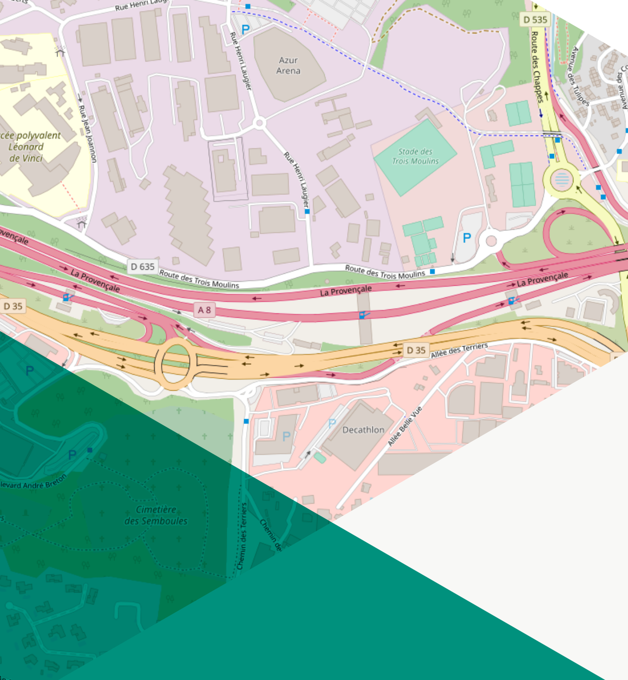


Karlsruhe Institute of Technology

P.E.Ts: Trajectory privacy

Patricia Guerra-Balboa

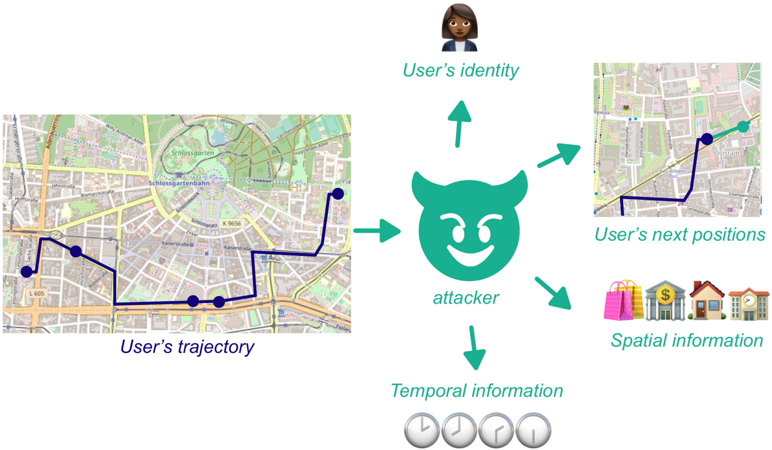
July 11, 2022

Overview

Motivation

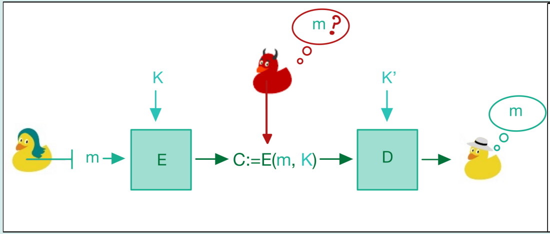
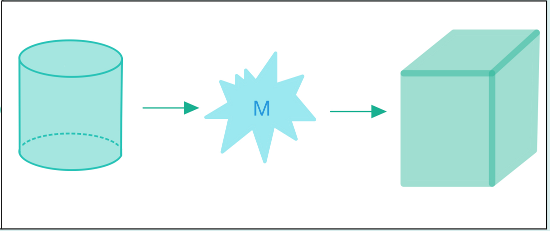
Motivation

Why do we Need to Anonymize Trajectory Data?

