Location Differential Privacy

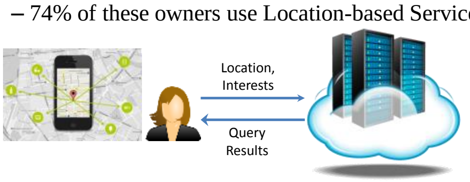
Zhipeng Cai

Outline

- Location Privacy
  - Motivation
  - Existing Notions of Privacy
- Geo-Indistinguishability
  - Definition
  - Characterization
  - Mechanism
  - Accuracy
- Hierarchical Location Publishing
  - Motivation
  - *PriLocation* Algorithm
  - Accuracy Analysis
  - Privacy Analysis

Location Privacy

- Ubiquitous Location-based Services (LBS)
  - 46% of the adult population in US own smartphones by 2012 [Pew Internet & American Life Project]
  - 74% of these owners use Location-based Services



Location Privacy

- Ubiquitous Location-based Services (LBS)

Location Privacy

- Privacy Issues Related to Locations
  - Individuals' locations themselves are sensitive information
  - Locations could be used to infer individuals' sensitive information
    - Home location, work location
    - Sexual preferences, political views, religious inclinations
    - Etc.

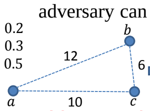
Location Privacy

- Privacy Issues Related to Locations
  - Monitoring and controlling of an individual's location has been considered as a form of slavery
  - Even lead to security issue to individuals

Location Privacy

- Existing Notions of Privacy
  - Expected Distance Error
    - A natural way to quantify the accuracy by which an adversary can guess the real location

$\Pr(a) = 0.2$
$\Pr(b) = 0.3$
$\Pr(c) = 0.5$



Three locations and their prior probability

Location\Guess	a	b	c
a	0.8	0.1	0.1
b	0.1	0.6	0.3
c	0.2	0.3	0.5

Adversary's guessing strategy

$\mathcal{K}$

Location Preserving Mechanism

Mechanism design: Maximizing the expected distance error

Expected Distance Error:

At a: $0.8 \times 0 + 0.1 \times 12 + 0.1 \times 10 = 2.2$

At b: $0.1 \times 12 + 0.6 \times 0 + 0.3 \times 6 = 3$

At c: $0.2 \times 10 + 0.3 \times 6 + 0.5 \times 0 = 3.8$

Overall: $0.2 \times 2.2 + 0.3 \times 3 + 0.5 \times 3.8 = 3.24$

Location Privacy

- Existing Notions of Privacy
  - Expected Distance Error
    - Inaccuracy estimation of adversary's side information leads to poorly designed mechanism

$\Pr(a) = 0.2$
$\Pr(b) = 0.3$
$\Pr(c) = 0.5$

Inaccurate prior probability

Location\Guess	a	b	c
a	0.8	0.1	0.1
b	0.1	0.6	0.3
c	0.2	0.3	0.5

Inaccurate adversary's guessing strategy

Poor M… esign

➢ Aim… ing the inaccurate expected distance error

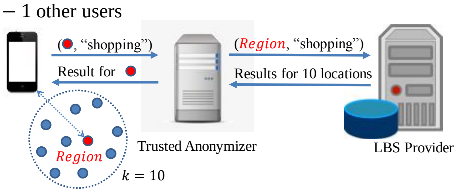
$\mathcal{K}$

Location Preserving Mechanism

Inaccurate expected distance error

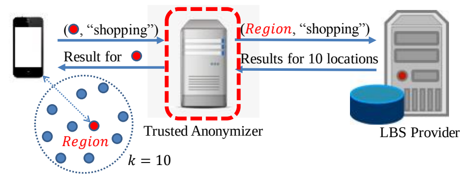
Location Privacy

- Existing Notions of Privacy
  - k-Anonymity (Cloaking)
    - The most widely used privacy notion for location-based systems
    - Protect user's identity by hiding a user among at least $k-1$ other users



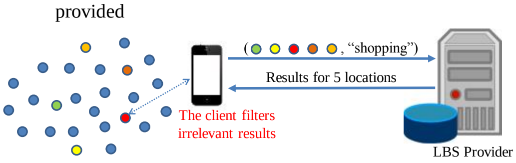
Location Privacy

- Existing Notions of Privacy
  - k-Anonymity (Cloaking)
    - Privacy breach
    - Performance bottleneck



Location Privacy

- Existing Notions of Privacy
  - k-Anonymity (Client-based Solution)
    - Generate $k-1$ dummy locations and inject them in the query reported to the LBS server
    - No meaningful indistinguishability among k objects is provided

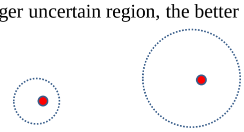


Location Privacy

- Existing Notions of Privacy
  - Differential Privacy
    - Modifying a single user's data have a negligible effect on the outcome
    - Not suitable for scenarios where only a single object (location) is involved

Location Privacy

- Existing Notions of Privacy
  - Other location-privacy metrics
    - Uncertain region: the real location is inside it, but the adversary does not know its exact position
    - Privacy is measured by the size of uncertain region
    - The larger uncertain region, the better privacy



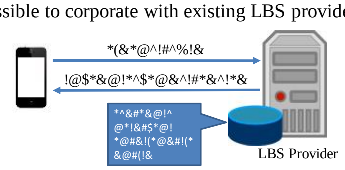
Poor Privacy Better Privacy

Location Privacy

- Existing Notions of Privacy
  - Other location-privacy metrics
    - The ratio between the inference accuracy before and after the application of mechanism
      - An optimal guess: pick the location with the largest probability
      - The inference accuracy before the application of mechanism
        $$acc = \max_{l \in L} \Pr(l)$$
      - The inference accuracy after the application of the mechanism with output r
        $$acc = \max_{l \in L} \Pr(l|r)$$
      - $privacy = \frac{\max_{l \in L} \Pr(l)}{\max_{l \in L} \Pr(l|r)}$ ← The larger, the better

