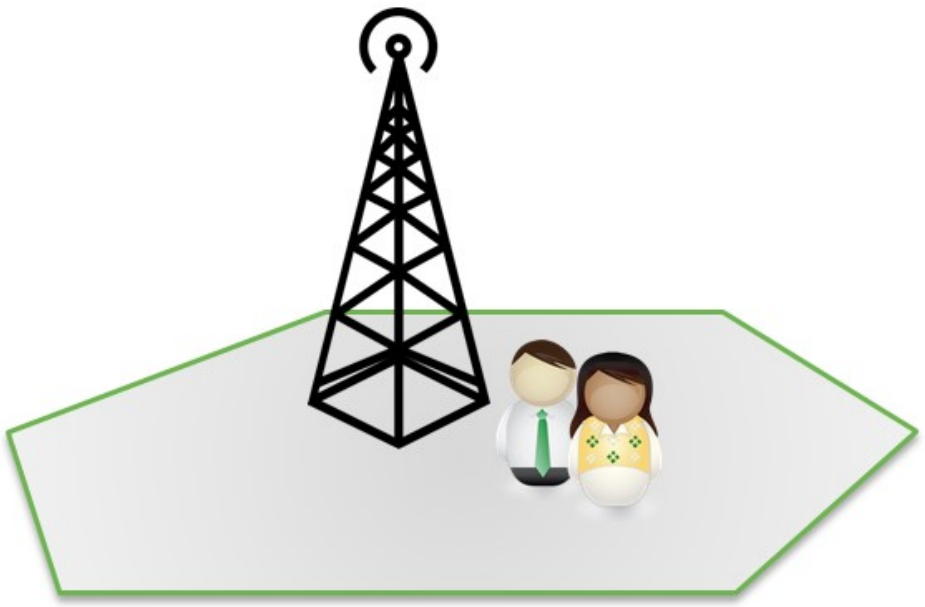
AUTORITEIT
PERSOONSGEGEVENS

Over de anonimiteit van geaggregeerde telecomlocatiedata

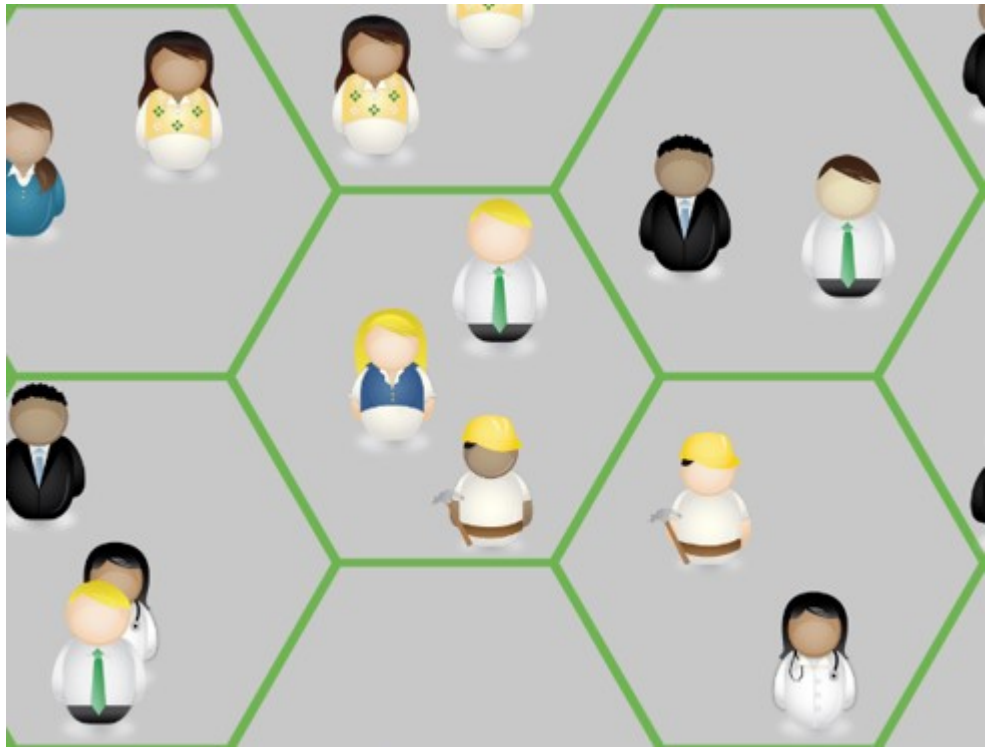
April 2020

Wat zijn geaggregeerde*) telecomlocatiedata?

