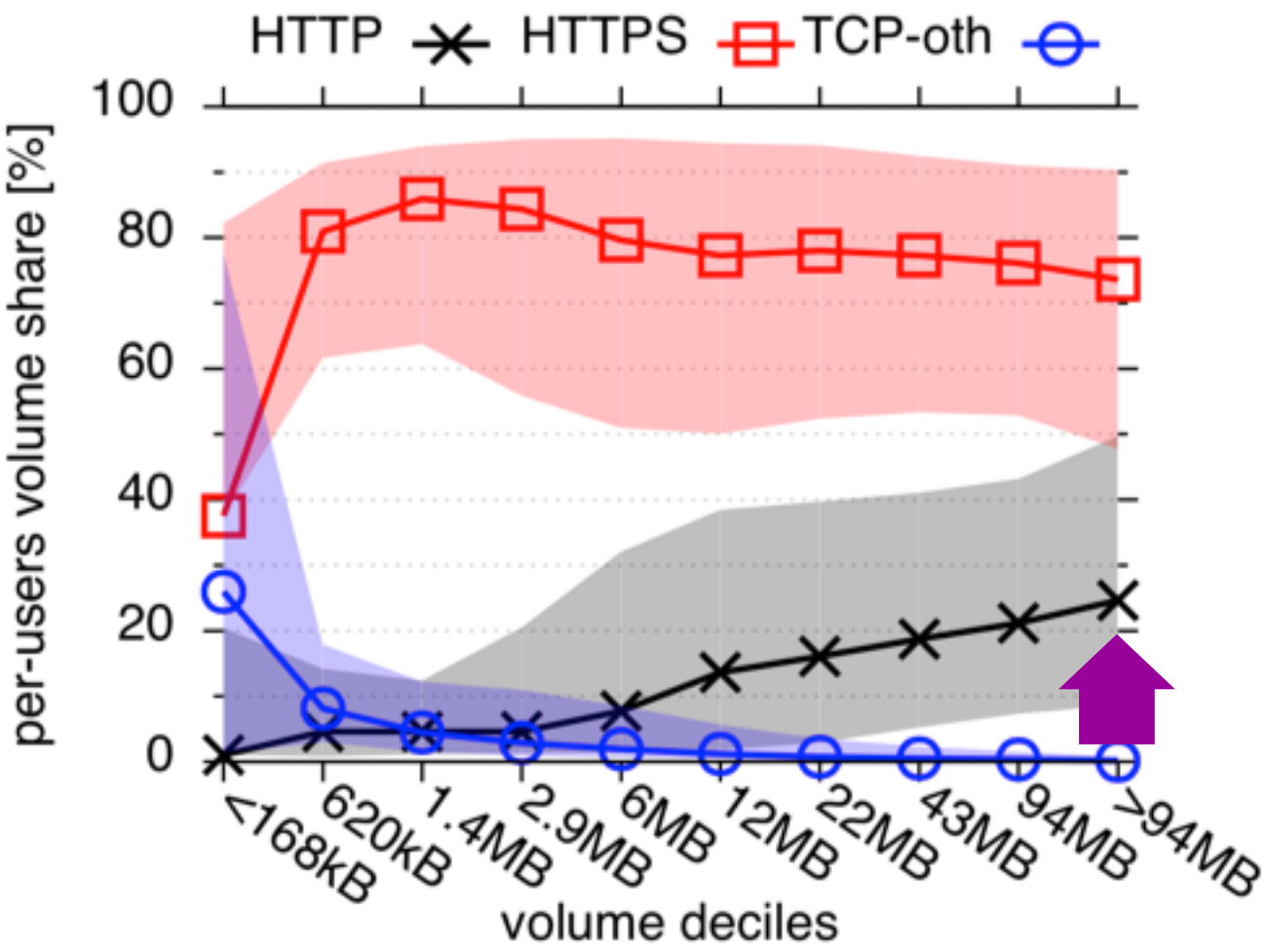
Notice also the increased "weight" of the remaining TCP traffic

