D4D-Senegal: The Second Mobile Phone Data for Development Challenge

Yves-Alexandre de Montjoye[1], Zbigniew Smoreda[2],
Romain Trinquart[2], Cezary Ziemlicki[2], Vincent D. Blondel[3]

[1]Media Lab, Massachusetts Institute of Technology, Cambridge, MA

[2]Orange Labs, France

[3]Université catholique de Louvain, Belgium

July 16, 2014

The D4D-Senegal challenge is an open innovation data challenge on anonymous call patterns of Orange's mobile phone users in Senegal. The goal of the challenge is to help address society development questions in novel ways by contributing to the socio-economic development and well-being of the Senegalese population. Participants to the challenge are given access to four mobile phone datasets. This paper describes the three datasets. The datasets are based on Call Detail Records (CDR) of phone calls and text exchanges between more than 9 million of Orange's customers in Senegal between January 1, 2013 to December 31, 2013. The datasets are: (a) antenna-to-antenna traffic for 1666 antennas on an hourly basis, (b) fine-grained mobility data on a rolling 2-week basis for a year with bandicoot behavioral indicators at individual level for about 300,000 randomly sampled users, (3) one year of coarse-grained mobility data at arrondissement level with bandicoot behavioral indicators

at individual level for about 150,000 randomly sampled users

Introduction

There are Big Hopes associated with Big Data: it has been dubbed the oil of the digital economy (*1*), the next big thing in medical care (*2*), and a vital tool for building smart cities (*3*). In science, the availability of large-scale behavioral datasets has even been compared to the invention of the microscope (*4*).

There is little doubt that impressive work has already been done by the computational social science and mobile phone research communities. Metadata has, for example, been used to better understand the propagation of malaria, to monitor poverty (*5, 6*), to analyse human mobility (*7*), and to study the structure of social communities at a national level (*8*). Big Data has, however, to be made more broadly available to further realize its promises. Understanding context remains critical, particularly for a sound interpretation and solution of practical questions. Development economists, urban planners, sociologists, and NGOs need to become familiar with this data. "Inanimate data can never speak for themselves, and we always bring to bear some conceptual framework, either intuitive and ill-formed, or tightly and formally structured, to the task of investigation, analysis, and interpretation" (*9*).

This is why, in 2012, Orange launched the Data For Development challenge in partnership with the University of Louvain and MIT. D4D-Cote d'Ivoire made five months of mobile phone metadata available (*10*). The results were impressive: 260 applications from around the world were submitted to access the data and, after three months, more than 80 research papers had been produced (*11*). These papers covered topics as diverse as optimizing bus routes, analyzing social divisions (*12*), and studying disease containment policies (*13*).

We are now launching, in collaboration with Sonatel Senegal, the second challenge: **D4D-Senegal** (*14*) where selected teams will have access to one year of metadata for up to 300,000 people across Senegal. This paper describes the data pre-processing and the three datasets that will be made available, as well as a set of research questions that have been suggested by local partner organizations. More details and the application to participate in the challenge are available at `http://www.d4d.orange.com`.

Data preprocessing

The Call Detail Records (CDR) have been collected for a year, from January 1 to December 31, 2013. The customer identifiers were anonymized by Sonatel before the data was transferred to Orange Labs who did the preprocessing.

The original dataset contained more than 9 million unique aliased mobile phone numbers. When preparing datasets, we retained only users meeting both of these criteria:

1. users having more than 75% days with interactions per given period (biweekly for the second dataset, yearly for the third dataset)

2. users having had an average of less than 1000 interactions per week. The users with more than 1000 interactions per week were presumed to be machines or shared phones.

For commercial and privacy reasons, we do not release the real geographical coordinates of the site where BTSs, the mobile network antennas, are located. Note that several BTS can be co-located. We assigned a new position to each site uniformly in its Voronoi cell (the region consisting of all points closer to that antenna than to any other) to make it harder to re-identify users (*15*). The `SITE_ARR_LATLON.csv` file contains the new, noisy, latitude and longitude of the site.

