Communication from Public

Name:

Date Submitted: 12/03/2023 10:18 PM

Council File No: 22-0392

Comments for Public Posting: No one wants digital billboards, especially not close to the Ballona Wetlands Reserve or close to the Tule Wetlands. The Audubon Society is against digital billboards (community Comment Letter file 10-26-22). The neighbors are against digital billboards (community Comment Letter files 9-28-22 by the Villa Marina Council and 10-24-22 by the Del Rey Residents Association). I am against digital billboards. Proposed Structures FF-29 and FF-30 are within a street's width of the Ballona Wetlands Reserve and near the Tule Wetlands. These 2 structures with 4 massive billboards will serve no purpose other than to raise money, and even that is questionable as I have not seen any complete cost assessment, including the millions (per the internet) to build the structures and the reduced revenue (I assume) from the billboards that are less than a mile from the proposed sites (why pay for advertising no one will see since they'll be watching the digital billboards?). Note that City Planning was apparently not allowed to evaluate the project, just tasked with writing the language to move it forward. I live near the proposed site of FF-29 and FF-30. I have no confidence that these proposed digital billboards will be seen by cars on the 90 freeway, but not seen from the residential units located parallel to the 90 and next to the Ballona Wetlands Reserve. I have no confidence that these proposed digital billboards will not adversely impact the wildlife in the area who do not recognize lines drawn on a map (remember the goose landing in Dodger stadium?). Digital billboards were banned. Let's keep it that way. Thank you, Pat Allinson Del Rey Resident

Communication from Public

Name:	Casey Maddren/Citizens for a Better Los Angeles
Date Submitted:	12/03/2023 11:41 PM
Council File No:	22-0392
Comments for Public Posting:	Citizens for a Better Los Angeles submits the attached comments to the PLUM Committee regarding the Transportation Communication Network.

CBLA

Citizens for a Better Los Angeles

December 3, 2023

Planning & Land Use Management Committee
Los Angeles City Hall
John Ferraro Council Chamber, Room 340
200 N. Spring St.
Los Angeles, CA 90012

Posted to council file via City Clerk Public Comment Form.
Sent via e-mail to LA City Councilmembers.

Re: Transportation Communication Network Ordinance/Digital Billboards
PLUM Hearing, December 5, 2023, Agenda Item 22
Council File: 22-0392
STRONGLY OPPOSED

Members of the Planning & Land Use Management Committee,

Citizens for a Better Los Angeles (CBLA) is a nonprofit public benefit corporation organized to protect the rights and promote the well-being of all people throughout Los Angeles County.

We are writing again to express our deep concern about the Transportation Communication Network and the associated Ordinance. This is a dangerous program, and the process used by the City and Metro to push the program forward has been seriously flawed, not least by the fundamental dishonesty regarding the TCN's origin. We're concerned by many aspects of the program, but primarily by: 1) The fact that it is in direct conflict with the right to privacy conferred by the California Constitution; 2) The TCN EIR's failure to adequately assess the program's numerous impacts, as well as its failure to accurately describe the program; 3) The potential for increased roadway fatalities in a city that has already seen a significant increase in traffic deaths.

We're submitting the additional arguments below in the hope that the Council will seriously consider the grave consequences of adopting this program, and vote to reject it.

Sincerely,
Casey Maddren
Citizens for a Better Los Angeles

Digital Out-of-Home Advertising Is Based on Collecting Personal Data

We have previously submitted comments regarding privacy concerns, including in our November 1, 2023 letter to the PLUM Committee. We submit the following information to further explain the reasons for our concern.

Over the past three decades, private corporations have built a vast data-collection network with almost no regulatory oversight. Most citizens are aware that tech companies, retailers and advertisers collect data from interactions with their personal devices, and in many cases those collecting data offer consumers a chance to opt-out if desired. However, what is less known is that companies utilizing digital out-of-home advertising are now collecting data from personal devices belonging to people in both private and public spaces. Citizens are not notified that their data is being collected and are given no opportunity to opt-out. This is especially concerning when we realize that data is being collected from minors who carry smartphones or tablets.

Advertisers then use the data collected from digital OOH assets and combine it with more data through relationships with data brokers. Advertisers and data brokers routinely claim that they collect no personally identifiable data and that all data is anonymized. This claim is misleading at best and completely dishonest at worst.

Capturing an IP address from someone's phone may not identify them by name, but because the IP address is a unique identifier, it's child's play to associate the IP address with an individual's name and a broad range of additional data. Phones also have mobile advertising IDs (or MAIDs) which are also unique identifiers that easily allow advertisers to identify an individual.

To talk about specifics, let's look at claims made by Clear Channel, one of the advertisers that Metro is already deriving revenue from under its billboard program. The following text comes from a Clear Channel web page. *[See Exhibit A.]*

Driving innovation for Pharma
https://clearchanneloutdoor.com/case-studies/driving-innovation-for-pharma/

Objective

A pharmaceutical company sought to build awareness and increase sales of a brand used to treat a specific, moderate-to-severe medical condition.

Solution
Clear Channel Outdoor (CCO) leveraged our proprietary CCO RADAR suite of data-driven solutions, and our best-in-class data partners, Veeva Crossix and LiveRamp, to onboard the brand's target audience into our RADARView platform for campaign planning. We further amplified the OOH campaign with a mobile retargeting campaign through RADARConnect. And finally, with RADARProof, we were able to measure brand awareness, doctor visits, and number of prescriptions issued.

Results
The OOH campaign successfully drove brand awareness and further retargeting through mobile, increased visitations to a specialist, and also drove lift in prescriptions for the overall category and for the advertised drug.

Please note the reference to "mobile retargeting". In some cases, this refers to sending ads or offers to consumers who have already visited a web site. But with digital OOH, consumers have not interacted with a web site, but have merely passed near a digital billboard. Please also note that Clear Channel states, "And finally, with RADARProof, we were able to measure brand awareness, doctor visits, and number of prescriptions issued." It's hard to understand how Clear Channel can measure doctor visits and prescriptions issued without tracking the movements and activities of individuals.

Clear Channel partners with Geopath to execute advertising campaigns. Clear Channel's web page explains how Geopath functions.

Geopath and audience measurement
https://clearchanneloutdoor.com/geopath-measurement/

Geopath enables media buyers, sellers, and advertisers to strategically plan and execute effective OOH advertising campaigns. Geopath's approach is centered around understanding consumers' journeys and exposure to OOH in the physical world, by observing and analyzing mobile location data from smartphone applications and connected cars. [Emphasis added.]

We also offer the following text from Geopath's own web site.

Geopath
https://geopath.org/

From Document: "Geopath Insights"
Section 1: The Evolution of Geopath Insights
https://geopath.org/wp-content/uploads/2019/11/Geopath-Standards-and-Best-Practices-Document__11-7-19_Section-1.pdf

[See Exhibit B.]

The following text is found within a graphic on page 4:

> *Nationwide Location and Movement Data*
> *Using mobile locations data, we identify movement and activity patterns that help us understand how and why people travel -- their pathway, mode, volume, frequency motivation and destination. [Emphasis added.]*
>
> *Connecting Trip Paths to OOH Media*
> *We then contextualize population movement alongside audited inventory to quantify audience exposure to metrics -- the proximity, dwell time, opportunity-to-see, and likelihood to see. [Emphasis added.]*

CBLA is not alone in its concerns about protecting privacy. Surveillance advertising is a problem that has caught the attention of members of the US Congress. In September of this year, Representatives Jan Schakowsky and Anna G. Eshoo and Senators Ron Wyden and Cory Booker introduced the Banning Surveillance Advertising Act. At the time the bill was introduced, Representative Eshoo remarked that, "[Surveillance advertising] is at the root of disinformation, discrimination, voter suppression, privacy abuses, and so many other harms."

Banning Surveillance Advertising Act, Press Release from US Rep. Jan Schakowsky
https://schakowsky.house.gov/media/press-releases/schakowsky-eshoo-wyden-booker-introduce-bill-ban-surveillance-advertising

Rather than acknowledging and addressing the potential privacy problems that could arise from a massive expansion of the number of digital billboards in Los Angeles, both Metro and the City of LA have chosen to ignore the issue completely.

Using "Anonymized" Data to Identify Individuals Is Easily Accomplished

While advertisers such as Clear Channel claim that they don't collect personally identifiable data, and insist that the data they do collect is anonymized, using a data set to identify individuals is simple with current technology. Re-identification is easily accomplished, and the practice is commonplace. As noted above, we don't understand how Clear Channel is able to present pharmaceutical clients with information about doctor visits and prescriptions issued as a result of an ad campaign without identifying individuals who have been exposed to their digital OOH ads.

Researchers who have studied re-identification find that, contrary to the claims made by some, it is not difficult to identify an individual, even in very large data sets. *[See Exhibit C.]*

"The risk of re-identification remains high even in country-scale location datasets", Ali Farzanehfar, Florimond Houssiau, Yves-Alexandre de Montjoye, 12 March 2021 https://www.sciencedirect.com/science/article/pii/S2666389921000143

The following excerpts provide a brief summary of the researchers' conclusions.

> *Here, we empirically measure, mathematically model, and provide a lower bound on the relationship between the size of a dataset and the risk of re-identification. Our results all show that re-identification risk decreases very slowly with increasing dataset size. Contrary to previous claims, people are thus very likely to be re-identifiable even in country-scale datasets.*
>
> *Taken together, these results show that the scale of a dataset does not prevent re-identification. Human mobility, much like a physical fingerprint, is highly unique and can be used to find a person across mobility datasets.*

The City of LA Should Be Acting as Lead Agency, Not Metro

In addition to the many ways in which the EIR for the TCN is flawed, we must ask why Metro has served as lead agency for environmental review rather than the City of LA? The TCN digital billboards will all be installed within the boundaries of the City of LA. The program's environmental impacts will be borne by the citizens of LA. Furthermore, Metro being lead agency keeps the City of LA from oversight of the program's costs, yet the City has a fiduciary duty to its citizens. Finally, the City of LA is required by its Charter to give notice of programs like TCN to neighborhood councils. By allowing Metro to serve as lead agency, it appears that the City is trying to circumvent this requirement. While many NCs have submitted comments, we are not aware of any formal effort by the City to invite review of the TCN by NCs.

Digital Billboards Are Not Permitted in Parcels Zoned for Public Facilities

Many parcels in the Program are zoned Public Facilities (PF) land where digital billboard/commercial advertising is not a permitted use. PF parcels are zoned for projects that serve a public purpose, for example, transit, schools, libraries and other projects that benefit the public. In response to the Executive Order enacted by Mayor Bass and as part of their housing plan, Metro adopted a report including 17 parcels identified for joint housing development projects, including FOUR parcels that are included in the TCN Program.

Public Facilities and other Metro surplus land should be reserved for uses that benefit the public. The zoning code defines what can occur on Public Facilities parcels, and billboards used mainly for advertising are not included. In fact, Metro's Vision 2028 Strategic Plan (adopted in 2018) encourages the development of affordable housing near transit (including on Public Facilities lots) to give more people, especially in low-income communities, better access to transit. In June 2021, Metro published an update to the Metro Joint Development Program (JD), a

real estate development program for properties owned by Metro (adopted on October 28, 2021), approving a ten-year Joint Development goal of completing 10,000 housing units, at least 5,000 of which would be income-restricted (https://la.urbanize.city/post/metro-wants-complete-10000-homes-agency-owned-land). The Joint Development program prioritizes the construction of housing on Metro property, but potential conflicts with the TCN were not disclosed or analyzed in Metro's EIR.