Telefónica

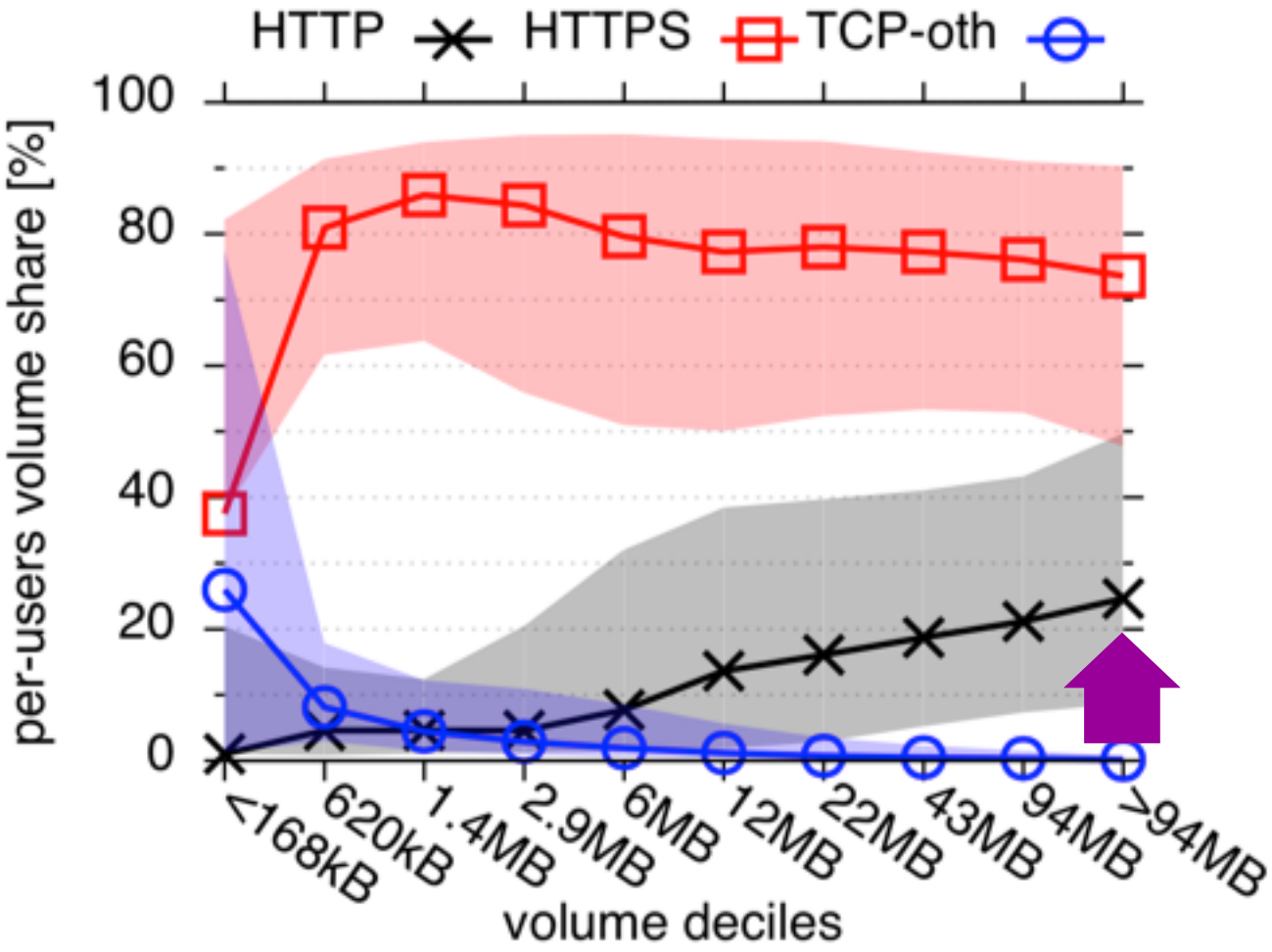
03 Quantifying the gaps

Higher user demand = more HTTP

Higher user demand = more HTTP



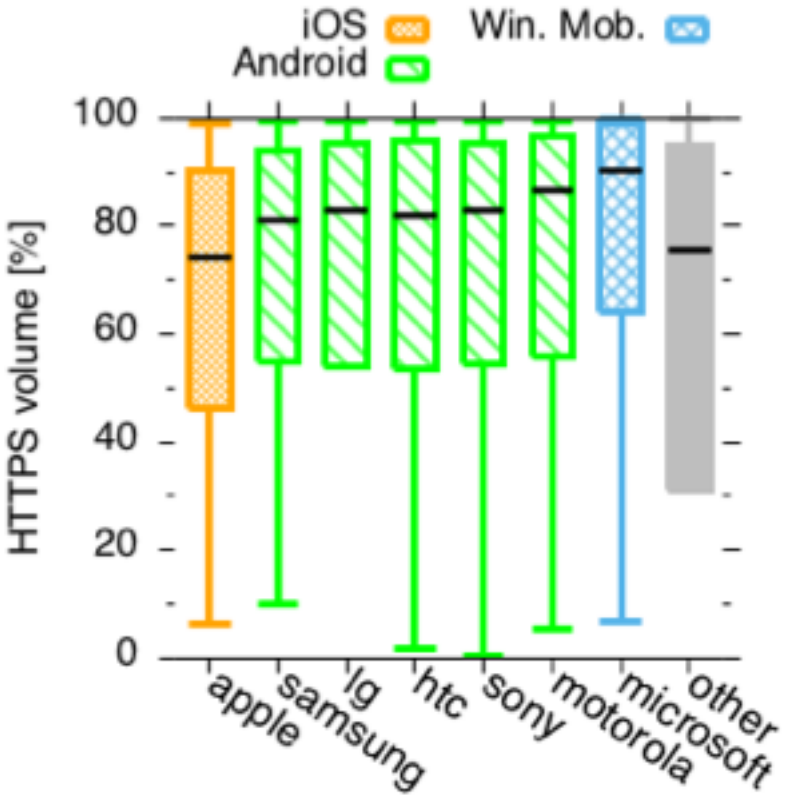
Higher user demand = more HTTP



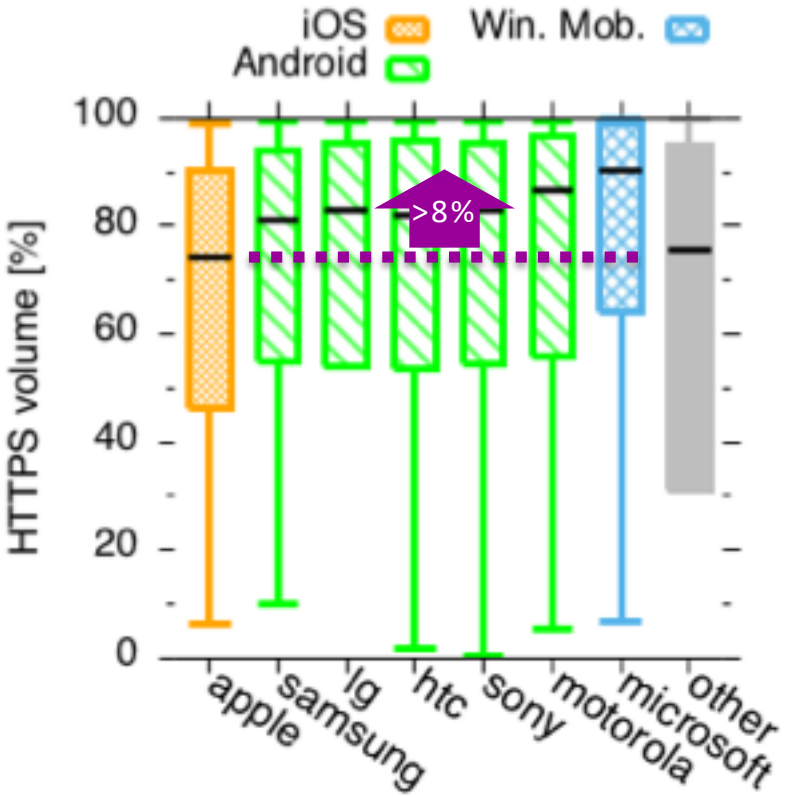
Using only HTTP can result in over representing "data hungry" users

