Datathon

SIMPLON.CO

mit media lab

Overview of the data

Zbigniew Smoreda, Cezary Ziemlicki
SENSE/Orange Labs

School / Datathon

Senegal data

- The datasets are based on Call Detail Records (CDR) of phone calls and text exchanges between more than 9 million of Sonatel's customers in Senegal
- collected for the period January 1 to December 31, 2013

datasets constructed

1. antenna-to-antenna traffic for 1,666 antenna towers on an hourly basis
2. fine-grained mobility data (antenna level) on a rolling 2-week basis with about 300,000 randomly sampled users
3. one year coarse-grained mobility data at 123 arrondissement level for about 150,000 randomly sampled users

2 & 3 with bandicoot behavioral indicators at individual level

one year of site-to-site traffic (voice)

for 1666 sites on an hourly basis

- 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 (sec)

one year of site-to-site traffic (voice)

for 1666 sites on an hourly basis

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

timestamp: day and hour consic
(24 hours fo
outgoing_site_id: id of site the c
incoming_site_id: id of site rece
number_of_calls: the total num
during this h
total_call_duration: the total du
two sites du

```
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
```

School / Datathon

one year of site-to-site traffic (sms)

(for 1666 sites on an hourly basis)

- 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

one year of site-to-site traffic (sms)

(for 1666 sites on an hourly basis)

- 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 (24 hours format)

outgoing_site_id: id of site the text c

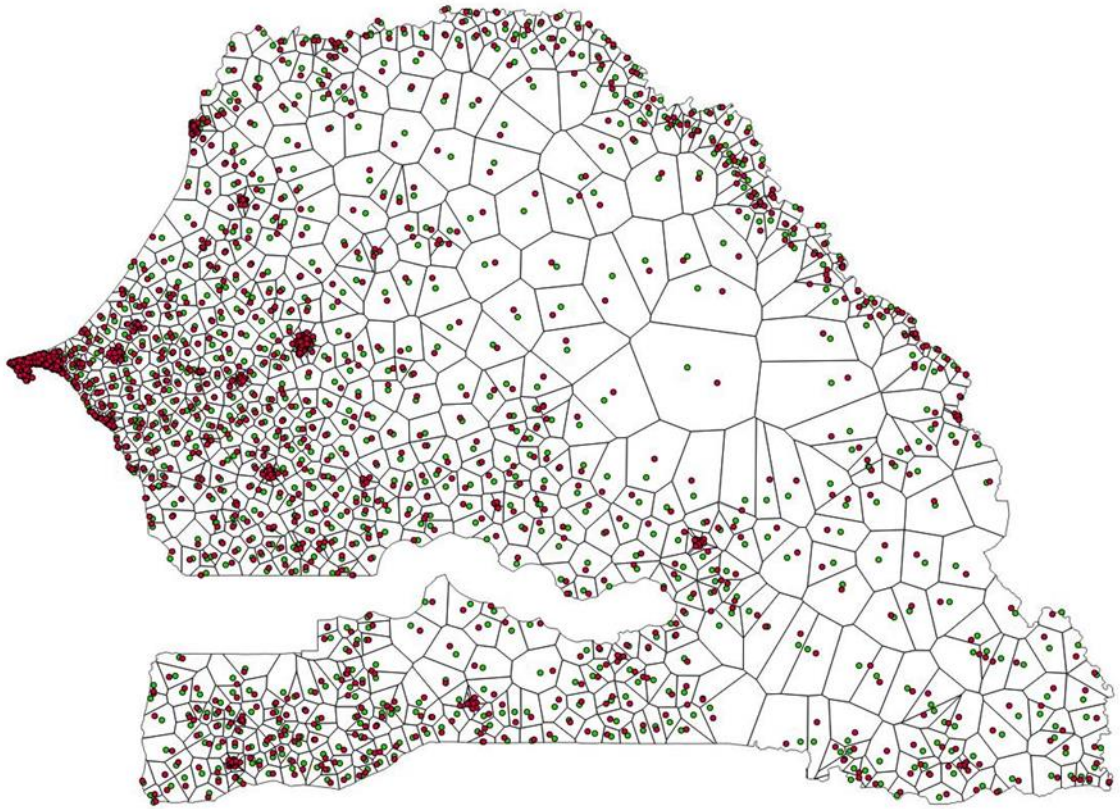
incoming_site_id: id of site receiving

number_of_sms: the total number c ces during this hour

```
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
```