Uw telefoon maakt verbinding met de dichtst-bijzijnde zendmast van uw mobiele aanbieder. Na verbinden is uw locatie bij benadering bekend.



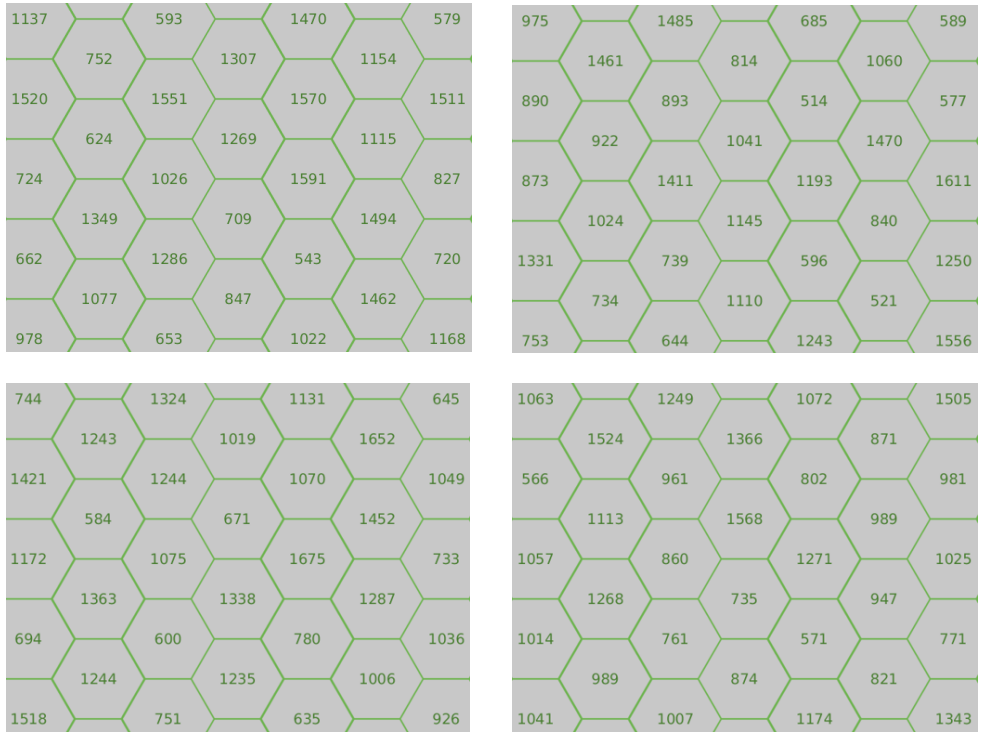
Het mobiele netwerk bestrijkt een groot aantal cellen. In iedere cel bevindt zich een groot aantal mensen.



*) Geaggregeerde data zijn samengevoegde data, bijvoorbeeld door onderliggende data te tellen, op te tellen of te middelen.

Wat zijn geaggregeerde telecomlocatiedata?

Mobiele aanbieders tellen het aantal mensen in een cel. Deze telling wordt periodiek gedaan, bijv. ieder uur of elk kwartier.



Deze tellingen worden verzameld en opgeslagen in een tabel. Deze tabel bevat de bedoelde geaggregeerde telecomlocatiedata.

Datum	Tijd	Cell ID	Aantal
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
…			

Dé grote vraag

Zijn deze tellingen anoniem?

of, met andere woorden,

aggregatie $\stackrel{?}{=}$ anonimiseren

Wanneer is data anoniem?