iOS devices consume much more HTTP

iOS devices consume much more HTTP

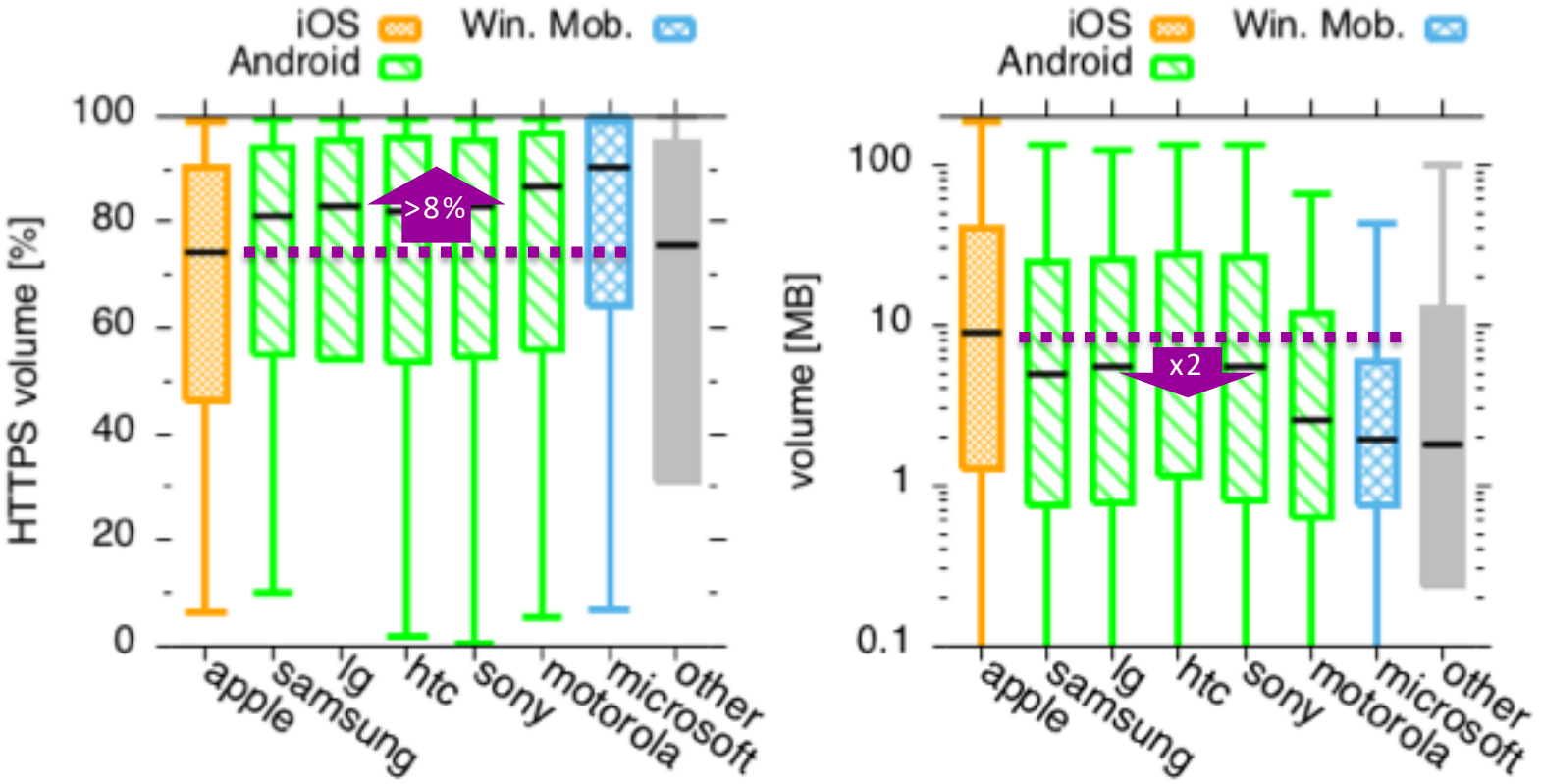


iOS devices consume much more HTTP



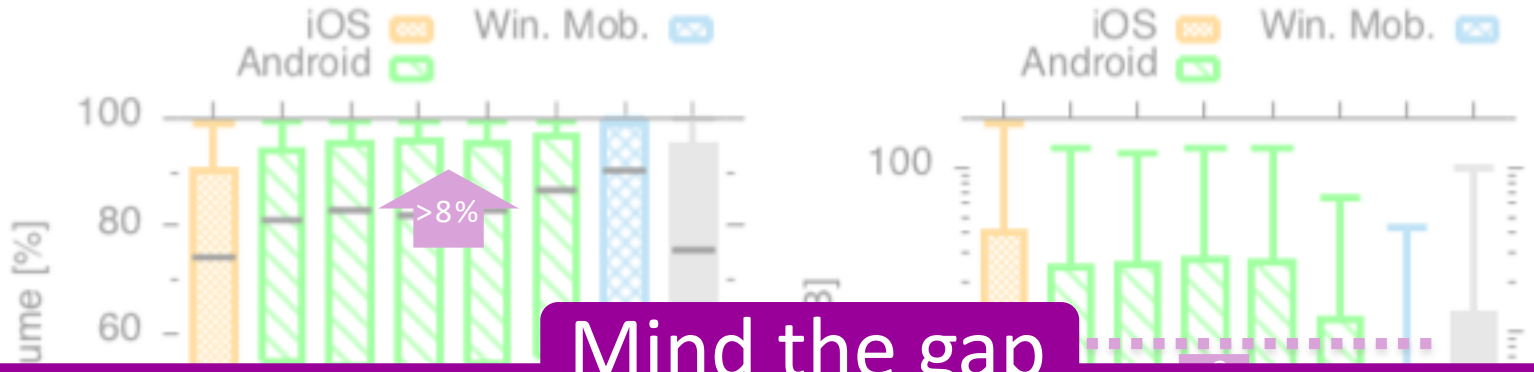
- Android devices consumes (as a median) >8% more HTTPS

