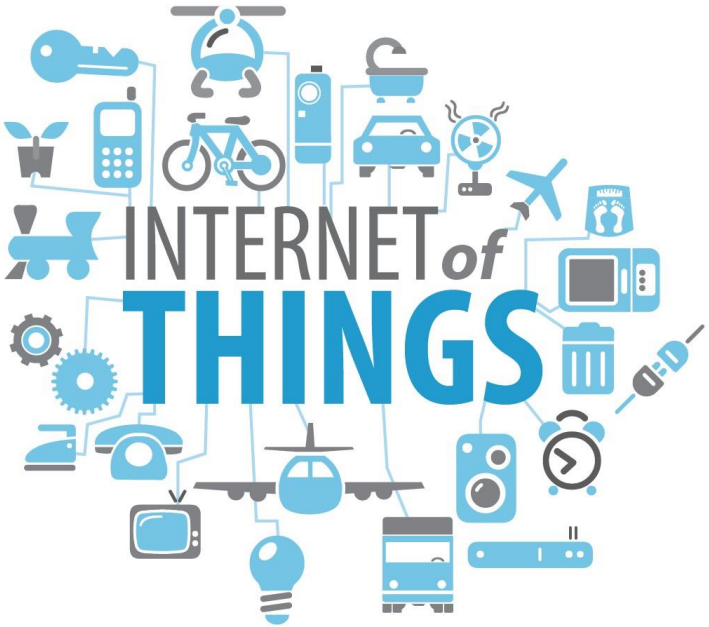
From Cells to Streets: Estimating Mobile Paths with Cellular-Side Data

Ilias Leontiadis, Antonio Lima, Haewoon Kwak, Rade Stanojevic, David Wetherall, Konstantina Papagiannaki

Motivation

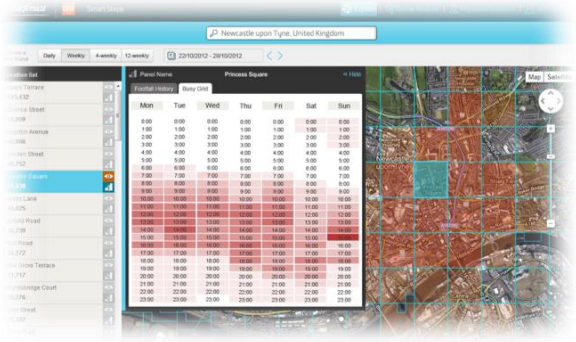
- Billions of connected devices in the near future.
- A lot of this devices are connected to cellular infrastructure
- As devices move from one location to another, information about the **set of associated sectors** can be logged for network management purposes.

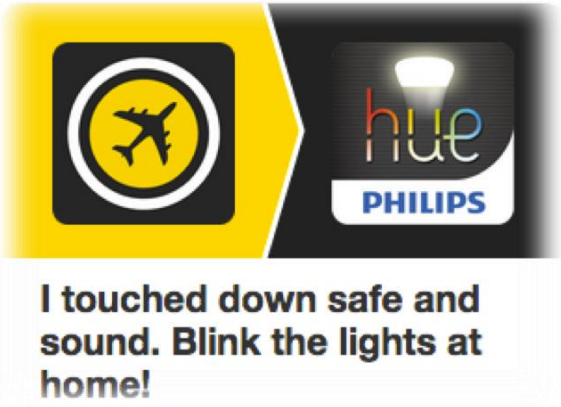


Can we use cellular information to infer the mobility patterns of individual devices ?

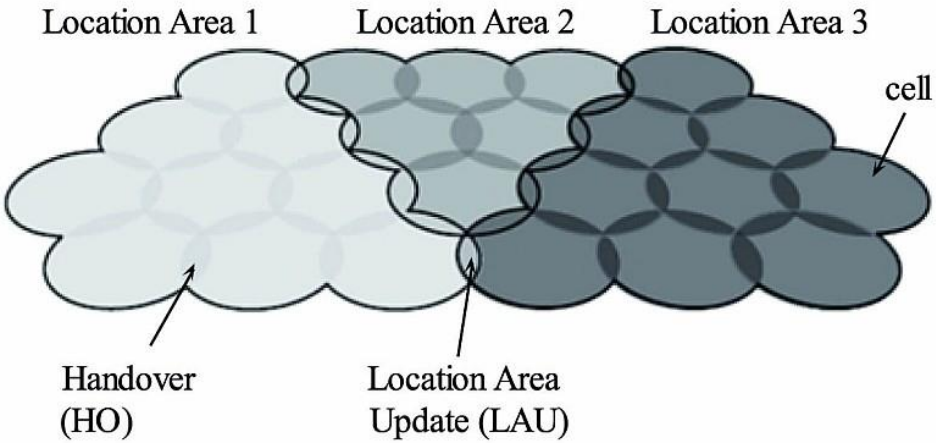
Possible Applications

- This information can support **location-based-services** for millions of users **at zero battery cost.**
  - § Business insights
  - § Travel logging
  - § Location Based Advertising
  - § Model human behavior
  - § Insurance telematics
  - § Traffic modeling and prediction
  - § Urban planning
  - § Public transport planning
  - § Event detection
  - § Car-pooling
  - § Fleet management
  - § IoT tracking
  - § User APIs (geofencing)





Challenges 1/3



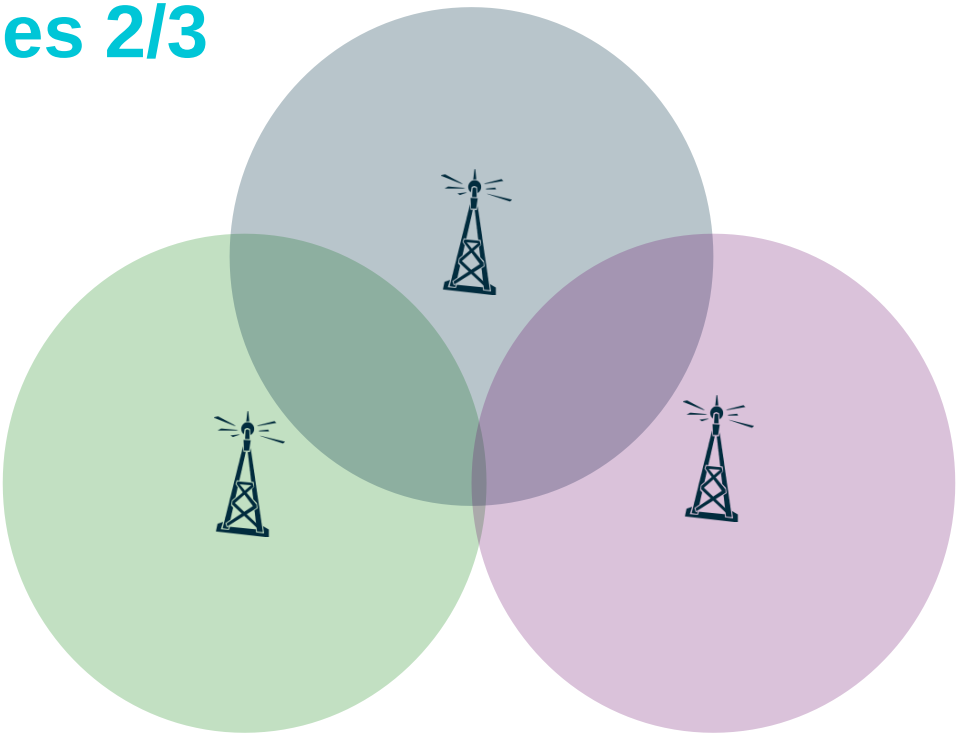
Network events are **sparse**

Information is only collected when a user:

- Makes/Receives a call
- Sends/Receives an SMS
- Switches between two Location Areas (LACs)
- Pinged by the network (every 3 hours)

Image Credit: http://www.traffic.bme.hu

Challenges 2/3



No trilateration possible

- Only associations to a single tower
- No RSSI or timing information

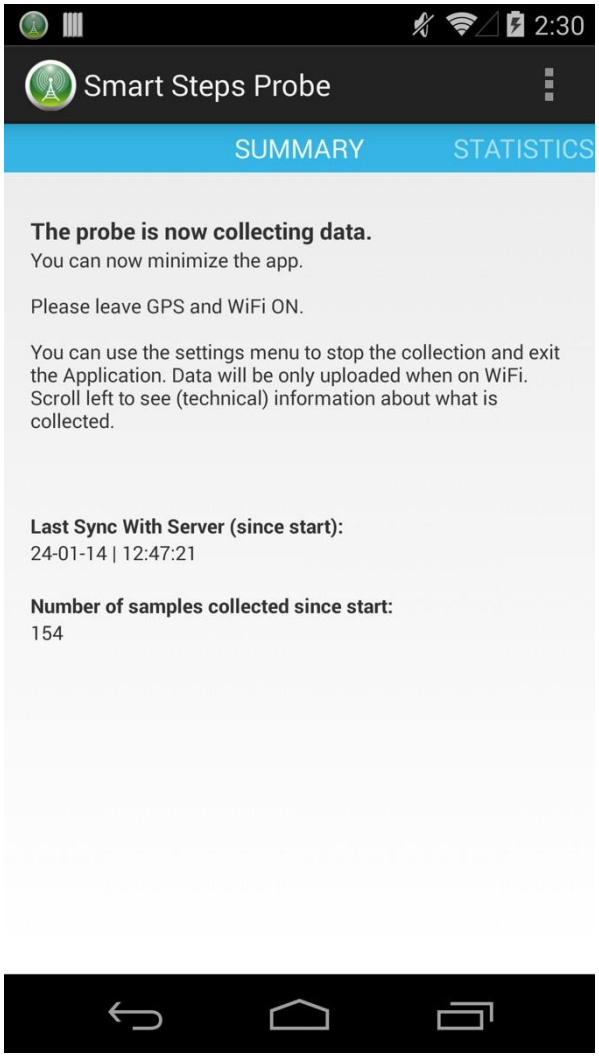
Challenges 3/3

There is no ground truth!