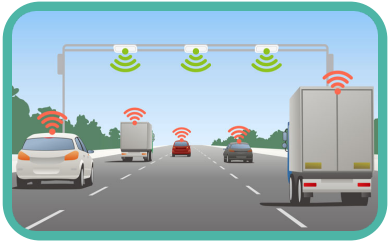


Motivation
Data Privacy

Privacy frameworks

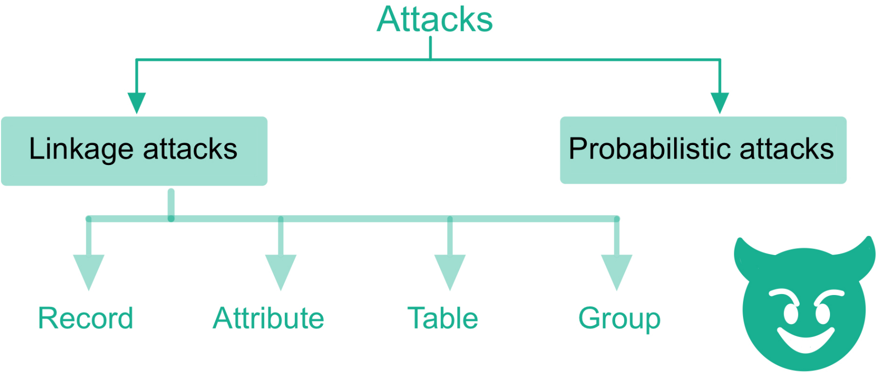
Cryptography	SDC
	
secure message delivery	information sharing with privacy

Motivation



Motivation

Why do we Need to Anonymize Trajectory Data?



Model

Model
Modeling Trajectories

Trajectories and Data Sets

Raw Trajectories

$$T = (x_1, y_1, t_1) \rightarrow \cdots \rightarrow (x_n, y_n, t_n)$$

Semantic Trajectories

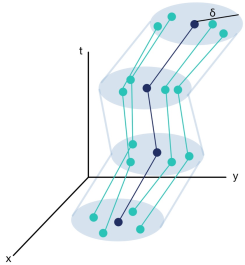
Adds semantic meaning

Privacy Notions in Trajectory Data

Privacy Notions in Trajectory Data

Syntactic Notions

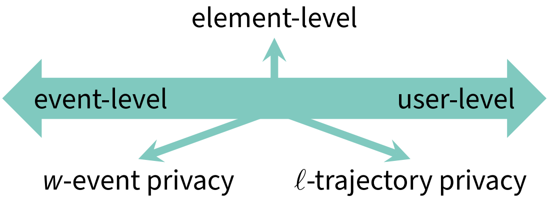
k-anonymity, *l*-diversity, *t*-closeness

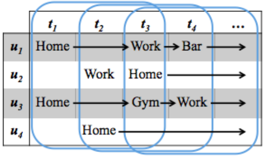


- (k, δ)-anonymity
- k^m-anonymity
- . . .

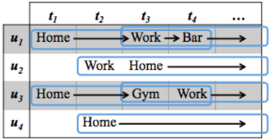
Semantic Notions

ϵ-differential privacy





	t_1	t_2	t_3	t_4	...
u_1	Home →		→ Work →	Bar →	→
u_2		Work	Home →		→
u_3	Home →		→ Gym →	Work →	→
u_4		Home →			→

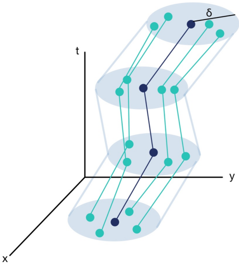


	t_1	t_2	t_3	t_4	...
u_1	Home →		→ Work →	Bar →	→
u_2		Work	Home →		→
u_3	Home →		→ Gym	Work →	→
u_4		Home →			→

Privacy Notions in Trajectory Data

Syntactic Notions

k-anonymity, *l*-diversity, *t*-closeness



- (k, δ)-anonymity
- k^m-anonymity
- . . .