Location Privacy

- Existing Notions of Privacy
  - Transformation-based approaches
    - Employing cryptographic techniques to data and query
    - Private information retrieval
    - Difficult to implement in mobile devices
    - Impossible to corporate with existing LBS providers



Geo-Indistinguishability [CCS 2013]

- Basic Idea
  - Differential privacy guarantees that for neighboring databases D and D'
    $$\frac{\Pr(\mathcal{M}(D) \in S)}{\Pr(\mathcal{M}(D') \in S)} \le e^{\varepsilon}$$
  - Geo-indistinguishability (gi) provide differential privacy to locations
    - Different locations could produce similar outputs
    - Make different locations indistinguishable

Geo-Indistinguishability

- Basic Idea
  - Can we make each pair of locations indistinguishable?

$$\frac{\Pr(\mathcal{K}(x)=z)}{\Pr(\mathcal{K}(x')=z)} \leq e^{\varepsilon}$$

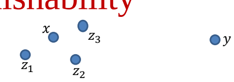
here x and x' are input locations, z is any output location, and ε is the privacy budget.

  - Any pair of locations x and x' are indistinguishable when ε is small
  - Strict privacy has been obtained
  - What about location utility?

Geo-Indistinguishability

- Basic Idea
  - Utility Point of View
    - Location-based services are usually used to search nearby services, points of interests and etc.
    - Suppose we are looking for a service at location x
    - To preserve our location privacy, we adopt a mechanism called $\mathcal{K}$ and report $\mathcal{K}(x) = z$ to the LBS provider
    - $\mathcal{K}(x) = z$, then z should not be far away from x, otherwise one can not obtain meaningful service at x

Geo-Indistinguishability



- Basic Idea
  - Suppose we have made each pair of locations indistinguishable with

$$e^{-\varepsilon} \leq \frac{\Pr(\mathcal{K}(x)=z)}{\Pr(\mathcal{K}(x')=z)} \leq e^{\varepsilon}$$

  - Consider two locations x and y at significant distance
    - If we have good utility at x, then $\mathcal{K}(x)$ should be nearby x (say z_1, z_2 and z_3) with large probability p
    - Then $\mathcal{K}(y) \in \{z_1, z_2, z_3\}$ with probability no smaller than $pe^{-\varepsilon}$, and $pe^{-\varepsilon} \rightarrow p$ when ε is small
    - So we can not obtain good utility at y

Geo-Indistinguishability

- Basic Idea
  - Geo-indistinguishability makes nearby locations hard to distinguish
  - Locations faraway from each other remain easy to distinguish
  - Privacy budget controls the level of privacy at each unit of distance

$$\frac{\Pr(\mathcal{K}(x)=z)}{\Pr(\mathcal{K}(x')=z)} \leq e^{\varepsilon} \Rightarrow \frac{\Pr(\mathcal{K}(x) = z)}{\Pr(\mathcal{K}(x') = z)} \leq e^{\varepsilon d(x,x')}$$

Geo-Indistinguishability

- Notation
  - ε: privacy budget, the level of privacy at one unit of distance
  - $\mathcal{X}$: the set of points of interests (locations)
  - $\mathcal{Z}$: the set of possible reported locations
  - π: the prior distribution on $\mathcal{X}$
  - $d(x, x')$: the Euclidean distance between locations x and x'

Geo-Indistinguishability

- Notation
  - $\mathcal{K}$: a mechanism $\mathcal{K}$ is a probabilistic function for selecting a reported value
  - $\mathcal{K}(x)$: the probabilistic distribution of reported location, given x
  - $\mathcal{K}(x)(Z)$: the probability that the reporting a location belongs to set $Z \subseteq \mathcal{Z}$, given x

$\mathcal{X} \backslash \mathcal{Z}$	z_1	z_2	z_3
x_1	0.7	0.2	0.1
x_2	0.15	0.75	0.1
x_3	0.2	0.7	0.1

A mechanism $\mathcal{K}$

Distribution of $\mathcal{K}(x_1)$

$\mathcal{K}(x_1)(Z) = 0.9$ where $Z = \{z_1, z_2\}$

Geo-Indistinguishability

- Notation
  - $d_{\mathcal{P}}(\sigma_1, \sigma_2)$: the multiplicative distance between two distributions σ_1 and σ_2 on some set $\mathcal{S}$
    - $d_{\mathcal{P}}(\sigma_1, \sigma_2) = \max_{S \subseteq \mathcal{S}} |\ln \frac{\sigma_1(S)}{\sigma_2(S)}|$
  - $Bayes(\pi, \mathcal{K}, Z)$: the posterior distribution on $\mathcal{X}$, given the observation Z produced by $\mathcal{K}$
    - $Bayes(\pi, \mathcal{K}, Z) = \frac{\mathcal{K}\ (x)(Z)\pi(x)}{\sum_{x' \in \mathcal{X}} \mathcal{K}\ (x')(Z)\pi(x')}$

Geo-Indistinguishability

- Original Definition of Geo-Indistinguishability
  - Given privacy budget $\varepsilon \geq 0$, a mechanism $\mathcal{K}$ satisfies ε-geo-indistinguishability if and only if for all $x, x' \in \mathcal{X}$:
    $$d_{\mathcal{P}}(\mathcal{K}(x), \mathcal{K}(x')) \leq \varepsilon d(x, x')$$
- Definition in dp fassion
  - Given privacy budget $\varepsilon \geq 0$, a mechanism $\mathcal{K}$ satisfies ε-geo-indistinguishability if and only if for all $x, x' \in \mathcal{X}, Z \subseteq \mathcal{Z}$:
    $$\mathcal{K}(x)(Z) \leq e^{\varepsilon d(x,x')} \mathcal{K}(x')(Z)$$

Geo-Indistinguishability

- Characterizations of Geo-Indistinguishability
  - Adversary's conclusions under hiding
    - $\phi: \mathcal{X} \rightarrow \mathcal{X}$: A hiding function
    - ϕ can be applied to the actual location before $\mathcal{K}$
      - $\phi(x) = y$
    - A mechanism $\mathcal{K}$ with hiding applied is $\mathcal{K} \circ \phi$
      - $\mathcal{K} \circ \phi(x) = \mathcal{K}(\phi(x)) = \mathcal{K}(y)$
    - $d(\phi)$: the maximum distance between the real and hidden location, that is

$$d(\phi) = \max_{x \in \mathcal{X}} d(x, \phi(x))$$

Can we improve the privacy of geo-indistinguishability using hiding?