antenna-towers coordinates

- The SITE_ARR_LATLON.csv file contains the **latitude and longitude of the site** (and the arrondissement code)



```
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
```

fine-grained mobility data (site level)

for 1666 sites

- The files SET2_P01.csv through SET2_P25.csv contain the **25 two-week periods** for 300,000 randomly selected users detailed records:

user_id: selected user random id

timestamp: 24 h format YYYY-MM-DD HH:M0:00

site_id: id of the antenna site

Note: The second digits of the minutes and all the seconds of the timestamps have been replaced with zeros;
Only users having more than 75% days with interactions in the considered period can be selected.

fine-grained mobility data (site level)

for 1666 sites

- The files SET2_P01.csv through SET2_P25.csv contain the **25 two-week periods** for 300,000 randomly selected users detailed records:

user_id: selected user random id

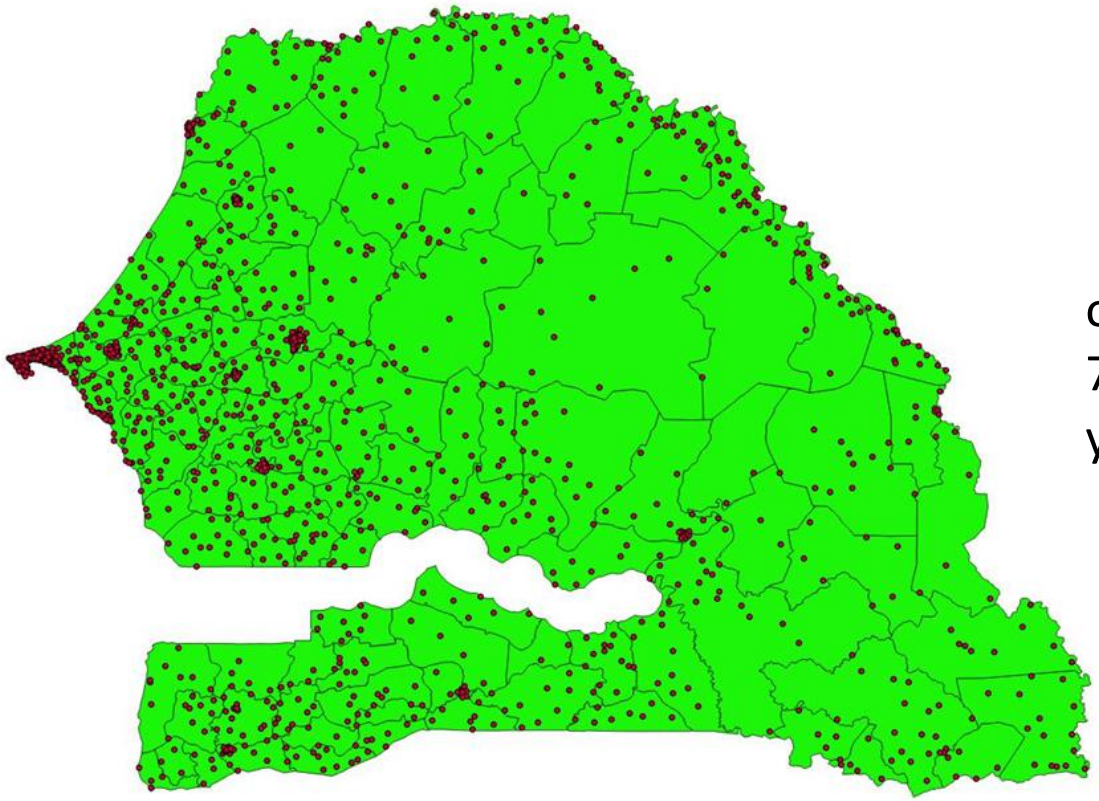
timestamp: 24 h format YYYY-MM-DD

site_id: id of the antenna site

```
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
```