WP216 stelt "*In de zin van Richtlijn 95/46/EG en andere ter zake dienende EU-rechtsinstrumenten wordt anonimisering bewerkstelligd door persoonsgegevens zodanig te verwerken dat elke mogelijkheid tot identificatie van betrokkenen onherroepelijk wordt uitgesloten. Daarbij moeten de voor de verwerking verantwoordelijken rekening houden met diverse factoren. Tevens dient te worden gekeken naar alle middelen 'waarvan mag worden aangenomen' dat zij 'redelijkerwijs' door degene die voor de verwerking verantwoordelijk is dan wel door enige andere derde in te zetten zijn om een persoon te identificeren.*" en "*…het resultaat van anonimisering als op persoonsgegevens toegepaste techniek volgens de huidige stand van de techniek even permanent moet zijn als uitwissing.*"

WP136 stelt: "*In overweging 26 van de richtlijn wordt bijzondere aandacht geschonken aan de term "identificeerbaar": "om te bepalen of een persoon identificeerbaar is, moet worden gekeken naar alle middelen waarvan mag worden aangenomen dat zij redelijkerwijs door degene die voor de verwerking verantwoordelijk is dan wel door enig ander persoon in te zetten zijn om genoemde persoon te identificeren". Dit houdt in dat een slechts hypothetische mogelijkheid om iemand te onderscheiden niet voldoende is om die persoon als "identificeerbaar" te beschouwen.*"

Gegevens kunnen worden beschouwd als anoniem als **voor welke partij dan ook**, met inzet van (voor het doel) **redelijke middelen**, het **onwaarschijnlijk** is hieruit personen te **identificeren**.

Intermezzo: kenmerken menselijke verplaatsing

Wetenschappelijk onderzoek toont aan dat slechts 4 locaties (op omgevingsniveau) voldoende zijn om een enkel individu te kunnen herleiden [ref 1],

- Bijvoorbeeld de omgeving rond uw werk, uw huis, uw ouderlijk huis en uw sportschool

Mensen zijn gewoontedieren; meestal

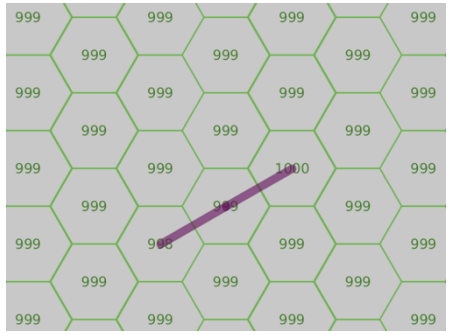
- slapen we thuis
- werken we op kantoor (wanneer er geen pandemie heerst)
- nemen we geen onzinnige omwegen als we forenzen
- spreken we op vaste momenten in de dag/week/maand af met vrienden

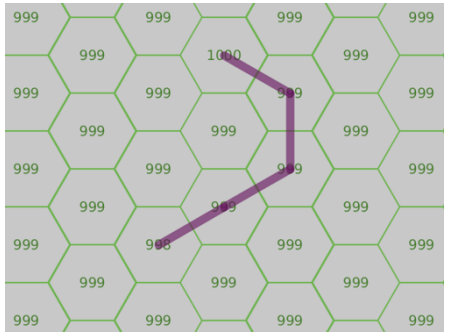
Onthoud:

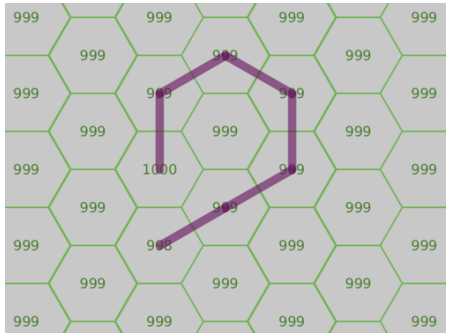
1. locatiedata is **zeer gevoelig** en
2. menselijke verplaatsingen zijn **niet willekeurig**