For example:

```
site_id,arr_id,lon,lat
1,2,-17.5251,14.74683
2,2,-17.5244,14.74743
3,2,-17.5226,14.7452
4,2,-17.5164,14.74673
```

Datasets

Simply anonymized mobile phone datasets have been shown to be re-identifiable. For instance, it is possible to find a user in a large-scale mobility data using only four spatio-temporal points and coarsening the data only makes it slightly harder (*16*).

To balance the potential of the data being broadly used with the risks of re-identification we provide three sampled and aggregated datasets for this challenge:

- **Dataset 1:** One year of site-to-site traffic for 1666 sites on an hourly basis,
- **Dataset 2:** Fine-grained mobility data (site level) on a rolling 2-week basis with bandicoot behavioral indicators at individual level for about 300,000 randomly sampled users meeting the two criteria mentioned before for each 2 week period,
- **Dataset 3:** One year of coarse-grained (123 arrondissement level) mobility data with bandicoot behavioral indicators at individual level for about 150,000 randomly sampled users meeting the two criteria mentioned before for a year,

Each dataset has been designed to balance utility with privacy, utility beeing the research that can be done with the data while privacy is the potential risk of re-identification of users. Datasets are thus either precise spatially and temporally but limited in the time they span (dataset 2), or aggregated geographically (dataset 3) or across users (dataset 1) but covering a longer period of time. Finally, precomputed indicators are provided to help inform behavioral research. Columns that might help re-identification have been 3-anonymized when binned to remove outliers (*17*).

Note that a fourth dataset of synthetic data will be made available in September and will be described in a future paper.

Individual indicators

Mobility datasets 2 and 3 are supplemented with behavioral indicators from (*18*) computed from metadata using the bandicoot toolbox (*19*).

The indicators we provide are:

- `active_days_callandtext_mean`
- `active_days_callandtext_sem`
- `duration_of_calls_mean_mean`
- `duration_of_calls_mean_sem`
- `entropy_of_contacts_call_mean`
- `entropy_of_contacts_call_sem`

- entropy_of_contacts_text_mean
- entropy_of_contacts_text_sem
- entropy_of_contacts_callandtext_mean
- entropy_of_contacts_callandtext_sem
- entropy_places_callandtext_mean
- entropy_places_callandtext_sem
- interactions_per_contact_callandtext_mean_mean
- interactions_per_contact_callandtext_mean_sem
- interactions_per_contact_call_mean_mean
- interactions_per_contact_call_mean_sem
- interevents_callandtext_mean_mean
- interevents_callandtext_mean_sem
- interevents_call_mean_mean
- interevents_call_mean_sem
- interevents_text_mean_mean
- interevents_text_mean_sem

Places are in this case sites and nocturnal is defined as 7pm to 7am. A full description of the indicators can be found on the bandicoot document in the data repository and the indicator files have been 3-anonymized on binned data on specific columns to remove outliers (*17*).

Dataset 1: Antenna-to-antenna traffic

This dataset contains the traffic between each site for a year.

The files `SET1V_M01.csv` through `SET1V_M12.csv` contain monthly voice traffic between sites and are structured as follow:

- **timestamp:** day and hour considered in format YYYY-MM-DD HH (24 hours format)
- **outgoing_site_id:** id of site the call originated from
- **incoming_site_id:** id of site receiving the call
- **number_of_calls:** the total number of calls between these two sites during this hour
- **total_call_duration:** the total duration of all calls between these two sites during this hour

For example:

```
timestamp, outgoing_site_id, incoming_site_id,...
...number_of_calls, total_call_duration
2013-04-01 00,2,2,7,138
2013-04-01 00,2,3,4,136
2013-04-01 00,2,4,7,121
2013-04-01 00,2,5,13,272
2013-04-30 23,1651,1632,1,3601
2013-04-30 23,1653,575,1,20
2013-04-30 23,1653,1653,2,385
2013-04-30 23,1659,608,1,3601
```

