Retour adres:

Faculteit Wiskunde en Informatica

Den Dolech 2, 5612 AZ Eindhoven
Postbus 513, 5600 MB Eindhoven

Datum
October 7, 2020

Referentie
07102020

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Faculteit Wiskunde en Informatica
Informatica
Prof. R.W. van der Hofstad
T +31 (0)40 247 2910
r.w.v.d.hofstad@tue.nl

www.tue.nl/~rhofstad

Aan: Vaste commissie voor Economische Zaken en Klimaat

Geachte voorzitter van de Vaste commissie voor Economische Zaken en Klimaat,

Op 15 oktober ben ik uitgenodigd als technisch expert op het gebied van privacy bij de rondetafeldiscussie met uw commissie. Deze schriftelijke inbreng bespreekt telecomdata, en hoe een consortium van wetenschappers en data bedrijven deze data wil gebruiken om COVID-19 te bestrijden. Ik heb, als Hoogleraar Kansrekening aan de TU Eindhoven, inzicht in hoe geanonimiseerde mobiliteitsinformatie gegenereerd kan worden, en hoe deze informatie kan helpen om COVID-19 te bestrijden.

Achtergrond van het gebruik van mobiliteitsdata
Begin maart, in de begindagen van de COVID-19 crisis, werd ik benaderd door het databedrijf Mezuro, met wie u ook spreekt. Mezuro had blootgelegd dat mobiliteit een enorme invloed heeft op de COVID-19 verspreiding. Met de gedetailleerde mobiliteitsinformatie van Mezuro zijn we samen gaan werken om nauwkeurige modellen te maken, om het voorspellend vermogen te vergroten. Binnen een week had mijn team een plan over hoe we dit konden aanpakken, en zijn we hard aan de slag gegaan, vooral omdat het een enorm urgent maatschappelijk probleem betrof. Het was ongekend hoezeer de wetenschap probeerde in te springen om de overheid te helpen om dit probleem op een zo goed mogelijke manier aan te pakken. Ons onderzoek draagt op fundamentele wijze bij aan de volksgezondheid. We hebben vanaf het begin gepoogd contact te leggen met meerdere ministeries, om hierin met hen samen op te trekken, helaas zonder reacties. Hierin werk ik samen met Prof. Heesterbeek (een befaamd epidemioloog), prof. Litvak (hoogleraar algoritmes voor netwerken), en Prof. Chavannes (hoogleraar e-Health).

We werken op basis van mobiliteitsdata verkregen door de Mezuro technologie. Het gaat hier volgens ons om geanonimiseerde data, waaruit persoonsgegevens niet herleidbaar zijn. Dit wordt bevestigd door Dr. Koot van Secura, één van de weinige Nederlandse experts op dit onderwerp, met wie u ook zult spreken. Een probleem dat zich heeft voorgedaan is dat de Autoriteit Persoonsgegevens (AP) eind april naar buiten bracht dat `anonimiseren van locatiegegevens bijna ondoenlijk is'. Hierdoor is er een impasse ontstaan: telecomoperators (Mobile Network Operators, MNOs) willen hierdoor (begrijpelijk) Mezuro geen toegang geven tot hun data uit 2020, en in de politiek wordt nu al behoorlijke tijd nagedacht over een noodwet om telecomdata beschikbaar te maken. Het resultaat is dat we sinds april 2020 nauwelijks zijn opgeschoten en in een impasse zijn beland. In de Bijlage van deze brief leg ik uit hoe ik concludeer dat de Mezuro data geen persoonsgegevens zijn. Helaas is dat stuk iets technischer dan de rest van deze brief.

Mijns inziens is een **noodwet niet nodig en ongewenst**. Er is een technologie die met een druk op de knop geschikte mobiliteitsinformatie, die voldoet aan de VGA richtlijnen, genereert. Op deze data kan de wetenschap en het RIVM haar werk baseren. Als op basis van een noodwet een andere technologie gebouwd moet worden, dan kost dit onnodig extra tijd, tijd die we niet hebben. Daarnaast is het in het kader van data minimalisatie onnodig en onwenselijk om persoonsgegevens te gebruiken om

ziekteverspreiding te voorspellen.