Geo-Indistinguishability

- Adversary's Conclusions Under Hiding
  - A mechanism $\mathcal{K}$ satisfies ε-gi, if and only if for all $\phi: \mathcal{X} \rightarrow \mathcal{X}$, all priors π, and all $Z \subseteq \mathcal{Z}$, the following condition holds: $d_{\mathcal{P}}(\sigma_1, \sigma_2) \leq 2\varepsilon d(\phi)$.
    - $\sigma_1 = Bayes(\pi, \mathcal{K}, Z)$
    - $\sigma_2 = Bayes(\pi, \mathcal{K} \circ \phi, Z)$
  - ➢ $d(\phi)$ should not be large due to utility consideration
  - ➢ Adversaries have similar inference no matter whether hiding is adopted
  - ➢ Hiding does not improve the privacy of gi
  - ➢ When $d(\phi)$ grows large, privacy is exchanged with utility, not improved by hiding

Geo-Indistinguishability

- Adversary's Conclusions Under Hiding
  - Proof Sketch
    - Suppose $\mathcal{K}$ satisfies ε-gi, for all $x \in \mathcal{X}$, any hiding function $\phi: \mathcal{X} \rightarrow \mathcal{X}$ and all $Z \subseteq \mathcal{Z}$, we analyze the ratio between $\mathcal{K}(x)$ and $\mathcal{K}(\phi\ (x))$ ← $\mathcal{K} \circ \phi\ (x)$

$d_{\mathcal{P}}(\mathcal{K}(x), \mathcal{K}(x')) \leq \varepsilon d(x, x')$

↓

$d_{\mathcal{P}}(\mathcal{K}(x), \mathcal{K}(\phi\ (x))) \leq \varepsilon d(x, \phi\ (x))$

↓

$d_{\mathcal{P}}(\mathcal{K}(x), \mathcal{K} \circ \phi\ (x)\ ) \leq \varepsilon d(\phi)$

↓

$$e^{-\varepsilon d(\phi)} \leq \frac{\mathcal{K}(x)(Z)}{\mathcal{K} \circ \phi\ (x)(Z)} \leq e^{\varepsilon d(\phi)}$$

Geo-Indistinguishability

- Adversary's Conclusions Under Hiding
  - Proof Sketch
    - Suppose $\mathcal{K}$ satisfies ε-gi, for all $x \in \mathcal{X}$, any hiding function $\phi: \mathcal{X} \rightarrow \mathcal{X}$ and all $Z \subseteq \mathcal{Z}$, we analyze the ratio between σ_1 and σ_2

$$\frac{\sigma_1}{\sigma_2} = \frac{\frac{\mathcal{K}\ (x)(Z)\pi(x)}{\sum_{x' \in \mathcal{X}} \mathcal{K}\ (x')(Z)\pi(x')}}{\frac{\mathcal{K} \circ \phi\ (x)(Z)\pi(x)}{\sum_{x' \in \mathcal{X}} \mathcal{K} \circ \phi\ (x')(Z)\pi(x')}} = \frac{\mathcal{K}\ (x)(Z)\pi(x)}{\mathcal{K} \circ \phi\ (x)(Z)\pi(x)} \times \frac{\sum_{x' \in \mathcal{X}} \mathcal{K} \circ \phi\ (x')(Z)\pi(x')}{\sum_{x' \in \mathcal{X}} \mathcal{K}\ (x')(Z)\pi(x')}$$

$$e^{-\varepsilon d(\phi)} \leq \frac{\mathcal{K}(x)(Z)}{\mathcal{K} \circ \phi\ (x)(Z)} \leq e^{\varepsilon d(\phi)} \left\{ \begin{array}{l} \mathcal{K}(x)(Z) \leq e^{\varepsilon d(\phi)} \mathcal{K} \circ \phi\ (x)(Z) \\ \mathcal{K} \circ \phi\ (x)(Z) \leq e^{\varepsilon d(\phi)} \mathcal{K}(x)(Z) \end{array} \right.$$

Geo-Indistinguishability

- Adversary's Conclusions Under Hiding
  - Proof Sketch

$\mathcal{K}(x)(Z) \leq e^{\varepsilon d(\phi)}\mathcal{K} \circ \phi\ (x)(Z)$

$\mathcal{K} \circ \phi\ (x)(Z)\ \leq e^{\varepsilon d(\phi)}\mathcal{K}(x)(Z)$

- $\frac{\sigma_1}{\sigma_2} = \frac{\mathcal{K}\ (x)(Z)\pi(x)}{\mathcal{K}\circ\phi\ (x)(Z)\pi(x)} \times \frac{\sum_{x'\in\mathcal{X}} \mathcal{K}\circ\phi\ (x')(Z)\pi(x')}{\sum_{x'\in\mathcal{X}} \mathcal{K}\ (x')(Z)\pi(x')}$

$$\leq \frac{e^{d(\phi)}\mathcal{K}\circ\phi\ (x)(Z)}{\mathcal{K}\circ\phi\ (x)(Z)} \times \frac{\sum_{x'\in\mathcal{X}} e^{d(\phi)}\mathcal{K}\ (x')(Z)\pi(x')}{\sum_{x'\in\mathcal{X}} \mathcal{K}\ (x')(Z)\pi(x')} = e^{2d(\phi)}$$