Metro Has Not Acknowledged this Project Goes Back More Than a Decade

Though it has been rebranded as the Transportation Communication Network, this program is actually an extension of Metro's Billboard Program, which has been in existence for more than a decade. Beyond that, Metro and the City have been planning to expand the program into the City of LA since at least 2016. This expansion depended on the City adopting a new ordinance allowing additional digital billboards, which is now before the Council disguised as the TCN Ordinance. In a January 23, 2023 letter to Metro's Board of Directors, CBLA cited the August 18, 2016 Metro Board Report which discusses the plan on page 6:

> *6. Los Angeles: All Vision and Metro staff have had preliminary discussions with the City of Los Angeles. The City is considering various options for the adoption of a new billboard ordinance. The City of Los Angeles Project offers Metro the greatest potential for new revenue from the conversion of static billboards to digital billboards.*

In addition to ignoring the reality that the TCN Program is an extension of Metro's long-running Billboard Program, Metro has failed to produce e-mails and other records related to the Program for the administrative record in pending litigation. Citizens for a Better Los Angeles and Coalition for a Beautiful Los Angeles have filed a lawsuit to stop the implementation of the TCN/Billboard Program (Los Angeles Superior Court, Case No.: 23STCP00670), but in preparing the administrative record, Metro has produced no records related to the Program dated prior to 2020. Metro has not produced records related to its initial communications with the City of LA regarding the program, nor has it included its agreements with Allvision (AKA All Vision) which manages the Billboard Program for Metro.

EXHIBIT A

Clear Channel Outdoor

Back to Case Studies

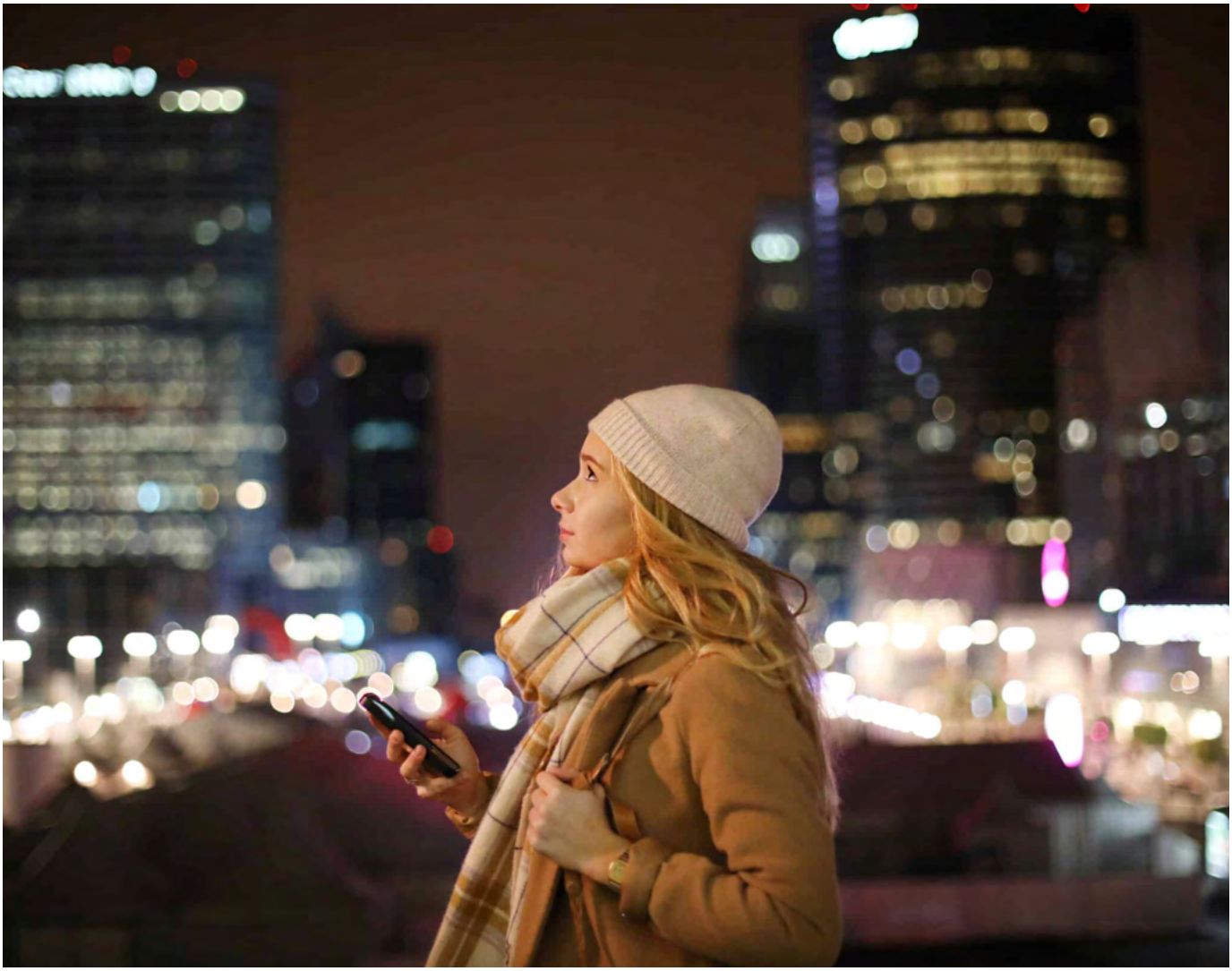
RADARProof

Audience Targeting, Brand Awareness, Doctor Visits & Script Lift

Driving innovation for Pharma

Download Case Study

- Integrated Pharma brand's target audiences for OOH planning
- Raised consideration intent by 50%
- Increased doctor visits by 20%
- Increased new prescriptions for the category by 75%
- Delivered a 76% lift in sales of the advertised brand



Objective

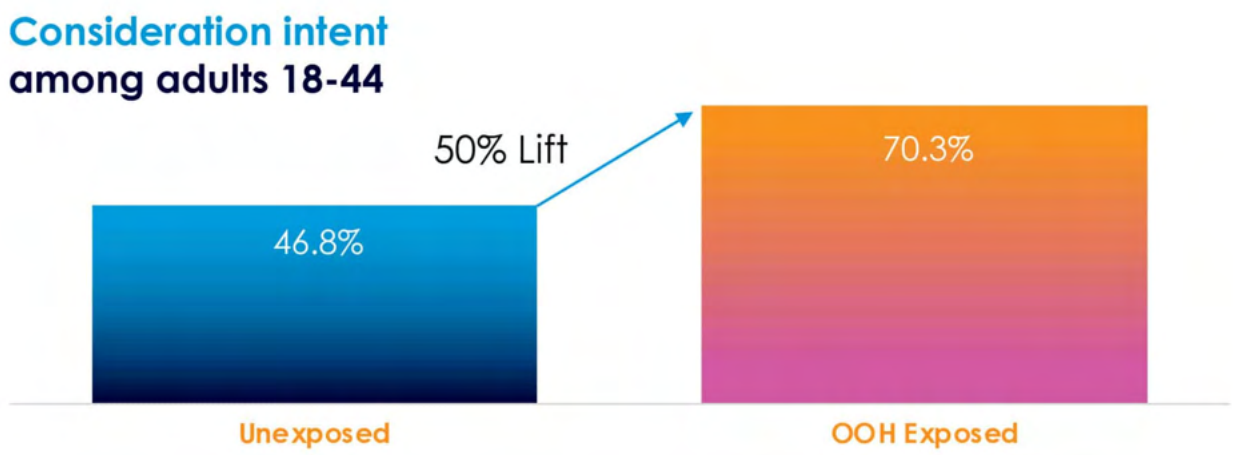
A pharmaceutical company sought to build awareness and increase sales of a brand used to treat a specific, moderate-to-severe medical condition.

Solution

Clear Channel Outdoor (CCO) leveraged our proprietary CCO RADAR suite of data-driven solutions, and our best-in-class data partners, Veeva Crossix and LiveRamp, to onboard the brand's target audience into our RADARView platform for campaign planning. We further amplified the OOH campaign with a mobile retargeting campaign through RADARConnect. And finally, with RADARProof, we were able to measure brand awareness, doctor visits, and number of prescriptions issued.

Results

The OOH campaign successfully drove brand awareness and further retargeting through mobile, increased visitations to a specialist, and also drove lift in prescriptions for the overall category and for the advertised drug.

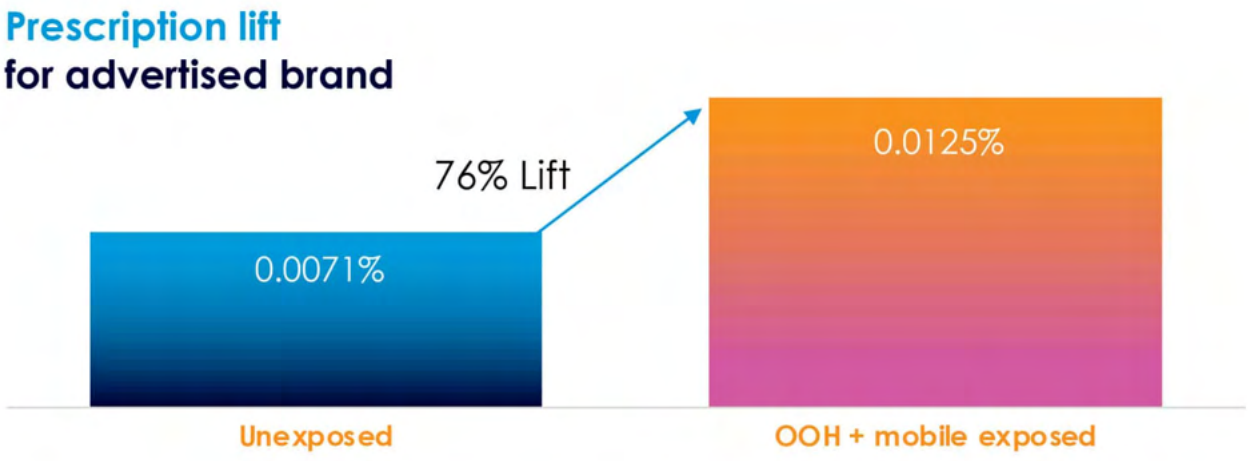


OOH drove 50% lift in consideration intent

CCO RADARProof is our campaign measurement and attribution solution that allows us to observe consumer devices exposed to an OOH campaign. In this pharma study, we learned that the campaign effectively reached the target audience and influenced their intention to consider the product. Among those exposed to the OOH campaign, consideration intent was lifted by 50%.

Visits to specialty doctors rose 20%

The data we extract from CCO RADAR products underscore the effectiveness of a combined OOH + mobile campaign to drive additional engagement. This OOH + mobile campaign drove up visits to specialist by 20% among exposed audiences.



OOH + mobile lifted conversions and drove up incremental sales

In addition, the data showed that audiences exposed to the OOH + mobile campaign helped to lift brand conversions by 76%, and pushed incremental sales up to 75%, in new patient prescriptions for the category.

Source: CCO RADAR; Kantar; Veeva Crossix, January, 2022

How can we help you?

We invite you to find out exactly what it means to GET MORE WITH US. Reach for expert help and smart, customized solutions. We're here to talk options, plan your campaign, or simply answer questions. Just fill out the form. We'll be in touch quickly.

First Name *

First Name

Last Name *

Last Name

Company Name *

Company

Job Title *

Job Title

Zip Or Postal Code *

Phone Number *

Zip Or Postal Code

Phone Number

Email *

Email

Message *

How can we help with your outdoor advertising needs?

I would like to receive Clear Channel Outdoor newsletters and updates

By submitting your information, you acknowledge our Privacy Statement and agree to our Terms of Use.

Submit

Connect with your audience. Drive measurable results. Get more with Clear Channel Outdoor.