iOS devices consume much more HTTP



- Android devices consumes (as a median) >8% more HTTPS
- …but iOS devices consume (as a median) x2 more data

iOS devices consume much more HTTP



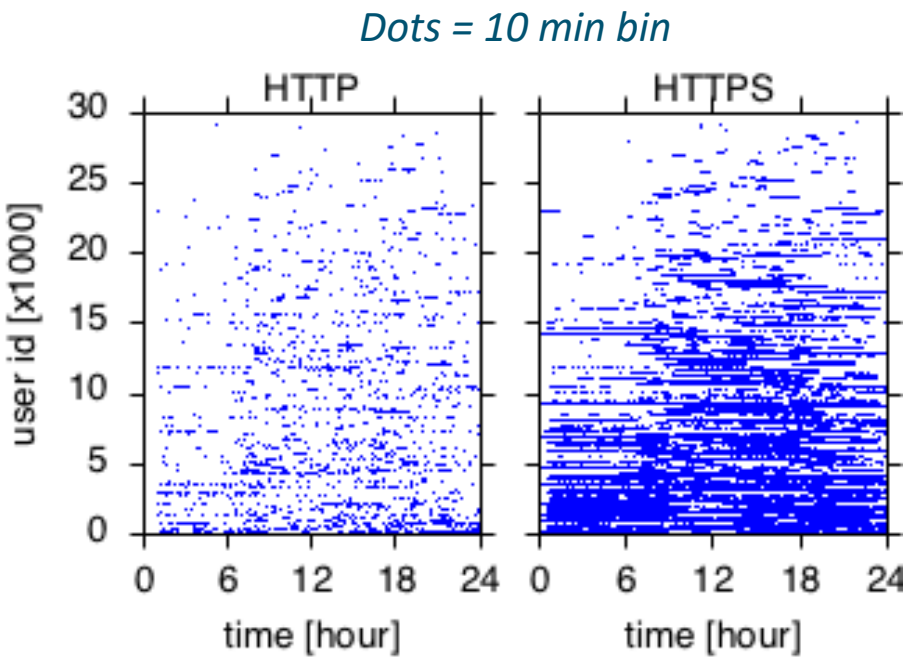
Mind the gap

1. Overall, (as expected) HTTP monitoring alone is not enough to study volume
2. Tangled relationship between device/OS, accessed services, and data usage

- Android devices consumes (as a median) >8% more HTTPS
- …but iOS devices consume (as a median) x2 more data

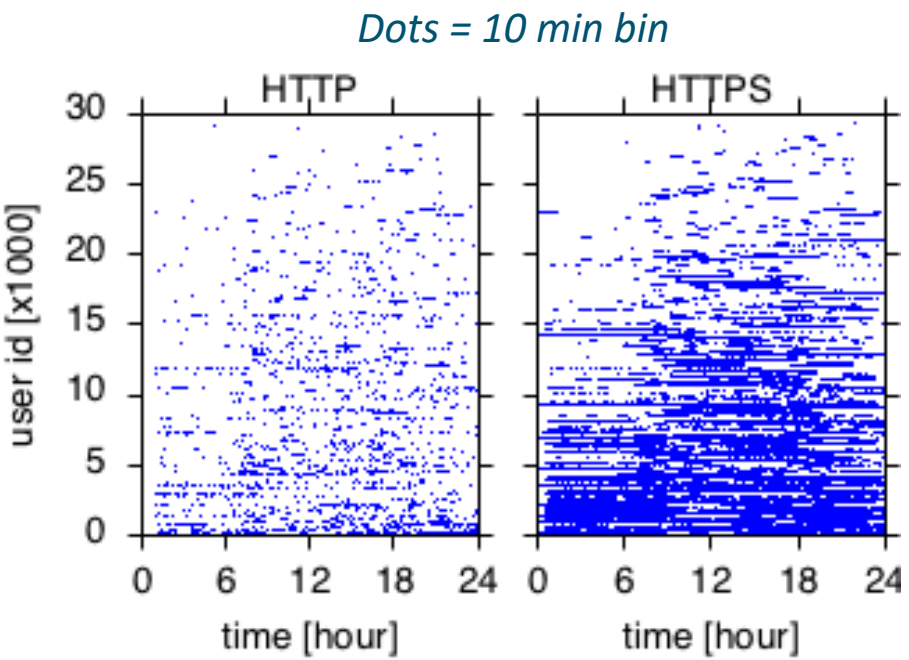
HTTP traffic is very sparse across the day

