Inference attacks on location data

Sébastien Gambs
Université du Québec à Montréal (UQAM), Canada

gambs.sebastien@uqam.ca

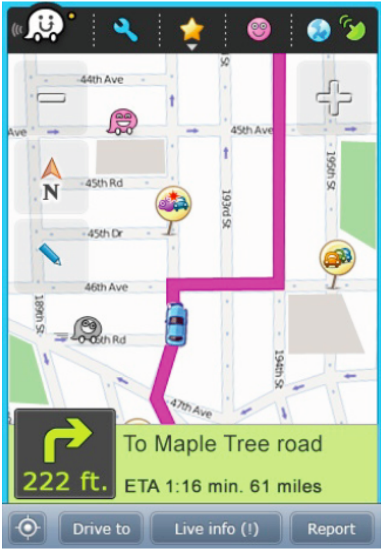
17 July 2017

Location-based services (LBSs)

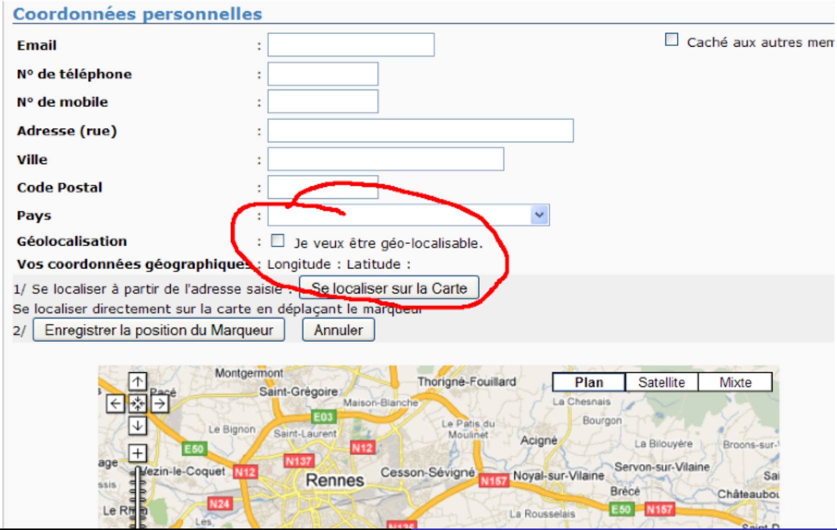
- Personalize the service provided to the user according to his current position.
- Example :

- Main types of LBS :
  1. LBS depending only from the individual position of the user.
  2. Collaborative LBS whose global output is a function of the locations of many users.
- Non-interactive scenario : sanitization of location data.

Example of LBS: Collaborative traffic monitoring



Location, a new type of personal data (INRIA Alumni, old version)



Coordonnées personnelles

□ Caché aux autres mem

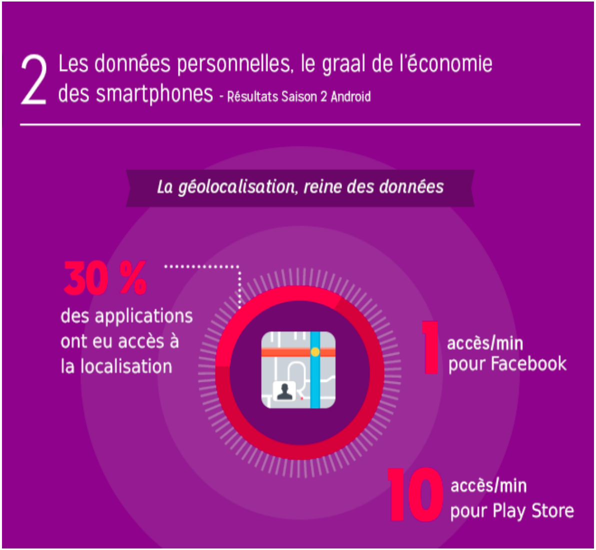
Email	:	
N° de téléphone	:	
N° de mobile	:	
Adresse (rue)	:	
Ville	:	
Code Postal	:	
Pays	:	
Géolocalisation	:	□ Je veux être géo-localisable.
Vos coordonnées géographiques	:	Longitude : Latitude :