Waarvoor is mobiliteitsdata nuttig en wat is de nuttigste vorm van data?
Ons onderzoeksconsortium gebruikt mobiliteitsdata om schattingen te produceren van het aantal bewegingen tussen gemeentes. We doen dit nu op basis van de Mezuro data uit 2019, omdat de data van 2020 vanwege bovenstaande impasse (nog?) niet beschikbaar is. Die data bevatten de aantallen bewegingen tussen de 1250 zogenaamde Mezuro gebieden, per uur. Hierbij worden aantallen kleiner dan 16 niet gerapporteerd. Deze bewegingen worden vervolgens in een epidemiologisch model meegenomen, waarbij de infectie zich verplaatst tussen de gebieden door de bewegingen van mensen ertussen. Dergelijke analyses zijn zeer relevant, omdat ze inzicht geven in het effect van, bijvoorbeeld, regionale lock-downs en ander maatregelen. Ik wil hier wel benadrukken dat de vorm waarin de data aangeleverd is cruciaal is. Aan te precieze data van de vorm wie precies waar was op regelmatige momenten in de tijd hebben we voor epidemiologische modellen weinig. Cruciaal is dat de data aangeeft dat de reizigers een bepaalde tijd hebben doorgebracht in een plek, ofwel, het moet oorsprong-destinatie data zijn. Immers, dat iemand met de auto door een gemeente reist, betekent niet dat er daar ook iemand besmet kan worden. Dat kan alleen als er enige tijd wordt doorgebracht. Die informatie dient met slimme big-data analyse uit de ruwe locatiedata te worden verkregen. Hiervoor dient een nieuw platform gebouwd te worden, of het al uitontwikkelde en geteste platform van Mezuro gebruikt te worden. Het laatste is sneller.

Ons onderzoeksvoorstel is inmiddels gehonoreerd door NWO ZonMw. Hierin gaan we deze data combineren met data uit de COVID-Radar app (waarin mensen op vrijwillige basis aangeven hoeveel contacten alsmede ziekteverschijnselen ze hebben), om een idee te krijgen in welke regio's de kans op besmettingen het grootst is. Dit doen we om beleidsmakers van advies te kunnen voorzien, door slimme dashboards te bouwen die risico's weergeven, in plaats van aantallen besmettingen en ziekenhuisopnames. Deze aantallen lopen immers inherent achter op de besmettingen, waardoor problemen te laat gesignaleerd worden.

Impasse doorbreken
Op dit moment maken wij in onze modellen gebruik van mobiliteitsinformatie uit 2019, wat niet optimaal is. Graag willen we aan de slag met de data van 2020, omdat met name deze inzicht geven in hoe mensen nu bewegen. Hiervoor is het zaak dat de overheid de MNOs de opdracht geeft om deze data toegankelijk te maken. We zijn ervan overtuigd dat die stap de impasse zal doorbreken.

De urgentie om deze impasse te doorbreken is zeer hoog, omdat de 2020 data inzichtelijk maakt hoe de Nederlandse bevolking heeft gereageerd op de, zeer ingrijpende, maatregelen en dringende adviezen van de overheid. Dat geeft inzicht in hoe de bevolking zal reageren op eventuele toekomstige maatregelen. Telecomdata blijft in principe 6 maanden beschikbaar bij de MNOs. We lopen dus het risico dat de data uit het begin van de COVID-19 crisis snel verloren zal gaan, wat een ramp zou zijn voor de wetenschap, maar ook voor beleidsmakers. Dit risico dient zonder dralen afgewend te worden.

Hoe nu verder?
Wij willen benadrukken dat voor ons de bescherming van de persoonlijke levenssfeer van personen een groot goed is dat ook wij willen beschermen. We hopen dat de politiek inziet dat er een data platform beschikbaar is dat meteen ingezet kan worden, dat voldoende informatie biedt voor de benodigde risico inschattingen, en dat geen persoonsgegevens bevat. Het vereist wel dat de overheid op een fundamenteel andere manier naar de materie gaat kijken, en **snel actie** onderneemt.

Met vriendelijke groet en hoogachting, mede namens Prof. dr. N. Litvak (TU/e en U Twente),

Prof. dr. R.W. van der Hofstad, Hoogleraar Kansrekening, Technische Universiteit Eindhoven

Bijlage: Argumenten waarom de data technologie van Mezuro geen persoonsgegevens betreft.

Het belangrijkste argument waarom locatiegegevens privacy gevoelige informatie zijn, wordt gegeven in een paper van Xu et al. uit 2017. Daarin wordt, terecht, gesteld dat het moeilijk is om locatiegegevens van mobiele telefoongebruikers goed te anonimiseren. Zelfs het aanleveren van alleen aantallen telefoongebruikers die verblijven in de telecomcellen (waarin dus niet verwezen wordt naar wie er precies zijn) maakt de data niet anoniem. Deze paper, die Prof. Litvak en ik goed hebben bestudeerd, is zeer interessant, omdat de conclusie misschien wat onverwacht is, en ingrijpende consequenties heeft op wat wel, en wat niet, privacy gevoelige data is. Deze paper vormt ook de reden waarom de Autoriteit Persoonsgegevens stelt dat `anonimiseren van locatiegegevens bijna ondoenlijk is'.