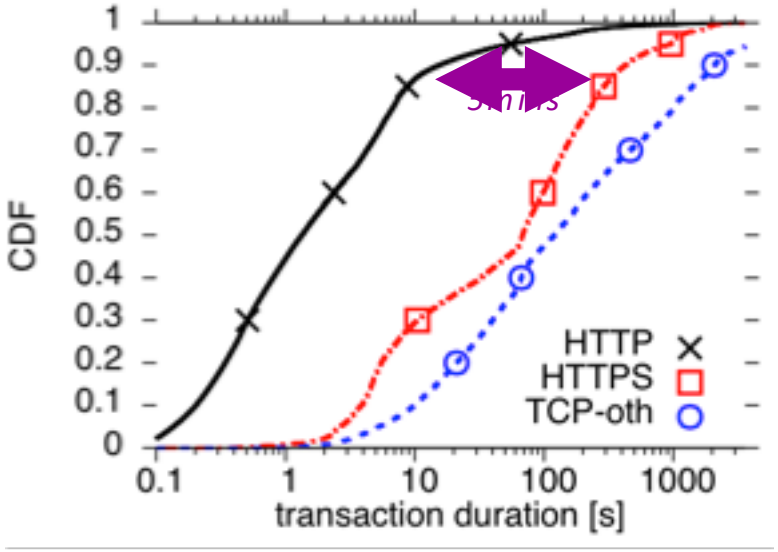
HTTP traffic is sparse across the day



HTTP traffic is very “occasional”

HTTP traffic is sparse across the day





HTTP traffic is very “occasional”

…and this is not because HTTPS transactions last significantly longer (90% of HTTPS trans are < 5mins)

HTTP traffic is sparse across the day

Dots = 10 min bin

Mind the gap

1. HTTP traffic is sparse, but mostly in the daily hours *(see paper)*
2. This (possibly) implies that is HTTP more related to user engagement while HTTPS has a component of background traffic

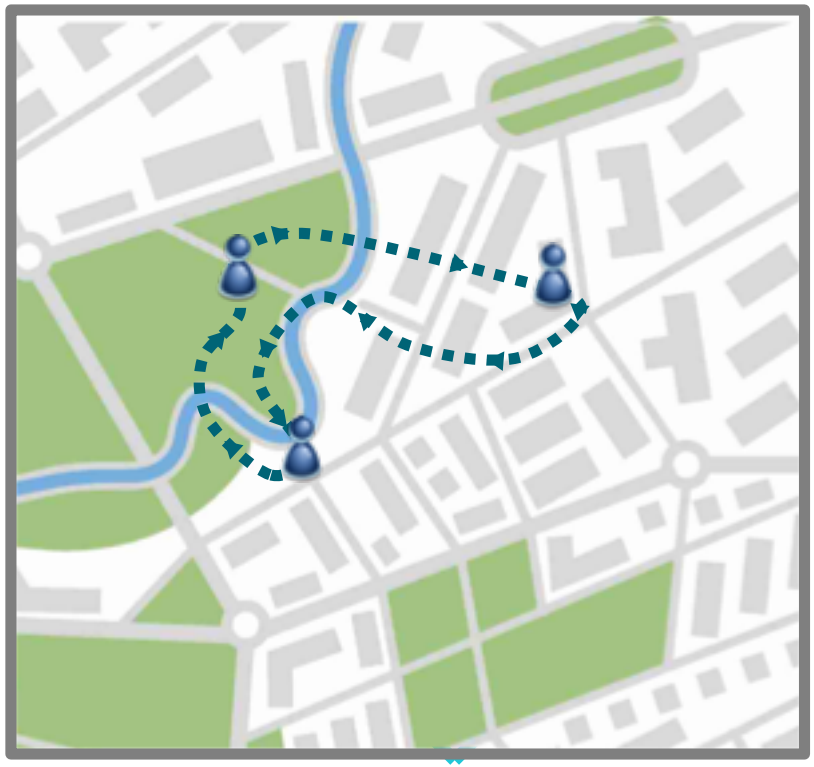
HTTP traffic is very “occasional”

…and this is not because HTTPS transactions last significantly longer (90% of HTTPS trans are < 5mins)

Is HTTP consumed in the same locations as for the overall traffic?

Is HTTP consumed in the same locations as for the overall traffic?

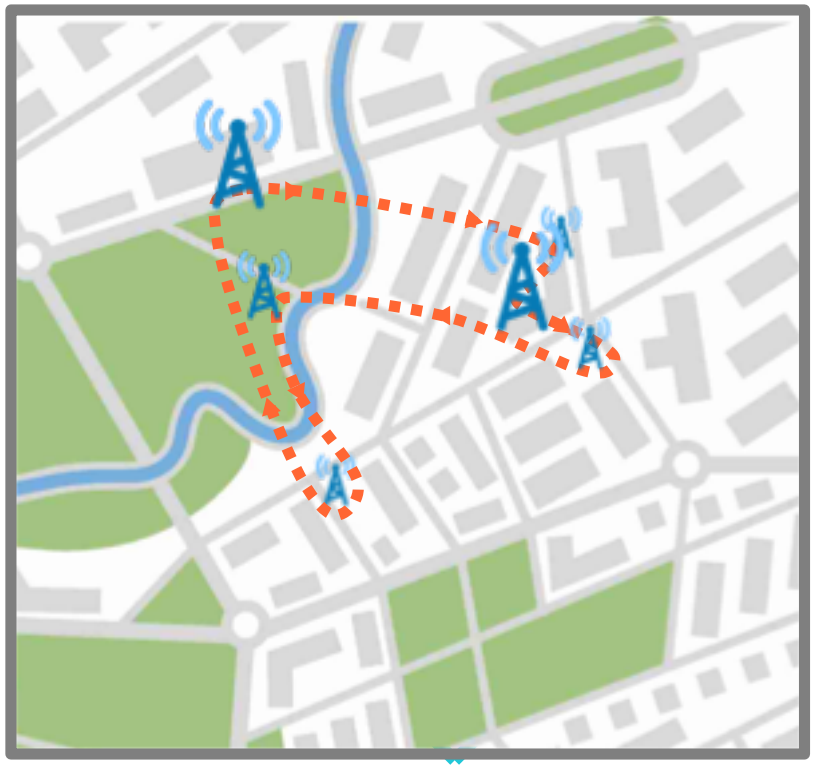
Recall: the dataset specifies in which sector a transaction took place



Is HTTP consumed in the same locations as for the overall traffic?