Semantic Notions

ϵ-differential privacy

element-level

event-level

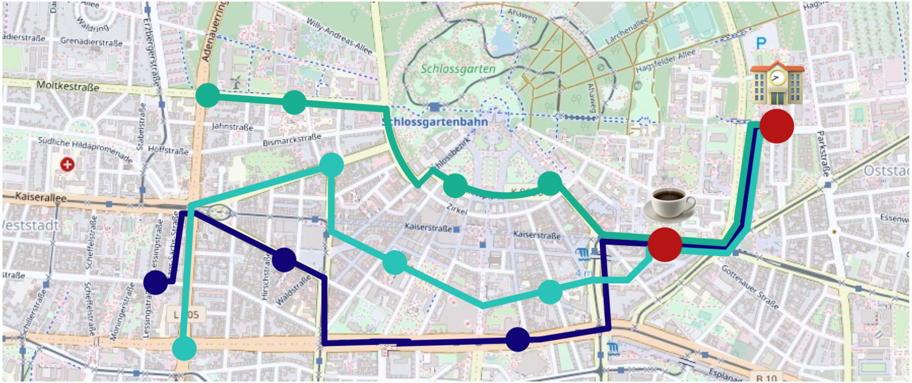
user-level

w-event privacy

ℓ-trajectory privacy

Privacy Notions in Trajectory Data

k-anonymity, *l*-diversity and *t*-closeness



Privacy Notions in Trajectory Data

k-anonymity, *l*-diversity and *t*-closeness

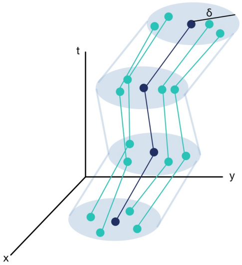
Privacy notion	RL	AL	TL	GL	PA
k-anonymity	✓				
l-diversity	✓	✓			
t-closeness	✓	✓	✓		✓

Table 1: RL = Record linkage, AL = Attribute linkage, TL = Table linkage, GL = Group linkage, PA = Probabilistic attack

Privacy Notions in Trajectory Data

Syntactic Notions

k-anonymity, *l*-diversity, *t*-closeness



- (k, δ)-anonymity
- k^m-anonymity
- ...

Privacy Notions in Trajectory Data

Syntactic Techniques

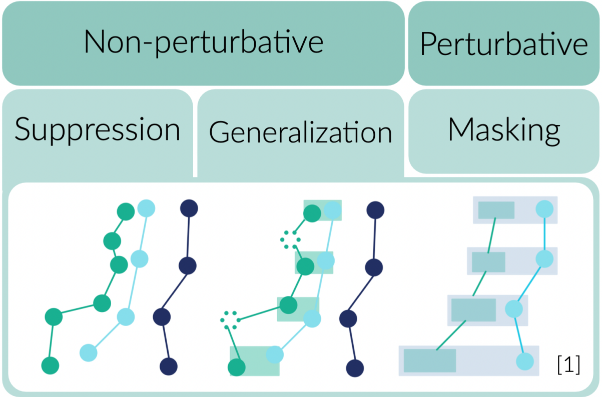


Figure 1: The main three techniques in syntactic anonymization

Privacy Notions in Trajectory Data
Syntactic Techniques Deficiencies: Suppression

- Drastic reduction of database
- Dangerous when used by itself

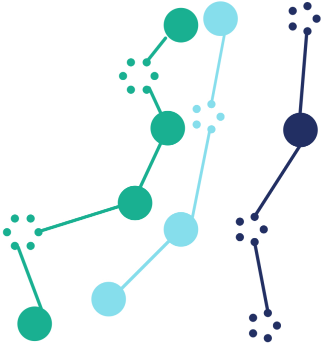


Figure 2: Suppression

Privacy Notions in Trajectory Data

Syntactic Techniques Deficiencies: Generalization

- Not generalizing all dimensions
- Inappropriate regions definition
- Background knowledge attacks
- Drastic reduction of precision
- Dangerous when used by itself