- We don’t know **the actual location** of the devices!
- We don’t even know in what granularity the data are gathered and how many and what type of **network events** are **missed**.

Collecting the Ground Truth

Android Probe



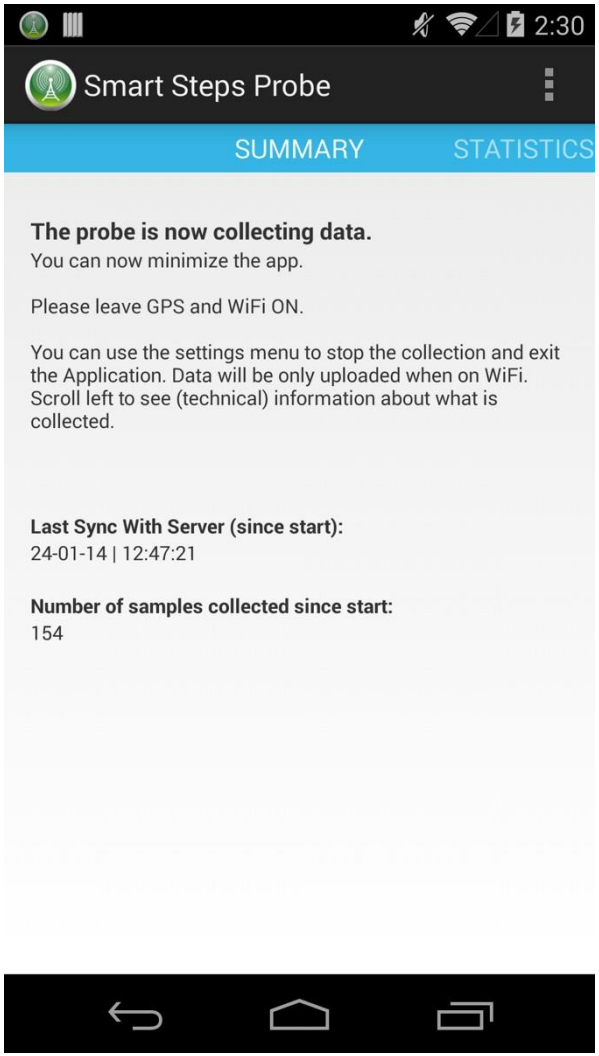
We build a probe for Android

Collected information

- Cellular information (sectors seen)
- GPS information

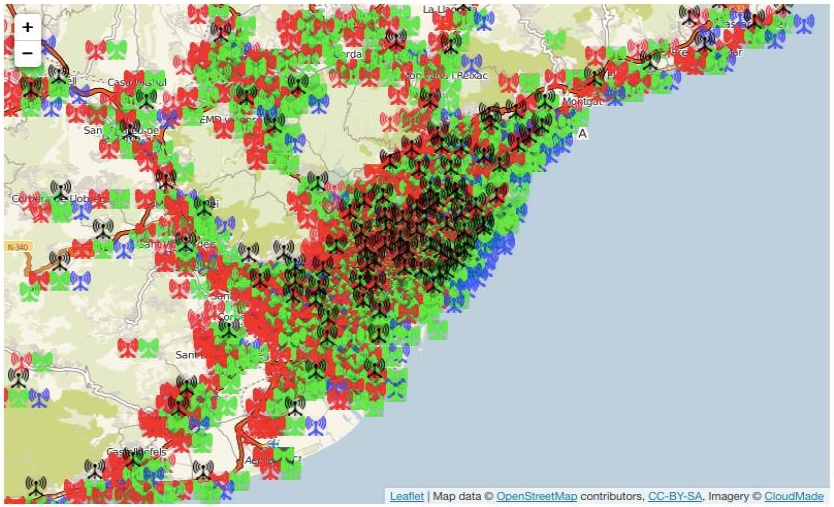
- **Study**
  - **30 Users**
  - **8 Months**

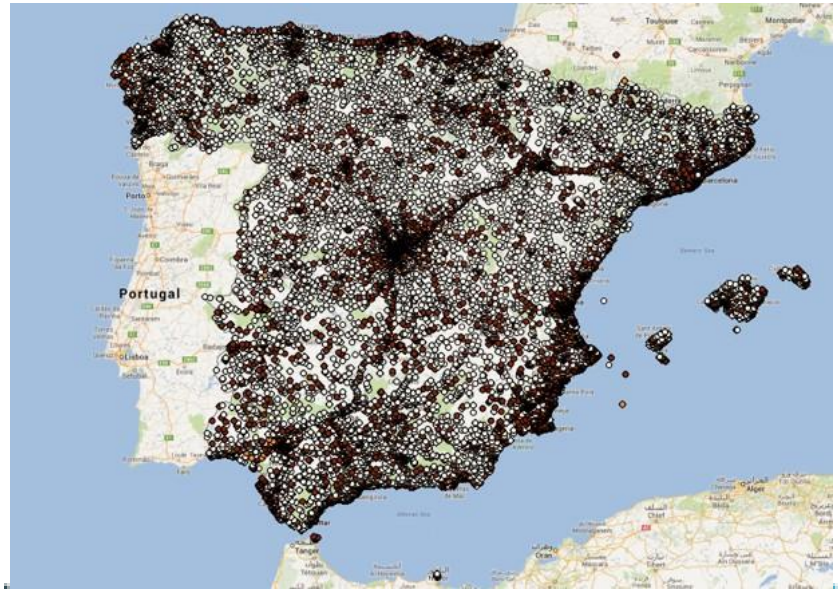
Android Probe



Type	Number
Total number of Cell samples	1,718,504
Total number of Hand-overs	433,031
Number of Distinct Sectors observed	15,455
Sectors in the operator's database	> 100,000
Number of Location samples (Netw. and GPS)	673,468
Number of GPS samples	259,032
Total time logged (all devices)	19,438 hours
Total time stationary	18,335 hours
Total time moving	1,102 hours
Number of trajectories (trips)	3,216
Total distance traveled	19,840 km

Tower Dataset

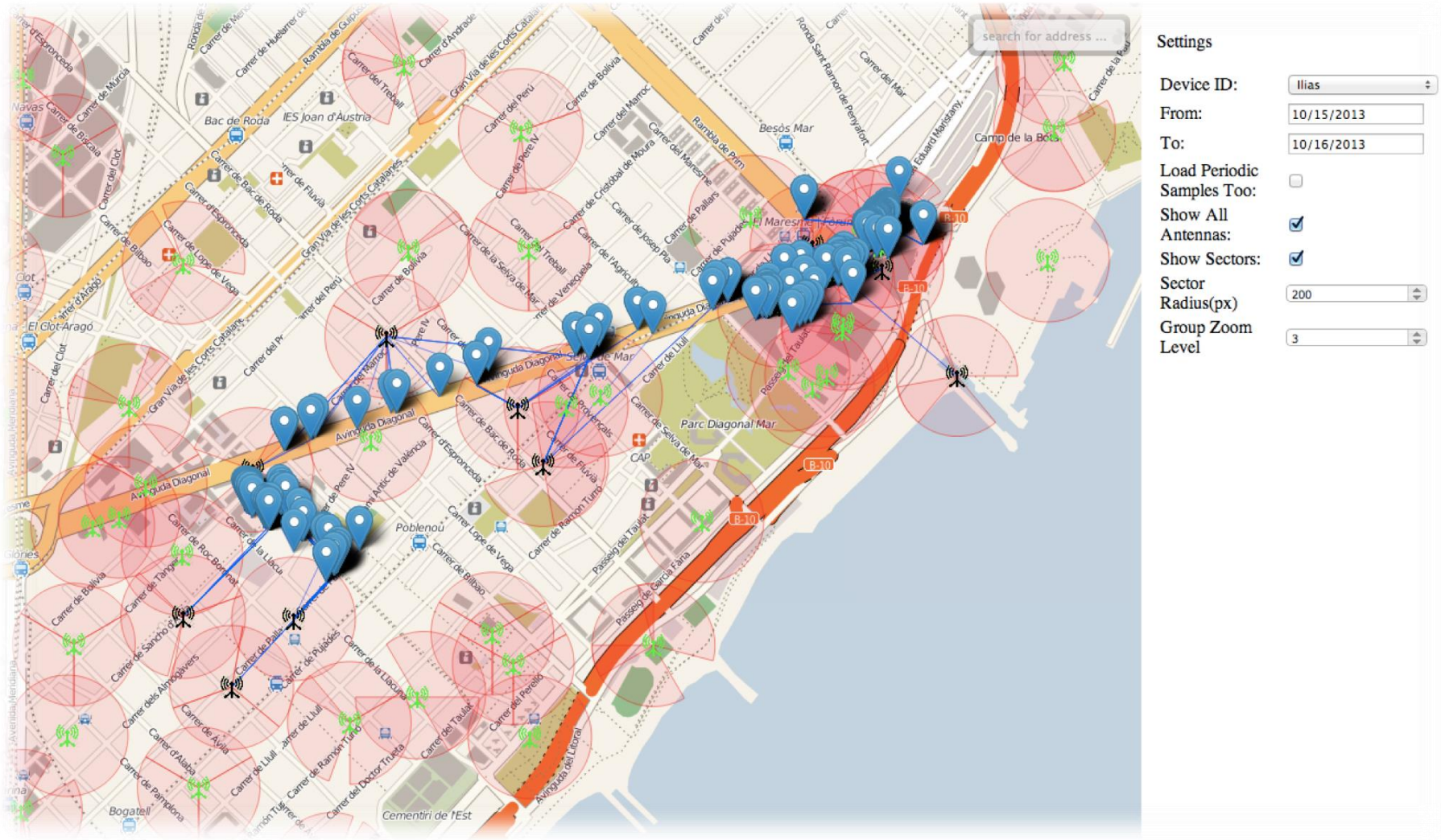




Tower dataset (>160K sectors)

- Location
- Orientation
- Beam Width
- Type
- Power

Example of collected information



Cell*: From Cells to Streets

Use mapping information (GIS) together with the network events to infer the path of mobile devices.

