Analysis and Modeling of Mobile Data Traffic in Mexico City

Eduardo Mucelli Rezende Oliveira[◇,×], Aline Carneiro Viana[×], K. P. Naveen[×], and Carlos Sarraute[*]

[◇]École Polytechnique, France [×]INRIA, France [*]Grandata Labs, Argentina

Introduction

- Worldwide smartphone sales are increasing mobile data traffic.
  - This creates a heavy load on cellular operator networks.
- Understanding mobile data traffic demands is crucial to design data offloading solutions.
- Smartphones provide a powerful and cost effective way to study mobile traffic behaviour on a large scale.

Objectives

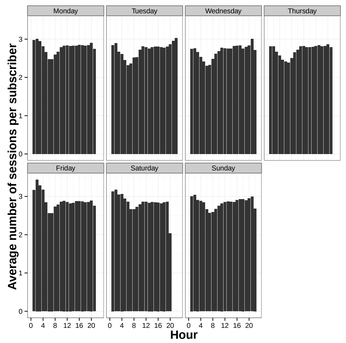
- Analyse **urban mobile data traffic** usage patterns.
- Create a **mobile data traffic simulator** capable of imitating activity patterns for different periods of the day.

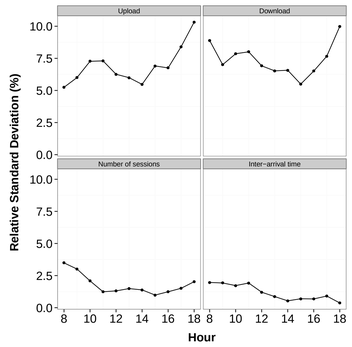
Outline of our Contribution

- Characterize **traffic dynamics** and its **temporal** variability.
- Find a set of **profiles** that best describe users' traffic demands.
- Model usage patterns for different profiles and periods of the day.
- Design and validate a **synthetic trace** generator.

Dataset Analysis

- Anonymized dataset collected by a major operator in Mexico City.
- Data traffic of 6.8 million subscribers.
- 1.05 billion sessions from July 1st to October 31st, 2013.
  - Session information: (1) upload and download volume, (2) session duration, (3) session timestamp.

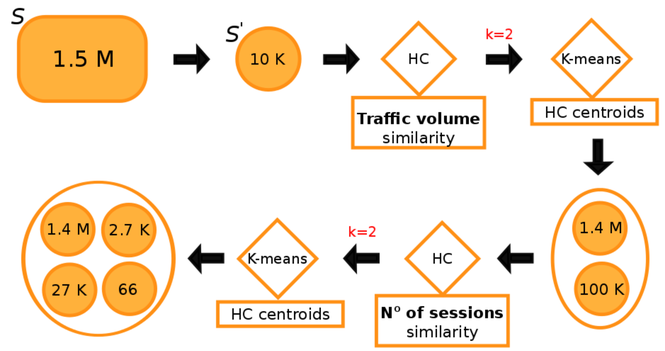




- Hourly dynamics (left) and Relative Standard Deviation (RSD) within the week (right) illustrate the temporal dynamics.
- Parameters from same hours on different days present less variability than parameters on different hours within the same day.

Subscriber Profiling Methodology

- Due to the routinary behavior, we use one day to model traffic behavior.
- Take random sample of subscribers $S' \subset S$.
- Profiling occurs in 4 stages:
  1. Build similarity graph.
  2. Perform Hierarchical Clustering (HC).
  3. Determine best number of cluster (k=2).
  4. Classify users in $S - S'$ using k-means.
- Stages performed in 2 rounds:
  1. First on **traffic volume** similarity graph.
  2. Then on **N° of sessions** similarity subgraphs.

