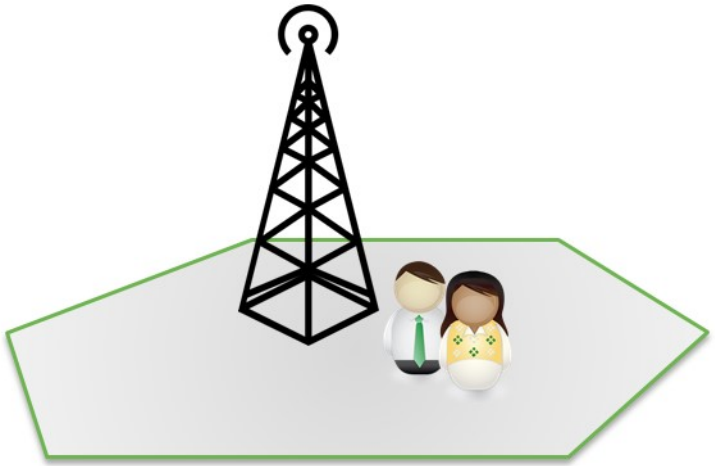
AUTORITEIT
PERSOONSGEGEVENS

On the anonimity of aggregated telco location data

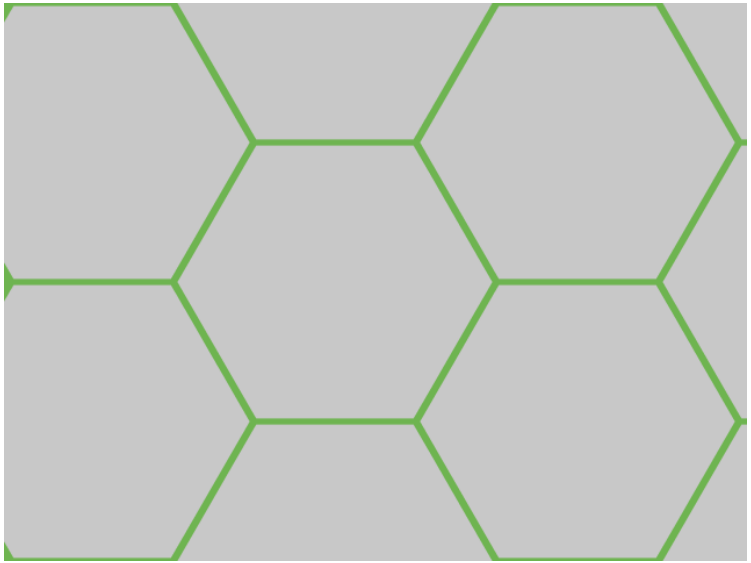
April 2020

What is aggregated telco location data?

Your phone connects to the nearest cell tower of your mobile operator. When using the network your approximate location is known.

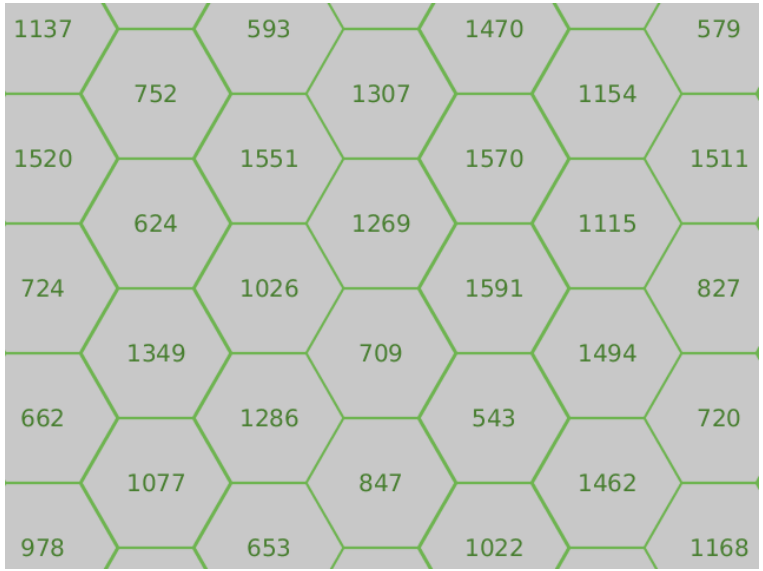


Many people at the same time connect to the mobile network. There are many cells, each containing many people.