1. Model the coverage area of the sectors.
2. Identify stationary and mobile segments.
3. Estimate location for a stationary sequence.
4. Identify high-probability areas of the path.
5. Estimate the path with map information.

Cell*: From Cells to Streets

Use mapping information (GIS) together with the network events to infer the path of mobile devices

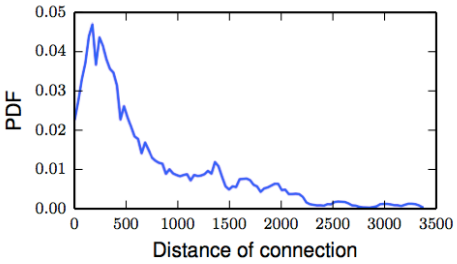
1. Model the coverage area of the sectors.
2. Identify stationary and mobile segments.
3. Estimate location for a stationary sequence.
4. Identify high-probability areas of the path.
5. Estimate the path with map information.

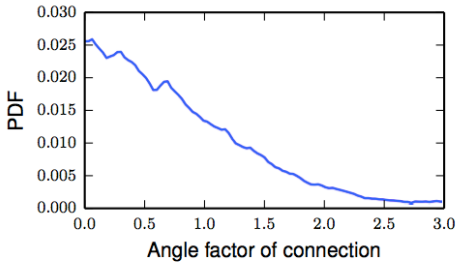
Model the coverage area of the sectors.

Number of ways to do this:

- GIS simulation tools
- Fingerprinting (data driven)
- Build a model
  - Distance ***d*** between the mobile and the cell tower
  - Polar angle factor **φ**
  - Type **P** (power/micro/pico)

Model the coverage area of the sectors.

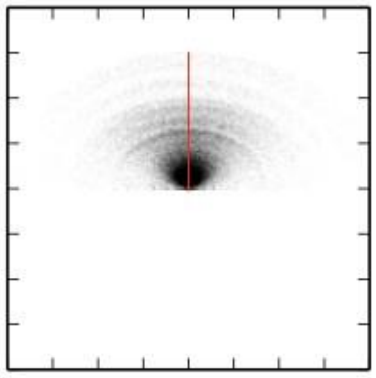


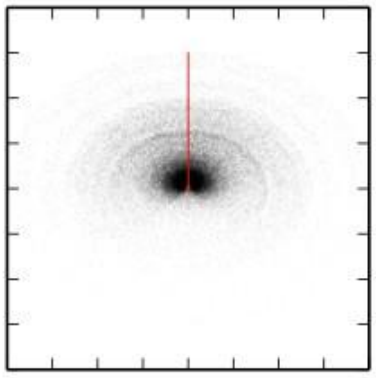


Use 1.7M cell samples collected.

Per sector type:

Calculate a 2-D PDF W_P (d, φ)



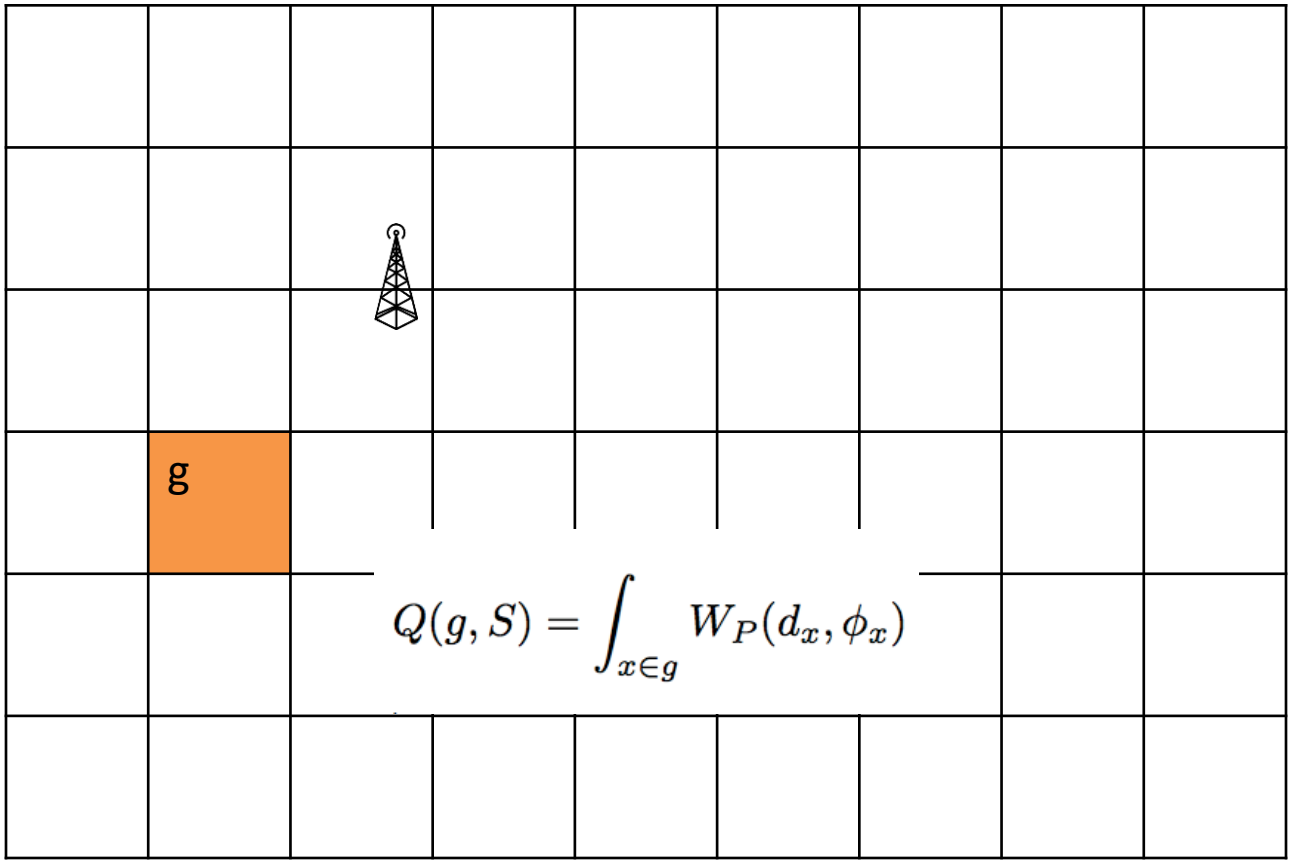


Split the area in **square grids** g and calculate the probability of a device in in a grid:

$$Q(g,S) = \int_{x \in g} W_P(d_x, \phi_x)$$

We used Monte-Carlo to approximate the calculation

Model the coverage area of the sectors.

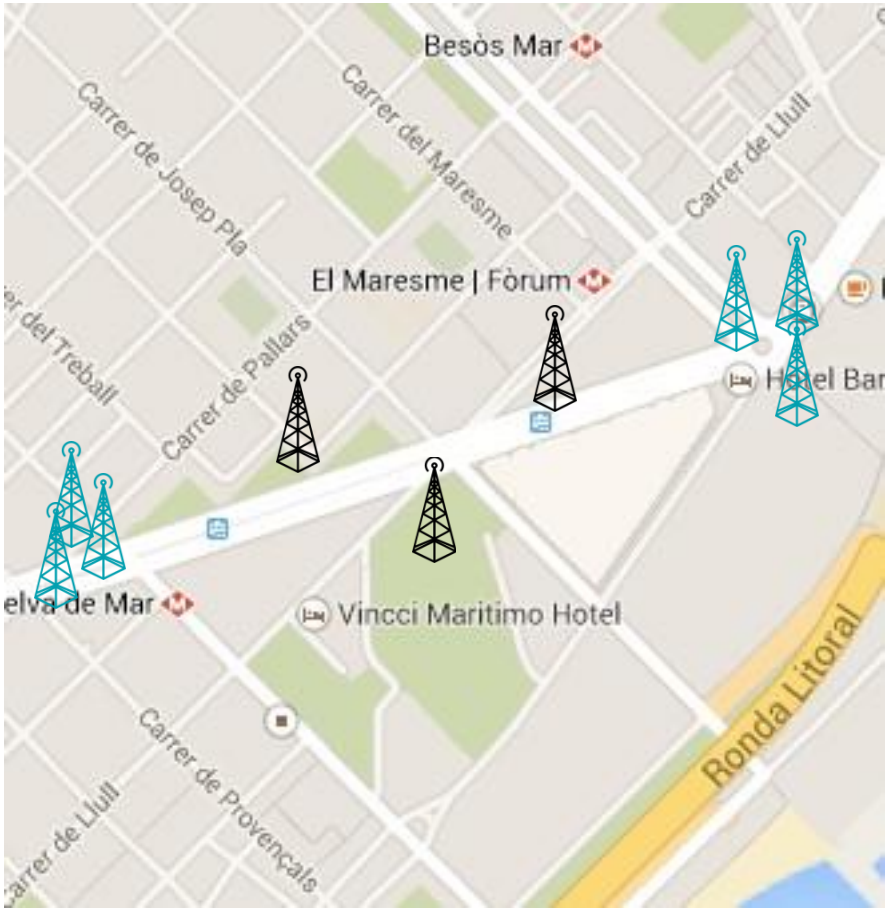


Cell*: From Cells to Streets

Use mapping information (GIS) together with the network events to infer the path of mobile devices

1. Model the coverage area of the sectors.
2. Identify stationary and mobile segments.
3. Estimate location for a stationary sequence.
4. Identify high-probability areas of the path.
5. Estimate the path with map information.