Contact Us

About Us

Solutions

Investors

Section 1:

The Evolution of Geopath Insights

Section 1

Table of Contents

A Quick Review:
The Evolution of Geopath Insights

The original TAB ratings put OOH on a level playing field with other media channels by allowing the industry to move from "showings" to measures more commonly used in other channels.

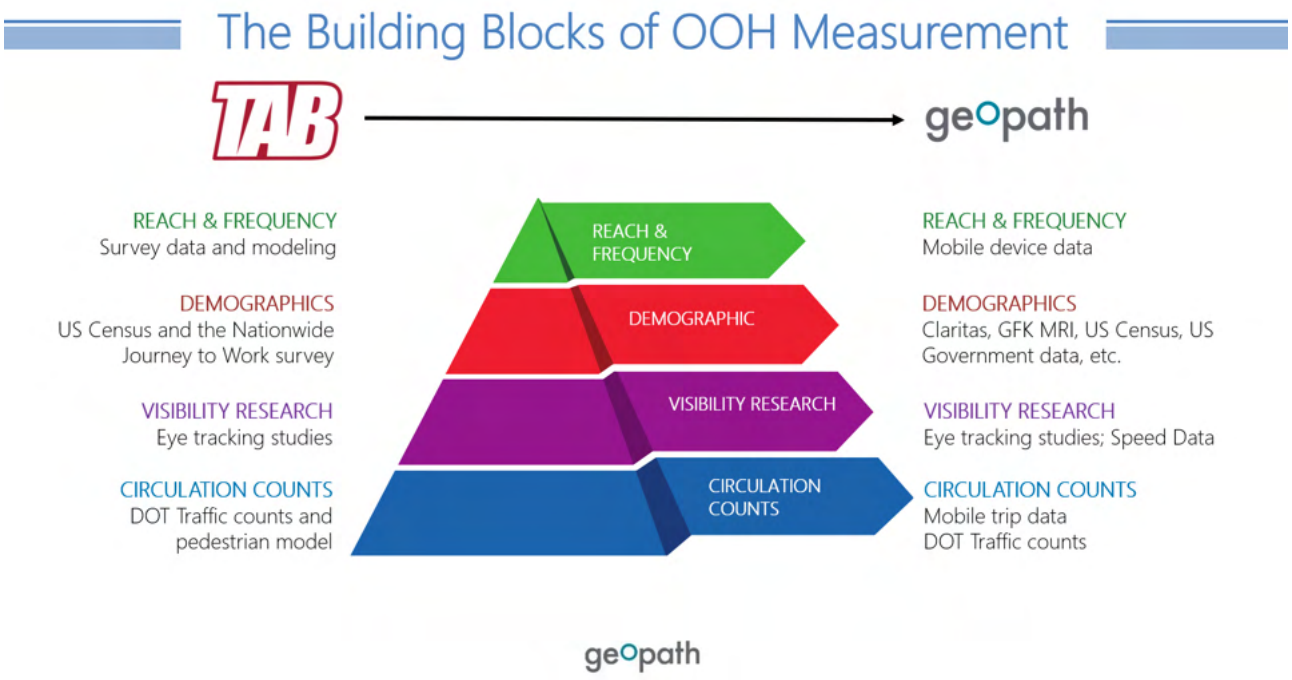
Weekly Impressions | TRPs | Reach | Frequency

Geopath is now enhancing its measures with greater granularity and precision, including additional audiences, increased geographic options, and more detailed analysis of inventory.

Audience | Coverage | Composition | Indices

The Building Blocks of OOH Measurement

Geopath Insights is built upon the same core building blocks. The key differences are the data inputs.



Contextualizing Audience and OOH Media

Geopath curates all of this aggregated and anonymized data from across all roadways and places in the US to create a fully contextualized movement matrix of the entire population. Only when the movements of the full population are understood can we fully compare all OOH media locations and understand the audiences viewing the media. The graphic below outlines how Geopath understands audience movement and connects it to OOH media to develop its measures.

Comparing Geopath Insights – Yesterday to Today

Distributive computing and the mobile data available have allowed us to enhance the capabilities available to our members, whether you are accessing our data via our API or the new Geopath Insights Suite (which will ultimately replace our legacy tools). On the next page is a comparison of the legacy capabilities to the enhanced capabilities that will ultimately be available through Geopath.

Measurement Enhancement:
Comparison of Legacy Data to Today

Audited Inventory Database	Legacy Systems/Data	Insights Today
Roadside	Yes	Yes
Transit - Place-Based	Yes	Yes
Transit - Fleet - Scheduled Routes	Yes	Yes
Place-Based	N/A	Yes
Fleet - Dynamic Routes	N/A	Under Development

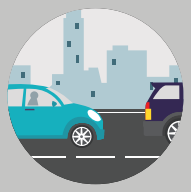
Audience Measurement Data	Legacy Systems/Data	Insights Today
Reporting Precision		
Single Location	Yes	Yes
Inventory Sets	Yes	Yes
Individual Ad-play	N/A	Yes
Geographic Resolution	**4,388**	**40,000+**
National	N/A	Yes
DMA (210)	Yes	Yes
CBSA (942)	Yes	Yes
County (3,236)	Yes	Yes
ZIP Code (32,336)	N/A	Yes
Custom (∞)	N/A	Yes
Temporal Resolution (time scale options)	**1**	**2,016**
Annual (1)	Yes	Yes
Seasonal (4)	N/A	Under Development
Monthly (12)	N/A	Under Development
Day of Week (7)	N/A	Yes
Hour of Day (24)	N/A	Yes
Audience Segments	**500**	**8,000+**
Census Demographics	Yes	Yes
Enhanced Demographics (Housing, Commute, Language)	N/A	Yes
Consumer Behaviors	N/A	Yes
Psychographics	N/A	Yes
Segmentation (PRIZM)	N/A	Yes
Segment Cross-Tabs	No	Under Development

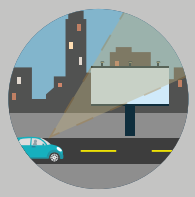
Technology Platforms	Legacy Systems/Data	Insights Today
Forecasting Tools		
Inventory Search	Yes	Yes
Audience Search	N/A	Yes
Market Planning	Yes	Yes
Measurement Tools		
Campaign Delivery	Yes	Yes
Historical Analysis	Yes	Under Development
API	Yes	Yes

How Impressions Have Evolved

While the data we harness from mobile devices and connected cars creates a more robust measurement system, increasing our understanding of the audience viewing inventory, there will be some changes in the impressions delivered by Geopath audited inventory.

Overall, there are seven key components that impact changes to OOH impressions. The following table provides an overview of each component, why it is important to Geopath's measurement, and what has changed.

Component	What Has Changed	Why It Matters
 VEHICULAR TRAFFIC COUNTS	Geopath is no longer solely reliant upon manually collected information from government resources for traffic counts. Mobile technology provides a better estimate of hourly traffic on roadways throughout the week and throughout the year Millions of traffic count locations can now be cross-referenced and aligned with mobile trip data and calculated for every unique road segment in the US by direction.	Traffic counts are the basic building block that allow Geopath to understand overall audience circulation. While a high traffic count may lead to higher impressions, other factors such as illumination, vehicle occupancy, and directionality all play a role.
 PERSONS PER VEHICLE	Mobile data, regional patterns, and trip purpose information now allow for variable occupancy • Every road segment in the country will have a unique vehicular occupancy calculation	Different markets have very different travel and transportation usage patterns. Markets with higher vehicle ownership have fewer people per car. The expected number of people in a vehicle is different depending on the trip purpose. Commuting trips have low occupancy, while shopping and leisure trips have high occupancy. Higher levels of occupancy have a positive impact on impressions as they lead to a higher number of "opportunities-to-see."

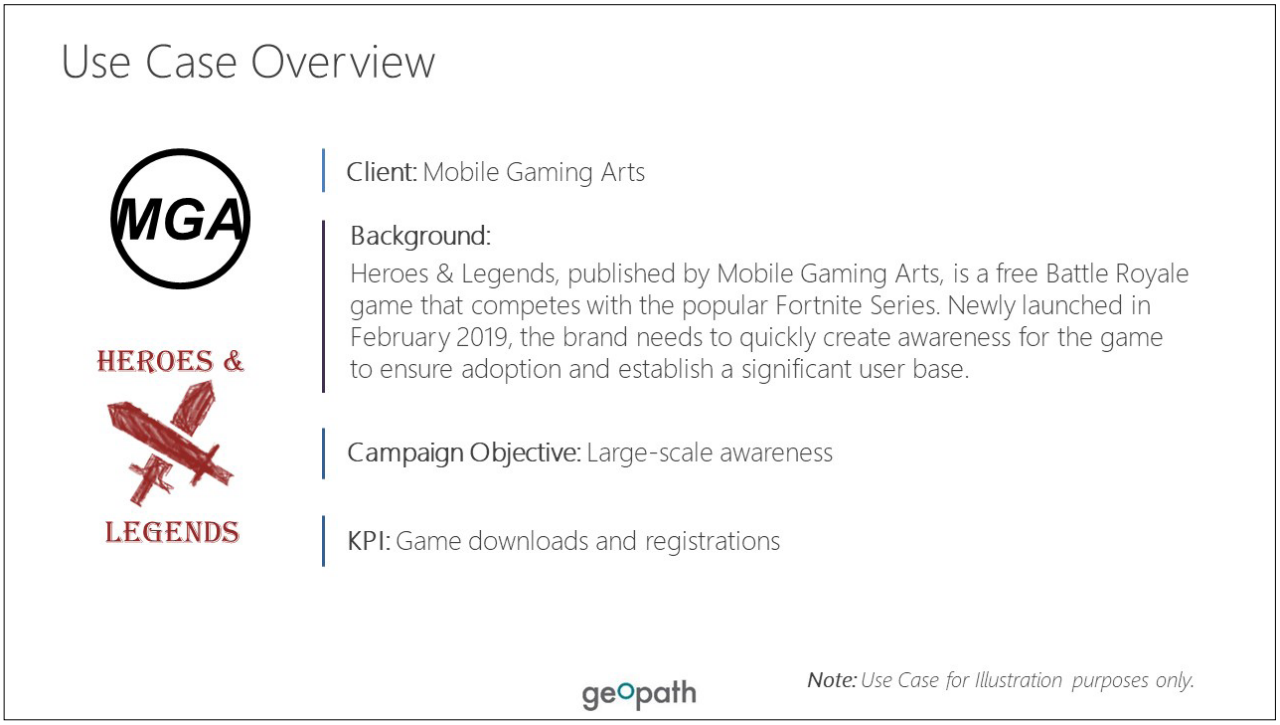
Component	What Has Changed	Why It Matters
 PEDESTRIAN TRAFFIC	Pedestrian pathways now have unique counts, factoring in mobile activity, employment density, business locations, and more. New default walking speed is 3.1 MPH (vs. 3.4 MPH).	Pedestrian traffic can make up the majority of audience in central business districts, commercial, entertainment, and tourism areas. The use of mobile applications for social, fitness, weather, and navigation has created a powerful resource to understand activity on a block by block level.
 ILLUMINATED CIRCULATION	Sunrise and sunset at the inventory location by season, in conjunction with illumination periods, are used to gauge visibility and circulation.	Many OOH assets rely upon ambient light for illumination. These units can only be seen by traffic during daylight hours. It is important to know the location of a unit within a time zone as the sunrise and sunset times can vary up to an hour. Daylight hours may change significantly throughout the year depending on latitude.
 VISIBILITY ADJUSTMENT	Angle to oncoming traffic taken into account, providing infinite permutations vs. LH/RH/ Center, Parallel/Perpendicular. Observed dwell time, degrees off-center (at optimal view), and apparent size (at optimal view) taken into account.	Visibility is dependent on several factors: How large does the media **APPEAR** within the audience's field of view? **WHERE** is the media within the audience's field of view? How much **TIME** does the audience have to see the media? Detailed road network information and inventory attributes enable precise visibility calculations.
 SPEED/ DWELL TIME	Hourly speed data for all US roadways.	Dwell time influences the likelihood of content being seen, as well as the number of spots that a single person has an opportunity to see. The greater the time that an audience dwells near an OOH media location: • the more likely they are to look at the unit • the more opportunities those audiences have to see multiple spots on the same unit Speed data from connected cars and navigation apps is available on more roadways than ever before.
 HOME LOCATION	Mobile device data from across the country for all trip purposes. Home locations aggregated by block group. All geographies accurately reflected in the in/ out of market impressions.	Mobile data enables Geopath to understand the home locations of the audience passing by all OOH media. Comprehensive coverage across the US allows Geopath to quantify out-of-market audiences, such as business travelers or tourists.