Recall: the dataset specifies in which sector a transaction took place

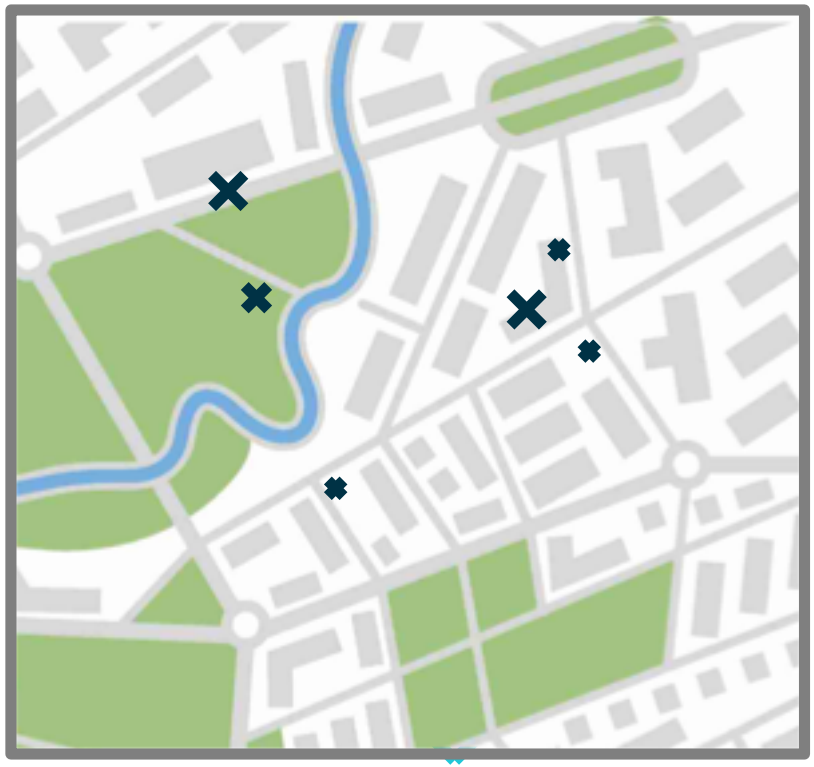
1. Approximate users location with towers position



Is HTTP consumed in the same locations as for the overall traffic?

Recall: the dataset specifies in which sector a transaction took place

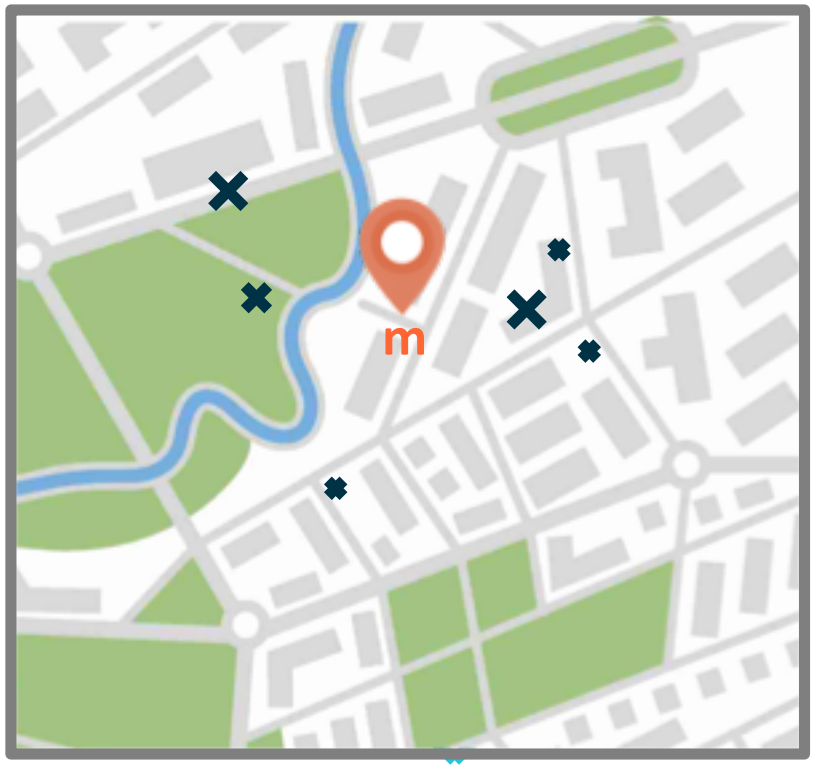
1. Approximate users location with towers position
2. We are not interested into tracing paths, but rather the trans for each tower



Is HTTP consumed in the same locations as for the overall traffic?

Recall: the dataset specifies in which sector a transaction took place

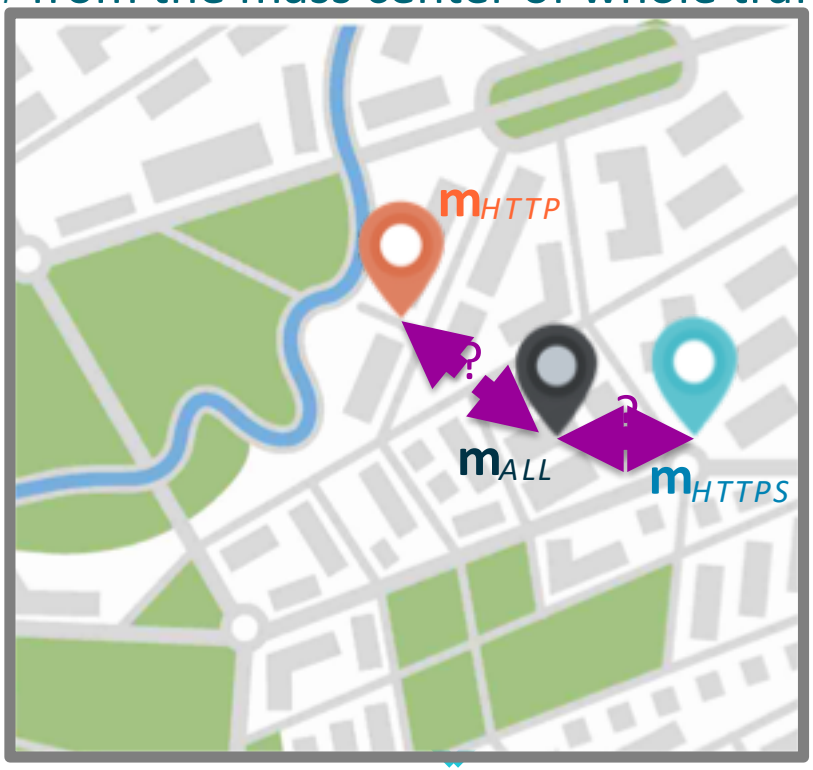
1. Approximate users location with towers position
2. We are not interested into tracing paths, but rather the trans for each tower
3. Define a mass center (*m*) as weighted average (with the transactions served) of towers location