Similarly, we can proof $\frac{\sigma_2}{\sigma_1} \leq e^{2d(\phi)}$

Geo-Indistinguishability

- Adversary's Conclusions Under Hiding
  - Proof Sketch
    - Next, we are to prove that given a mechanism $\mathcal{K}$, if for all $\phi: \mathcal{X} \to \mathcal{X}$, all priors π, and all $Z \subseteq \mathcal{Z}$, $d_{\mathcal{P}}(\sigma_1, \sigma_2) \leq 2\varepsilon d(\phi)$ holds, then $\mathcal{K}$ satisfies ε-gi

$d_{\mathcal{P}}(\sigma_1, \sigma_2) \leq 2\varepsilon d(\phi)$ ⇒ $\frac{\sigma_1}{\sigma_2} \leq e^{2d(\phi)}$ for any x

    - For any pair of locations $x_1, x_2 \in \mathcal{X}$, we construct a hiding function $\phi_{x_1,x_2}: \mathcal{X} \to \mathcal{X}$ and a prior π_{x_1,x_2}
    - Then we take the constructed ϕ_{x_1,x_2} and π_{x_1,x_2} into the presentation of $d_{\mathcal{P}}(\sigma_1, \sigma_2)$

Geo-Indistinguishability

- Adversary's Conclusions Under Hiding
  - Proof Sketch
    - For any pair of locations $x_1, x_2 \in \mathcal{X}$, we construct a hiding function $\phi_{x_1,x_2}: \mathcal{X} \to \mathcal{X}$ as follow:
      - $\phi_{x_1,x_2}(x_1) = x_2$
      - $\phi_{x_1,x_2}(x_2) = x_1$
      - $\phi_{x_1,x_2}(y) = y$ for any $y \in \mathcal{X}/\{x_1, x_2\}$
      - Then we have $d(\phi_{x_1,x_2}) = d(x_1, x_2)$

Geo-Indistinguishability

- Adversary's Conclusions Under Hiding
  - Proof Sketch
    - For any pair of locations $x_1, x_2 \in \mathcal{X}$, we construct a prior π_{x_1,x_2} on $\mathcal{X}$ as follow:
      - $\pi_{x_1,x_2}(x_1) = \frac{1}{n}$ where n can be any positive number that $n > 1$
      - $\pi_{x_1,x_2}(x_2) = 1 - \frac{1}{n}$
      - $\pi_{x_1,x_2}(y) = 0$ for any $y \in \mathcal{X}/\{x_1, x_2\}$
    - When $n \to +\infty$
      - $\pi_{x_1,x_2}(x_1) \to 0^+$
      - $\pi_{x_1,x_2}(x_2) \to 1$

Geo-Indistinguishability

- Adversary's Conclusions Under Hiding
  - Proof Sketch

➢ $\frac{\sigma_1}{\sigma_2} = \frac{\mathcal{K}\ (x)(Z)\pi(x)}{\mathcal{K}\circ\phi\ (x)(Z)\pi(x)} \times \frac{\sum_{x'\in\mathcal{X}} \mathcal{K}\circ\phi\ (x')(Z)\pi(x')}{\sum_{x'\in\mathcal{X}} \mathcal{K}\ (x')(Z)\pi(x')} \leq e^{2\varepsilon d(\phi)}$ holds for for all $\phi: \mathcal{X} \to \mathcal{X}$, all priors π, all $Z \subseteq \mathcal{Z}$ and any $x \in \mathcal{X}$

➢ Take ϕ_{x_1,x_2} and π_{x_1,x_2} into the above inequation, and let $x = x_1$

First term $\boxed{\frac{\mathcal{K}\ (x)(Z)\pi(x)}{\mathcal{K}\circ\phi\ (x)(Z)\pi(x)}} \Rightarrow \boxed{\frac{\mathcal{K}\ (x_1)(Z)\pi_{x_1,x_2}(x_1)}{\mathcal{K}\circ\phi_{x_1,x_2}(x_1)(Z)\pi_{x_1,x_2}(x_1)}}$

➢ $\frac{\mathcal{K}\ (x_1)(Z)\pi_{x_1,x_2}(x_1)}{\mathcal{K}\circ\phi_{x_1,x_2}(x_1)(Z)\pi_{x_1,x_2}(x_1)} = \frac{\mathcal{K}\ (x_1)(Z)\pi_{x_1,x_2}(x_1)}{\mathcal{K}\ (x_2)(Z)\pi_{x_1,x_2}(x_1)} = \frac{\mathcal{K}\ (x_1)(Z)}{\mathcal{K}\ (x_2)(Z)}$

Geo-Indistinguishability

- Adversary's Conclusions Under Hiding
  - Proof Sketch

Second term $\boxed{\frac{\sum_{x'\in\mathcal{X}} \mathcal{K}\circ\phi\ (x')(Z)\pi(x')}{\sum_{x'\in\mathcal{X}} \mathcal{K}\ (x')(Z)\pi(x')}} \Rightarrow \boxed{\frac{\sum_{x'\in\mathcal{X}} \mathcal{K}\circ\phi_{x_1,x_2}(x')(Z)\pi_{x_1,x_2}(x')}{\sum_{x'\in\mathcal{X}} \mathcal{K}\ (x')(Z)\pi_{x_1,x_2}(x')}}$

➢ $\frac{\sum_{x'\in\mathcal{X}} \mathcal{K}\circ\phi_{x_1,x_2}(x')(Z)\pi_{x_1,x_2}(x')}{\sum_{x'\in\mathcal{X}} \mathcal{K}\ (x')(Z)\pi_{x_1,x_2}(x')}$

$= \frac{\mathcal{K}\circ\phi_{x_1,x_2}(x_1)(Z)\pi_{x_1,x_2}(x_1)+\mathcal{K}\circ\phi_{x_1,x_2}(x_2)(Z)\pi_{x_1,x_2}(x_2)}{\mathcal{K}\ (x_1)(Z)\pi_{x_1,x_2}(x_1)+\mathcal{K}\ (x_2)(Z)\pi_{x_1,x_2}(x_2)}$