coarse-grained individual mobility data at arrondissement level

- **One year** of coarse-grained (**123 arrondissement level**) mobility data for about 150,000 randomly sampled users



only users having more than 75% days with interactions yearly were selected

School / Datathon

coarse-grained individual mobility data at arrondissement level

- The files SET3_M01.csv through SET3_M12.csv contain data of 150,000 randomly sampled users month by month:

  user_id: selected user random id

  timestamp: 24 h format YYYY-MM-DD HH:M0:00

  arrondissement_id: arrondissement code

Note: The second digits of the minutes and all the seconds of the timestamps have been replaced with zeros

coarse-grained individual mobility data at arrondissement level

- The files SET3_M01.csv through SET3_M12.csv contain data of 150,000 randomly sampled users month by month:

user_id: selected user random ic
timestamp: 24 h format YYYY-MI
arrondissement_id: arrondissem

```
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
```

geographic information

- The arrondissement shapefile is provided (SHAPEFILE_SENEGAL.zip) as well as a summary table (senegal_arr_centroids.csv) with geometrical center of the arrondissement coordinates, its code, and the names of region, department, municipality/city and arrondissement

```
Centroid_Lon,Centroid_Lat,arr_id,reg,dept,cav,arr
-17.446025848,14.747966068,1,DAKAR,DAKAR,VILLE DE DAKAR,PARCELLES ASSAINIES
-17.485963486,14.736288483,2,DAKAR,DAKAR,VILLE DE DAKAR,ALMADIES
-17.440683318,14.714470627,3,DAKAR,DAKAR,VILLE DE DAKAR,GRAND DAKAR
-17.449171084,14.681392965,4,DAKAR,DAKAR,VILLE DE DAKAR,DAKAR PLATEAU
```

supplementary individual indicators

- Mobility datasets (2 & 3) are supplemented with **behavioral indicators** computed from CDRs using the bandicoot toolbox
- The indicators are computed, for every user:
  - for fine-grained mobility over the course of the two weeks - INDICATORS_SET2_P01.csv through INDICATORS_SET2_P25.csv
  - on a monthly basis for coarse-grained mobility data - INDICATORS_SET3_M01.csv through INDICATORS_SET3_M12.csv

behavioral indicators
(http://bandicoot.mit.edu/docs/)

- active_days_callandtext_mean (user's active days – mean)
- active_days_callandtext_sem (user's active days – standard error of the mean)
- 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
- …..

Virtual machines

- note the data files are compressed to save workspace

```
datathon2@vc1-vm1:~$ ll /data/
total 24
drwxr-xr-x  6 root root 4096 Mar 27 15:37 ./
drwxr-xr-x 23 root root 4096 Mar 23 10:54 ../
drwxr-xr-x  2 root root 4096 Mar 23 10:55 ContextData/
drwxr-xr-x  2 root root 4096 Mar 27 12:23 SET1/
drwxr-xr-x  2 root root 4096 Mar 27 15:41 SET2/
drwxr-xr-x  2 root root 4096 Mar 27 12:43 SET3/
```

```
datathon2@vc1-vm1:~$ ll -h /data/ContextData/
total 1.4M
drwxr-xr-x 2 root root 4.0K Mar 23 10:55 ./
drwxr-xr-x 6 root root 4.0K Mar 27 15:37 ../
-rw-r--r-- 1 root root 150K Mar 23 10:55 bandicoot_v01.pdf
-rw-r--r-- 1 root root 200K Mar 23 10:55 D4D_Senegal.pdf
-rw-r--r-- 1 root root 9.9K Mar 23 10:55 senegal_arr_centroids.csv
-rw-r--r-- 1 root root 3.6K Mar 23 10:55 SENEGAL_ARR_V2.csv
-rw-r--r-- 1 root root 994K Mar 23 10:55 Shapefile_Senegal_V2.zip
-rw-r--r-- 1 root root  46K Mar 23 10:55 SITE_ARR_LONLAT.CSV
```