Is HTTP consumed in the same locations as for the overall traffic?

Recall: the dataset specifies in which sector a transaction took place

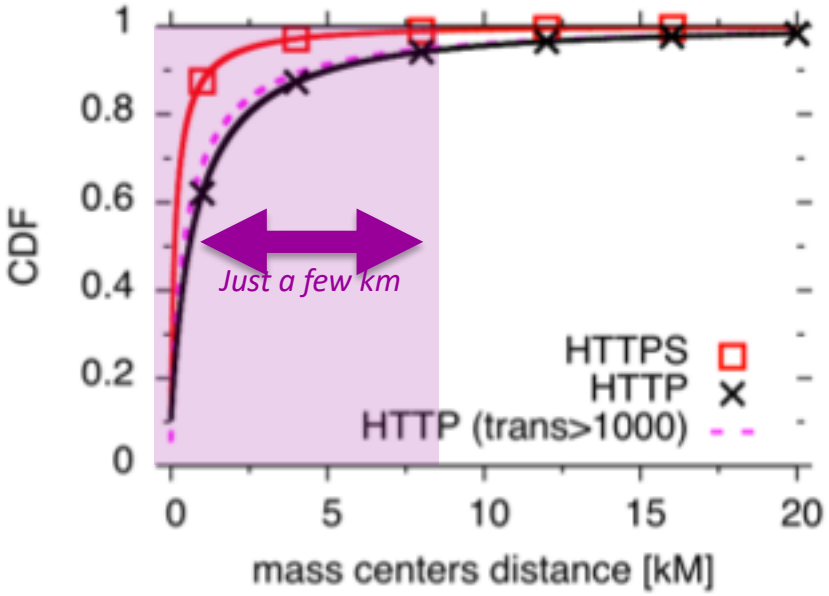
1. Approximate users location with towers position
2. We are not interested into tracing paths, but rather the trans for each tower
3. Define a mass center (*m*) as weighted average (with trans served) of towers location
4. How far is m_{HTTP} from the mass center of whole traffic? What about HTTPS?



Is HTTP consumed in the same locations as for the overall traffic?

Recall: the dataset specifies in which sector a transaction took place

1. Approximate users location with towers position
2. We are not interested into tracing paths, but rather the trans for each tower
3. Define a mass center (*m*) as weighted average (with trans served) of towers location
4. How far is m_{HTTP} from the mass center of whole traffic? What about HTTPS?



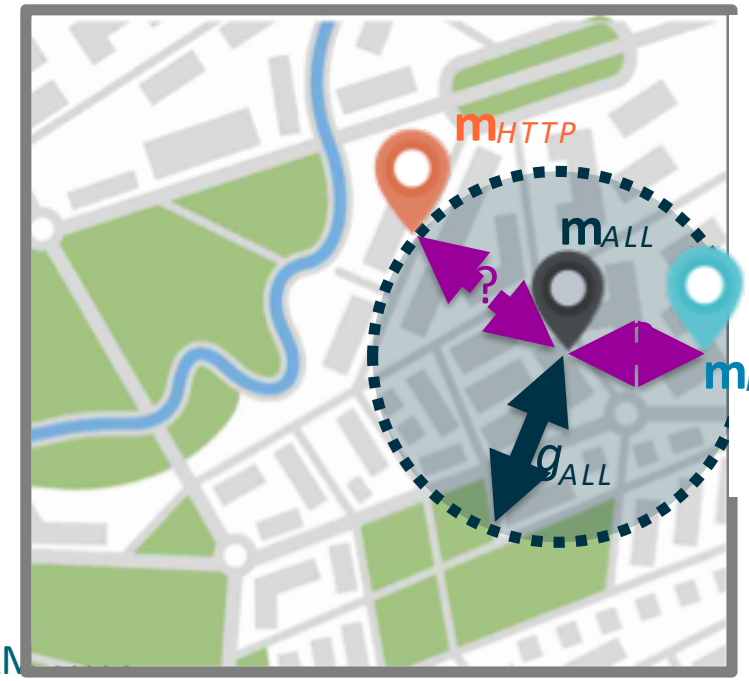
Once again, HTTP is much sparser than HTTPS

…but a few km are significant wrt the size of the whole area in which the user is moving?

Is HTTP consumed in the same locations as for the overall traffic?