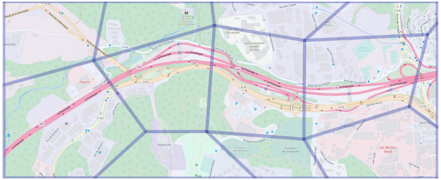


Figure 3: Generalization

Privacy Notions in Trajectory Data

Syntactic Techniques Deficiencies: Masking

- Unpredictable biases
- Impossible trajectories

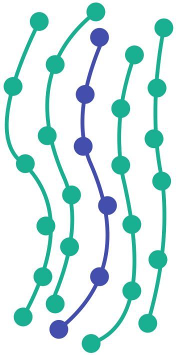


Figure 4: Dummy generation

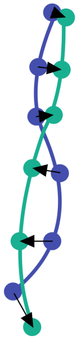


Figure 5: Noise addition

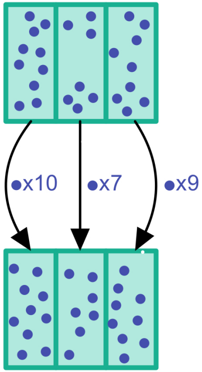


Figure 6: Condensation

Privacy Notions in Trajectory Data

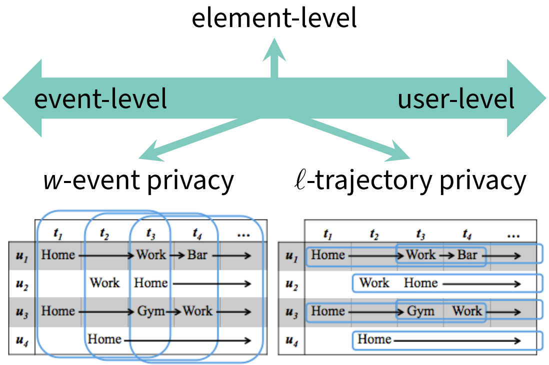
Syntactic Notions

k-anonymity, *l*-diversity, *t*-closeness

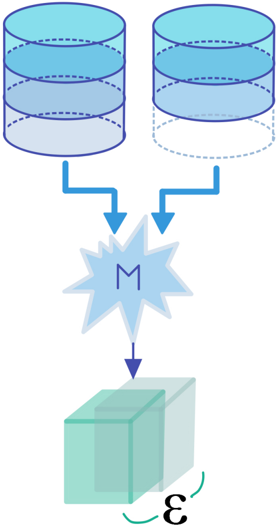
- (k, δ)-anonymity
- k^m-anonymity
- ...

Semantic Notions

ϵ-differential privacy



Privacy Notions in Trajectory Data

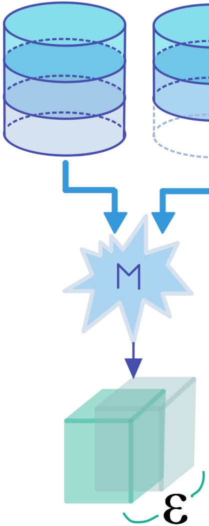


ϵ-Differential Privacy

A randomized algorithm M is said to be ϵ-*differentially private* if for all *neighboring* databases D, D' and all $\mathcal{S} \subseteq Range(M)$,

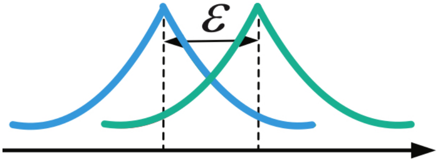
$$\mathbb{P}\{M(D) \in \mathcal{S}\} \leq e^{\epsilon}\, \mathbb{P}\{M(D') \in \mathcal{S}\}.$$

Differential Privacy

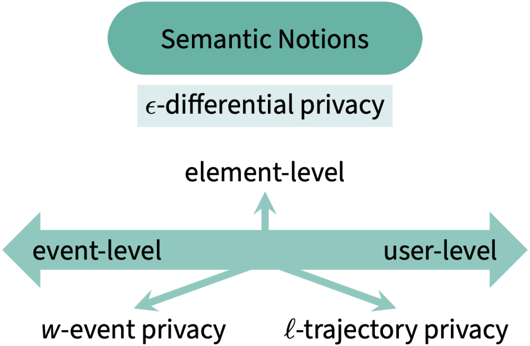


Privacy Loss (by observing r)

$$\mathcal{L}^r_{M(D)||M(D')} = ln\left(\frac{\mathbb{P}(M(D)=r)}{\mathbb{P}(M(D')=r)}\right)$$

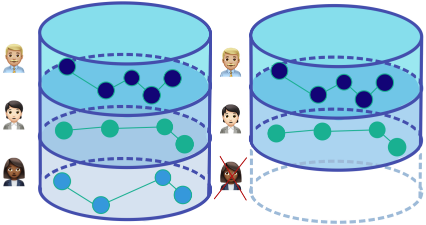


Privacy Notions in Trajectory Data



	t_1	t_2	t_3	t_4	...
u_1	Home →	→	Work →	Bar →	→
u_2		Work	Home →	→	→
u_3	Home →	→	Gym →	Work →	→
u_4		Home →	→	→	→

