Location Privacy.
Where do we stand and where are we going?

Fernando Pérez-González
Signal Theory and Communications Department
Universidad de Vigo - SPAIN
Why do we like location based apps?
Google maps

The driver was following Google map.
Foursquare

The location of our Bat Cave is meant to be secret, so STOP CHECKING IN!
Facebook place tips

![Cartoon of a conversation where one person says, "WOW!!! GOLDEN EARRINGS! THANK YOU DARLING! THAT’S SO CUTE! BUT WHO HAS GOT THE SECOND PAIR YOU’VE BOUGHT?"

![Facebook Bluetooth Beacon]
Waze
And, of course...
How can you be geolocated? (without you fully knowing)
IP-based Geolocation

Source: GeoIPTool
Meta-data based Geolocation
Landmark recognition Geolocation

Legend of 5 goats | Guangzhou.chn.info
www.guangzhou.chn.info/overview/legend-of-5goats.html
There are many goat statues in Guangzhou and the Statue of the Five Goats is the most impressive, and now the one which were built in Yuexiu Park in 1959 ...

Five Rams Statue, Guangzhou 24 Insider Tips, Photos and Reviews
https://www.virtualtourist.com/places/Things_To_Do-Guangzhou-Five_Rams_Statue-BR...
Five Rams Statue reviews and photos from real travelers and locals in Guangzhou, ... The statue of five goats was originally created by Guangzhou sculptors Yin ...
Biometric geolocation
Credit card usage Geolocation

Mastercard under fire for tracking customer credit card purchases to sell to advertisers

- Credit card firm refuses to reveal 'proprietary' technique that allows it to anonymously track customers and target them with online ads
- Privacy campaigners accuse firm of 'treating details of our personal behaviour like their own property'
- System tracks information about the date, time, amount and merchant
- Credit card firm says system is only operational in US

By MARK PRIGG

Mastercard has come under fire for tracking its US customer’s purchases and selling the data to advertisers.
The credit card company’s MasterCard Advisors Media Solutions Group boasts it can target the most affluent customers and tell advertisers who is most likely to buy their products.
The firm does this by tracking a consumer's credit card details - although it says their identity...
Triangulation and other geolocation techniques
Signal strength-based triangulation

Source: The Wrongful Convictions Blog
Signal strength-based triangulation

Source: The Wrongful Convictions Blog
Multilateration: Time Difference of Arrival (TDOA)

Source: [Fujii et al. 2015]
Wardriving geolocation (Wigle)

Source: Wigle.net
Electrical Network Frequency Geolocation

Audio

ENF Extraction (filtering)

Estimation of ENF Fluctuations

Video

ENF Extraction (luminosity)

Time: 3:09:30 - 3:18:50
Place: North America

Time: 9:19:30 - 9:28:50
Place: Europe
Why is it dangerous?
Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI
December 24, 2012

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell’s screen, just a few miles away, was $14.29.
Buster busted!
Raising awareness about over-sharing

Check out our guest blog post on the CDT website.

Check your own Twitter timeline for checkins

Are you curious if people can see your checkins?
Enter your Twitter username and find out.

Your Twitter username

More Info
- Home
- Why

Made Possible By
- Foursquare
- Twitter
- @boyvanamstel
6 months in the life of Malte Spitz (2009-2010)

Source: http://www.zeit.de/datenschutz/malte-spitz-data-retention
Are we concerned about it?
Are people really concerned about location privacy?

• Survey by Skyhook Wireless (July 2015) of 1,000 Smartphone app users.
• 40% hesitate or don’t share location with apps.
• 20% turned off location for all their apps.
• Why people don’t share location?
  • 50% privacy concerns.
  • 23% don’t see value in location data.
  • 19% say it drains their battery.
• Why people turn off location?
  • 63% battery draining.
  • 45% privacy.
  • 20% avoid advertising.
How much is geolocation data worth?
How much value do we give to location data? [Staiano et al. 2014]

Many participants opted-out of revealing geolocation information.