see also: http://www.d4d.orange.com/en/partners-resources/resources

Virtual machines

datathon2@vc1-vm1:~$ ll -h /**data/SET1**
total 8.8G
drwxr-xr-x 2 root root 4.0K Mar 27 12:23 ./
drwxr-xr-x 6 root root 4.0K Mar 27 15:37 ../
-rw-r--r-- 1 root root 170M Mar 23 10:56 SET1S_01.CSV.gz
-rw-r--r-- 1 root root 144M Mar 23 10:56 SET1S_02.CSV.gz
-rw-r--r-- 1 root root 153M Mar 23 10:56 SET1S_03.CSV.gz
…
-rw-r--r-- 1 root root 513M Mar 23 11:01 SET1V_01.CSV.gz
-rw-r--r-- 1 root root 471M Mar 23 11:01 SET1V_02.CSV.gz
-rw-r--r-- 1 root root 521M Mar 23 11:02 SET1V_03.CSV.gz
….

datathon2@vc1-vm1:~$ ll -h /**data/SET2**
total 3.6G
drwxr-xr-x 2 root root 4.0K Mar 27 15:41 ./
drwxr-xr-x 6 root root 4.0K Mar 27 15:37 ../
-rw-r--r-- 1 root root 936 Mar 27 15:37 INDICATORS_SET2_HEADERS.CSV
-rw-r--r-- 1 root root 20M Mar 27 15:37 INDICATORS_SET2_P01.CSV.gz
-rw-r--r-- 1 root root 20M Mar 27 15:37 INDICATORS_SET2_P02.CSV.gz
-rw-r--r-- 1 root root 20M Mar 27 15:37 INDICATORS_SET2_P03.CSV.gz
…
-rw-r--r-- 1 root root 121M Mar 27 15:38 SET2_P01.CSV.gz
-rw-r--r-- 1 root root 117M Mar 27 15:38 SET2_P02.CSV.gz
-rw-r--r-- 1 root root 120M Mar 27 15:38 SET2_P03.CSV.gz
….

School / Datathon

Virtual machines

```
datathon2@vc1-vm1:~$ ll -h /data/SET3
total 2.1G
drwxr-xr-x 2 root root 4.0K Mar 27 12:43 ./
drwxr-xr-x 6 root root 4.0K Mar 27 15:37 ../
-rw-r--r-- 1 root root  287 Mar 23 11:06 INDICATORS_SET3_HEADERS.CSV.gz
-rw-r--r-- 1 root root 9.8M Mar 23 11:06 INDICATORS_SET3_M01.CSV.gz
-rw-r--r-- 1 root root 9.7M Mar 23 11:06 INDICATORS_SET3_M02.CSV.gz
-rw-r--r-- 1 root root 9.8M Mar 23 11:06 INDICATORS_SET3_M03.CSV.gz
…
-rw-r--r-- 1 root root 152M Mar 23 11:08 SET3_M01.CSV.gz
-rw-r--r-- 1 root root 135M Mar 23 11:08 SET3_M02.CSV.gz
-rw-r--r-- 1 root root 153M Mar 23 11:09 SET3_M03.CSV.gz
-rw-r--r-- 1 root root 150M Mar 23 11:09 SET3_M04.CSV.gz
-rw-r--r-- 1 root root 162M Mar 23 11:09 SET3_M05.CSV.gz
-rw-r--r-- 1 root root 157M Mar 23 11:10 SET3_M06.CSV.gz
…
```

- note the data files are compressed to save workspace

merci