A one-page infographic of the above table, as well as additional information on the new methodology, how it has evolved, and its impact to impressions, can be found in the **geekOUT Library** on the Geopath website. We recommend that everyone download the above table for easy reference as it will be helpful in answering questions that may come up from clients in regard to the changes.

For a deeper discussion of the above table, a webinar covering ***How Impressions are Evolving*** is available on the Geopath YouTube Channel.

An Illustrative Use Case

The following use case is provided to help illustrate the new capabilities available through Geopath Insights, and provide context for the standards and protocols outlined in the remaining document. The example looks at how audience and location can impact the inventory selected for an overall plan, and how this has changed.



While multiple DMAs are included in the request, this example will focus on the Atlanta DMA. The same process would hold for the other markets.

Use Case Details

MGA

HEROES & LEGENDS

Media Target: 50 Weekly TRPs in each market

Markets:
- Atlanta
- Chicago
- Los Angeles
- New York
- San Francisco

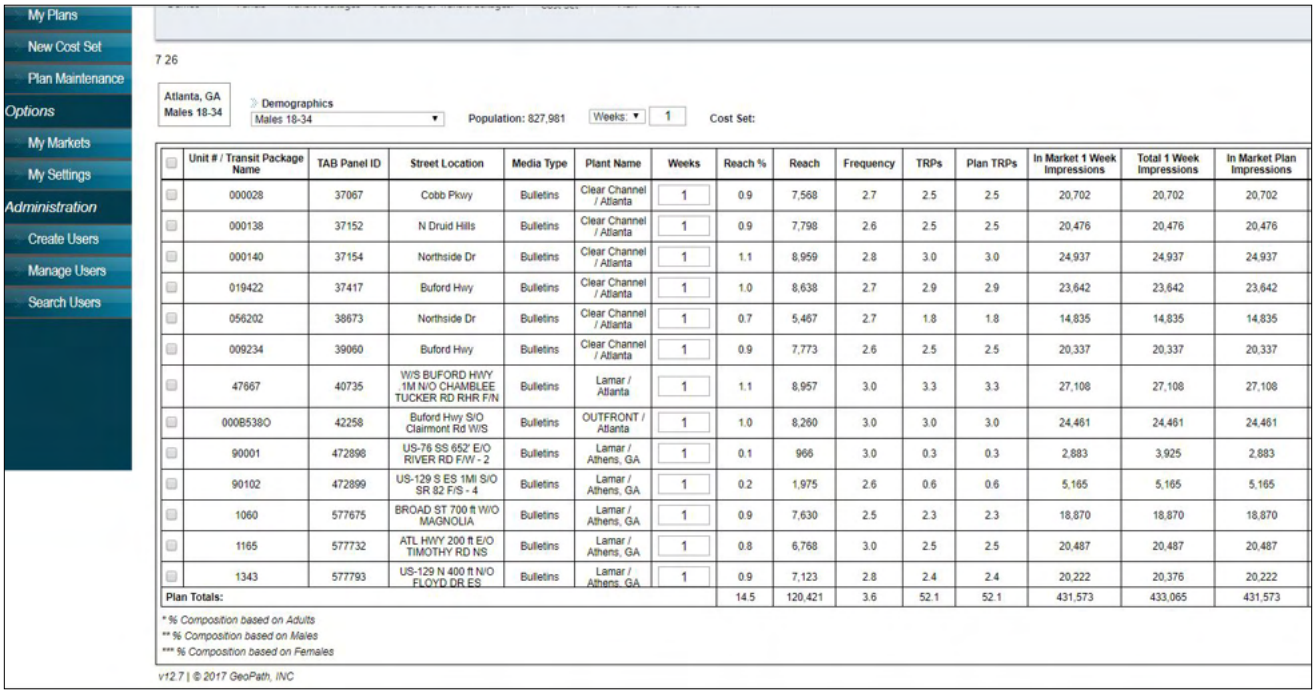
Target:
- *Primary Target: Gamers that download/use mobile games* [NEW]
- Secondary Target: Early tech adopters/influencers [NEW]
- Alternate: Males, age 18-34 in large metropolitan areas

Formats: Large format bulletins (non-digital)

geopath

Note: *Use Case for Illustration purposes only.*

Previously, an agency or operator responding to a proposal like the one outlined below would only have been able to respond to the demographic target, leaving the primary audience request unanswered.

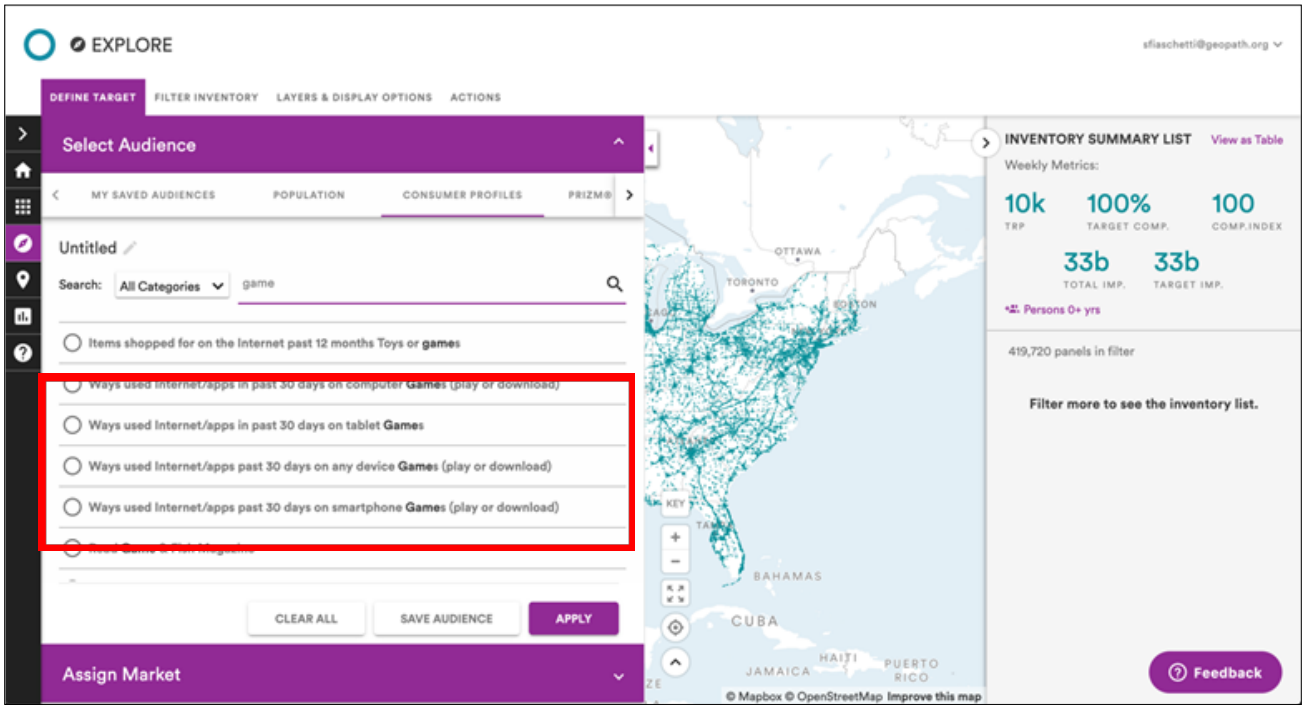


Unit # / Transit Package Name	TAB Panel ID	Street Location	Media Type	Plant Name	Weeks	Reach %	Reach	Frequency	TRPs	Plan TRPs	In Market 1 Week Impressions	Total 1 Week Impressions	In Market Plan Impressions
000028	37067	Cobb Pkwy	Bulletins	Clear Channel / Atlanta	1	0.9	7,568	2.7	2.5	2.5	20,702	20,702	20,702
000138	37152	N Druid Hills	Bulletins	Clear Channel / Atlanta	1	0.9	7,798	2.6	2.5	2.5	20,476	20,476	20,476
000140	37154	Northside Dr	Bulletins	Clear Channel / Atlanta	1	1.1	8,959	2.8	3.0	3.0	24,937	24,937	24,937
019422	37417	Buford Hwy	Bulletins	Clear Channel / Atlanta	1	1.0	8,638	2.7	2.9	2.9	23,642	23,642	23,642
056202	38673	Northside Dr	Bulletins	Clear Channel / Atlanta	1	0.7	5,467	2.7	1.8	1.8	14,835	14,835	14,835
009234	39060	Buford Hwy	Bulletins	Clear Channel / Atlanta	1	0.9	7,773	2.6	2.5	2.5	20,337	20,337	20,337
47667	40735	W/S BUFORD HWY 1M N/O CHAMBLEE TUCKER RD RHR F/N	Bulletins	Lamar / Atlanta	1	1.1	8,957	3.0	3.3	3.3	27,108	27,108	27,108
000B538O	42258	Buford Hwy S/O Clairmont Rd W/S	Bulletins	OUTFRONT / Atlanta	1	1.0	8,260	3.0	3.0	3.0	24,461	24,461	24,461
90001	472898	US-76 SS 652' E/O RIVER RD F/W - 2	Bulletins	Lamar / Athens, GA	1	0.1	966	3.0	0.3	0.3	2,883	3,925	2,883
90102	472899	US-129 S ES 1MI S/O SR 82 F/S - 4	Bulletins	Lamar / Athens, GA	1	0.2	1,975	2.6	0.6	0.6	5,165	5,165	5,165
1060	577675	BROAD ST 700 ft W/O MAGNOLIA	Bulletins	Lamar / Athens, GA	1	0.9	7,630	2.5	2.3	2.3	18,870	18,870	18,870
1165	577732	ATL HWY 200 ft E/O TIMOTHY RD NS	Bulletins	Lamar / Athens, GA	1	0.8	6,768	3.0	2.5	2.5	20,487	20,487	20,487
1343	577793	US-129 N 400 ft N/O FLOYD DR ES	Bulletins	Lamar / Athens, GA	1	0.9	7,123	2.8	2.4	2.4	20,222	20,376	20,222
Plan Totals:						14.5	120,421	3.6	52.1	52.1	431,573	433,065	431,573