Avg. daily value of location info: 3 €

Strong correlation between the amount traveled and the value given to location data.
Earn money as you share data

- GeoTask
- £1 PayPal cash voucher per 100 days of location data sharing (£0.01/day)

*Financial Times* in 2013: advertisers are willing to pay a mere $0.0005 per person for general information such as their age, gender and location, or $0.50 per 1,000 people.
Pay as you drive

- Formula can be a function of the amount of miles driven, or the type of driving, age of the driver, type of roads used...
- Up to 40% reduction in the cost of insurance.
BIA/Kelsey projects U.S. location-targeted mobile ad spending to grow from $9.8 billion in 2015 to $29.5 billion in 2020.

That’s $90 per person year!!!!
SAP, Germany, estimates wireless carrier revenue from selling mobile-user behavior data in $5.5 billion in 2015 and predicts $9.6 billion for 2016.
How about anonymization/pseudonymization?
Anonymity

Problems:

• Difficult authentication and personalization.
• Operating system or apps may access location before anonymization.
Pseudonimity

Problems:

• Operating system or apps may access location data before pseudonymization.

• Deanonymization.
Deanonymization based on home location [Hoh, Gruteser 2006]

• Data from GPS traces of larger Detroit area (1 min resolution).
• No data when vehicle parked.
• K-means algorithm for clustering locations + 2 heuristics:
  • Eliminate centroids that don’t have evening visits.
  • Eliminate centroids outside residential areas (manually).

Source: [Hoh, Gruteser 2006]
Deanonymization based on home location [Krummer 2007]

- 2-week GPS data from 172 subjects (avg. 6 sec resolution).
- Use heuristic to single out trips by car.
- Then use several heuristics: destination closest to 3 a.m. is home; place where individual spends most time is home; center of cluster with most points is home.
- Use reverse geocoding and white pages to deanonymize. Success measured by finding out name of individual.
- Positive identification rates around 5%.
- Even noise addition with std=500 m gives around 5% success when measured by finding out correct address.
Mobile trace uniqueness [de Montjoye et al 2013]

• Study on 15 months of mobility data; 0.5M individuals.
• Dataset with hourly updates and resolution given by cell carrier antennas, only 4 points suffice to identify 95% of individuals.
• Uniqueness of mobility traces decays as 1/10th power of their resolution.

Source: [de Montjoye et al. 2013]
Location privacy protection mechanisms
Location white lies

Source: Caro Spark (CC BY-NC-ND)
Location based privacy mechanisms

Input location \( X \)

Output pseudolocation \( Z \)

Source: Motherboards.org
Location privacy protection mechanisms (LPPMs)

• $Z = \varphi(X)$

• The mechanism may be deterministic (e.g., quantization) or stochastic (e.g., noise addition).

• Function $\varphi(\cdot)$ may depend on other contextual (e.g., time) or user-tunable (e.g., privacy level) parameters.

• When the mechanism is stochastic, there is an underlying probability density function, i.e.,

$$f(Z \mid X)$$
Hiding
Perturbation: (indepedent) noise addition
Perturbation: quantization
Obfuscation
Spatial Cloaking
How to commit the perfect murder
Space-time Cloaking
User-centric vs. Centralized LPPM
User-centric vs. Centralized LPPM
Utility vs. Privacy

• In broad terms:
Very nice, but…

• There are two main problems:
  How do we measure utility?
  How do we measure privacy?
How to measure utility?
How to measure utility?
How to measure utility?