$= \frac{\mathcal{K}(x_2)(Z)\frac{1}{n}+\mathcal{K}\ (x_1)(Z)\frac{n-1}{n}}{\mathcal{K}(x_1)(Z)\frac{1}{n}+\mathcal{K}\ (x_2)(Z)\frac{n-1}{n}}$

➢ When $n \to +\infty$, $\frac{\mathcal{K}(x_2)(Z)\frac{1}{n}+\mathcal{K}\ (x_1)(Z)\frac{n-1}{n}}{\mathcal{K}(x_1)(Z)\frac{1}{n}+\mathcal{K}\ (x_2)(Z)\frac{n-1}{n}} \to \frac{\mathcal{K}\ (x_1)(Z)}{\mathcal{K}\ (x_2)(Z)}$

Geo-Indistinguishability

- Adversary's Conclusions Under Hiding
  - Proof Sketch
    - Put the first term and second term together ($n \to +\infty$)

$$\frac{\mathcal{K}\ (x_1)(Z)\pi_{x_1,x_2}(x_1)}{\mathcal{K}\circ\phi_{x_1,x_2}(x_1)(Z)\pi_{x_1,x_2}(x_1)} \frac{\sum_{x'\in\mathcal{X}} \mathcal{K}\circ\phi_{x_1,x_2}(x')(Z)\pi_{x_1,x_2}(x')}{\sum_{x'\in\mathcal{X}} \mathcal{K}\ (x')(Z)\pi_{x_1,x_2}(x')}$$

$$= \left(\frac{\mathcal{K}\ (x_1)(Z)}{\mathcal{K}\ (x_2)(Z)}\right)^2 \leq e^{2\varepsilon d(\phi)} = e^{2\varepsilon d(x_1,x_2)}$$

Then we have $\frac{\mathcal{K}\ (x_1)(Z)}{\mathcal{K}\ (x_2)(Z)} \leq e^{\varepsilon d(x_1,x_2)}$

That is for any $x_1, x_2 \in \mathcal{X}$, we get $d_{\mathcal{P}}(\mathcal{K}(x_1), \mathcal{K}(x_2)) \leq \varepsilon d(x_1, x_2)$

Geo-Indistinguishability

- Characterizations of Geo-Indistinguishability
  - Knowledge of an informed attacker
    - Suppose the adversary already knows $x \in N \subseteq \mathcal{X}$
    - $d(N) = \max_{x,x'\in N} d(x, x')$
  - A mechanism $\mathcal{K}$ satisfies ε-gi if and only if for all $N \subseteq \mathcal{X}$, all priors π on $\mathcal{X}$, and all $Z \subseteq \mathcal{Z}$:

$$d_{\mathcal{P}}(\pi(x|N), Bayes(\pi, \mathcal{K}, Z|N)) \leq d(N)$$

Geo-Indistinguishability

- Characterizations of Geo-Indistinguishability
  - Knowledge of an informed attacker
    - The user's location remains private, regardless the adversary's prior knowledge of N
    - The knowledge obtained by learning the mechanism result is bounded by $d(N)$
    - When $d(N)$ is small, the adversary could no longer improve the accuracy of guessing
    - When $d(N)$ is small, the adversary could improve the accuracy of guessing, however this is due to the demand of location utility

Geo-Indistinguishability

- Knowledge of an informed attacker
  - Proof Sketch
    - Suppose $\mathcal{K}$ satisfies ε-gi, lets analyze the ratio between $\pi(x|N)$ and $Bayes(\pi, \mathcal{K}, Z|N)$ and the vice
    - $\frac{\pi(x|N)}{Bayes(\pi,\mathcal{K},Z|N)} = \frac{\pi(x|N)}{\frac{\pi(x|N)\mathcal{K}(x)(Z)}{\sum_{x' \in N}\pi(x'|N)\mathcal{K}(x')(Z)}} = \frac{\sum_{x' \in N}\pi(x'|N)\mathcal{K}(x')(Z)}{\mathcal{K}(x)(Z)} \leq \frac{\sum_{x' \in N}\pi(x'|N)e^{d(x,x')}\mathcal{K}(x)(Z)}{\mathcal{K}(x)(Z)} \leq \max_{x' \in N} e^{d(x,x')} \leq e^{d(N)}$

$\mathcal{K}(x')(Z) \leq e^{d(x,x')}\mathcal{K}(x)(Z)$

Geo-Indistinguishability

- Knowledge of an informed attacker
  - Proof Sketch
    - $\frac{Bayes(\pi,\mathcal{K},Z|N)}{\pi(x|N)} = \frac{\mathcal{K}(x)(Z)}{\sum_{x' \in N}\pi(x'|N)\mathcal{K}(x')(Z)} = \frac{\sum_{x' \in N}\pi(x'|N)\mathcal{K}(x)(Z)}{\sum_{x' \in N}\pi(x'|N)\mathcal{K}(x')(Z)} \leq \frac{\sum_{x' \in N}\pi(x'|N)\mathcal{K}(x')(Z)e^{d(x,x')}}{\sum_{x' \in N}\pi(x'|N)\mathcal{K}(x')(Z)} = e^{d(x,x')} \leq e^{d(N)}$
    - Then we conclude that

      $d_{\mathcal{P}}(\pi(x|N), Bayes(\pi, \mathcal{K}, Z|N)) \leq d(N)$

$\mathcal{K}(x)(Z) \leq e^{d(x,x')}\mathcal{K}(x')(Z)$

Geo-Indistinguishability

- Knowledge of an informed attacker
  - Proof Sketch
    - Given that for all $N \subseteq \mathcal{X}$, and all $Z \subseteq \mathcal{Z}$:

      $d_{\mathcal{P}}(\pi(x|N), Bayes(\pi, \mathcal{K}, Z|N)) \leq d(N)$
    - We employ contradiction for the other direction of proof
    - Suppose $\mathcal{K}$ does not satisfy ε-gi, then there exist $x, y \in \mathcal{X}$ and $Z \subseteq \mathcal{Z}$, so that $d_{\mathcal{P}}(\mathcal{K}(x), \mathcal{K}(y)) > \varepsilon d(x, y)$
      - $\frac{\mathcal{K}(x)(Z)}{\mathcal{K}(y)(Z)} > e^{d(x,y)}$ or $\frac{\mathcal{K}(y)(Z)}{\mathcal{K}(x)(Z)} > e^{d(x,y)}$
      - With no loss of generality, let $\frac{\mathcal{K}(x)(Z)}{\mathcal{K}(y)(Z)} = r > e^{d(x,y)}$

Geo-Indistinguishability

- Knowledge of an informed attacker
  - Proof Sketch
    - Let $N = \{x, y\}$, and $\pi(x|N) < \frac{r - e^{d(x,y)}}{(r-1)e^{d(x,y)}}$, then we get the following condition:
    - $\frac{Bayes(\pi, \mathcal{K}, Z|N)}{\pi(x|N)} = \frac{\mathcal{K}(x)(Z)}{\sum_{x' \in N} \pi(x'|N)\mathcal{K}(x')(Z)} = \frac{r}{\pi(x|N)r + \pi(y|N)} > e^{d(x,y)} = e^{d(N)}$
    - The contradiction illustrates that $\mathcal{K}$ satisfies ε-gi

$r > e^{d(x,y)} > 1$

Geo-Indistinguishability

- Characterizations of Geo-Indistinguishability
  - Abstracting from side information
    - Prior distribution of locations are not involved in the definition of gi
    - Location is protected by gi under all prior instead of a specific prior
    - The above two characterizations also adopt to all prior

Geo-Indistinguishability

- Mechanism
  - Step 1: achieving ε-gi in a continuous plane
  - Step 2: achieving ε-gi in a discrete domain
  - Step 3: achieving ε-gi in a truncated region

Geo-Indistinguishability

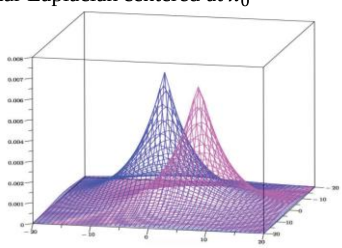
- Step 1: Achieving ε-gi in a Continuous Plane
  - Planar Laplacian centered at x_0
    - Given $\varepsilon \in \mathbb{R}^+$, and the actual location $x_0 \in \mathbb{R}^2$, the probability density function of planar Laplacian centered at x_0 , on any other point $x \in \mathbb{R}^2$, is:

$$D_\varepsilon(x_0)(x) = \frac{\varepsilon^2}{2\pi} e^{-\varepsilon d(x_0, x)}$$

    - $\frac{\varepsilon^2}{2\pi}$ is a normalization factor

Geo-Indistinguishability

- Step 1: Achieving ε-gi in a Continuous Plane
  - Planar Laplacian centered at x_0



The pdf of two planar Laplacians, centered at $(-2,-4)$ and $(5,3)$ with $\varepsilon = 1/5$

Geo-Indistinguishability

- Step 1: Achieving ε-gi in a Continuous Plane
  - Mechanism
    - Given the actual location $x_0 \in \mathbb{R}^2$, parameter $\varepsilon \in \mathbb{R}^+$, draw a random point x to achieve ε-gi according to the probability density function:

$$D_\varepsilon(x_0)(x) = \frac{\varepsilon^2}{2\pi} e^{-\varepsilon d(x_0,x)}.$$

Why does the above mechanism work?

Geo-Indistinguishability

- Step 1: Achieving ε-gi in a Continuous Plane
  - Proof of the Correctness for the Mechanism
    - For any $x, x' \in \mathcal{X}$ and $z \in \mathcal{Z}$, we have that

$$\frac{D_\varepsilon(x)(z)}{D_\varepsilon(x')(z)} = \frac{\varepsilon^2}{2\pi} e^{-\varepsilon d(x,z)} / \frac{\varepsilon^2}{2\pi} e^{-\varepsilon d(x',z)} = e^{-\varepsilon(d(x,z)-d(x',z))}$$

    - Due to triangle inequality, we have that

$$\frac{D_\varepsilon(x)(z)}{D_\varepsilon(x')(z)} = e^{-\varepsilon(d(x,z)-d(x',z))} \le e^{-\varepsilon d(x,x')}$$

    - That is to say the mechanism satisfies ε-gi

How to efficiently draw a random point?

Geo-Indistinguishability

- Step 1: Achieving ε-gi in a Continuous Plane
  - Calculating $D_\varepsilon(r,\theta)$
    - The pdf only depends on the distance from x_0
    - Switch the Cartesian system to polar coordinates

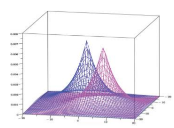
$$D_\varepsilon(x_0)(x) = \frac{\varepsilon^2}{2\pi} e^{-\varepsilon d(x_0,x)} \Longrightarrow D_\varepsilon(r,\theta) = \frac{\varepsilon^2}{2\pi} r e^{-\varepsilon r}$$

    - r is the distance of x from x_0
    - θ is the angle that the line xx_0 forms with respect to the horizontal axis of the Cartesian system

The two variables r **and** θ **are independent!**

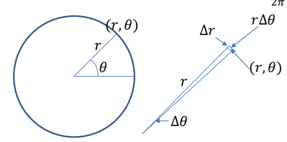
Geo-Indistinguishability

- Step 1: Achieving ε-gi in a Continuous Plane
  - Calculating $D_\varepsilon(r,\theta)$
    - A 2-d probability density function is a cap, the volume under which is 1
    - The volume at each point is decided by the pdf (f)
      - Cartesian system: $f(x,y)\Delta x\Delta y$ Volume assigned to (x,y)
      - Polar coordinates: $f(r,\theta)\Delta r\Delta\theta$ Volume assigned to (r,θ)
    - How to calculate $D_\varepsilon(r,\theta)$?
      - Calculate the volume at point (r,θ)
      - Remove terms Δr and $\Delta\theta$