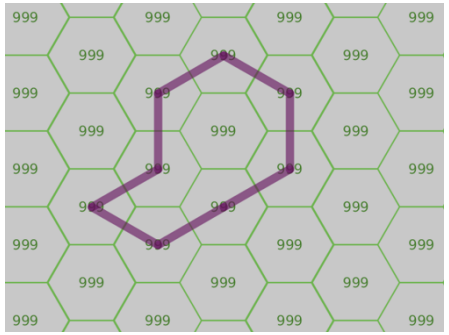
Voorbeelden van verplaatspatronen

Stel, iemand laat de hond uit. Als op dat moment niemand anders zich verplaatst gaat deze persoon niet langer op in de massa.

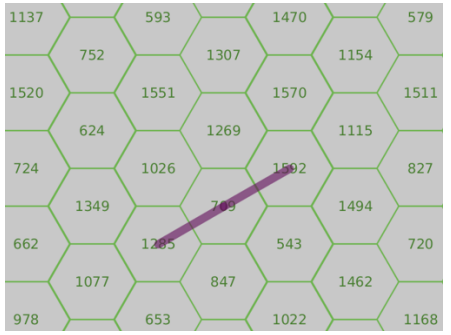


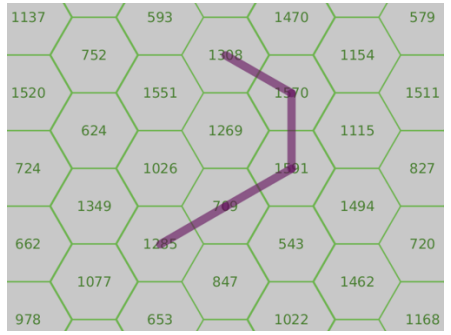


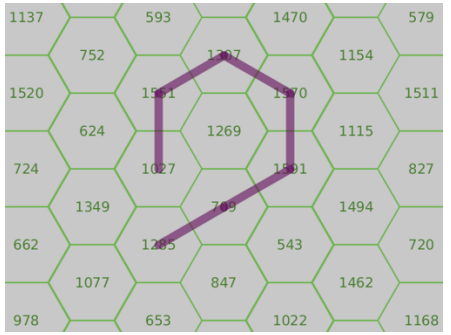


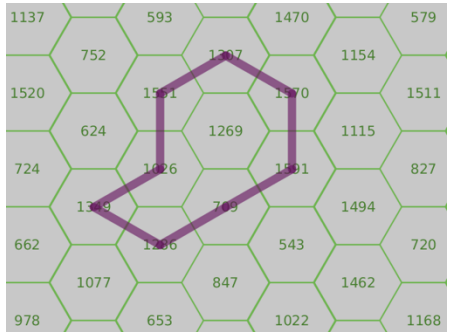


Op rustige momenten, zoals de vroege uurtjes van de nacht, worden alle verschillen door slechts enkele mensen veroorzaakt.