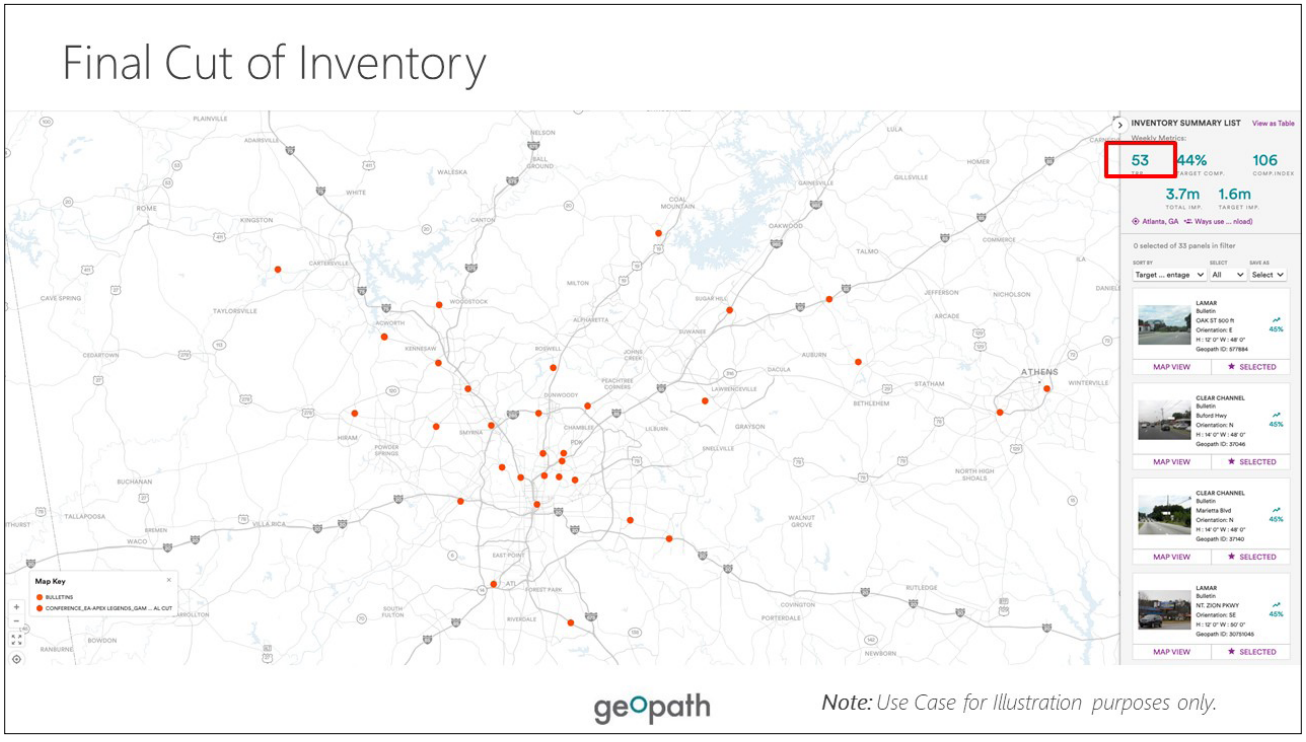
* % Composition based on Adults
** % Composition based on Males
*** % Composition based on Females

v12.7 | © 2017 GeoPath, INC

Now, the industry no longer needs to only focus on demographic-based targets. There are more than 8,000 audience targets available in the new Geopath Insights dataset. Given all the available audiences, it is critical that the desired audiences are clearly communicated among all parties involved (agency, operator, and/or advertiser).

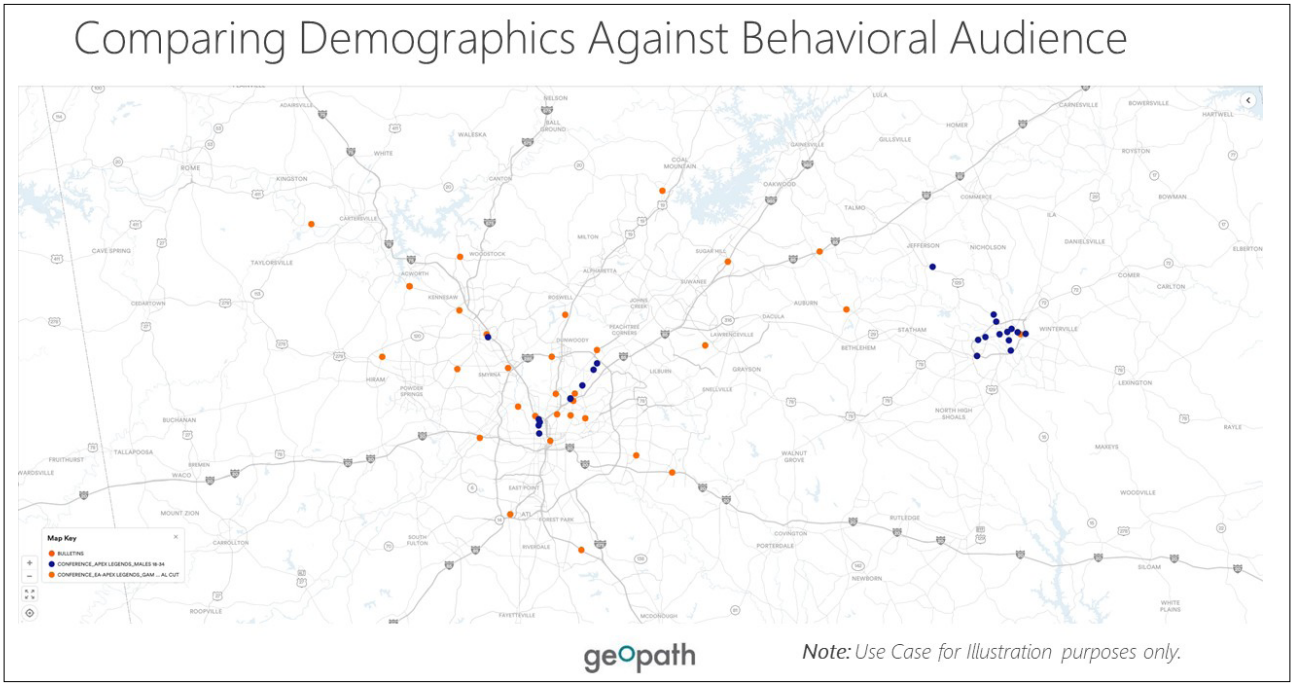


After narrowing down the inventory based on geographic distribution, as well as efficiency at reaching the desired target, the following plan was identified. The plan includes 33 units across multiple operators and slightly exceeds the 50 TRP minimum requested.



So, what does this mean?

As you can see on the maps, the plans are very different geographically (blue dots = traditional demo-based audience / orange dots = behavior target plan). The expanded capabilities available through the new Geopath Insights allow us to fundamentally change the conversation from one based on demographics, to one that includes audience behaviors. Ultimately allowing us to more efficiently meet advertisers' needs.



However, it also means that as an industry we need to be aligned on how we communicate the information needed and establish a set of protocols for how this information will be used. The following document provides a set of guidelines to use as a starting point.

For a deeper discussion of the above use case, as well as other use case examples, you can go to the OOH Office Hours Section of the Geopath website, and/or the Geopath YouTube channel.

Article

Patterns

The risk of re-identification remains high even in country-scale location datasets

Highlights

- Re-identification risk is statistically modeled and shown to decrease slowly with dataset size
- With increasing dataset size, the unicity decrease is lower-bounded and convex
- Previous estimates of unicity unrealistically underestimated the risks
- Individuals are likely re-identifiable in country-size location data and other high-dimensional datasets

Authors

Ali Farzanehfar, Florimond Houssiau, Yves-Alexandre de Montjoye

Correspondence

demontjoye@imperial.ac.uk

In Brief

Researchers have claimed that individuals could not be re-identified in large-scale location datasets, making them safe. We here empirically measure and mathematically model the relationship between the size of a dataset and the risk of re-identification. Our results show that the risk decreases slowly with dataset size, making even large country-scale datasets very likely to be re-identifiable.

Farzanehfar et al., 2021, Patterns *2*, 100204
March 12, 2021
https://doi.org/10.1016/j.patter.2021.100204

CellPress

Patterns

CellPress
OPEN ACCESS

Article

The risk of re-identification remains high even in country-scale location datasets

Ali Farzanehfar,[1] Florimond Houssiau,[1] and Yves-Alexandre de Montjoye[1,2,*]
[1]Department of Computing, Imperial College London, London SW7 2AZ, UK
[2]Lead contact
*Correspondence: demontjoye@imperial.ac.uk
https://doi.org/10.1016/j.patter.2021.100204

THE BIGGER PICTURE Data about us are being collected in many different ways, when we use our bank cards, use our phones, browse the web, or even drive our cars. These datasets contain detailed information about our lives. For each person, a dataset might contain thousands to tens of thousands of records. Previous research has shown that knowing just a few points about a target can single out the vast majority of people in location datasets. However, some had argued the risk of re-identification becomes negligible if we look at large-scale datasets containing tens of millions of people.

Here, we empirically measure, mathematically model, and provide a lower bound on the relationship between the size of a dataset and the risk of re-identification. Our results all show that re-identification risk decreases very slowly with increasing dataset size. Contrary to previous claims, people are thus very likely to be re-identifiable even in country-scale datasets.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Although anonymous data are not considered personal data, recent research has shown how individuals can often be re-identified. Scholars have argued that previous findings apply only to small-scale datasets and that privacy is preserved in large-scale datasets. Using 3 months of location data, we (1) show the risk of re-identification to decrease slowly with dataset size, (2) approximate this decrease with a simple model taking into account three population-wide marginal distributions, and (3) prove that unicity is convex and obtain a linear lower bound. Our estimates show that 93% of people would be uniquely identified in a dataset of 60M people using four points of auxiliary information, with a lower bound at 22%. This lower bound increases to 87% when five points are available. Taken together, our results show how the privacy of individuals is very unlikely to be preserved even in country-scale location datasets.

INTRODUCTION

Throughout our day, we interact with many digital services when using our phone, paying with our credit card, or using public transport with a smart card. This results in our location data being collected broadly, sometimes on the scale of countries. For instance, Vodafone UK collects location trajectories of 20M citizens[1]—a third of the population—while up to 5 million people use London's subway daily.[2]

Location data have been used extensively in research. In urban planning, mobility data can be used to monitor urban activity[3] and help design better cities.[4] In epidemiology, it has been used to monitor and mitigate the spread of infectious diseases such as Ebola and COVID-19.[5–10] In computational social science, it has allowed us to gain unprecedented insights into the spatial distribution of poverty,[11] and even to study the impact of mass employment layoffs on society.[12] Further, the use of location data has withstood scrutiny into potential biases in their collection mechanisms.[13]

Despite this, the large-scale collection and use of location data has raised serious privacy concerns. It consists of fine-grained records of where we are and how we move around, and was considered sensitive by 82% of Americans in a recent survey.[14] Location data can furthermore be used to predict individuals' income,[11,15] their home and work locations,[16–21] when they sleep and wake up,[22–26] their gender and age,[27] their personality,[28] who their friends are,[29,30] and where they tend to socialize.[31]

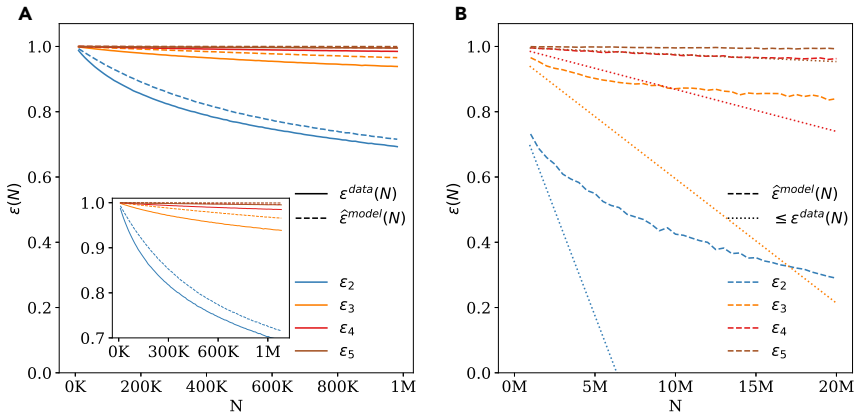


Figure 1. The relationship between unicity and dataset size

(A) Empirical (solid lines) and estimated (dashed lines) unicity decreases slowly with the size of the dataset. Inset: close up of the region $\epsilon \geq 0.7$.