The files `SET1S_M01.csv` through `SET1S_M12.csv` contain monthly text traffic between sites and are structured as follow:

- **timestamp:** day and hour considered in format YYYY-MM-DD HH (24 hours format)

- **outgoing_site_id:** id of site the text originated from
- **incoming_site_id:** id of site receiving the text
- **number_of_sms:** the total number of texts between these two sites during this hour

For example:

```
timestamp, outgoing_site_id, incoming_site_id, number_of_sms
2013-05-01 00,2,12,6
2013-05-01 00,2,14,1
2013-05-01 00,2,21,1
2013-05-01 00,2,28,9
2013-05-31 23,1653,190,2
2013-05-31 23,1653,314,3
2013-05-31 23,1653,367,8
2013-05-31 23,1653,520,1
2013-05-31 23,1653,558,2
```

Note that calls spanning multiple time slots are considered to be in the time slot they started in and only calls or texts between Sonatel customers are taken into account.

The latitude and longitude of the sites is provided in `SITE_ARR_LATLON.csv`.

Dataset 2: Fine-grained mobility

This second dataset contains the trajectories at site level of about 300,000 randomly selected users meeting the two criteria mentioned before over two-week periods. The site locations are provided in `SITE_ARR_LATLON.csv`.

The files `SET2_P01.csv` through `SET2_P25.csv` contain the `user_id`, `timestamp`, and `site_id` for each of the 25 two-week periods. The second digits of the minutes and all the seconds of the timestamps have been replaced with zeros (format YYYY-MM-DD HH:M0:00) For each period, a new sample of about 300,000 users was selected and their `user_id` scrambled. Note that this mean that even if a user were to appear in two periods, he would have a different id, and vice versa, the same id in two periods does not mean that it is the same person.

For example:

```
user_id,timestamp,site_id
1,2013-03-18 21:30:00,716
1,2013-03-18 21:40:00,718
1,2013-03-19 20:40:00,716
1,2013-03-19 20:40:00,716
1,2013-03-19 20:40:00,716
1,2013-03-19 20:40:00,716
1,2013-03-19 21:00:00,716
1,2013-03-19 21:30:00,718
1,2013-03-20 09:10:00,705
1,2013-03-21 13:00:00,705
```

The indicators are computed, for every user, over the course of the two week, and are available in the files `INDICATORS_SET2_P01.csv` through `INDICATORS_SET2_P25.csv`.

Dataset 3: Coarse-grained mobility

This third dataset contains the trajectories at arrondissement level of 146,352 randomly selected users meeting the two criteria mentioned before on a yearly basis.

```
user_id,timestamp,arrondissement_id
37509,2013-01-29 15:00:00,3
84009,2013-01-14 07:00:00,3
84009,2013-01-14 07:00:00,3
84009,2013-01-14 07:00:00,3
80150,2013-01-27 16:50:00,3
52339,2013-01-09 19:50:00,48
52339,2013-01-06 17:50:00,48
52339,2013-01-13 15:40:00,48
52339,2013-01-03 19:00:00,48
52339,2013-01-07 01:30:00,48
```

The files `SET3_M01.csv` through `SET3_M12.csv` contain the `user_id`, `timestamp`, and `arrondissement_id` month by month. The second digits of the minutes and all the seconds of the timestamps have been replaced with zeros (format YYYY-MM-DD HH:M0:00)

The indicators are computed, for every user, on a monthly basis. They are available in the files INDICATORS_SET3_M01.csv through INDICATORS_SET3_M12.csv.

The arrondissement shapefile is provided (SHAPEFILE_SENEGAL.zip) as well as a summary table (SENEGAL_ARR.csv).