Geo-Indistinguishability

- Step 1: Achieving ε-gi in a Continuous Plane
  - Calculating $D_\varepsilon(r,\theta)$
    - Calculation of the volume at point (r,θ) (take it as a bar)
    - Point (r,θ) can be taken as a rectangle with length and width as $r\Delta\theta$ and Δr approximately
    - The height of the bar is $D_\varepsilon(x_0)(x) = \frac{\varepsilon^2}{2\pi}e^{-\varepsilon d(x_0,x)} = \frac{\varepsilon^2}{2\pi}e^{-\varepsilon r}$ $x\sim(r,\theta)$
    - The volume of the bar is $r\Delta\theta \times \Delta r \times \frac{\varepsilon^2}{2\pi}e^{-\varepsilon r} = D_\varepsilon(r,\theta)\Delta\theta\Delta r$



$D_\varepsilon(r,\theta)=\frac{\varepsilon^2}{2\pi}re^{-\varepsilon r}$

Geo-Indistinguishability

- Step 1: Achieving ε-gi in a Continuous Plane
  - Draw random variables r and θ according to

$$D_\varepsilon(r,\theta) = \frac{\varepsilon^2}{2\pi}re^{-\varepsilon r}$$

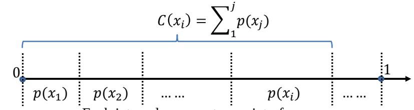
  - The two margins of r and θ are:
    - $D_{\varepsilon,R}(r) = \int_0^{2\pi} D_\varepsilon(r,\theta)d\theta = \varepsilon^2 re^{-\varepsilon r}$
    - $D_{\theta,R}(\theta) = \int_0^{\infty} D_\varepsilon(r,\theta)dr = \frac{1}{2\pi}$
  - $D_{\theta,R}(\theta)$ is constant, thus draw θ from a uniform distribution with range $[0,2\pi)$

How to draw a value of r from $D_{\varepsilon,R}(r)$?

Geo-Indistinguishability

- Step 1: Achieving ε-gi in a Continuous Plane
  - Draw a sample from a discrete distribution $p(x)$ (its cumulative distribution is $c(x)$)

$$C(x_i) = \sum_1^j p(x_j)$$

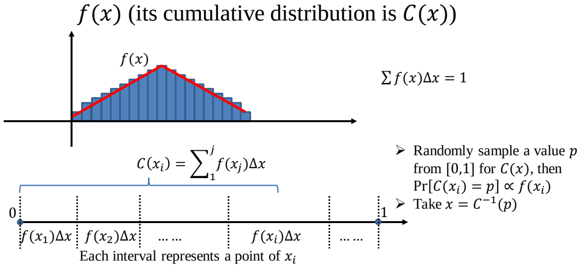


Each interval represents a point of x_i

- ➢ Randomly sample a value p from [0,1] for c(x), then $\Pr[p\ falls\ into\ the\ interval\ of\ x_i] = p(x_i)$
- ➢ Randomly draw a value p from [0,1], and take $x = \min_{c(x_i)\geq p} x_i$

Geo-Indistinguishability

- Step 1: Achieving ε-gi in a Continuous Plane
  - Draw a sample from a continuous distribution $f(x)$ (its cumulative distribution is $C(x)$)

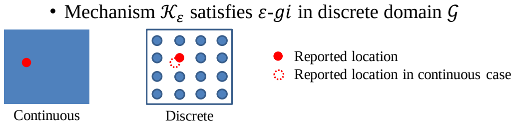


Geo-Indistinguishability

- Step 1: Achieving ε-gi in a Continuous Plane
  - The cumulative distribution function of $D_{\varepsilon,R}(r)$
    - $C_\varepsilon(r) = \int_0^r D_{\varepsilon,R}(\rho)d\rho = 1 - (1 + \varepsilon r)e^{-\varepsilon r}$
  - Draw the value of r
    - Draw a random number p with uniform probability in range $[0,1)$
    - Set $r = C_\varepsilon^{-1}(p) = -\frac{1}{\varepsilon}\left(W_{-1}\left(\frac{p-1}{e}\right) + 1\right)$
      - W_{-1} is the Lambert W function (the -1 branch)
  - Build the point x with drawn θ and r

Geo-Indistinguishability

- Step 2: Achieving ε-gi in a Discrete Domain
  - Mechanism $\mathcal{K}_\varepsilon$: given the actual location x_0, report the point x in a discrete domain $\mathcal{G}$ as follow:
    - Draw a point (r, θ) as that of Step 1 (which satisfies ε-gi in continuous plane)
    - Remap (r, θ) to the closest point x in $\mathcal{G}$.
  - Property of Mechanism $\mathcal{K}_\varepsilon$
    - Mechanism $\mathcal{K}_\varepsilon$ satisfies ε-gi in discrete domain $\mathcal{G}$



Geo-Indistinguishability

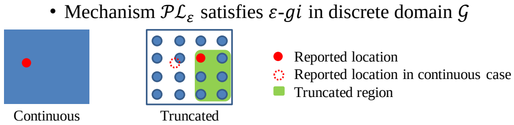
- Step 2: Achieving ε-gi in a Discrete Domain
  - Proof Sketch for the Property of Mechanism $\mathcal{K}_\varepsilon$
    - Let $R(g) = \{z \in Z | g \textit{ is the closest point to } z, g \in \mathcal{G}\}$
    - For all $x, x' \in \mathcal{X}$, all $G \subseteq \mathcal{G}$, we analyze $\frac{\mathcal{K}_\varepsilon(x)(G)}{\mathcal{K}_\varepsilon(x')(G)}$

$$\frac{\mathcal{K}_\varepsilon(x)(G)}{\mathcal{K}_\varepsilon(x')(G)} = \frac{\sum_{z \in R(g), g \in G} \mathcal{K}(x)(z)}{\sum_{z \in R(g), g \in G} \mathcal{K}(x')(z)}$$
$$\leq \frac{\sum_{z \in R(g), g \in G} e^{\varepsilon d(x,x')} \mathcal{K}(x')(z)}{\sum_{z \in R(g), g \in G} \mathcal{K}(x')(z)}$$
$$= e^{\varepsilon d(x,x')}$$