(B) The estimated unicity remains high even in large datasets. This is confirmed by the lower bound results (dotted lines). Taken together, these results strongly suggest that unicity remains high even in country-scale datasets.

Unicity has been proposed as a measure for the risk of re-identification in anonymous datasets and was used to show how four points of auxiliary information (places and times where someone was) are enough to uniquely identify 95% of people in a large-scale location dataset.[32] These four points of auxiliary information could be in the form of geo-tagged "tweets," online check-ins, or information obtained by more traditional means, such as observing someone making a call. Unicity (ϵ_p) is defined as the fraction of trajectories that are unique based on knowledge of p randomly chosen points in a given trajectory. Unicity has since been used to quantify re-identification risk across a number of domains, including the mobility of vehicles,[33] apps downloaded by smartphones over time,[34,35] smart cards used in public transport,[24] credit card transaction histories,[36] and location data from mobile phones in a number of countries.[32,37,38] A range of studies have furthermore exploited the unicity of datasets to re-identify people. Narayanan and Shmatikov famously showed that close to 90% of people could be re-identified in the Netflix dataset,[39] while Riederer and colleagues used the unicity of traces to match the same individual across multiple datasets.[40]

Researchers and industry practitioners have, however, argued that these high unicity numbers are an artifact of the small size of the datasets considered, and are overestimating the risk of re-identification.[41–43] For instance, Riederer et al.[40] relied on a location dataset of 1.7k people, while other case studies report unicity on dataset sizes ranging from several thousands (respectively 12k and 55k)[33,34] to over 1 million people (1.5M).[32] Examining a published study,[36] El Emam et al. estimated that the unicity of a dataset of ≈20M trajectories will be as low as 1% given four points of auxiliary information, the conclusion being that privacy was preserved in such large datasets.[42]

We here (1) study 3 months of location data and show empirically that unicity decreases slowly with the size of the dataset, (2) approximate this decrease with a simple statistical model taking into account three population-wide marginal distributions along with the underlying geography, and (3) prove that the decrease in unicity is a convex function of the dataset size and obtain a linear lower bound on unicity. We finally perform a sensitivity analysis suggesting that the decrease in unicity is agnostic to broad perturbations in the input distributions. These results disprove previous claims, instead showing that unicity is likely to remain high even in country-scale datasets.

RESULTS

Our experiments are performed on a dataset of call detail records containing the location of 1M individuals over 3 months. Each record contains a unique user ID, an hourly time stamp, and an antenna ID, which relates to a location (see Supplemental information for more details). We formally model this dataset as a sequence, $D = (D_1, \ldots, D_N)$, populated with user time/location traces of the type $D_i = (X_i, C_i)$. X_i and C_i are lists of positions (antennas) and times (hours) representing the spatial and temporal components of a user's location trace.

Using this dataset, we empirically study the decrease in unicity with the dataset size by randomly sampling individuals from our original dataset and measuring the unicity of the sample as we increase its size (see Experimental procedures for details). We use the formal definition of unicity and the estimation algorithm S2 from de Montjoye et al.[36] In line with previous work, we use the subscript p in $\epsilon_p(N)$ to indicate the number of points of auxiliary information used in the computation of unicity.

Figure 1A shows that unicity empirically decreases slowly with the size of the dataset. With three points of auxiliary information, unicity (solid orange line) goes down from $\epsilon_3(100\text{K}) = 0.98$ in a dataset of 100,000 people to $\epsilon_3(1\text{M}) = 0.93$ in a dataset of a million people. With two points (solid blue line) this decreases slightly faster, reaching $\epsilon_2(1\text{M}) = 0.69$, while unicity with four points or more (solid red and brown lines) decreases very slowly with $\epsilon_4(1\text{M}) = 0.98$. These results show that, while the size of the dataset has an impact on unicity, the decrease in unicity is slow.

To further study how unicity decreases with dataset size and whether it decreases sufficiently in population-scale datasets, we propose a simple statistical model taking into account three population-wide marginal distributions—circadian (P_C), frequency (P_F), and activity (P_A)—along with the network of mobile phone antennas in a country. Using solely these quantities, the model is able to replicate the observed decrease in unicity with dataset size.

Figure 2 displays the information extracted from the dataset, three distributions, and the antenna network. (P_C) characterizes the circadian cycle, the overall likelihood of a record to occur at a given time in a week. The existence of circadian cycles is well documented in the computational social science literature,[22,23,25,26] and we use their empirical form in the model. The frequency distribution, (P_F), is the relative overall likelihood of a location to be visited. This distribution too has been studied before and has been widely shown to be well approximated by a power-law distribution,[44–48] as is also the case here (Figure 2B, $R^2 = 0.99$). The activity distribution, (P_A), captures the number

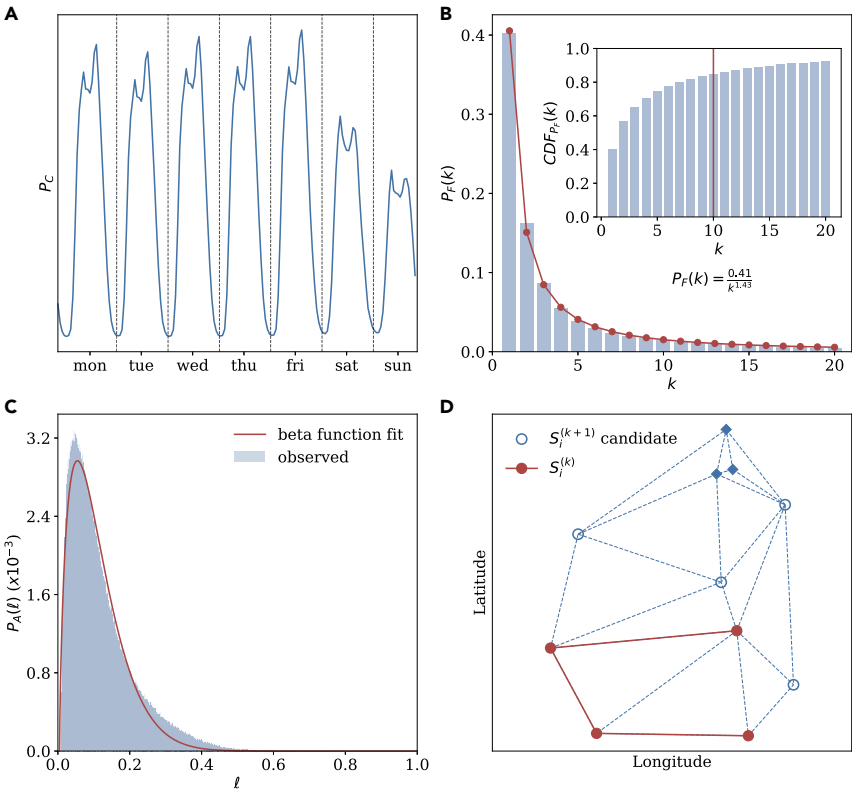


Figure 2. Inputs to the unicity model

(A) The circadian distribution, P_C.

(B) The frequency distribution, P_F, along with a power law fit (solid line, $R^2 = 0.99$). The inset displays the cumulative distribution with 85% of activity captured by the top 10 locations.

(C) The activity distribution, P_A, indicating the distribution of the number of records per trajectory along with a β distribution fit (solid line, $R^2 = 0.98$).

(D) Illustration of the sub-graph sampling method used to generate an antenna set S_i where $S_i^{(k)} \in S_i$. The underlying antenna network is represented by dotted lines. The filled nodes (circles) correspond to locations already selected, while the hollow nodes are potential locations that could be selected next ($S_i^{(k+1)}$ candidates) (see Supplemental information for detailed algorithm). Remaining locations are represented by filled diamonds.

of records $\ell_{(i)} = |D_i|$ that appear in each user trace. We approximate it here with a β distribution ($\alpha = 1.72$, $\beta = 14.7$, $R^2 = 0.98$). Finally, S_i is the set of locations visited by person i. It is a sub-graph sampled from the Delaunay tessellation of the antenna coordinates ($\mathcal{L}$) in the dataset (see Supplemental information for the detailed algorithm).

In short, for each user, our model samples a list of 10 connected antennas $(S_1, \ldots, S_{10})$ on the network and an activity (number of records in the user's trace), $A \sim P_A$. Each record's timestamp C and position X is then sampled according to the circadian distribution $C \sim P_C$ and $X = S_K$, $K \sim P_F$. This model is formally defined in the Experimental procedures.

Figure 1A shows that our simple statistical model closely follows the empirical measure of unicity from 1 to 1M people (dashed and solid lines). Using the model, we then study how unicity is likely to evolve as the size of the dataset increases to 20M people (Figure 1B). For $N = 20$M, our model estimates unicity with three points to be close to $\widehat{\epsilon_3}(20\text{M}) = 0.93$, while knowing one more point would increase this to the region of $\widehat{\epsilon_4}(20\text{M}) = 0.99$. This is a stark difference with the linear extrapolation made by El Emam,[42] who reports a unicity of 0.01 with four points (we replicate El Emam's method in the discussion and display our results for up to 60M people in the Supplemental information).

The model provides good evidence that unicity is likely to remain high even in datasets as large as 20M people. For further evidence, we prove that the decrease in unicity with increasing dataset size follows a convex form, and use this result to provide a lower bound on unicity in large datasets. We show in the Supplemental information that the unicity of a dataset of size N can be expressed as a sum of convex functions of N, and is thus convex.

This builds on two assumptions: (1) there exists an underlying trajectory distribution T_X from which all trajectories $D_i \in D$ are sampled and (2) all trajectories are independent of one another, $D_i \perp\!\!\!\perp D_j$. The first assumption states that an underlying distribution for trajectories exists. Such a distribution would also capture correlations between individuals on a large scale (e.g., commuting patterns, cities, weekends). The second assumption presumes that the correlation between specific individuals is negligible when estimating unicity of large datasets.

A direct consequence of unicity being a strictly decreasing convex function is that it will be lower bounded by its linear tangent (treating unicity as a function of a real-valued N):

$$\epsilon(D(N)) \geq \epsilon(D(N')) + (N - N') \cdot \left.\frac{d\epsilon}{dN}\right|_{N=N'}. \qquad \text{(Equation 1)}$$

Re-arranged and expressed for discrete values, this gives a lower bound for unicity:

$$\epsilon(D(N')) - \epsilon(D(N)) \leq (N - N') \cdot (\epsilon(D(N' - 1)) - \epsilon(D(N'))). \qquad \text{(Equation 2)}$$

Using the tangent to the empirical unicity curves estimated by discrete difference over the range of $N \in [0.9\text{M}, 1\text{M}]$, we obtain a lower bound of 0.73 for $\epsilon_4(20\text{M})$ and 0.9 for $\epsilon_5(20\text{M})$ (Figure 1B, dotted lines).

Our results show that unicity decreases slowly with the size of the dataset and that it, very likely, remains high even in population-scale datasets. This refutes previous claims that privacy is preserved in population-scale datasets, instead showing the risk of re-identification to be high. Modern location datasets have a great potential to improve our society, for example, by training AI algorithms, but robust privacy engineering solutions are needed to use them safely.