Real position
	pseudolocation
A note about distances
Adversarial definition of privacy [Shokri et al 2011-]

• Assume stochastic mechanism for the user \( f(Z \mid X) \).
• Adversary constructs a (possibly stochastic) estimation remapping \( r(\hat{X} \mid Z) \).
• Prior \( \pi(X) \) assumed available to the adversary.
• \( d_p(x, \hat{x}) \) : Distance between \( \hat{x} \) and \( x \).
• \( d_q(x, z) \) : Distance between \( x \) and \( z \).
Adversarial definition of privacy [Shokri et al 2011-]

• Establish a cap on average utility loss: \( E\{d_q(X, Z)\} \leq QL \)
• This is a Stackelberg game in which the user chooses first and the adversary plays second.
• Find optimal adversarial ‘remapping’:
  \[
  r^*(\hat{X} \mid Z) = \arg \min \ E\{d_p(\hat{X}, X) \mid Z\}
  \]
• Optimal remapping depends on \( f(Z \mid X) \) and \( \pi(X) \).

\[
E\{d_p(\hat{X}, X) \mid Z\} = \sum_{X, \hat{X}} r(\hat{X} \mid Z)f(X \mid Z)d_p(\hat{X}, X)
\]

where

\[
f(X \mid Z) = \frac{f(Z \mid X) \cdot \pi(X)}{f(Z)}
\]
Example: uniform noise addition

\[
f(Z = z \mid X)
\]

\[
f(Z \mid X = x)
\]
Adversarial definition of privacy [Shokri et al 2011-]

• When for a given $Z$ there are several minimizers $\hat{X}$ the function $r^*(\hat{X} | Z)$ becomes stochastic.

• The user now must maximize privacy:

$$\max E\{d_p(\hat{X}, X)\} = \max \sum_{Z, X, \hat{X}} r^*(\hat{X} | Z)f(Z | X)\pi(X)d_p(X, \hat{X})$$

• Which is achieved for some mechanism $f^*(Z | X)$

• Privacy is defined as $E\{d_p(\hat{X}, X)\}$ after solving this maxmin problem.
An interesting result

• When \( d_p = d_q \):

\[
\begin{align*}
    f^* (Z = z \mid X) &= \arg \min E\{d_p(z, X)\} \\
    r^* (\hat{X} \mid Z = z) &= \delta(\hat{X} - z)
\end{align*}
\]

i.e. do nothing!

• When \( d_p = d_q = d_2 \) the following identity must hold

\[
z = E\{X \mid Z = z\}
\]

• When both user and adversary play optimally:

Privacy=Utility Loss
The Utility Loss-Privacy plane

Achievable region
Optimal Adversary

P=UL

Adv. Strategy 1
Adv. Strategy 2
Adv. Strategy 3
Adv. Strategy 4

Achievable region
Optimal Mechanism

Utility Loss
Privacy
What’s wrong with priors?

• Is it realistic to assume that the adversary knows the prior?
• Adversary no longer plays optimally with the ‘wrong’ prior.
• Shokri’s privacy definition is prior-dependent.
• Definition of differential privacy is prior-independent:

\[
\log(\Pr\{A(D_1) \in S\}) \leq \varepsilon + \log(\Pr\{A(D_2) \in S\})
\]

- Two databases \( D_1, D_2 \) differing in a single element.
- \( A \): randomized algorithm.
- \( S \): set of possible subsets of \( \text{im}(A) \).
Geoindistinguishability [Chatzikokolakis et al 2013-]

• A mechanism is \( \varepsilon \)-geo-indistinguishable iff:

\[
| \log(f(z \mid X = x)) - \log(f(z \mid X = x')) | \leq \varepsilon \cdot d_p(x, x')
\]

for all \( x, x', z \).

• Differential privacy corresponds to \( d_p = \) Hamming distance.

• Definition is prior-independent.

• Guarantees a small leakage of information BUT is no defense against EVERY adversary: with proper side information, adversary can learn a lot!
Uniform mechanisms do not provide geo-ind

\[ f(Z \mid X = x') \]

\[ \log(f(z \mid X = x) - \log(f(z \mid X = x')) \geq \infty \]
Laplacian mechanism

• Laplacian distribution in polar coordinates:

\[
f(z \mid X = x) = \frac{\varepsilon}{2\pi} e^{-\varepsilon \cdot d_2(x, z)}
\]

• Then,

\[
|\log f(z \mid X = x) - \log f(z = X = x')| = | \varepsilon \cdot d_2(z, x') - \varepsilon \cdot d_2(z, x) | \leq \varepsilon \cdot d_2(x, x')
\]

Triangle inequality

• The Laplacian mechanism satisfies the geo-ind condition.
Laplacian mechanism
Optimal mechanisms for geo-ind

• Minimize quality loss (i.e., $E\{d_q (X, Z)\}$) subject to $\varepsilon$–geo-ind constraint.