Identify stationary and mobile segments.



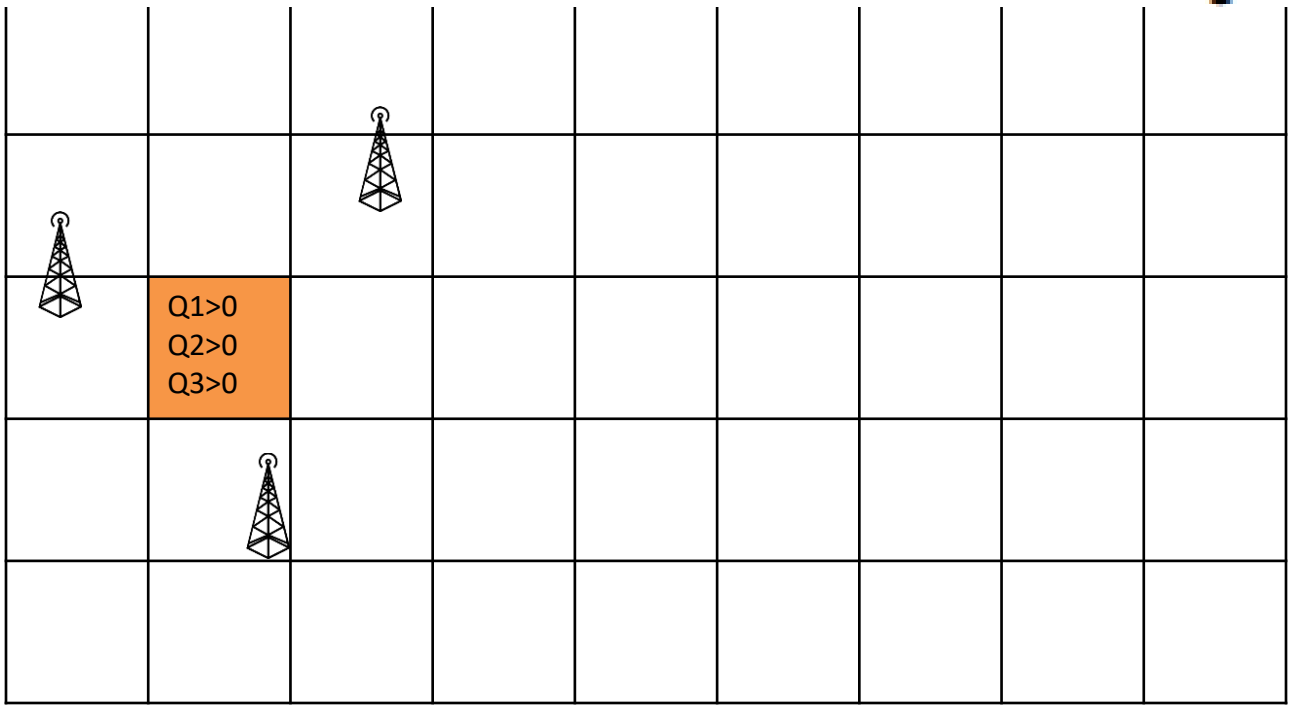
For each device, we get a sequence of sectors: $S_1, S_2 \ldots S_n$

We need to split the continuous set of of sectors into:

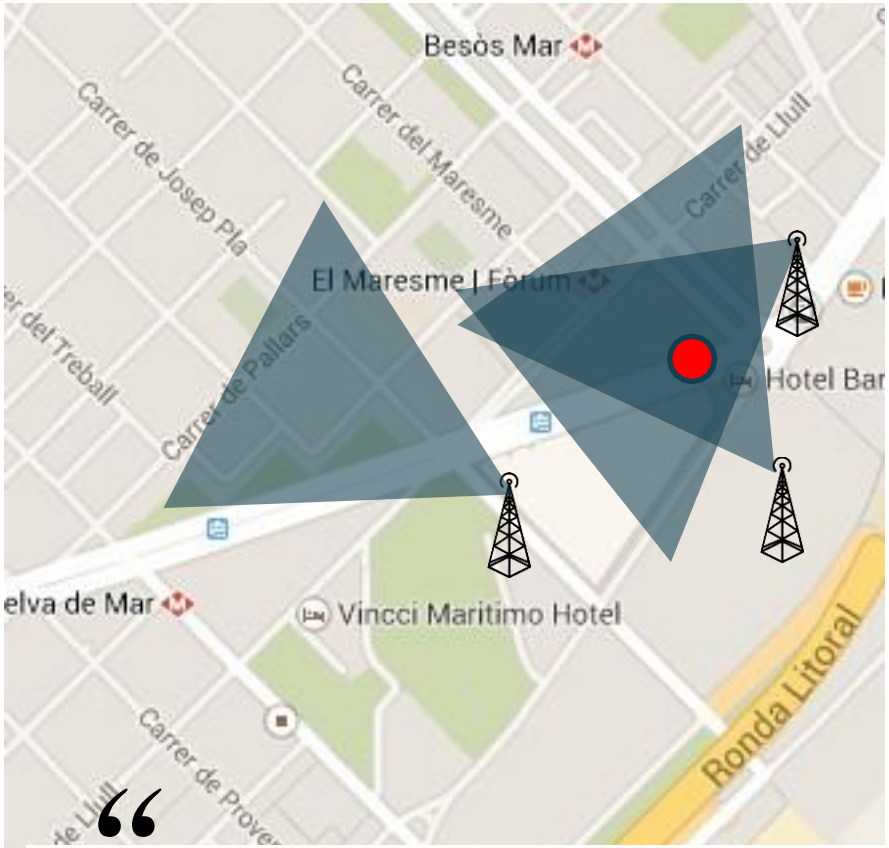
- **Stationary segments:** device stays in **same logical location >15m**
- **Mobile segments:** between stationary endpoints

Model the coverage area of the sectors.

We call two sectors S_1 and S_2 *adjacent* if there is a square g in the grid such that $Q(g, S_1) > 0$ and $Q(g, S_2) > 0$. That is, a mobile can connect to both sectors without moving. We say a



Identify stationary and mobile segments.



Problem: While stationary a device can connect to various sectors!

- Median is 4 unique sectors!

Solution:

- We calculate the probability *Q* of any two towers overlapping.
- We merge sectors when Q > 0
- When a device actually moves the first non-overlapping tower will trigger the mobile sequence.

> “sequence of sectors $S_1, \ldots, S_n$ is *stationary* if every two sectors are adjacent and the duration of the sequence as determined by its timestamps is greater than τ. That is, the mobile can connect to all ”

Cell*: From Cells to Streets

Use mapping information (GIS) together with the network events to infer the path of mobile devices

1. Model the coverage area of the sectors.
2. Identify stationary and mobile segments.
3. Estimate location for a stationary sequence.
4. Identify high-probability areas of the path.
5. Estimate the path with map information.

Identify stationary and mobile segments.

Use the **multiple** observed sectors during stationary periods to improve accuracy

Use the **intersection**

Problem: it doesn't always work!

Identify stationary and mobile segments.

Use the **multiple** observed sectors during stationary periods to improve accuracy

Use the **intersection**

Problem: it doesn't always work!