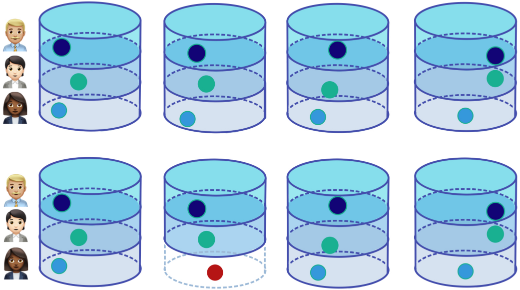
	t_1	t_2	t_3	t_4	...
u_1	Home →	→	Work →	Bar →	→
u_2		Work	Home →	→	→
u_3	Home →	→	Gym	Work →	→
u_4		Home →	→	→	→



Privacy Notions in Trajectory Data

Event-level

$$
S: \begin{array}{cccc}
t_1 & t_2 & \dots & t_m \\
(u_1, p_1^1) & (u_1, p_2^1) & \dots & (u_1, p_m^1) \\
(u_2, p_1^2) & \color{teal}{(u_2, p_2^2)} & \dots & (u_2, p_m^2) \\
\vdots & \vdots & & \vdots \\
(u_n, p_1^n) & (u_n, p_2^n) & \dots & (u_n, p_m^n)
\end{array}
\;\color{teal}{\Longrightarrow}\;
S': \begin{array}{cccc}
t_1 & t_2 & \dots & t_m \\
(u_1, p_1^1) & (u_1, p_2^1) & \dots & (u_1, p_m^1) \\
(u_2, p_1^2) & \color{teal}{(u_2, \hat{p})} & \dots & (u_2, p_m^2) \\
\vdots & \vdots & & \vdots \\
(u_n, p_1^n) & (u_n, p_2^n) & \dots & (u_n, p_m^n)
\end{array}
$$



Privacy Notions in Trajectory Data

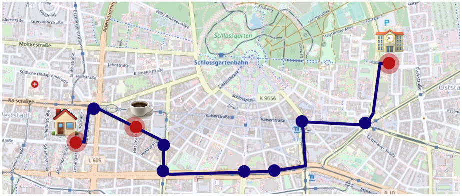
Event-level

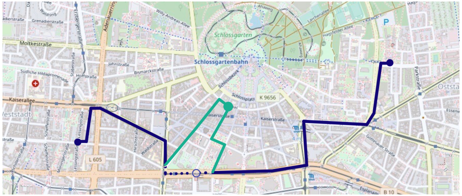
Event-neighborhood

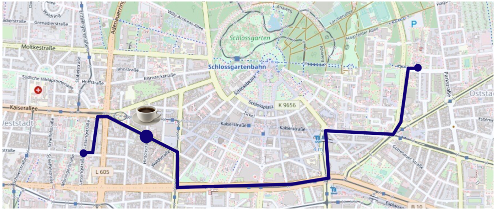
Two finite streams S and S' of symbols drawn from the discrete universe $\mathcal{X}$ are called *event-neighbors*, if and only if there exists $a, b \in \mathcal{X}$ such that if we change the instance of a in S to b we get S'.