• Fact: $\varepsilon$–geo-ind constraint is kept under any adversarial remapping $r(\hat{X} | Z)$

• Optimal mechanism is then

$$f^* (Z | X) = \arg \min E\{d_q (Z, X)\}$$

where

$$E\{d_q (X, Z)\} = \sum_{X,Z} f (Z | X) \pi (X) d_q (X, Z)$$

• The optimal adversarial remapping would find

$$r^* (\hat{X} | Z) = \arg \min E\{d_p (\hat{X}, X) | Z\}$$
Optimal mechanisms for geo-ind

- If $d_p = d_q$, the adversary does nothing. Minimization of the QL has been already done by the mechanism!!
- But if the adversary does nothing, Privacy=QL.
- The operating value thus depends on $\varepsilon$ (the smaller, the larger the privacy).
Where are we going?
Sensitivity [Bertino et. al 2010]
Sensitivity

• The mechanism should weigh the importance given by the user to each location.
• This can be specified semantically by defining categories.
• Sensitivity of a region: prob. that the user, known to be in that region, is actually in a sensitive place.
• For other mechanisms: open problem.
Graph-based models
Graph-based models
Graph-based models
Graph-based models

• A trace is a path together with time \( \{X_i, t_i\}_{i=1}^N \).

• Common assumption for an adversary: the true trace can be described through a Markov chain.

• Prior transition probabilities between states can be estimated if training traces are (at least partially) available.
Graph-based models

• Shokri et al.’s approach: depending on what the adversary wants to learn, apply a different method.

• **Maximum likelihood**: find the most likely trace given the observed trace

\[
\arg \max_{\{X_i, t_i\}_{i=1}^N} f(\{X_i, t_i\}_{i=1}^N \mid \{Z_i, t_i\}_{i=1}^N)
\]

• Dynamic programming (e.g., Viterbi algorithm) can be used.
Graph-based models

- **Distribution estimation**: estimate the probabilities of all traces using the Metropolis-Hastings algorithm.
Graph-based models

- **Location estimation**: find the most likely node at time $t_k$

$$\arg \max_{X_k} f(X_k \mid \{Z_i, t_i\}_{i=1}^N)$$

- Can be solved using the backward-forward algorithm to recursively compute the probabilities.
Privacy as a zero-sum game

Privacy + Utility = constant

Privacy

Utility

Loss
Adding a new dimension: bandwidth

\[ n = 3 \implies 8 \text{ dummies} \]

Utility Loss \( \propto d_2(s) = d_2(S)/3 \)
Privacy \( \propto d_2(S) \)
The Utility Loss-Privacy-Bandwidth region

- Achievable region
  - Optimal Adversary
- Privacy gain due to dummying
- Achievable region
  - Optimal Mechanism

- Service provider utility loss
  - P=UL
- User utility loss
  - P=3 UL

- BW is now 9 times larger
Space-time cloaking

Utility Loss $\propto$ area
Privacy $\equiv k - \text{anonymity} \propto \text{area} \times \text{time} \times \text{pop. density}$
Delay $\propto$ time
Privacy-preserving queries

Retrieval in Encrypted Domain

Encrypted query

Encrypted reply
Thanks!

fperez@gts.uvigo.es
www.gpsc.uvigo.es
What utility? An example

Utility \equiv 1 / d_{max} \propto 1 / \sqrt{\text{area}}

Privacy \equiv k - anonymity \propto \text{area} \times \text{time} \times \text{pop. density}
But delay also counts…

Utility vs Privacy graph showing three curves for delays of 5 min, 10 min, and 15 min.
What utility? Another example

• Space-time slicing
• Is this related to bandwidth?
• Space-time slicing
• Is this related to bandwidth?