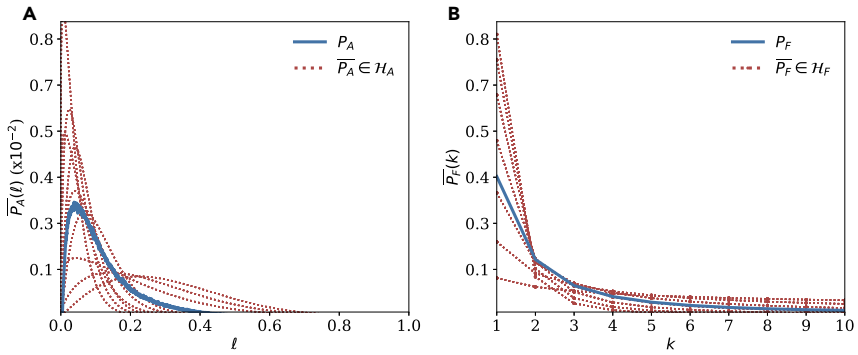


Figure 3. Range of distributions studied for the sensitivity analysis

The ranges of perturbed activity $\overline{P_A}$ (A) and frequency $\overline{P_F}$ (B) distributions are displayed (dotted lines) along with their empirical forms (solid lines).

DISCUSSION

Taken together, these results show that the scale of a dataset does not prevent re-identification. Human mobility, much like a physical fingerprint, is highly unique and can be used to find a person across mobility datasets.

Legally, the European Union (EU) General Data Protection Regulation sets a high threshold for what constitutes anonymous data, namely that the individual should not be identifiable taking into account both the "available technology at the time of the processing" but also future "technological developments" (Recital 26). The Article 29 Working Party, the predecessor to the European Data Protection Board, in its guidance sets out three criteria to assess whether a dataset is anonymous, singling out, linkability, and inference[49] with the former two being directly applicable here. As an example, the Centre for Humanitarian Data of the United Nations (UN OCHA) adopted 5% as a threshold for what constitutes an acceptable re-identification risk.[50] Even our lower bound of 22% far exceeds this liberal threshold.

Finally, here we study the unicity of location datasets with a spatial resolution of $\approx$1 km^2 and a temporal resolution of an hour. Fine-grained GPS data are likely to lead to even higher values of unicity, and previous research has shown that, in general, de-identification methods do not meaningfully reduce the risk of re-identification. For instance, research[32,34] has shown that reducing the spatial and temporal resolution of the data further only slowly decreases the risk, while another study[51] concluded that location data "show poor anonymizability [as measured by k-anonymity], i.e., require important spatial and temporal generalization in order to slightly improve user privacy".

Ensuring that these data can be accessed and used broadly is of paramount importance, but this should not come at the expense of people's privacy. A range of privacy engineering techniques allowing data to be used while giving individuals strong privacy guarantees have been developed and are starting to be used.[52–54] As standards for anonymization are being redefined, in the EU and around the world, it is essential for them to emphasize the strong limits of de-identification, possibly banning the uncontrolled release of individual-level de-identified data, and to give guidance on the use of modern privacy-engineering solutions.

In the next three sections we discuss the underlying assumptions of the unicity model and some considerations regarding the sensitivity of our results and, finally, include a discussion on previous estimates of unicity.

Assumptions underpinning the simple unicity model

We here evaluate the four assumptions underpinning the simple unicity model we present.

First, the model treats each of the four inputs in Figure 2 as independent of one another. Considering them, or some of them, jointly might further improve the model. This would, however, also increase its complexity and, therefore, its sensitivity to small changes in the data. Although further exploration would be interesting, we consider that the simple model approximates the decrease in unicity with increasing dataset size well enough to support our conclusion that unicity is unlikely to be low even in population-scale datasets.

Second, our model uses input distributions extracted from a dataset of 1M people to study the unicity of datasets with up to 60M people (see Supplemental information). This assumes that these distributions estimated from a smaller sample are representative of the larger sample (i.e., the estimation of the distributions has converged). We show that this is a reasonable assumption by instantiating our model $\mathcal{M}$ with distributions extracted from samples of sizes significantly smaller than 1M, and showing that the unicity results remain largely unchanged (Figure S5 in Supplemental information). We also perform a sensitivity analysis to evaluate the impact of broad variations on these input distribution on our results (see next subsection).

Third, the model assumes each trajectory to contain at most one unique location. This allows for the mean frequency distribution (P_F) to be used in the modeling process (Figure 2B). As seen in the inset of Figure 2B, more than 85% of the activity in the average trajectory is captured by the top 10 locations visited. Furthermore, we find that P_F changes only slightly when the number of unique locations is altered, and that our conclusions are not influenced by this choice.

Finally, our model assumes that the set of locations appearing in each trajectory can be described by a connected planar subgraph of the underlying antenna network. We believe this to be a reasonable assumption, as previous work suggests that subgraphs spanned by each trajectory in human mobility are highly localized, with the distribution $P(r_g)$ of the radius of gyration—a metric for how far people tend to travel on average—following a power law with increasing radius.[55]

Sensitivity analysis

Our simple statistical model for unicity takes as input three distributions. However, these distributions may vary depending on specifics of the dataset, such as the country where it was collected or the sources of location information. Here we perform a sensitivity analysis to ensure the robustness of our model to even broad changes to the distributions.

We first perturb the P_A and P_F distributions (Figure 3) around their empirical forms using a scaled earth mover's distance as

Table 1. Summary of unicity results at $N = 20M$ as per the sensitivity analysis

	ϵ_2	ϵ_3	ϵ_4	ϵ_5
Mean	0.307	0.735	0.876	0.935
Standard deviation	0.175	0.216	0.159	0.113
Minimum	0.071	0.260	0.431	0.544
Maximum	0.704	0.997	1	1

the guiding metric (see Supplemental information for details). The P_C distribution, on the other hand, has been shown to be very stable across datasets[22–26] and we thus keep it constant throughout our analysis.

These distributions are combined to produce 63 different instantiations of the unicity model (Figure S2). Table 1 summarizes the unicity values for models using the broad range of distributions in Figure 3, at a dataset size of 20M trajectories (see Supplemental information for 60M results). Note that the lowest unicity values across all instantiations of the model are still high, with $Min(\epsilon_4(20M)) = 43.1\%$ and would still be considered as putting people's privacy at risk.

Further, we study how certain aspects of human mobility contribute to unicity. Starting from empirical user location traces $D_i = (X_i, C_i)$, first, we find that removing the association between times (C_i) and locations (X_i), by shuffling the vectors and recombining them, only slightly affects unicity values (Figure S4A). Specifically, consider a dataset D' composed of trajectories $D\prime_i = (X\prime_i, C\prime_i)$ such that:

$$X\prime_i = \sigma_i(X_i),$$

$$C\prime_i = \pi_i(C_i),$$

where σ_i and π_i refer to random permutations of the spatial and temporal components of D_i. This only marginally affects unicity, showing that unicity does not depend on the specific places being visited at specific times, as long as those times and places appear in the trace with their respective frequencies independently.

Second, we replace the set of locations in each trajectory with uniformly picked locations. Instead of using the sub-graph sampling method displayed in Figure 2D, we populate each S_i with antennas picked from the entire set of locations $\mathcal{L}$ uniformly at random. We find that this leads to unicity being overestimated (Figure S4D).

Third, replacing P_C or P_F with uniform distributions (Figures S4B and S4C) or attempts to model unicity using a simple combinatorial model (Figure S3) also cause the model to overestimate unicity. These demonstrate the importance of all three distributions and the underlying geography to correctly capture the unicity of mobility datasets.

This analysis, combined with the relative simplicity and generality of the unicity model, strongly suggest that our results would generalize to any location dataset. Likewise, the strong underlying combinatorial effect that underpins unicity combined with previous research[34–36] suggests that unicity will similarly decrease slowly in other types of high-dimensional data.

El Emam's method

El Emam[42] proposed a method (hereafter the EE method) to estimate the uniqueness of a population-size (N) dataset given the unicity $\epsilon(m)$ of a smaller sample dataset of size m. Using this method, he estimates that the uniqueness for a population of size $N = 22 \cdot 10^6$ is about 1%, given a uniqueness of 90% of a sample of size $m = 22 \cdot 10^6$ of the same dataset. This estimate forms the basis for his claim that uniqueness is low in large-scale datasets.

We here show that the EE method (1) is unrealistic and (2) provably gives the lowest possible estimate for the risk in the larger dataset, and that (3) by using our dataset, we observe that the real empirical unicity is significantly higher than the upper bound given by the EE method.

First, the method is unrealistic, as it effectively generates a dataset D of size N where a fraction α of records are unique, while all the other records are identical to exactly one and only one other record. The parameter α is selected such that the expected estimated uniqueness on a sample of size m, which we denote by $\nu_D(m)$, is equal to the empirical unicity. This assumes that users in the real mobility dataset are either unique or exact duplicates of another user.

Second, we prove in the Supplemental information that the risk estimated by the EE method will be lower or equal to the risk of *any* other dataset of size N, as this estimate is an *affine* function of m. In other terms, this method will *always* return the absolute lowest possible estimate of the risk.

Third, we apply the EE method to our dataset and show that its estimate of the risk is significantly lower than the real empirical value, leading to the risk of re-identification being strongly underestimated. For a dataset of 200,000 people, we empirically observe an $\epsilon_2(200K) = 0.86$. Using this number, El Emam's method would estimate the risk of a larger 1M person dataset to be $\epsilon_2(1M) = 0.3$, while the correct empirical value is ≈ 0.7.

Taken together, our results cast serious doubt on the validity of the EE method to carry out risk assessments.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Yves-Alexandre de Montjoye (demontjoye@imperial.ac.uk).

Materials availability

There are no physical materials associated with this study.

Data and code availability

Due to reasons of confidentiality and user privacy, we cannot share the raw data. However, we can make available all the input distributions and raw empirical results upon request for purposes of reproducibility.

The code used for all experiments is available at: github.com/computationalprivacy/scaling-unicity.

The unicity model in detail

We propose a simple statistical model $\mathcal{M}$ taking into account three population-wide distributions: activity (P_A), circadian (P_C), and frequency (P_F). This model samples location traces for each user independent of other users to estimate unicity of a dataset of size N. These location traces are then grouped together to compute unicity.

Formally, the model $\mathcal{M}$ can be written as:

$$\mathcal{M}(P_A, P_C, P_F, \mathcal{L}, N) = D = (D_1, \ldots, D_N). \quad \text{(Equation 3)}$$

Each $D_i \in D$ is a location trace for a unique user, represented as a list of L_i records $(X_i^{(j)}, C_i^{(j)})_{j=1}^{L_i}$. The length L_i of trace D_i is sampled from the empirical activity distribution P_A:

$$\mathbb{P}[L_i = \ell] = P_A(\ell). \quad \text{(Equation 4)}$$

The timestamps of each record in a trace, $(C_i^{(j)})_{j=1}^{L_i}$, are sampled independent of the empirical circadian distribution P_C:

$$\mathbb{P}\left[C_i^{(j)} = c\right] = P_C(c) \quad \forall j \in \{1, \ldots, L_i\}. \quad \text{(Equation 5)}$$

For the spatial component, for each user, a connected sub-graph S_i of size 10 is first sampled from the Delaunay tessellation of the antenna coordinates $\mathcal{L}$. This sub-graph is then randomly ordered as a list, which we denote by $S_i = (S_i^{(k)})_{k=1}^{10}$ with a slight abuse of notations. Finally, the locations of the records $X_i^{(j)} \in X_i$ are sampled independent of S_i according to the empirical frequency distribution P_F:

$$\mathbb{P}\left[X_i^{(j)} = S_i^{(k)}\right] = P_F(k) \quad \forall j \in \{1, \ldots, L_i\}. \quad \text{(Equation 6)}$$

Note that when the size of the dataset N sampled by our model $\mathcal{M}$ increases, this corresponds to sampling more individuals from the same underlying geography. This is what we mean throughout this work when we increase the size of the dataset, e.g., in unicity curves (Figure 1): we consider the dataset to be a growing sample from the same underlying population.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100204.