What is aggregated telco location data?

Mobile operators count how many people are in a cell. This count is updated periodically, e.g. every hour or every 15 minutes.



These counts are collected and stored. The resulting list is the aggregated telco data.

Date	Time	Cell ID	Count
2020-04-12	12:00-13:00	1	2611
2020-04-12	12:00-13:00	2	1239
…			
2020-04-12	13:00-14:00	1	2576
2020-04-12	13:00-14:00	2	1435
…			

The Big Question

Are these counts anonymous?

or, in other words,

aggregation $\stackrel{?}{=}$ anonymisation

When is data anonymised?

WP216 states "*In the light of Directive 95/46/EC and other relevant EU legal instruments, anonymisation results from processing personal data in order to irreversibly prevent identification. In doing so, several elements should be taken into account by data controllers, having regard to all the means "likely reasonably" to be used for identification (either by the controller or by any third party).*" and "… the outcome of anonymisation as a technique applied to personal data should be, in the current state of technology, as permanent as erasure …"

WP136 states: "Recital 26 of the Directive pays particular attention to the term "identifiable" when it reads that "*whereas to determine whether a person is identifiable account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person.*" This means that a mere hypothetical possibility to single out the individual is not enough to consider the person as "identifiable"."

So data is considered anonymous if for **any third party** that invests **reasonable effort** it is **unlikely** to succeed in **singling out and re-identifying** persons from that dataset.

Intermezzo: human trajectory specifics

Scientific research shows that it only takes 4 locations (on neighbourhood level) to single out a human being [ref 1],

- For example the neighbourhood of your work, home, mom's house and gym

Humans are creatures of habit. We tend to:

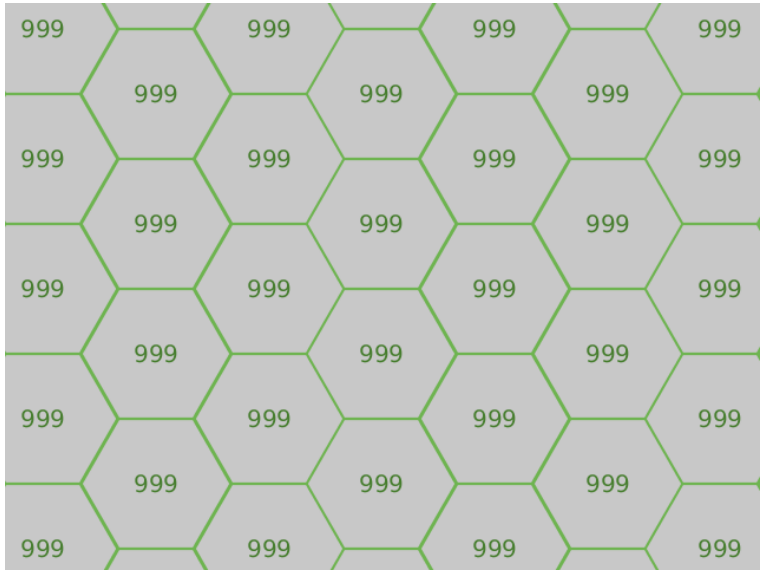
- sleep at home
- work at work (in non-pandemic times)
- take no unnecessary detours when we commute
- always meet with friends at the same moment in the day/week/month

Take-away:

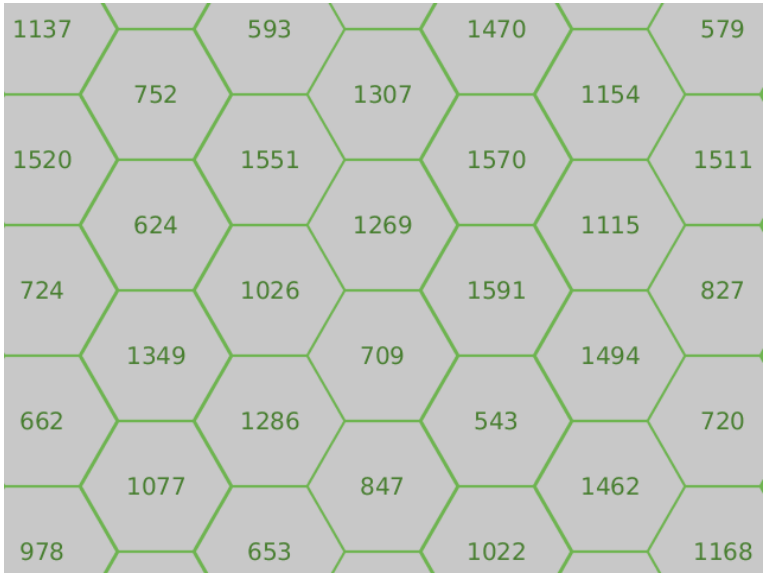
1. location data is very **sensitive** and
2. human trajectories are **not random**

Trajectory examples

Let´s walk the dog. And suppose everybody else stays where they are. You are no longer lost in the crowd.



In quiet moments like the early hours of the night all differences are caused by a few people.



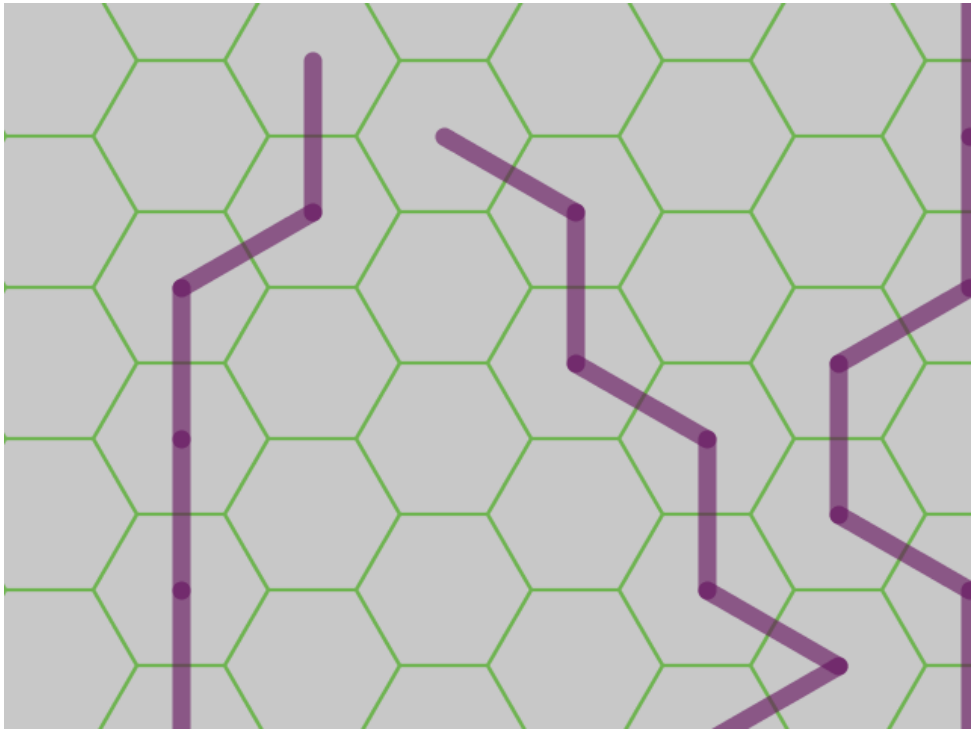
Brute forcing aggregated telco location data

Brute forcing means trying many combinations. In this case, many possible trajectories that, when summed, result in the same cell counts as the original aggregated telco data.