1/ Se localiser à partir de l'adresse saisie : Se localiser sur la Carte

Se localiser directement sur la carte en déplaçant le marqueur

2/ Enregistrer la position du Marqueur Annuler

Location, a data massively collected



2 Les données personnelles, le graal de l'économie des smartphones - Résultats Saison 2 Android

La géolocalisation, reine des données

30 % des applications ont eu accès à la localisation

1 accès/min pour Facebook

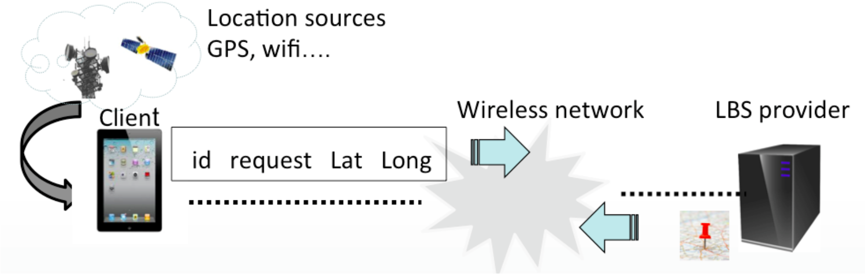
10 accès/min pour Play Store

What the General Data Protection Regulation says about anonymized data

> *"To ascertain whether means are reasonably likely to be used to identify the individual, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration both available technology at the time of the processing and technological development."*

Consequence : evaluation of risk of de-anonymization should take into account the ressources needed to conduct the re-identification and should be done on a regular basis.

System model of LBS



(from Maria Louisa Damiani)

- Possible approach : provide users with the possibility to define rules specifying with whom, how and at which level of granularity their location data is shared.
- Trust issue : the LBS provider has to be trusted in respecting and enforcing these rules.

Adversary model

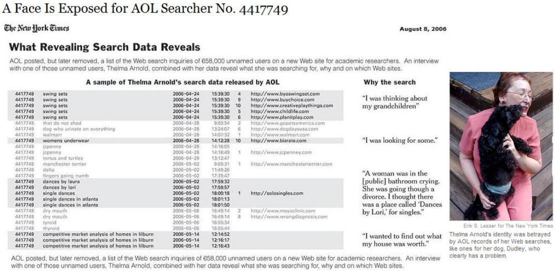
- To provide strong privacy guarantees the LBS provider should also be considered as a possible adversary (in addition of an external attacker).
- Main reason : Once the location data has been collected it is very difficult to control how it is used and disseminated.
- Main privacy issue : richness of the data in terms of its *inference potential* (points of interests and their semantics, mobility prediction, inference of demographic attributes).

Types of location data

- Abstract example of location data:
  $< id, coordinates, timestamp, other\_information >$
- More concretely:
  - *id* = identifier of the user or a device that belongs to him,
  - *coordinates* = latitude and longitude (ex: GPS), identifier of an area (ex: cellular antenna) or a particular place (ex: name of a subway station),
  - *timestamp* = time and day,
  - *other_information* = strength of the signal, estimation of the uncertainty on the position, . . .
- Remark : the collect of mobility can be *frequent* (ex: each minute) or *sporadic* (ex: when the user performs a geolocated query).

Pseudonymization is not an alternative to anonymization

Replacing the name of a person by a pseudonym $\not\Rightarrow$ preservation of the privacy of this individual



A Face Is Exposed for AOL Searcher No. 4417749

The New York Times

August 8, 2006

What Revealing Search Data Reveals

AOL posted, but later removed, a list of the Web search inquiries of 658,000 unnamed users on a new Web site for academic researchers. An interview with one of those unnamed users, Thelma Arnold, combined with her data reveal what she was searching for, why and on which Web sites.