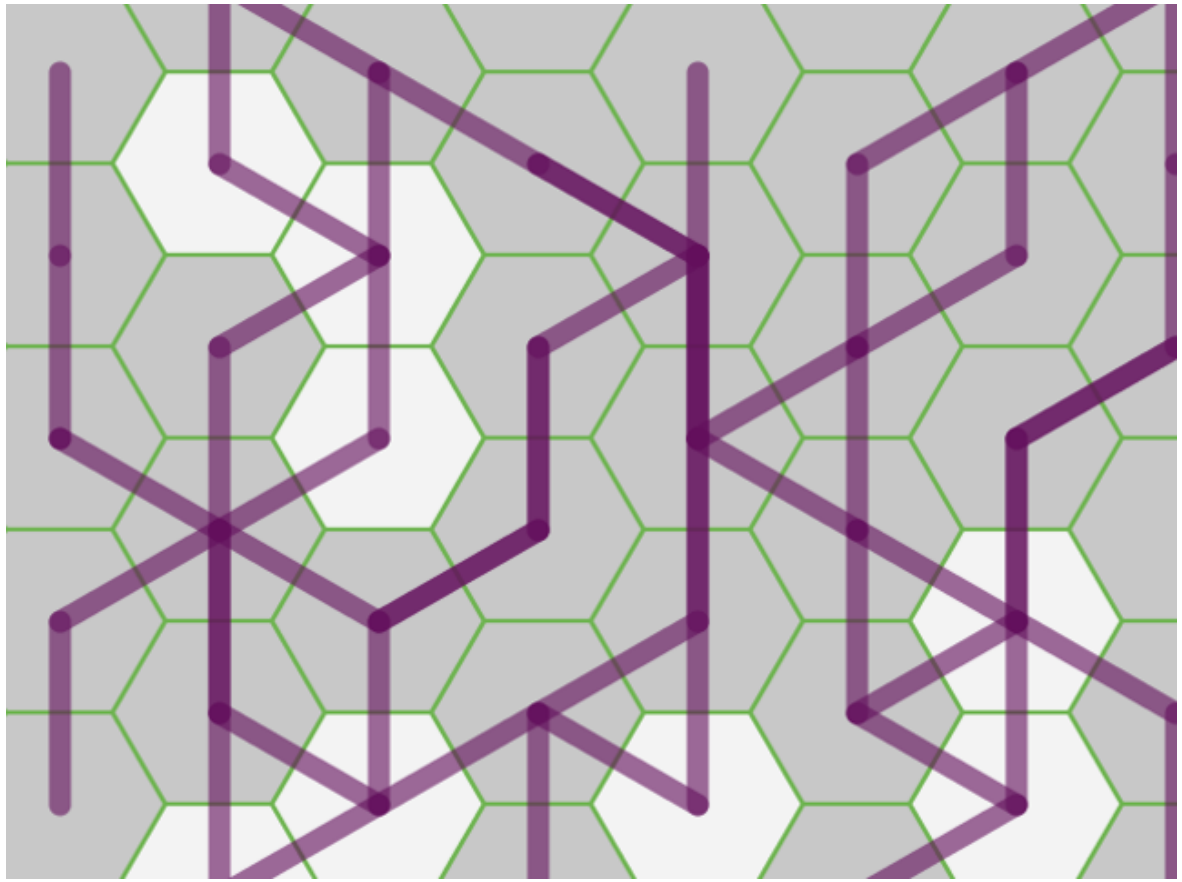
Terugrekenen van geaggregeerde locatiedata

‘Brute forcing’ is het op grote schaal proberen van heel veel combinaties. In dit geval vele mogelijke patronen die, wanneer opgeteld, resulteren in hetzelfde aantal celtellingen als op de lijst met de oorspronkelijk verzamelde data.

Zoals we eerder zagen zijn gebruikte patronen niet willekeurig; mensen slapen thuis, werken op kantoor etc.

Werkt dit? Ja. Onderzoekers hebben dit [ref2] en gerelateerde aanvallen [ref3] aangetoond.

Hoezo dan? De geaggregeerde locatiedata bevat nog steeds voldoende informatie over dag- en nachtritmes om menselijke verplaatsingen te reconstrueren.



Via herleiding naar identificatie

De patronen die het resultaat zijn van de ‘brute force attack’ leiden naar unieke individuen; afzonderlijke personen hebben afzonderlijke verplaatspatronen. Dit is herleiding.

Verplaatspatronen tonen huis, kantoor, afgelegde routes (soms inclusief het vervoersmiddel), sportverenigingen etc. Door in de echte wereld iemand op te zoeken en fysiek te benaderen is het mogelijk om tenminste enkele personen met dit verplaatspatroon te identificeren.

Met behulp van andere databronnen is het mogelijk om op grotere schaal personen te identificeren. Denk bijvoorbeeld aan informatie van de overheid, camerabeelden of bronnen zoals genoemd op de volgende slide.