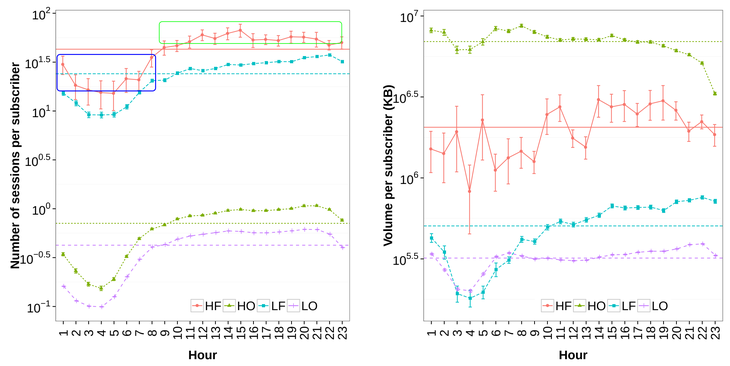


Resulting Subscriber Profiles

- Four profiles are obtained, which describe the typical data traffic of subscribers.

	Light		**H**eavy	
Volume	29 KB to 20 GB		21 GB to 625 GB	
N° of users	1489242		27659	
	Occasional	**F**requent	**O**ccasional	**F**requent
N° of sessions	1 to 278	279 to 8737	1 to 495	538 to 1670
N° of users	1486496	2746	27593	66

- There are **peak** and **non-peak** hours in the traffic demands.



Measurement-driven Traffic Modeling

- Characterize 4 user profiles for peak and non-peak hours.
- Estimate the distributions that **best fit** each **parameter** on each **profile** in **peak** and **non-peak** hours.

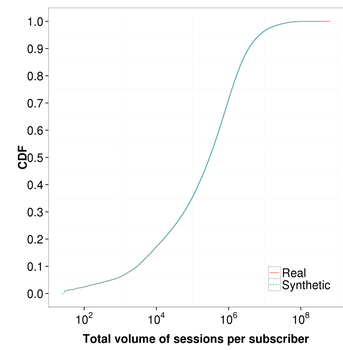
Hour	*Profile*	*Distribution*	*Parameters*
Peak	HO	Gamma	$\alpha = 1.3060$, $\beta = 0.001$, $x_0 = 1$
	HF	Log-normal	$\sigma = 4.0106$, $\mu = 1.2114$, $x_0 = 10$
	LO	Gamma	$\alpha = 1.2799$, $\beta = 0.001$, $x_0 = 0.5$
	LF	Weibull	$k = 0.9173$, $\lambda = 135$, $x_0 = 3.9$
Non-Peak	HO	Gamma	$\alpha = 1.2679$, $\beta = 0.001$, $x_0 = 1$
	HF	Log-normal	$\sigma = 3.8552$, $\mu = 0.9196$, $x_0 = 4.3$
	LO	Gamma	$\alpha = 1.2799$, $\beta = 0.001$, $x_0 = 0.5$
	LF	Log-normal	$\sigma = 4.1174$, $\mu = 1.0291$, $x_0 = 3$

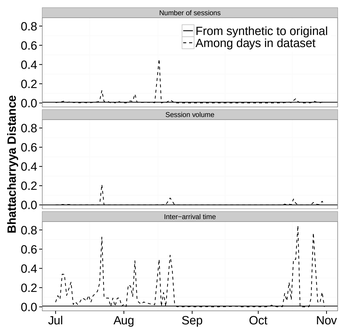
Generation of Synthetic Trace

- Assign users to profiles according to the profiles population.
  - e.g. LO users have 0.97 probability.
- For each user profile and hour:
  - Sample n° of sessions according to the distribution.
  - Sample IAT according to the distribution.
  - Sample volume according to the distribution.

Synthetic Traffic Model Evaluation

- We show the CDFs of total volume per subscriber, for the real trace and the synthetic trace (left).





- We used **Bhattacharyya distance** d to measure similarity between distributions (right).
  - Distances between original day D and the remaining days in dataset (dashed lines).
  - Distance between original day D and synthetic day D' (solid line).
  - We verified that $d(D, D')$ is within the 95% confidence interval of the distances $d(D, E)$ for $E \in \mathbb{D}$ (where $\mathbb{D}$ is the set of days in the dataset).
- The synthetic traffic is **consistent** with the real trace.