A sample of Thelma Arnold's search data released by AOL

4417749	swing sets	2006-04-24	15:39:30	4	http://www.byoswingset.com
4417749	swing sets	2006-04-24	15:39:30	9	http://www.buychoice.com
4417749	swing sets	2006-04-24	15:39:30	10	http://www.creativeplaythings.com
4417749	swing sets	2006-04-24	15:39:30	5	http://www.childlife.com
4417749	swing sets	2006-04-24	15:39:30	6	http://www.planitplay.com
4417749	that do not shed	2006-04-28	9:05:54	2	http://www.gopetsamerica.com
4417749	dog who urinate on everything	2006-04-28	13:24:07	6	http://www.dogdaysusa.com
4417749	walmart	2006-04-28	14:07:32	1	http://www.walmart.com
4417749	womens underwear	2006-04-28	14:12:28	10	http://www.bizrate.com
4417749	jcpenny	2006-04-28	14:16:05		
4417749	jcpenny	2006-04-28	14:16:49	1	http://www.jcpenney.com
4417749	tortus and turtles	2006-04-29	13:12:47		
4417749	manchester terrier	2006-05-02	9:05:31	1	http://www.manchesterterrier.com
4417749	delta	2006-05-02	11:49:26		
4417749	fingers going numb	2006-05-02	17:35:47		
4417749	dances by laura	2006-05-02	17:59:32		
4417749	dances by lori	2006-05-02	17:59:57		
4417749	single dances	2006-05-02	18:00:18	1	http://solosingles.com
4417749	single dances in atlanta	2006-05-02	18:01:13		
4417749	single dances in atlanta	2006-05-02	18:01:50		
4417749	dry mouth	2006-05-06	16:49:14	2	http://www.mayoclinic.com
4417749	dry mouth	2006-05-06	16:49:14	8	http://www.wrongdiagnosis.com
4417749	tyroid	2006-05-06	16:55:34		
4417749	thyroid	2006-05-06	16:55:44		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:14:52		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:16:17		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:16:43		

Why the search

"I was thinking about my grandchildren"

"I was looking for some."

"A woman was in the [public] bathroom crying. She was going though a divorce. I thought there was a place called 'Dances by Lori,' for singles."

"I wanted to find out what my house was worth."

AOL posted, but later removed, a list of the Web search inquiries of 658,000 unnamed users on a new Web site for academic researchers. An interview with one of those unnamed users, Thelma Arnold, combined with her data reveal what she was searching for, why and on which Web sites.

Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

(Extract from an article from the New York Times, 6 August 2006)

Same phenomenon is true for location data. Example: if the granularity is too small, the pair home-work becomes unique for a large fraction of the population (Colle and Kartridge 09).

Inference attack

- Joint work with Marc-Olivier Killijian (LAAS-CNRS) and Miguel Núñez del Prado (now Universidad del Pacifico, previously LAAS-CNRS).
- Inference attack : the adversary takes as input a location dataset (and possibly some background knowledge) and tries to infer some personal information regarding individuals contained in the dataset.
- Main objective : to quantify the privacy risks linked to the disclosure of location data.
- We may not even be able to model this *a priori* knowledge.
- Remark: maybe my data is private today but it may not be so in the future due to the public release of some other data.

Possible objectives of an inference attack on location data

1. Identification of important places, called *Point of Interests* (POI), characterizing the interests of an individual.

- Example: home, place of work, gymnasium, political headquarters, medical center, . . .

2. Prediction of the movement patterns of an individual, such as his past, present and future locations.
3. Linking the records of the same individual contained in the same dataset or in different datasets (either anonymized or under different pseudonyms).
4. Discover social relations between individuals.

- Example: people that are in the vicinity of each other on a frequent basis.

5. Prediction of demographic attributes.

Identification of home and place of work

Suppose that you have access to the GPS traces of the car of an individual in which the name of the person has been replaced by a pseudonym randomly generated.

Heuristic to identify the home :

- Choose the last stop before midnight.

Heuristic to identify the place of work :

- Choose the most “stable” location during the day.

Reverse geocoding : maps the coordinates of a location to a physical address.

⇒

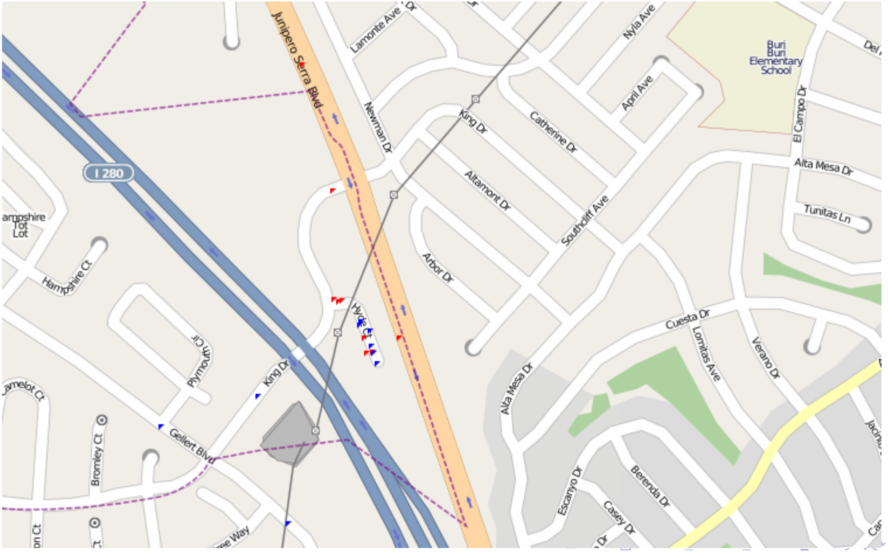
Yellow Pages : associate a physical address with a list of possibles candidates.