‘Bring your own data’

Wanneer er al data met tijden en (impliciete) locaties voorhanden is dan zal het proces van de-anonimiseren nog gemakkelijker gaan. Dit komt doordat aan de hand daarvan al een schatting gemaakt kan worden van bepaalde patronen/trajecten, waardoor de ‘brute force’ berekening minder werk heeft. De kans dat iemand kan worden geïdentificeerd wordt verder vergroot wanneer deze data ook unieke identificatoren bevat zoals kaartcodes of klantnummers.

Interessante bronnen van data die helpen om de verzamelde telecomdata te de-anonimiseren zijn bijvoorbeeld:

- Betaal- of creditcardtransacties
- (Openbaar) vervoergegevens
- Gegevens uit klanten-/loyaltyprogramma’s
- Metadata van (digitale) fotoalbums (EXIF)
- Gegevens verzameld door smartphone apps
- Openbare registers

De kosten van een ‘brute force’ aanval

Voorbeeld: in Nederland

- Aantal cellen: ongeveer 3200
- Aantal mensen met een smartphone: ongeveer 15 miljoen

Het ‘brute force’ proces is vooral afhankelijk van werkgeheugen, dus als we weten hoeveel geheugen het proces gebruikt kunnen we iets zeggen over de kosten. Het totaal benodigde werkgeheugen staat gelijk aan het aantal te proberen trajecten maal de grootte van een enkel traject. Als de betreffende data ieder uur opnieuw wordt verzameld, bestaat een traject dus uit 24 achtereenvolgende cellen per dag en 24 x 30 = 720 achtereenvolgende cellen per maand. Een enkel traject kan worden opgeslagen in 1kByte.

Het totaal benodigd geheugen is:

- naïeve implementatie: 15 miljoen x 1kByte = 15GB
- geoptimaliseerde implementatie waarschijnlijk in < 6GB

Dit is goed te doen met een Game PC van rond de €800. Een vergelijkbare cloudgebaseerde machine kan worden gehuurd voor ongeveer €0,40/uur en heeft misschien 1 of 2 weken nodig.

Is geaggregeerde telecomlocatiedata anoniem?

Zoals we eerder zagen:

> Gegevens kunnen worden beschouwd als anoniem als **voor welke partij dan ook**, met inzet van (voor het doel) **redelijke middelen**, het **onwaarschijnlijk** is hieruit personen te **identificeren**.

Middelen bestaan uit tijd en geld. Omdat onderzoekers al hebben bewezen dat dit werkt, is het creëren van een werkende oplossing slechts een kwestie van volhouden, niet van geniaal zijn. De kosten kunnen als relatief laag worden beschouwd, vooral in verhouding tot de opbrengsten.

Het herleiden van personen is te realiseren met behulp van een brute force aanval zoals eerder beschreven. Van daaruit is het mogelijk een persoon te identificeren, vooral bij combinatie met andere (openbare) gegevens. Het gebruik van eigen bronnen met tijd- en locatiegebonden data maken zowel de brute force proces als het identificeren eenvoudiger.
En, er bestaan veel van dit soort bronnen.

Dus nee, het aggregeren van telecomlocatiedata (in uren) is niet voldoende.

Nog enkele overwegingen

- Anonimiseren is lastig. Anonimiseren van locatiegegevens is bijna ondoenlijk.

- De beschreven aanval maakt gebruik van de tijd- en locatiegerelateerde informatie die aanwezig is in de geaggregeerde data. Voorkom het vrijgeven van persoonsgegevens die vaker dan 1x per dag worden bijgewerkt en gebruik robuuste anonimiseringstechnieken. Raadpleeg experts om te bepalen of een verzameling data inderdaad anoniem is. Bij twijfel: behandel de dataverzameling niet als anoniem maar als persoonsgegevens.

- In werkelijkheid heeft elke zendmast 3 antennes die elk een gebied eromheen beslaan. Uw globale locatie wordt dus niet alleen bepaald door de zendmast in de buurt, maar ook door de richting van de antenne(s). Hierdoor is de inschatting van uw locatie tot op 50 meter nauwkeurig.

- Met geavanceerdere programmeertechnieken kan er meer data worden verwerkt.

Referenties

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