And remember, the trajectories that are tried are not random; people tend to sleep at home, work at work etc.

Does this actually work? Yes, researchers have shown this [ref 2] and related attacks [ref 3].

Why does it work? Because the aggregated data still contains enough information on day-time and night-time patterns to find people's trajectories.



From singling out to identification

The trajectories that are the result of the brute force attack lead to singling out; separate persons are represented by separate trajectories.

Trajectories indicate people's homes, area of work, routes (sometimes including manner of transportation), places of leisure, etc. By going out in the field and physically approaching somebody it is possible to identify at least some of the people that are behind these trajectories.

With access to other data sources it is possible to identify persons on a much larger scale. Think of government data, camera footage or data sources mentioned on the next page.

Bring your own data

If you are already in possession of data with times and location, the de-anonymisation process gets easier. This is due to the fact that you can already estimate many possible trajectories, so the resulting brute force job becomes smaller. Also, the chance that you can re-identify persons increases, especially if you also have identifiers like card codes, user ID’s etc.

Interesting data sources to help with de-anonimising aggregated telco data are for example:

- Credit/debit card transactions
- (Public) transportation data
- Customer loyalty program data
- Metadata (EXIF) from photo galleries
- Mobile app usage data
- Public registries

Cost of brute force attack

Case: the Netherlands

- Number of cells: approx 3200
- Number of people with a smartphone: approx 15 million

The brute force process is memory-bound, so when we know how much memory it takes we can make a statement on the cost. The total memory required equals the number of trajectories times the size of a single trajectory. Assume data is aggregated once every hour, then trajectories are 24 “cell slots” long each day. One month of data consists of 24 x 30 = 720 slots. A single trajectory can be stored in 1kByte.

Total memory required:

- naïve implementation: 15 million x 1kByte = 15GB
- optimized implementation (e.g. with smaller data encodings) probably in < 6GB

This is a proper match for a regular €800 off-the-shelf gaming PC. A cloud-based machine capable of doing this can be rented for around €0,40/hour and will need perhaps weeks.

Is hourly aggregated telco location data anonymous?

As we saw earlier:

> So data is considered anonymous if for **any third party** that invests **reasonable effort** it is **unlikely** to succeed in **singling out and re-identifying** persons from that dataset.

Effort can be time and money. Because researchers have already shown this to work, creating a working solution yourself is a matter of perseverance, not genius. The cost can be considered low.

Singling out persons is doable by brute forcing as described earlier. From there, identification can be done, especially when combining the outcome with other (public) data. Bringing your own collection of time- and location based data makes both the brute force and the re-identification easier. And there exist many sources that allow for this.

So no, aggregating telco location data to hours does not make it anonymous.

More Things to Consider

- Anonimisation is hard. Anonimisation of location data is almost undoable.
- The described de-anonimisation attack benefits from the spatial and temporal information still present in the aggregated dataset. Avoid releasing data of time slots less than a day (so an attacker cannot reconstruct daily and nightly patterns) and combine this with robust anonimisation techniques. Then, call in experts and assess the anonymity of your dataset. If in doubt, don’t call the data anonymous but handle it as personal data.
- In reality, every cell tower has 3 antennas, each covering a sector. So your approximate location not only depends on the cell you are in, but also of the direction that antenna is for you. This makes the estimation of your location accurate up to 50 meters.
- More data can be handled by applying more advanced programming concepts.

References

1. Unique in the Crowd: The privacy bounds of human mobility
   https://www.nature.com/articles/srep01376

2. Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated Mobility Data
   https://arxiv.org/abs/1702.06270

3. Related articles through Google Scholar:
   https://scholar.google.com/scholar?q=related:AOpuqg_9oxoJ:scholar.google.com/&scioq=Trajectory+Recovery+From+Ash:+User+Privacy+Is+NOT+Preserved+in+Aggregated+Mobility+Data