ACKNOWLEDGMENTS

The authors would like to thank Shubham Jain for their comments on the codebase, and Ana-Maria Cretu, Andrea Gadotti, Shubham Jain, Thibaut Lienart, Axel Oehmichen, Luc Rocher, and Arnaud Tournier for their invaluable comments on the manuscript. We acknowledge support from the Agence Française de Développement as part of its financial assistance to the OPAL project.

AUTHOR CONTRIBUTIONS

A.F. designed and performed the experiments, built the models, helped with the mathematical results, and drafted the manuscript. F.H. derived the mathematical results, advised on model construction, and revised the manuscript. Y-A.d.M. designed the experiments and revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

Received: August 24, 2020
Revised: November 27, 2020
Accepted: January 7, 2021
Published: February 12, 2021

REFERENCES

1. Vodafone. (2018). Vodafone UK's company history and achievements. https://www.vodafone.co.uk/about-us/company-history/.
2. Lomas, N. (2017). How "anonymous" wifi data can still be a privacy risk (TechCrunch). http://tcrn.ch/2ywXGdy.
3. Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., and Tatem, A.J. (2014). Dynamic population mapping using mobile phone data. Proc. Natl. Acad. Sci. U S A *111*, 15888–15893.
4. Ratti, C., Frenchman, D., Pulselli, R.M., and Williams, S. (2006). Mobile landscapes: using location data from cell phones for urban analysis. Environ. Plann. B Plann. Des. *33*, 727–748.
5. Wesolowski, A., Eagle, N., Tatem, A.J., Smith, D.L., Noor, A.M., Snow, R.W., and Buckee, C.O. (2012). Quantifying the impact of human mobility on malaria. Science *338*, 267–270.
6. Gomes, M.F., Pastore, Y., Piontti, A., Rossi, L., Chao, D., Longini, I., Halloran, M.E., and Vespignani, A. (2014). Assessing the International spreading risk associated with the 2014 west African Ebola outbreak. PLoS Currents *6*, ecurrents.outbreaks.cd818f63d40e24aef769d-da7df9e0da5.
7. Mari, L., Bertuzzo, E., Righetto, L., Casagrandi, R., Gatto, M., Rodriguez-Iturbe, I., and Rinaldo, A. (2012). Modelling cholera epidemics: the role of waterways, human mobility and sanitation. J. R. Soc. Interface *9*, 376–388.
8. Bajardi, P., Poletto, C., Ramasco, J.J., Tizzoni, M., Colizza, V., and Vespignani, A. (2011). Human mobility networks, travel restrictions, and the global spread of 2009 H1n1 pandemic. PLoS One *6*, e16591.
9. Merler, S., and Ajelli, M. (2009). The role of population heterogeneity and human mobility in the spread of pandemic influenza. Proc. Biol. Sci. *277*, 557–565.
10. Aktay, A., Bavadekar, S., Cossoul, G., Davis, J., Desfontaines, D., Fabrikant, A., Gabrilovich, E., Gadepalli, K., Gipson, B., Guevara, M., et al. (2020). Google COVID-19 community mobility reports: anonymization process description (version 1.0). arXiv, preprint arXiv:2004.04145.
11. Steele, J.E., Sundsøy, P.R., Pezzulo, C., Alegana, V.A., Bird, T.J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.A., Iqbal, A.M., et al. (2017). Mapping poverty using mobile phone and satellite data. J. R. Soc. Interface *14*, 20160690.
12. Toole, J.L., Lin, Y.-R., Muehlegger, E., Shoag, D., González, M.C., and Lazer, D. (2015). Tracking employment shocks using mobile phone data. J. R. Soc. Interface *12*, 20150185.
13. Wesolowski, A., Eagle, N., Noor, A.M., Snow, R.W., and Buckee, C.O. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. J. R. Soc. Interface *10*, 20120986.
14. Madden, M., Rainie, L., Zickuhr, K., Duggan, M., and Smith, A. (2014). Public Perceptions of Privacy and Security in the Post-Snowden Era, *12* (Pew Research Center).
15. Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science *350*, 1073–1076.
16. Li G., Yu L., Ng W.S., Wu W., and Goh S.T. Predicting Home and Work Locations Using Public Transport Smart Card Data by Spectral Analysis. In 2015 IEEE 18th International Conference on Intelligent Transportation Systems, pages 2788–2793, Gran Canaria, Spain, September 2015. IEEE.
17. Ashbrook D. and Starner T. Learning significant locations and predicting user movement with GPS. In Proceedings. Sixth International Symposium on Wearable Computers, pages 101–108, Seattle, WA, USA, 2002. IEEE.
18. Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. (2011). Identifying important places in people's lives from cellular network data. In Pervasive Computing, Volume 6696 of Lecture Notes in Computer Science, K. Lyons, J. Hightower, and E.M. Huang, eds. (Springer Berlin Heidelberg), pp. 133–151.
19. Mahmud, J., Nichols, J., and Drews, C. (2014). Home location identification of Twitter users. ACM Trans. Intell. Syst. Technol. *5*, 47.
20. Li R., Wang S., Deng H., Wang R, and Chen-Chuan Chang K. Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1023–1031, New York, NY, USA, 2012. ACM.
21. Cho E., Myers S.A., and Leskovec J. Friendship and Mobility: User Movement in Location-based Social Networks. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1082–1090, New York, NY, USA, 2011. ACM.
22. Monsivais, D., Ghosh, A., Bhattacharya, K., Dunbar, R.I.M., and Kaski, K. (2017). Tracking urban human activity from mobile phone calling patterns. PLoS Comput. Biol. *13*, e1005824.

23. Monsivais, D., Bhattacharya, K., Ghosh, A., Dunbar, R.I.M., and Kaski, K. (2017). Seasonal and geographical impact on human resting periods. Sci. Rep. *7*, 10717.

24. Kondor, D., Hashemian, B., de Montjoye, Y.-A., and Ratti, C. (2020). Towards matching user mobility traces in large-scale datasets. In IEEE Transactions on Big Data, *6*, p. 1, 714-726.

25. Hasan S., Zhan X., and Ukkusuri S.V. Understanding Urban Human Activity and Mobility Patterns Using Large-scale Location-based Data from Online Social Media. In Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing, UrbComp '13, pages 6:1–6:8, New York, NY, USA, 2013. ACM.

26. Ahas, R., Aasa, A., Silm, S., and Tiru, M. (2010). Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data. Transport. Res. C Emerg. Tech. *18*, 45–54.

27. Felbo, B., Sundsøy, P., Pentland, A., Lehmann, S., and de Montjoye, Y.-A. (2017). Modeling the temporal nature of human behavior for demographics prediction. In Machine learning and knowledge discovery in databases, volume 10536 of lecture notes in computer science (Springer), pp. 140–152.

28. de Montjoye, Y.-A., Quoidbach, J., Robic, F., and Pentland, A.S. (2013). Predicting personality using novel mobile phone-based metrics. In International conference on social computing, behavioral-cultural modeling, and prediction (Springer), pp. 48–55.

29. Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. Proc. Natl. Acad. Sci. U S A *104*, 7332–7336.

30. Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., De Menezes, M.A., Kaski, K., Barabási, A.-L., and Kertész, J. (2007). Analysis of a large-scale weighted network of one-to-one human communication. New J. Phys. *9*, 179.

31. Krumme, C., Llorente, A., Cebrian, M., Pentland, A.S., and Moro, E. (2013). The predictability of consumer visitation patterns. Sci. Rep. *3*, 1645.

32. de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., and Blondel, V.D. (2013). Unique in the crowd: the privacy bounds of human mobility. Sci. Rep. *3*, 1376.

33. Pellungrini, R., Pappalardo, L., Pratesi, F., and Monreale, A. (2017). A data mining approach to assess privacy risk in human mobility data. ACM Trans. Intell. Syst. Technol. *9*, 31.

34. Achara, J.P., Acs, G., and Castelluccia, C. (2015). On the Unicity of Smartphone Applications (ACM Press), pp. 27–36.

35. Sekara, V., Mones, E., and Jonsson, H. (2018). Temporal limits of privacy in human behavior. arXiv, preprint arXiv:1806.03615.

36. de Montjoye, Y.-A., Radaelli, L., Singh, V.K., and Pentland, A.P. (2015). Unique in the shopping mall: on the reidentifiability of credit card metadata. Science *347*, 536–539.

37. Xu, Y., Belyi, A., Bojic, I., and Ratti, C. (2018). Human mobility and socioeconomic status: analysis of Singapore and Boston. Comput. Environ. Urban Syst. *72*, 51–67.

38. Deußer C., Passmann S., and Strufe T. Browsing unicity: On the limits of anonymizing web tracking data. In 2020 IEEE Symposium on Security and Privacy (SP), pages 777–790. IEEE, 2020.

39. Narayanan, A., and Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets (IEEE), pp. 111–125.

40. Riederer, C., Kim, Y., Chaintreau, A., Korula, N., and Lattanzi, S. (2016). Linking users across domains with location data: Theory and validation. In Proceedings of the 25th International Conference on World Wide Web (International World Wide Web Conferences Steering Committee), pp. 707–719.

41. Snchez, D., Martnez, S., and Domingo-Ferrer, J. (2016). Comment on "Unique in the shopping mall: on the reidentifiability of credit card metadata". Science *351*, 1274.

42. El Emam, K. (2015). On Re-identification: Not Really Unique in the Shopping Mall.

43. Barth-Jones, D., El Emam, K., Bambauer, J., Cavoukian, A., and Malin, B. (2015). Assessing data intrusion threats. Science *348*, 194–195.

44. Pappalardo, L., and Simini, F. (2018). Data-driven generation of spatio-temporal routines in human mobility. Data Min. Knowl. Discov. *32*, 787–829.

45. Gonzlez, M.C., Hidalgo, C.A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. Nature *453*, 779.

46. Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S., and Baronchelli, A. (2018). Evidence for a conserved quantity in human mobility. Nat. Hum. Behav. *2*, 1.

47. Song, C., Koren, T., Wang, P., and Barabási, A.-L. (2010). Modelling the scaling properties of human mobility. Nat. Phys. *6*, 818–823.

48. Hasan, S., Schneider, C.M., Ukkusuri, S.V., and González, M.C. (April 2013). Spatiotemporal patterns of urban human mobility. J. Stat. Phys. *151*, 304–318.

49. Article 29 Data Protection Working Party (2014). Opinion 05/2014 on Anonymisation Techniques (European Commission).

50. Centre for Humanitarian Data of the United Nations Office for the Coordination of Humanitarian Affairs (2019). Guidance Note Series on Data Responsibility on Humanitarian Action. Note 1: Statistical Disclosure Control (United Nations).

51. Gramaglia, M., and Fiore, M. (2014). On the anonymizability of mobile traffic datasets. arXiv, preprint arXiv:1501.00100.

52. Oehmichen, A., Jain, S., Gadotti, A., and de Montjoye, Y.-A. (2019). Opal: high performance platform for large-scale privacy-preserving location data analytics. In 2019 IEEE International Conference on Big Data (Big Data) (IEEE), pp. 1332–1342.

53. Mir, D.J., Isaacman, S., Cáceres, R., Martonosi, M., and Wright, R.N. (2013). Dp-where: Differentially private modeling of human mobility. In 2013 IEEE international conference on big data (IEEE), pp. 580–588.

54. Francis, P., Probst Eide, S., and Munz, R. (2017). Diffix: high-utility database anonymization. In Annual Privacy Forum (Springer), pp. 141–158.

55. Gonzalez, M.C., Hidalgo, C.A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. Nature *453*, 779.