Identification of POIs through clustering algorithm (SPRINGL'10)

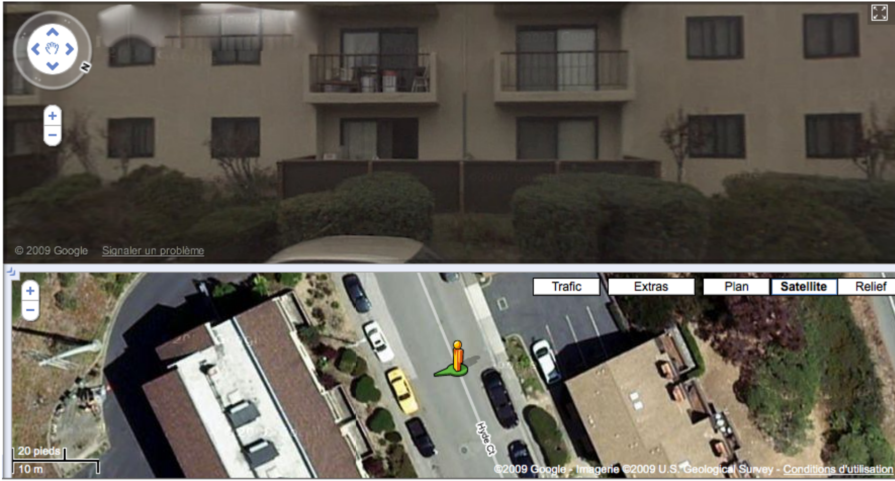
- Clustering : form of unsupervised learning that aims at grouping together objects that are similar (*intra-similarity*) while putting in separate clusters objects that are different (*inter-dissimilarity*).
- Inference attack :
  1. Delete all mobility traces in which the person is in movement.
  2. Run a clustering algorithm on the remaining traces in order to discover significant clusters.
  3. Return as POI the median of each cluster.

  Validation issue : how to evaluate the quality of the POIs returned if we do not have access to the "ground truth"?

Identification of the house of a taxi



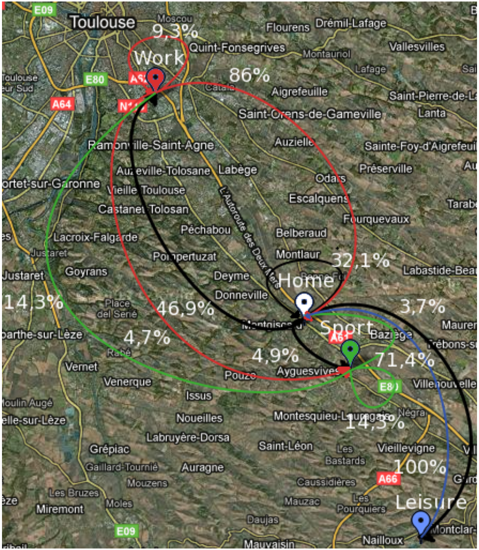
Identification of the house of a taxi (view from GoogleMaps and StreetView)



Mobility Markov chain (TDP'11)

- Objective : to represent in a compact way the mobility behaviour of an individual.
- The states of the chain are POIs and a transitions represents the probability from moving from one POI to another.
- Construction :
  - Remove all moving traces.
  - From the resulting traces, extract the POIs by running a clustering algorithm.
  - Label each trace with the corresponding POI and compute the transitions probabilities.
- Temporal variant of the model (DYNAM'11): decompose the time into slices, the label of a stage corresponds to POI/time slice.