The summary table contains:

- **ARR_ID:** the arrondissement_id
- **REG:** the name of the region
- **DEPT:** the name of the department
- **ARR:** the name of the arrondissement

For example:

```
ARR_ID,REG,DEPT,ARR
1,DAKAR,DAKAR,PARCELLES ASSAINIES
2,DAKAR,DAKAR,ALMADIES
3,DAKAR,DAKAR,GRAND DAKAR
4,DAKAR,DAKAR,DAKAR PLATEAU
5,DAKAR,GUEDIAWAYE,GUEDIAWAYE
6,DAKAR,PIKINE,PIKINE DAGOUDANE
```

Contextual data

- GIS shapefiles for Senegal: Administrative divisions of Senegal shapefiles provided by the ADSN are included in the data package SHAPEFILE_SENEGAL.zip
- Weather data: http://www.wunderground.com/weather-forecast/Senegal.html
- Demographic and socio-economic data: http://donnees.ansd.sn/en/BulkDownload
- Import/Export data: http://atlas.media.mit.edu/explore/tree_map/hs/export/sen/all/show/2010/
- More references at: http://www.d4d.orange.com/en/partners-resources/resources

Research collaboration

We strongly encourage developing a scientific collaboration between challenge participants and local teams. In both its production and interpretation, data is always the result of contingent and contested social practices, the knowledge of national political, cultural and socio-economic context is essential to ask sound research questions and to develop valid results interpretations. In order to facilitate this collaboration, the D4D team provide you with a collaborative space on the Sparkboard platform `http://d4d.sparkboard.com`, feel free to announce your project and to specify what kind of competencies and collaboration you would be interested in.

References and Notes

1. A. Pentland (2011).
2. R. Steinbrook, *New England Journal of Medicine* **358**, 1653 (2008).
3. MIT City Science, `http://cities.media.mit.edu/`. [Online; accessed 16-July-2014].
4. D. Lazer, *et al.*, *Science (New York, NY)* **323**, 721 (2009).
5. A. Wesolowski, *et al.*, *Science* **338**, 267 (2012).
6. N. Eagle, M. Macy, R. Claxton, *Science* **328**, 1029 (2010).
7. M. C. Gonzalez, C. A. Hidalgo, A.-L. Barabasi, *Nature* **453**, 779 (2008).
8. S. Sobolevsky, *et al.*, *PloS one* **8**, e81707 (2013).
9. P. Gould, *Annals of the Association of American Geographers* **71**, 166 (1981).
10. V. D. Blondel, *et al.*, *arXiv preprint arXiv:1210.0137* (2012).
11. D4D-Cote d'Ivoire – Book of abstract, `http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf`. [Online; accessed 16-July-2014].
12. M. Berlingerio, *et al.*, *Machine Learning and Knowledge Discovery in Databases* (Springer, 2013), pp. 663–666.

13. A. Lima, M. De Domenico, V. Pejovic, M. Musolesi, *arXiv preprint arXiv:1306.4534* (2013).

14. D4D-Senegal, `http://www.d4d.orange.com/`. [Online; accessed 16-July-2014].

15. Y.-A. de Montjoye, Privacy tools, `https://github.com/yvesalexandre/privacy-tools`. [Online; accessed 16-July-2014].

16. Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel, *Nature SRep* **3** (2013).

17. L. Sweeney, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10**, 571 (2002).

18. Y.-A. de Montjoye, J. Quoidbach, F. Robic, A. S. Pentland, *Social Computing, Behavioral-Cultural Modeling and Prediction* (Springer, 2013), pp. 48–55.

19. Bandicoot, a Python toolbox for mobile phone metadata., `http://bandicoot.mit.edu/`. [Online; accessed 16-July-2014; contact:bandicoot@media.mit.edu].

20. The authors would like to thanks Kevin Mustelier and Luc Rocher for their help with the Bandicoot toolbox