Solution:

- Use a **time-weighted** sum of the coverage areas

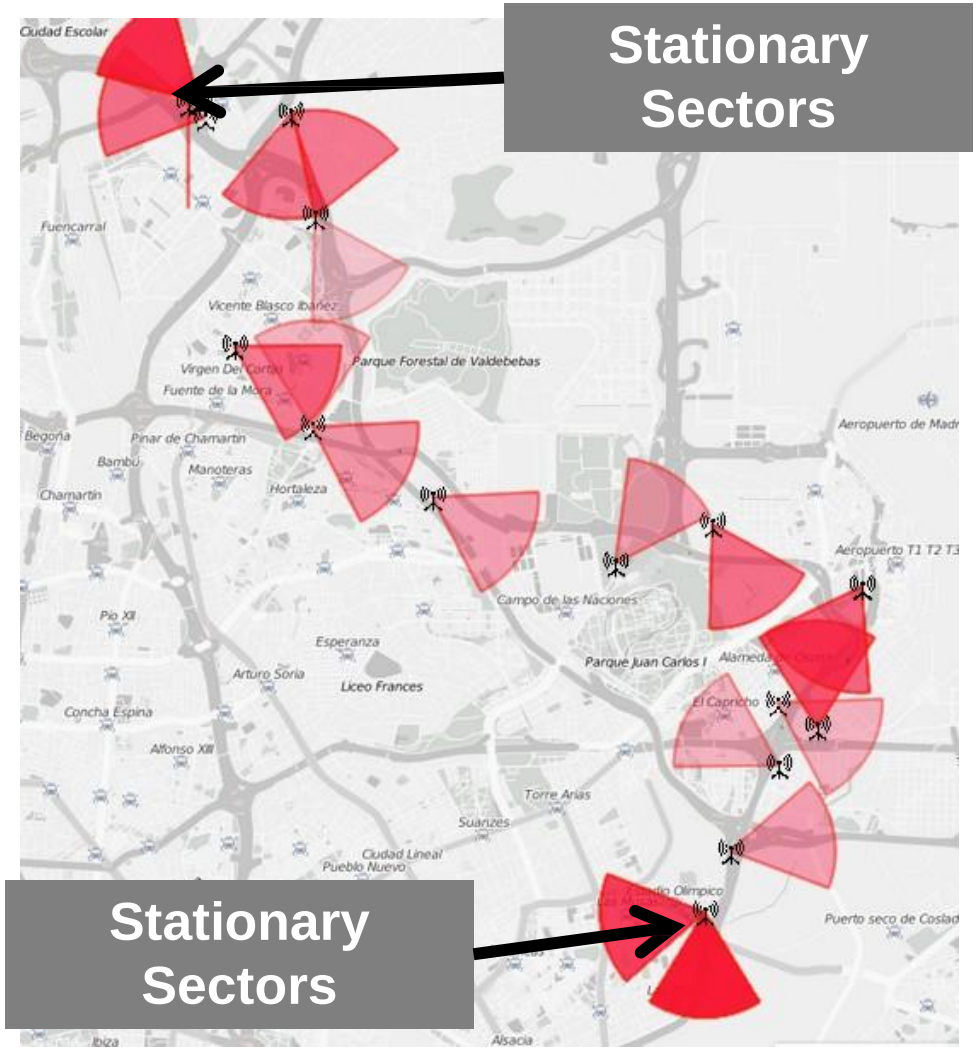
$$g^* = \arg\max_g \sum_{i=1}^{n} Q(g, S_i)\tau_i.$$

Cell*: From Cells to Streets

Use mapping information (GIS) together with the network events to infer the path of mobile devices

1. Model the coverage area of the sectors.
2. Identify stationary and mobile segments.
3. Estimate location for a stationary sequence.
4. Identify high-probability areas of the path.
5. Estimate the path with map information.

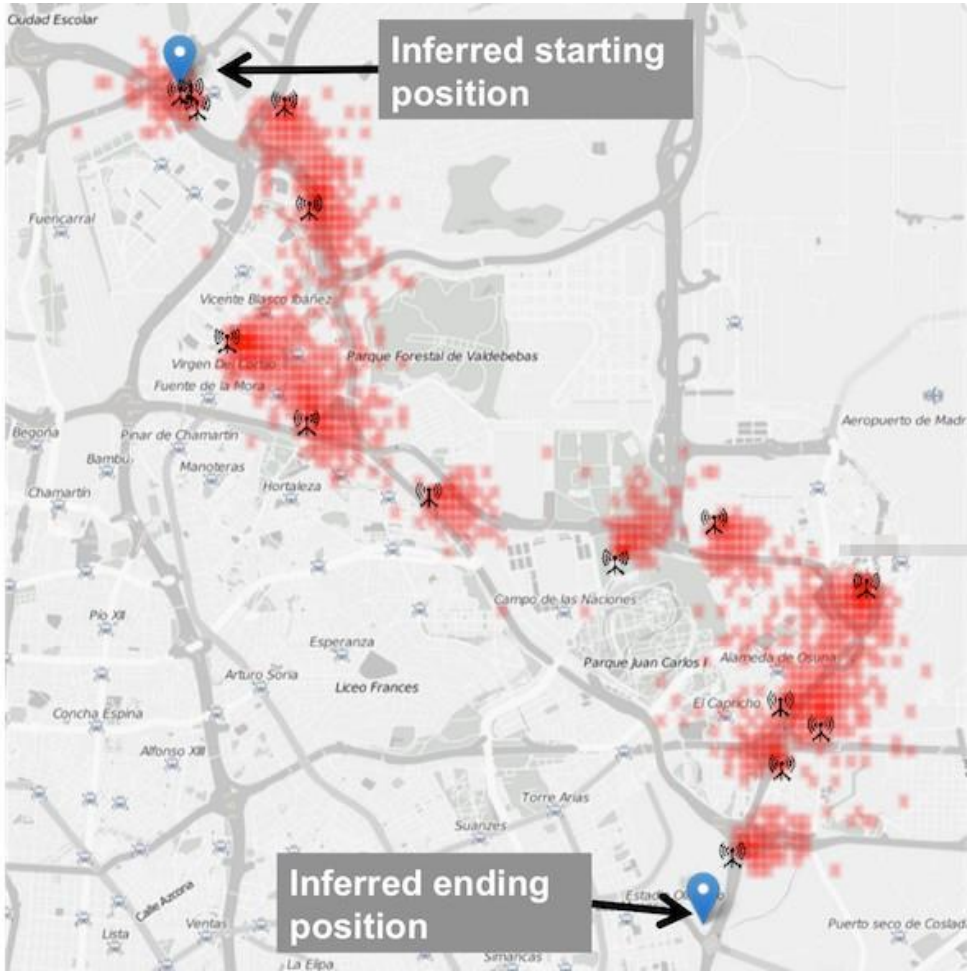
Identify high-probability areas of the path.



After splitting the day in stationary and non-stationary periods we focus on **mobile periods**.

- The device might connect to multiple ***intermediate sectors.***
- The intermediate sectors à **areas** where the device was with **high probability**

Identify high-probability areas of the path.



- Super-Impose the coverage areas of all these sectors

$$\rho(g) = \sum_{i=1}^{n} Q(g, S_i).$$

- We used **Monte-Carlo approximation** but an analytic model is possible

Cell*: From Cells to Streets

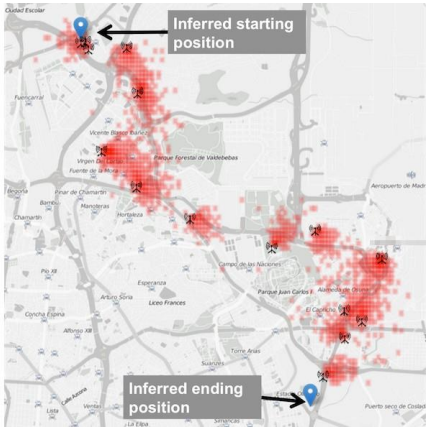
Use mapping information (GIS) together with the network events to infer the path of mobile devices

1. Model the coverage area of the sectors.
2. Identify stationary and mobile segments.
3. Estimate location for a stationary sequence.
4. Identify high-probability areas of the path.
5. Estimate the path with map information.

Estimate the path with map information.

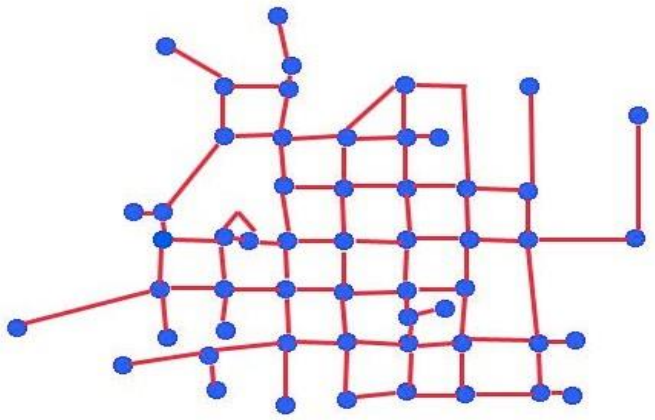
Intuition

- People are likely to follow shortest paths from one location to the other
- But not always [1]
  - § 60% of the users take the exact shortest
  - § 90% of cases people choose routes that have small deviations, of at most 5 minutes away from the optimal one.



[1] S. Zhu and D. Levinson. Do people use the shortest path? an empirical test of wardrops first principle. In *91th annual meeting of the Transportation Research Board, Washington*, volume 8, 2010.