Privacy Notions in Trajectory Data

Event-level







Privacy Notions in Trajectory Data
Location-Based Notions

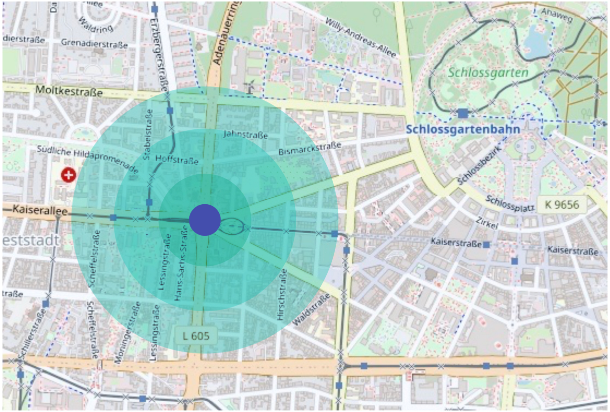
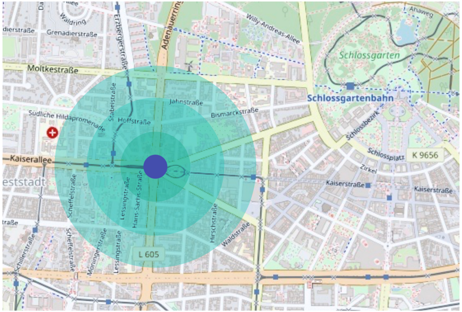
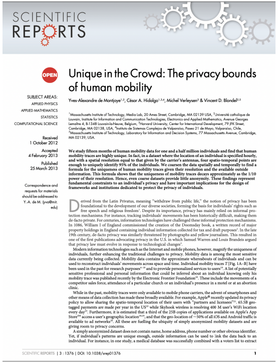


Figure 7: Geo-indistinguishability: $\mathbb{P}\{M(x) \in S\} \leq \mathrm{e}^{\epsilon d(x,x')} \cdot \mathbb{P}\{M(x') \in S\}$

Privacy Notions in Trajectory Data

Location-Based Notions





SCIENTIFIC REPORTS

OPEN

Unique in the Crowd: The privacy bounds of human mobility

SUBJECT AREAS:
APPLIED PHYSICS
APPLIED MATHEMATICS
STATISTICS
COMPUTATIONAL SCIENCE

Received
1 October 2012

Accepted
4 February 2013

Published
25 March 2013

Correspondence and requests for materials should be addressed to Y.-A. de M. (yva@mit.edu)

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a formula for the uniqueness of human mobility traces given their resolution and the available outside information. This formula shows that the uniqueness of mobility traces decays approximately as the 1/10 power of their resolution. Hence, even coarse datasets provide little anonymity. These findings represent fundamental constraints to an individual's privacy and have important implications for the design of frameworks and institutions dedicated to protect the privacy of individuals.

Derived from the Latin Privatus, meaning "withdraw from public life," the notion of privacy has been foundational to the development of our diverse societies, forming the basis for individuals' rights such as free speech and religious freedom[1]. Despite its importance, privacy has mainly relied on informal protection mechanisms. For instance, tracking individuals' movements has been historically difficult, making them de-facto private. For centuries, information technologies have challenged these informal protection mechanisms. In 1086, William I of England commissioned the creation of the Doomsday book, a written record of major property holdings in England containing individual information collected for tax and draft purposes[2]. In the late 19th century, de-facto privacy was similarly threatened by photographs and yellow journalism. This resulted in one of the first publications advocating privacy in the U.S. in which Samuel Warren and Louis Brandeis argued that privacy law must evolve in response to technological changes[3].

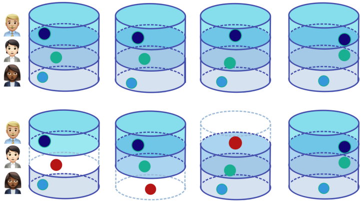
Modern information technologies such as the Internet and mobile phones, however, magnify the uniqueness of individuals, further enhancing the traditional challenges to privacy. Mobility data is among the most sensitive data currently being collected. Mobility data contains the approximate whereabouts of individuals and can be used to reconstruct individuals' movements across space and time. Individual mobility traces T [Fig. 1A–B] have been used in the past for research purposes[4–18] and to provide personalized services to users[19]. A list of potentially sensitive professional and personal information that could be inferred about an individual knowing only his mobility trace was published recently by the Electronic Frontier Foundation[20]. These include the movements of a competitor sales force, attendance of a particular church or an individual's presence in a motel or at an abortion clinic.