Example of mobility Markov chain



Understanding the semantic of a POI through the structure of the mobility model

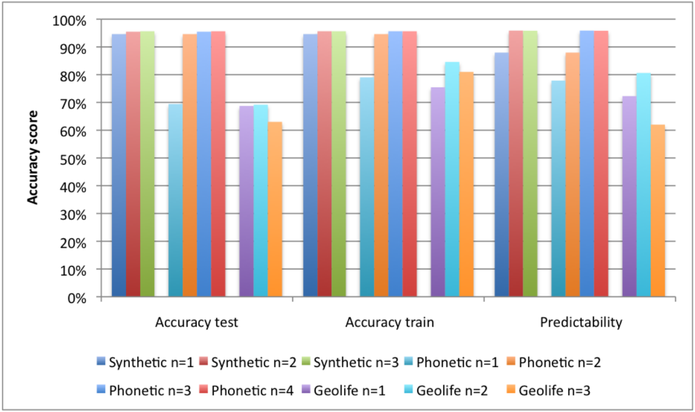
- Main idea : the structure of the mobility Markov chain gives information on the semantic of a particular POI.
- Example 1 : “home” is generally the state of the Markov chain that has the highest number of incoming transitions.
- Can also be inferred by finding the POI that has the highest probability in the stationary vector.
- Example 2 : “work” is in general of the state of the Markov chain that is the ending point of the transition leaving from “home” having the highest probability.

Predicting the next location (MPM'12)

- Prediction technique : from the actual location, choose te transition leaving from this POI that has the highest probability and predicts the corresponding POI.
- Evaluation method : splitting of the mobility traces between a training set and a testing set (50%-50%).
- The mobility Markov chain is learnt from the training set and his prediction rate is evaluated in the testing set.
- Variant of the method : to remember the *n* last visited states (instead of simply the current one).
- Example : a user has visited "work" and then "supermarket", which POI is the one visited next by the user?

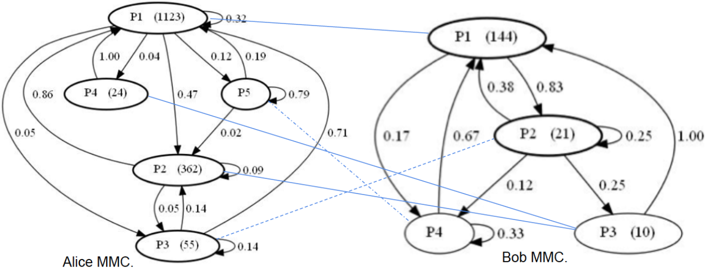
Experimental results

- The prediction method was tested on 3 mobility datasets (synthetic, Phonetic, Geolife) with n varying between 1 and 3 (best prediction rate obtained for $n = 2$).
- Results : success rate of the prediction between 70 and 95%.

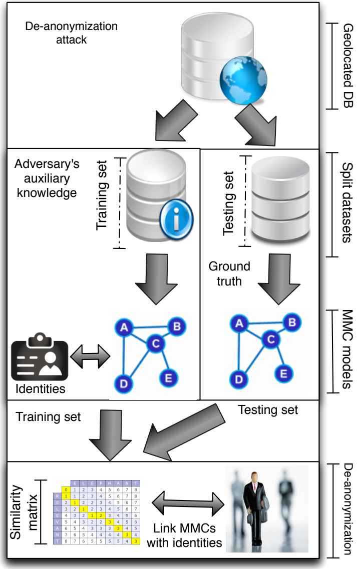


De-anonymization attack via MMC (Trustcom'13)

- Objective : find an individual in an anonymized mobility dataset.
- Assumption : the adversary has been able to observe in the past the mobility of the some individuals present in the dataset.
- Main idea : to compute a distance metric between 2 MMCs quantifying the difference between two mobility behaviours.



