Poster: On the Quest for Representative Mobile Datasets

Guangshuo CHEN
INRIA and Ecole Polytechnique, France
guangshuo.chen@inria.fr

Aline C. Viana
INRIA, France
aline.viana@inria.fr *

ABSTRACT

Mobile datasets are often incomplete or have heterogeneous spatiotemporal resolutions, *e.g.* a dataset is often aggregated or in lack of fields. We create a reliable dataset describing mobile data traffic in individual's spatiotemporal view. We focus on individuals having enough geographical information and merge their call records from one dataset with the data traffic records extracted from another dataset. The resulting dataset contains data session records associated with time and location fields.

Categories and Subject Descriptors

C.2.1 [**Computer-Communications Networks**]: Network Architecture and Design—*Wireless communication*

Keywords

Human mobility, Dataset

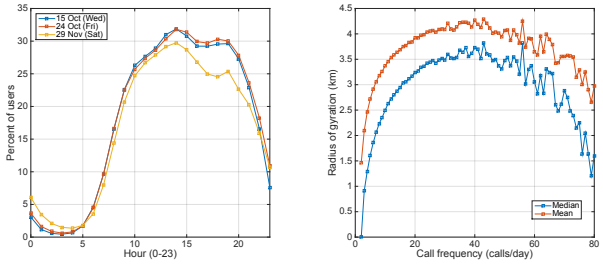
1. INTRODUCTION

We have two anonymized datasets gotten from a major cellular operator in Mexico. The dataset of call traffic, named as *mobility_trace*, contains call detailed records (CDRs) of $437,542$ mobile devices located in the urban area of Mexico city and have the following fields: *origin, destination, time, location* and *duration*. The dataset of data traffic, named as *web_trace*, includes information on data sessions created by devices appearing in the *mobility_trace*. A session record has the following fields: *subscriber, time*, and *volume*.

Note that the *web_trace* lacks mobility information, *i.e.* the location field, which makes impossible to perform any spacial analysis. We have to infer locations of a device at starting time points of sessions by extracting mobility information from the *mobility_trace* where locations only appears when a call is made or received. Nevertheless, people spend much more time using their phones to generate data traffic than to make/receive calls [1]. A large amount of users are expected to have more session records than call records at the same time. Hence we can not simply merging two datasets. Some few users are highly active in terms of calls. We will rely our analysis on the sampling of these users.

2. METHODOLOGY

Our final goal is to create a dataset of session records with geographical information. We attempt to achieve this goal by estimating locations for all sessions. Our approach has three steps: (1) time period selection, (2) user sampling, and (3) location estimation.



(a) (b)

Figure 1: (a) Percentage of users having calls per hour. (b) Call frequency versus (median and mean) radius of gyration.

Time period selection: According to [2], it is a good practice for mobility estimation to use CDRs showing at least 0.5 calls/hour. We see that there are more devices having calls in working hours and evenings than early in the morning (from $8am$ to $11pm$), and very few users having continuously calls after midnight (from $0am$ to $7am$), as shown in Fig.1(a). There are considerable number of users who use their mobile phone in working hours. Considering that we would like to cover the working time period in Mexico (*i.e.* from $8am$ to $6pm$ [1]), we focus our analysis on the time period from $8am$ to $9pm$.

User sampling: We sample users by (1) eliminating ones having an abnormal behavior on making/receiving calls and (2) selecting ones having a certain call frequency in the study time period. We find users who make/receive calls in extreme high frequency, that is considered as an abnormal behavior and may be justified by the presence of mobile calling centers in the dataset. As shown in Fig.1(b), we observe that devices having extremely frequent calls can no longer be distinguished by their routinary mobile behaviors. Therefore we set a probabilistic threshold to eliminate outliers: A device having more than 40 calls a day will be eliminated with 90% of probability. We select users who having at least 0.5 calls/hour [2], for whom we conclude that the time interval between a data session generation and a call is at most 1 hour.

Location estimation: For the sampled users, the interval between a session and a call is at most 1 hour. When generating a session, the device is approximated to be located at where it has a phone call most recently.

After these 3 steps, we finally obtain the dataset having all the fields of *web_trace* and the field of location in $(latitude, longitude)$.

*This work was supported by the EU FP7 ERANET program under grant CHIST-ERA-2012 MACACO.

S3'15, September 11, 2015, Paris, France.
ACM ISBN 978-1-4503-3701-4/15/09.
DOI: http://dx.doi.org/10.1145/2801694.2802147.

3. REFERENCES

[1] E. Mucelli, A. C. Viana, K. P. Naveen, and C. Sarraute, "Measurement-driven mobile data traffic modeling in a large metropolitan area," in *IEEE PerCom*, 2015.

[2] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of Predictability in Human Mobility," *Science*, vol. 327, pp. 1018–1021, Feb. 2010.