Shortest Path



A map is essentially a weighted graph

- Weights
  - § Speed limits
  - § Street type
  - § Historical or real-time information

Routingà Shortest weighted path (usually in time)

- § Dijkstra
- § A*
- § Contraction Hierarchies

Intuition:

Modify the weights so that streets within the coverage of observed sectors have lower weights

Weight Calculation

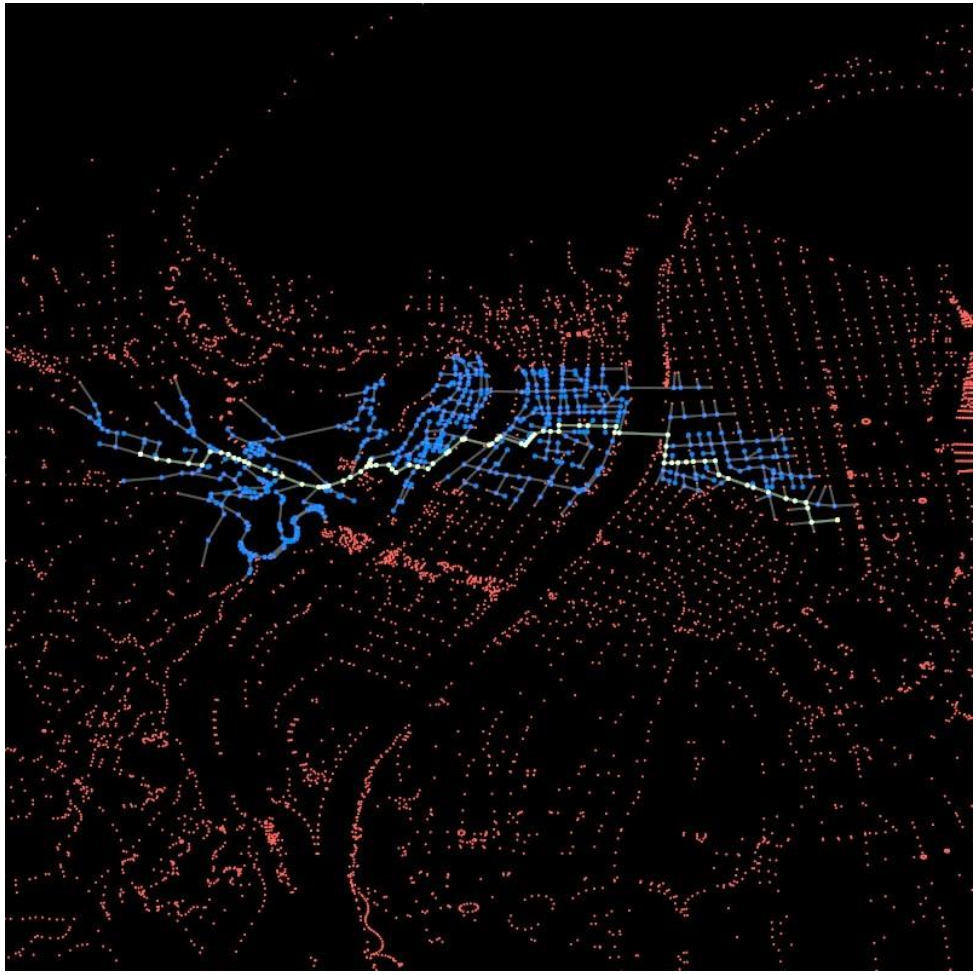
- We used Open Street Maps
- Initial weight, W(e) based on the expected time to traverse the edge

Modifying the default weights:

- The original weight W(e) of the segment e is then adjusted based on the probabilities ρ calculated in the previous step.

$$\bar{W}(e) = \frac{W(e)}{\frac{1}{l}\sum_{i=1}^{l}\rho(g_i) + \epsilon \max \rho(g)}$$

Cell* Routing

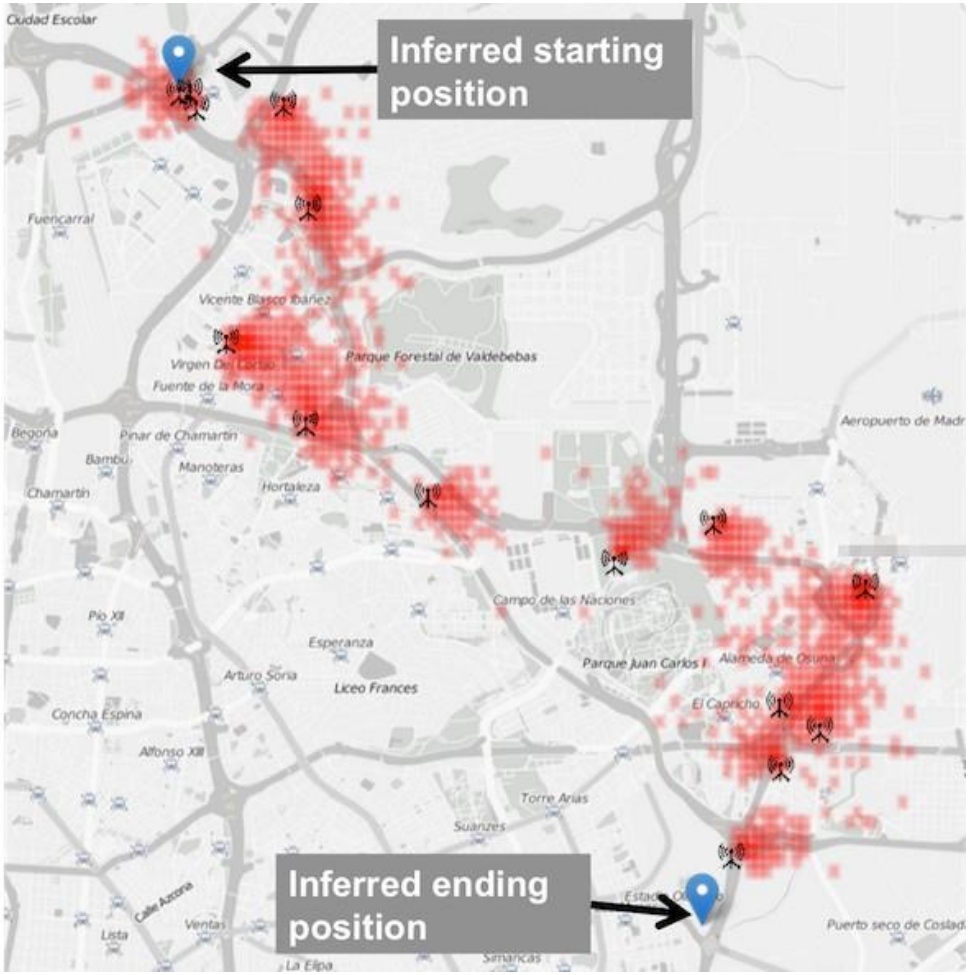


We implemented A* to run on the modified graph

A* is a heuristic based search algorithm

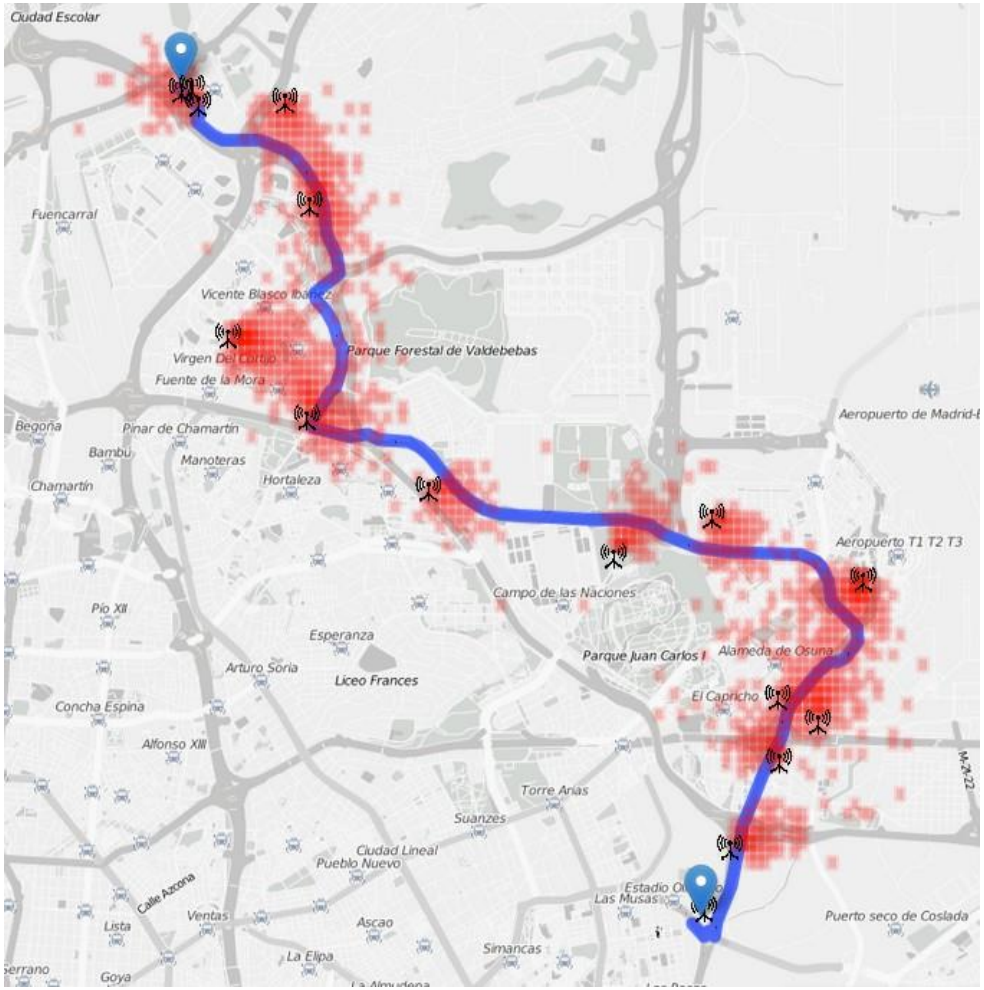
- Finds a path of the lowest expected total cost or distance
- To speed-up, it uses a knowledge-plus-heuristic cost
  - H(x) = time traveled + time remaining (straight line)