While in the past, mobility traces were only available to mobile phone carriers, the advent of smartphones and other means of data collection has made these broadly available. For example, Apple® recently updated its privacy policy to allow sharing the spatio-temporal location of their users with "partners and licensees"[21]. 65.5B geo-tagged payments are made per year in the US[22] while Skyhook wireless is resolving 400 M user's WiFi location every day[23]. Furthermore, it is estimated that a third of the 25B copies of applications available on Apple's App Store[SM] access a user's geographic location[24,25], and that the geo-location of ~50% of all iOS and Android traffic is available to ad networks[26]. All these are fuelling the ubiquity of simply anonymized mobility datasets and are giving room to privacy concerns.

A simply anonymized dataset does not contain name, home address, phone number or other obvious identifier. Yet, if individual's patterns are unique enough, outside information can be used to link the data back to an individual. For instance, in one study, a medical database was successfully combined with a voters list to extract

Privacy Notions in Trajectory Data

w-event privacy



	t_1	t_2	t_3	t_4	t_5	t_6	t_7
u_1	$home_1$	$home_1$	$work$	$work$	gym	$home_1$	$home_1$
u_2	$casino$	$casino$	$work_2$	$casino$	$casino$	$casino$	$casino$
u_3	$home_3$	$work_3$	$work_3$	$work_3$	$work_3$	$home_3$	$home_3$

w-event neighborhood

Let $w \in \mathbb{Z}^+$. $D_t = \{S_1, \ldots, S_t\}$ and $D_t' = \{S_1', \ldots, S_t'\}$ are *w-neighboring*, if, for all $i \leq t$, S_i and S_i' are either equal or we obtain one from the other by changing an entry of S_i, and all i, j corresponding to the latter case verify that $|i - j| < w$.

Privacy Notions in Trajectory Data

w-event privacy

user	t_1	t_2	t_3	t_4	t_5	t_6	t_7
u_1	home	home	work	work	gym	home	home
u_2	casino🎰	casino🎰	work	work	work	casino🎰	casino🎰
u_3	home	work	work	work	home	home	home

user	t_1	t_2	t_3	t_4	t_5
u_1	home				bar
u_2		bar			home
u_3	home	office	gym		home

Privacy Notions in Trajectory Data

ℓ-trajectory privacy

	user	time	loc
D_1	u_1	t_1	home
	u_3	t_1	home
D_2	u_2	t_2	bar
	u_3	t_2	office
D_3	u_3	t_3	gym
D_5	u_1	t_5	bar
	u_2	t_5	home
	u_3	t_5	home

user	t_1	t_2	t_3	t_4	t_5
u_1	home				bar
u_2		bar			home
u_3	home	office	gym		home

user	t_1	t_2	t_3	t_4	t_5
u_1	home				bar
u_2		bar			home
u_3	home	office	gym		home

1-trajectory

3-trajectory

2-trajectory

Privacy Notions in Trajectory Data

Element-level

Database

$D = \{x^{(u)}\}_{u=1}^{n}$

$x^{(u)} = \{x_1^{(u)}, \ldots, x_{m(u)}^{(u)}\}$

Clusters

$\mathcal{X}$ Universe $\longrightarrow \{C_1, \ldots, C_k\}$

Distance between users

$$d_{user}(x, x') := \sum_{k=1}^{K} 1_{\{\{x_i : x_i \in c_k\} \neq \{x'_i : x'_i \in c_k\}\}}$$

Privacy Notions in Trajectory Data

Element-level

user	t_1	t_2	t_3	$x \in C_1$	$x \in C_2$
u_1	café			{café}	$\emptyset$
u_2	café	café		{café, café}	$\emptyset$
u_3	café	café	café	{café, café, café}	$\emptyset$
u_4	café	café	home	{café, café}	$\{home\}$
u_5	café	home		{café}	$\{home\}$

$d(u_1, u_2) = 1$

$d(u_1, u_3) = 1$

$d(u_1, u_4) = 2$

$d(u_4, u_5) = 1$

Privacy Notions in Trajectory data

Type of privacy	Difference between neighboring databases
User-level	A user's whole trajectory
Event-level	A spatio-temporal point visited by a user (an event)
***w*-event**	A window of events over w consecutive timesteps
ℓ-trajectory	A sequence of ℓ consecutive spatio-temporal points from a single user
Element-level	A user's set of points belonging to the same unique cluster(*)

Table 2: Granularity notions and their concept of neighborhood.(*)unbounded notion