Recall: the dataset specifies in which sector a transaction took place

1. Approximate users location with towers position
2. We are not interested into tracing paths, but rather the trans for each tower
3. Define a mass center (m) as weighted average (with trans served) of towers location
4. How far is m_{HTTP} from the mass center of whole traffic? What about HTTPS?
5. Normalize distances wrt gyration



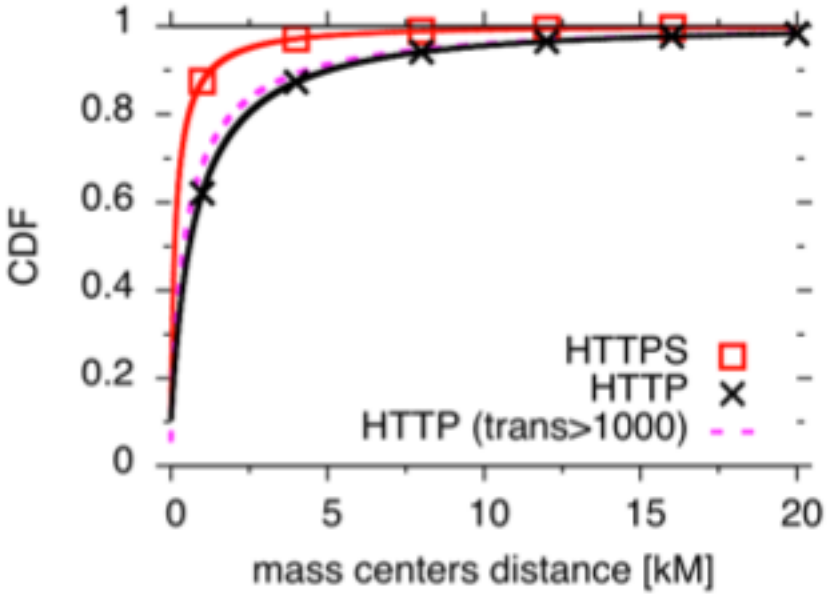
$$g_{ALL} = \sqrt{\sum_i \frac{distance^2(m_{ALL}, tower_i)}{N}}$$

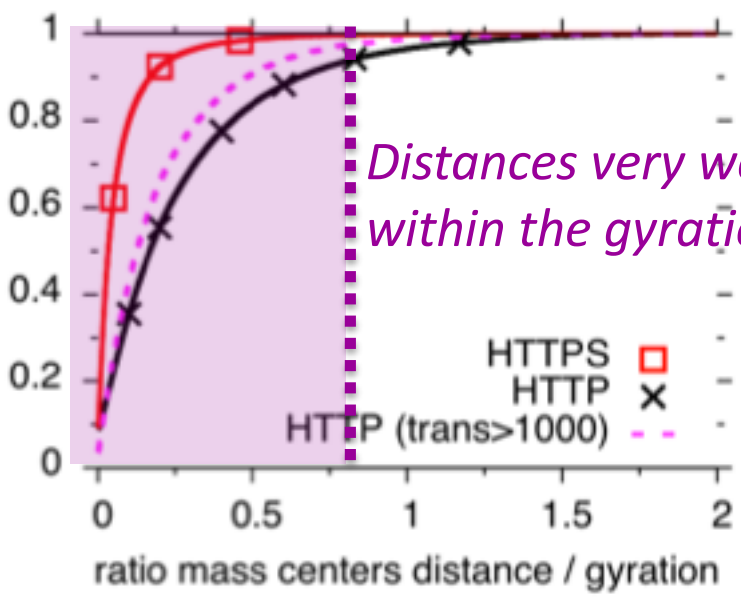
gyration radius is like a standard deviation of the mobility

Is HTTP consumed in the same locations as for the overall traffic?

Recall: the dataset specifies in which sector a transaction took place

1. Approximate users location with towers position
2. We are not interested into tracing paths, but rather the trans for each tower
3. Define a mass center (*m*) as weighted average (with trans served) of towers location
4. How far is m_{HTTP} from the mass center of whole traffic? What about HTTPS?
5. Normalize distances wrt gyration





Distances very well contained within the gyration radius

Is HTTP consumed in the same locations as for the overall traffic?

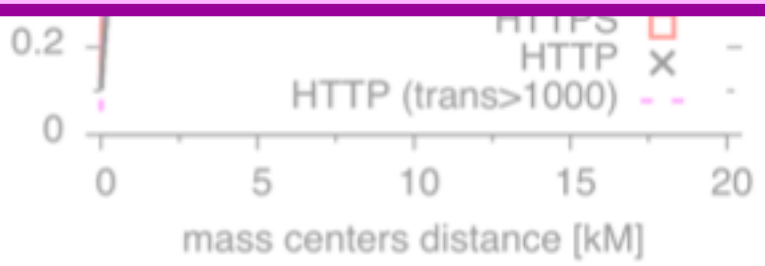
SPACE

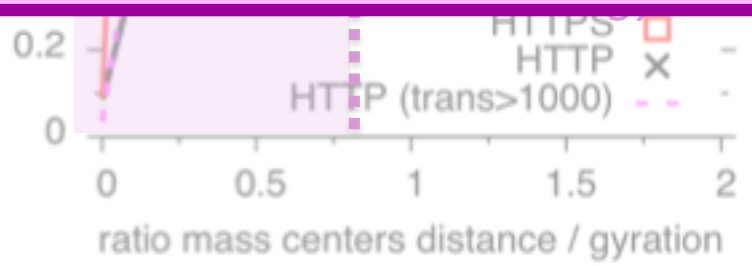
Recall: the dataset specifies in which sector a transaction took place

1. Approximate users location with towers position
2. We are not interested in tracing paths, but rather the trans for each tower

Mind the gap

1. HTTP spatial gap is smaller than for volume and time
2. Hence HTTP traffic alone is still sufficient to understand where users consume content
3. This is (possibly) due to the fact that users do not explicitly select to user HTTP or HTTPS, i.e., mobility is more related to users behavior





Telefónica

More analysis in the paper

- Analysis of other TCP volume
- Characterization of HTTP/HTTPS accessed services
- Day/night patterns
- Characterization of volume consumed per-tower
- Etc.

Future directions

Overall, some ”gaps” impact more than others, and there are entangled relationships that need further analysis

- Extend analysis using longer period of time
- Extend (HTTPS) traffic classification
- Compare HTTP / HTTPS QoE metrics
- Integrate information related to tariffs

Alessandro Finamore …….<alessandro.finamore@telefonica.com>