Cell* Routing



We run A* on the modified graph

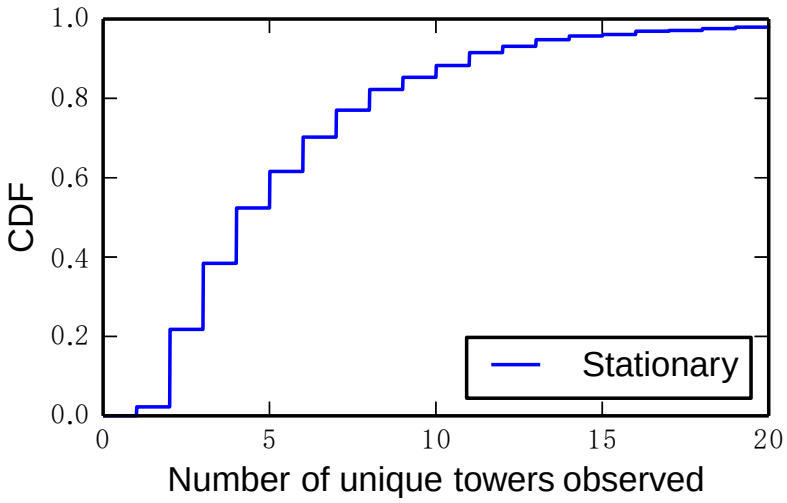
Routing

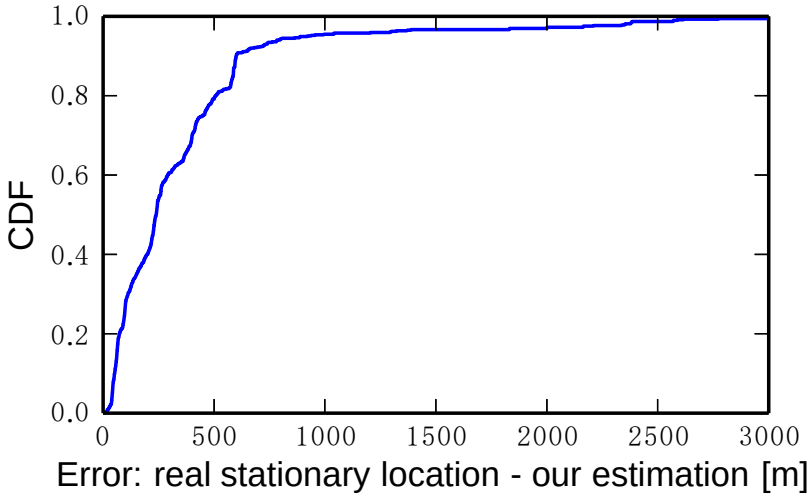


We run A* on the modified graph

Accuracy of stationary points

Accuracy of stationary points





In a **stationary** period:

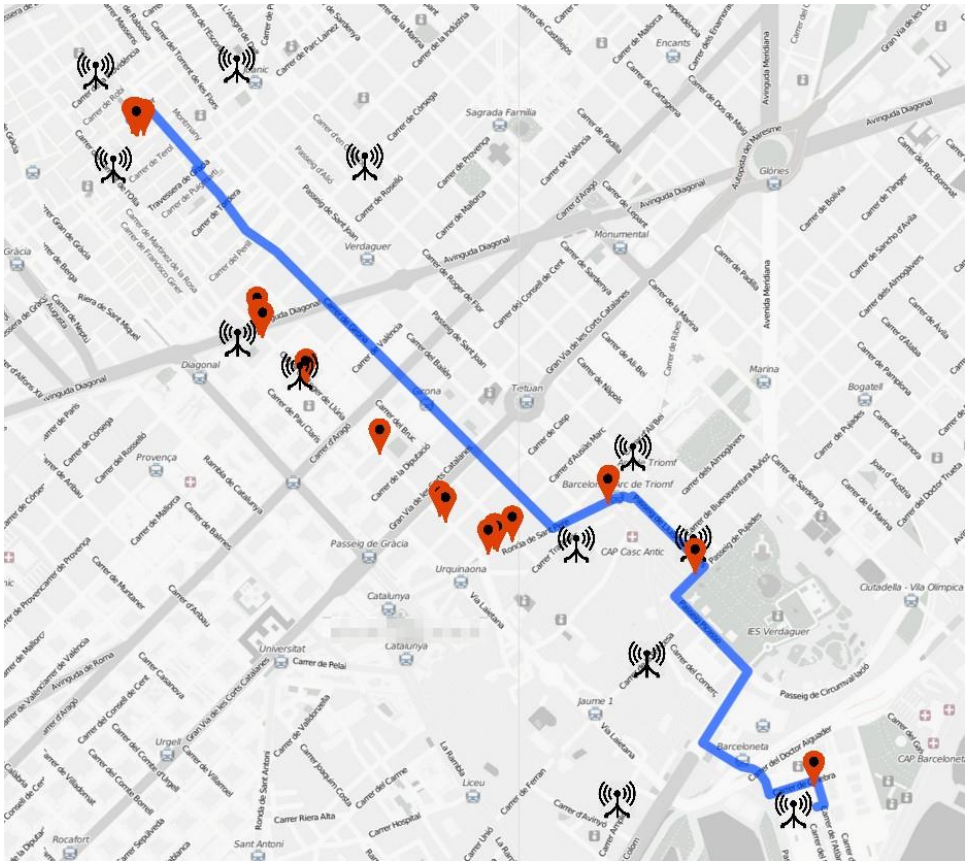
- A Median of 4 unique sectors.
- 15% of the cases more than 10

Accuracy:

- Median accuracy of 230 meters
- Using a single tower: 480 meters

Accuracy of mobile paths

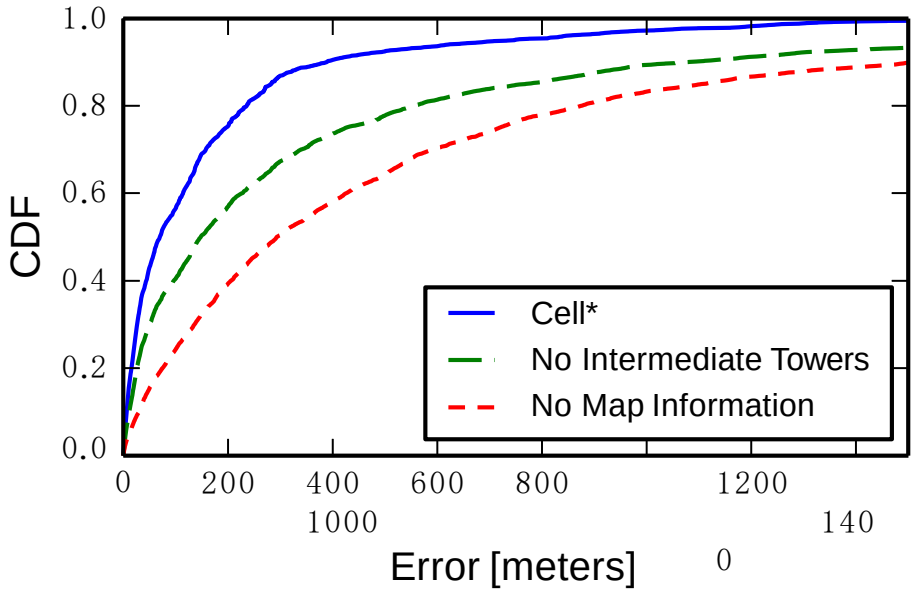
Accuracy of estimated paths



Error:

- Distribution of distances between GPS points and the estimated path.

Accuracy of estimated paths



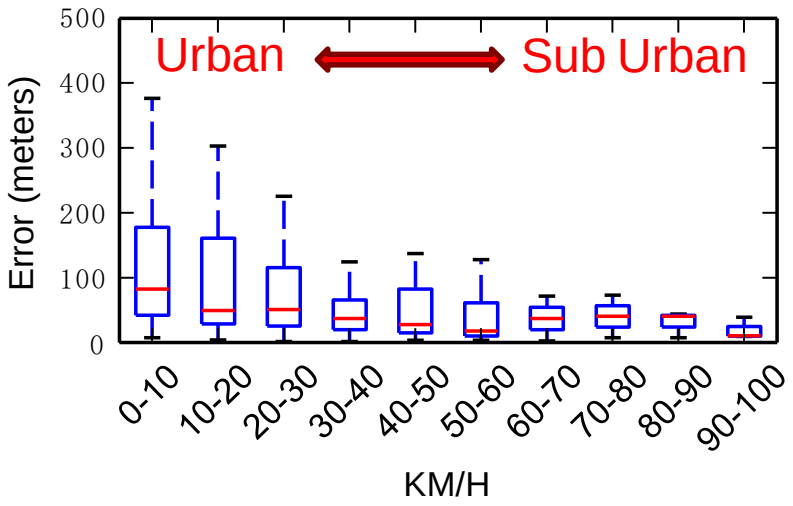
Compare

- Cell*
- Shortest Path (no intermediate)
- No Map (route of connected towers)

Accuracy:

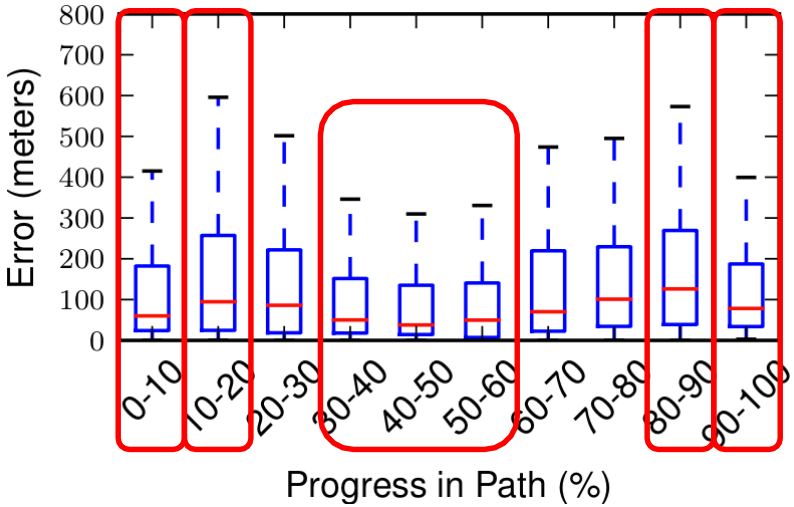
- Median error 70m
- 90% less than 400m
  - Shortest path > 1km
  - No mapping > 3 km

Type of paths



Error v.s. average speed:

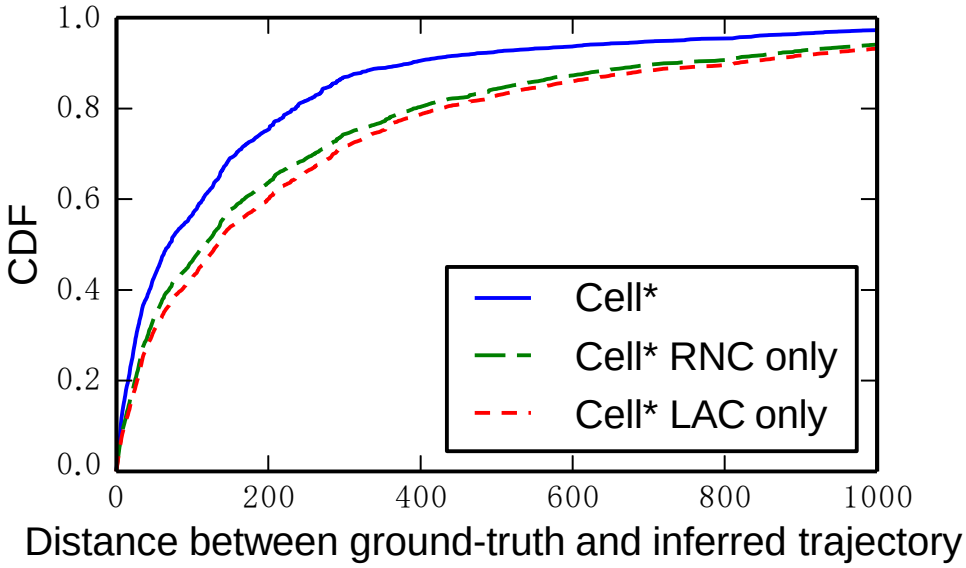
- Higher error in slow-speed urban environments.
  - Means of transport
  - Habitual effects
  - Traffic
  - People just walking around
- Great accuracy in rural areas *despite lower density of towers*!



Error v.s. progress in path:

- Accurate at stationary points
- Very accurate in the middle of the path
- Less accurate near the end points

Lower granularity



When less information is kept (as it happens now):

- Median error **129m** (was 70m)
- Still better than
  - Having no intermediate towers: 160m
  - **No map: 300m**

Conclusions

- It is possible to use network events to infer mobile paths
  - The places that the device was stationary can be inferred within 200m
  - In most cases the paths of devices can be estimated within an average city block
  - Cell* is more accurate in long sub-urban and rural commutes

- This may allow a new set of location-based applications to operate with passively collected, low-energy-cost information
  - Aggregated information (e.g., urban planning, road traffic information)
  - Individual (e.g., APIs)

+使用了扇区方向信息

-实际情况下，覆盖范围的计算方式可能有问题，因为基站密度差异很大

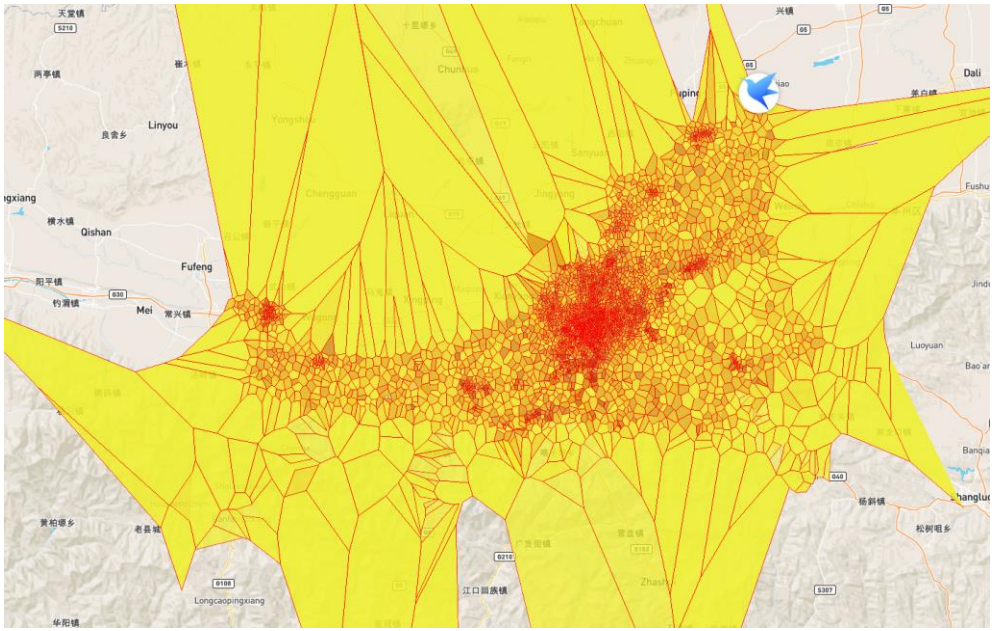
Per sector type:

Calculate a 2-D PDF $W_P(d, \varphi)$

Split the area in **square grids** g and calculate the probability of a device in in a grid:

$$Q(g, S) = \int_{x \in g} W_P(d_x, \phi_x)$$

We used Monte-Carlo to approximate the calculation



Thanks!

QUESTIONS ?

ILIAS.LEONTIADIS@TELEFONICA.COM

Types of area [edit]

Location area [edit]

A "location area" is a set of base stations that are grouped together to optimise signalling. Typically, tens or even hundreds of base stations share a single Base Station Controller (BSC) in GSM, or a Radio Network Controller (RNC) in UMTS, the intelligence behind the base stations. The BSC handles allocation of radio channels, receives measurements from the mobile phones, controls handovers from base station to base station.

To each location area, a unique number called a "location area code" is assigned. The location area code is broadcast by each base station, known as a "base transceiver station" BTS in GSM, or a Node B in UMTS, at regular intervals.

If the location areas are very large, there will be many mobiles operating simultaneously, resulting in very high paging traffic, as every paging request has to be broadcast to every base station in the location area. This wastes bandwidth and power on the mobile, by requiring it to listen for broadcast messages too much of the time. If on the other hand, there are too many small location areas, the mobile must contact the network very often for changes of location, which will also drain the mobile's battery. A balance has therefore to be struck[*citation needed*].

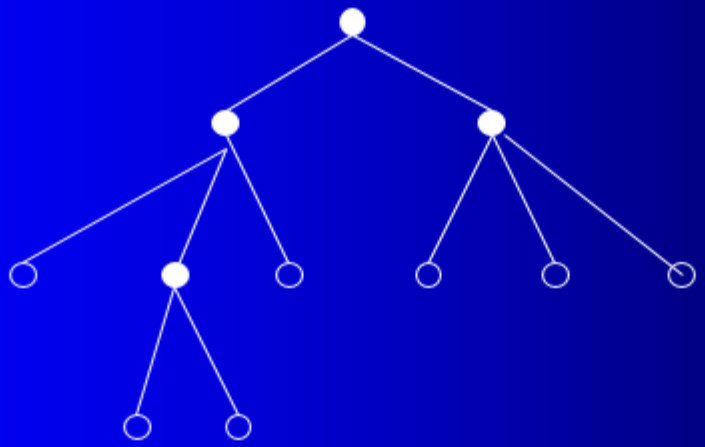
Routing area [edit]

The routing area is the packet-switched domain equivalent of the location area. A "routing area" is normally a subdivision of a "location area". Routing areas are used by mobiles which are GPRS-attached. GPRS is optimized for "bursty" data communication services, such as wireless internet/intranet, and multimedia services. It is also known as GSM-IP ("Internet Protocol") because it will connect users directly to Internet Service Providers

The bursty nature of packet traffic means that more paging messages are expected per mobile, and so it is worth knowing the location of the mobile more accurately than it would be with traditional circuit-switched traffic. A change from routing area to routing area (called a "Routing Area Update") is done in an almost identical way to a change from location area to location area. The main differences are that the "Serving GPRS Support Node" (SGSN) is the element involved.

基本思想

- 定义一个评价函数f，对当前的搜索状态进行评估，找出一个最有希望的节点来扩展。



53