Laten we daarom kort samenvatten wat de paper precies laat zien. In de paper wordt uitgegaan van locatiedata van gebruikers, geaggregeerd naar aantallen. Om precies te zijn, er wordt verondersteld dat we op regelmatige momenten precies weten hoeveel gebruikers er in welke cellen zitten. Op basis hiervan kunnen Xu et al. (2017) een statistische schatting maken van de paden die gebruikers volgen, en omdat de startpunten van deze paden veelal het thuisadres van deze gebruikers zijn, is het dus mogelijk om persoonsgegevens terug te halen. Hier wil ik benadrukken dat het schatten van deze paden mogelijk is omdat de bewegingspatronen van mensen zeer voorspelbare patronen volgen. Bijvoorbeeld, mobiele telefoongebruikers bewegen over het algemeen heel weinig in de nacht. Ook zijn bewegingen continu, zodat als een gebruiker van A naar C gaat, deze gebruiker hoogstwaarschijnlijk ook door het tussenliggende B gaat. De sleutel in de methode van Xu et al. (2017) is dat de exacte aantallen van gebruikers in elke telecom cel bekend zijn op een aantal tijden. Hieruit worden de volgende verblijfsplekken geschat op basis van een aantal (ook in onze ogen vrij onschuldige) aannames. Tot slot wordt de methode getest op echte data sets waarvan er een ground truth is en die op precies de beschreven manier geaggregeerd worden. Hieruit blijkt dat de methode behoorlijk goed werkt, en de gevolgde paden van gebruikers in vele tientallen procenten correct schat.

Waarom kan de methode van Xu et al. (2017) niet toegepast worden op de Mezuro data?

Prof. Litvak en ik hebben nogmaals, vanuit een wiskundige invalshoek, de Mezuro dataset bekeken om te kijken of de technieken van Xu et al. (2017) erop van toepassing zijn. Wij wij denken sterk dat dit niet zo is. In het bijzonder kunnen wij naar eer en geweten niet bedenken hoe enige natuurlijke persoon uit deze informatie te herleiden zou zijn. Laat ik nu uitleggen hoe Prof. Litvak en ik tot deze conclusie komen.

De Mezuro data is op een aantal sleutelaspecten anders dan wat door Xu et al. (2017) wordt aangenomen. Allereerst, en volgens ons het belangrijkste, is dat de Mezuro dataset verplaatsingen geeft tussen oorsprong en destinatie. Ofwel, hoeveel mensen zijn van een regio naar een andere gegaan, en zijn daar minstens 30 minuten verbleven, per uur. We weten dus op geen enkel moment precies hoeveel gebruikers zich bevinden in de Mezuro gebieden (dit is een door Mezuro bedachte geografische indeling van Nederland dat een gemiddelde grootte heeft van 33 km2 en 13.000 inwoners). In meer detail, als iemand van Utrecht naar Eindhoven reist, en daar enige tijd doorbrengt, dan weten we niet of, en zo ja op welk moment, hij of zij zich in Den Bosch bevond. Xu et al. (2017) nemen aan dat de data precies aangeeft hoeveel mensen er in welk gebied zijn, en die data hebben we dus niet tot onze beschikking. Dat betekent dat we geen startpunt hebben van waaruit we individuele paden van gebruikers kunnen berekenen. Immers, de methode van Xu et al. (2017) gebruikt continuïteit van bewegingspatronen, en daarvoor is het zaak om tussenliggende cellen te kennen op de momenten van rapportage. Daarnaast zijn nog drie relevante anonimiseringsstappen gedaan door Mezuro:

1. Aantallen worden alleen gerapporteerd als deze tenminste 16 gebruikers betreft, dus vóór ophoging, hetgeen het terugrekenen naar personen complexer maakt. Deze bewerkingsstap heet 16-anonimity, en is een sleutelstap om data te anonimiseren.
2. De technologie van Mezuro maakt gebruik van maar één MNO, met een bereik van ruwweg een derde van de Nederlandse bevolking, zodat aantallen opgehoogd moeten worden om een representatief beeld te geven voor Nederland. Deze ophoging wordt per dag en per Mezuro gebied bepaald en varieert afhankelijk van o.a. het aantal actieve gebruikers. Alleen de opgehoogde getallen worden geleverd aan Mezuro. De ophoog factoren worden nooit uitgeleverd aan klanten van Mezuro, hetgeen het terugrekenen naar personen nog complexer maakt.
3. De plaatsbepaling in welke cel een mobiel bevindt wordt vastgelegd op basis van een door de MNO theoretisch vastgelegd gebied met een waarschijnlijkheid van 90% waarna deze wordt geaggregeerd naar een Mezuro gebied. Alleen het Mezuro gebied wordt gebruikt alvorens te aggregeren.