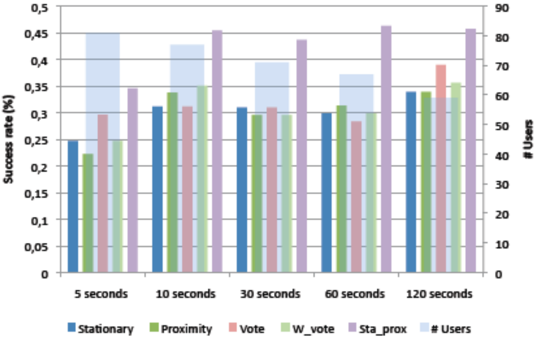
Overview of the de-anonymization attack



		E	L	E	P	H	A	N	T
	0	1	2	3	4	5	6	7	8
R	1	1	2	3	4	5	6	7	8
E	2	1	2	2	3	4	5	6	7
L	3	2	1	2	3	4	5	6	7
E	4	3	2	1	2	3	4	5	6
V	5	4	3	2	2	3	4	5	6
A	6	5	4	3	3	3	3	4	5
N	7	6	5	4	4	4	4	3	4
T	8	7	6	5	5	5	5	4	3

De-anonymization attack via MMC

- Design of different distance metrics (geometrical, topological, logical) between MMCs and different way to combine the predictor



- Best de-anonymization rate : 45% (by combining 2 pred.).
- Remark : the possibility of repeatedly de-anonymizing users is much more damaging to privacy that showing the uniqueness of the characteristics of an individual at a particular occasion.

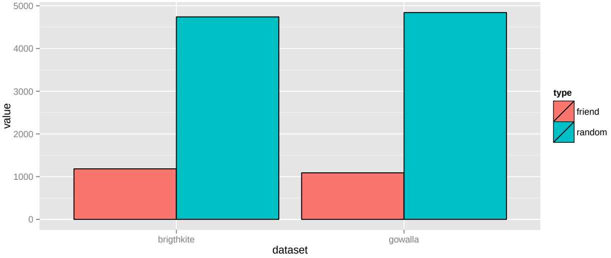
Privacy analysis of large geolocated datasets (HPDIC'13)

- Main issue : to be able to perform a privacy analysis on large scale datasets.
- Approach taken : develop implementation of the inference attacks based on the *MapReduce paradigm* (joint work with Izabela Moïse).

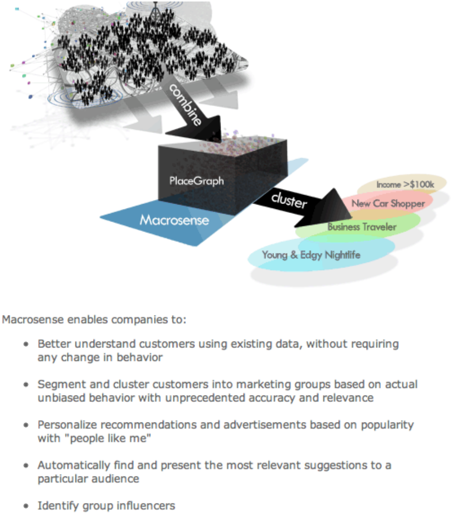
- Implementation based using Hadoop, the open-source version of MapReduce.
- Can be used to analyze large scale datasets in a matter of minutes.

Reconstruction of the social graph (work in progress)

- Main objective : to be able to predict that two persons share a social link based on their mobility data.
- Example : predicting that two persons share a social link based on their check-ins in Foursquare.
- Approach taken : create a predictor for a social link based on the distance between mobility models.



Use of inference attacks for profiling



Macrosense enables companies to:

- Better understand customers using existing data, without requiring any change in behavior
- Segment and cluster customers into marketing groups based on actual unbiased behavior with unprecedented accuracy and relevance
- Personalize recommendations and advertisements based on popularity with "people like me"
- Automatically find and present the most relevant suggestions to a particular audience
- Identify group influencers

Sense Networks, acquired by YP (Yellow Pages) in January 2014

Defining and quantifying location privacy

- Back to the fundamental question : what does it mean to have a "good" location privacy?
- To be hidden inside a crowd gathered in a small area?

- To be alone in a desert?
- To have a behavior indistinguishable from those of a non-negligible number of other individuals?
- To be unlinkable between different positions?
- Proposed answer : to prevent the inference of sensitive information from the location data revealed (rather than focusing on protecting the location itself).

Example of Call Details Records

One year after the 'Data For Development Côte d'Ivoire Challenge', Sonatel and the O
Group are now launching a 'D4D Challenge' in Senegal. For this purpose, they are ma
available to the scientific community, in the exclusive context of the D4D Competition,
statistical databases and samples extracted from the mobile network management sig

This is the end !

Thanks for your attention.
Questions?