Similarly, we can proof $\frac{\mathcal{K}_\varepsilon(x')(G)}{\mathcal{K}_\varepsilon(x)(G)} \leq e^{\varepsilon d(x,x')}$

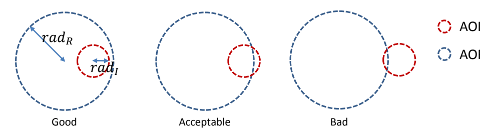
Geo-Indistinguishability

- Step 3: Achieving ε-gi in a Truncated Region
  - Mechanism $\mathcal{PL}_\varepsilon$: given the actual location x_0, report the point x in a finite discrete set of $\mathcal{G}$ as follow:
    - Draw a point (r, θ) as that of Step 1 (which satisfies ε-gi in continuous plane)
    - Remap (r, θ) to the closest point x in $\mathcal{G}$.
  - Property of Mechanism $\mathcal{PL}_\varepsilon$
    - Mechanism $\mathcal{PL}_\varepsilon$ satisfies ε-gi in discrete domain $\mathcal{G}$



Geo-Indistinguishability

- Accuracy
  - Can we get all the query results?
    - $\mathcal{B}(x, r)$ be the circle with center x and radius r
    - Area of Interest (AOI): we expect results in AOI
      - $\mathcal{B}(x, rad_I)$, x is the actual location
    - Area of Retrieval (AOR): server returns results in AOR
      - $\mathcal{B}(z, rad_R)$, z is the reported location



Geo-Indistinguishability

- Accuracy
  - Enlarging AOR to fully contain AOI may lead to privacy breach
    - The adversary is sure the true location lies in AOR
  - Enlarging AOR leads to additional bandwidth consumption
    - More searching results are returned to the mobile user
  - We should tolerate the incomplete results, and analyze the accuracy

Geo-Indistinguishability

- Accuracy
  - Abstraction of an LBS Application
    - $(\mathcal{K}, rad_R)$: $\mathcal{K}$ is a mechanism satisfying ε-gi, and rad_R is the radius of AOR
      - Given the actual location x, we report z according to $\mathcal{K}(x)$. Then the LBS server searches $\mathcal{B}(z, rad_R)$
  - Definition of LBS Application Accuracy
    - An LBS application $(\mathcal{K}, rad_R)$ is (c, rad_I)-accurate iff for all locations x we have that $\mathcal{B}(x, rad_I)$ is fully contained in $\mathcal{B}(\mathcal{K}(x), rad_R)$ with probability at least c.

Geo-Indistinguishability

$C_\varepsilon(r) = \int_0^r D_{\varepsilon,R}(\rho)d\rho = 1 - (1+\varepsilon r)e^{-\varepsilon r}$

$C_\varepsilon(r) = c$ then $C_\varepsilon^{-1}(c) = r$

- Accuracy
  - Achieving (c, rad_I)-Accurate LBS Application
    - The LBS application$(\mathcal{PL}_\varepsilon, rad_R)$ is (c, rad_I)-Accurate if $rad_R \geq rad_I + C_\varepsilon^{-1}(c)$.
  - Proof Sketch
    - Suppose the actual location is x and $\mathcal{PL}_\varepsilon$ reports z
    - $\Pr[d(x,z) \leq C_\varepsilon^{-1}(c)] = c$
    - $\Pr[rad_I + d(x,z) \leq rad_I + C_\varepsilon^{-1}(c)] = c$
    - It suffices to set $rad_R = rad_I + C_\varepsilon^{-1}(c)$ for achieving (c, rad_I)-Accurate

Geo-Indistinguishability

- Achieving Geo-Indistinguishability with Optimal Utility [CCS 2014]
  - Geo-indistinguishability provides guaranteed privacy for locations

Can you find an alternative way to satisfy gi and optimize utility at the same time?

  - How to measure the utility?
  - The standard Planar Laplace Mechanism provides no optimization towards utility

Geo-Indistinguishability

- Achieving gi with Optimal Utility [CCS 2014]
  - How to measure the service quality in LBS
    - Service quality could be measured by the actual location x and the reported location z
    - If x is close to z, users could get good service
    - If x is far away from z, users could not get service around x at all
    - Good service quality means good utility

Geo-Indistinguishability

- Achieving gi with Optimal Utility [CCS 2014]
  - Quality metric and privacy metric
    - Quality metric $d_Q: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a user specified distance function of locations
    - Privacy metric $d_{\mathcal{X}}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a user specified distance function of locations
    - d_Q and $d_{\mathcal{X}}$ could be initialized using Euclidean distance

Geo-Indistinguishability

- Achieving gi with Optimal Utility [CCS 2014]
  - Definition of $\varepsilon d_{\mathcal{X}}$-private
    - Given a location set $\mathcal{X}$, a privacy parameter ε and a privacy metric $d_{\mathcal{X}}$, for all $x, x', z \in \mathcal{X}$, a mechanism $\mathcal{K}$ is $\varepsilon d_{\mathcal{X}}$-private if and only if

$$\mathcal{K}(x)(z) \le e^{\varepsilon d_{\mathcal{X}}(x,x')}\mathcal{K}(x')(z).$$

When $d_{\mathcal{X}}$ is restricted to the Euclidean distance, $\varepsilon d_{\mathcal{X}}$-private reduces to ε-gi

Geo-Indistinguishability

- Achieving gi with Optimal Utility [CCS 2014]
  - Service quality of a mechanism
    - Given a prior π, a location set $\mathcal{X}$, a mechanism $\mathcal{K}$ and a quality metric d_Q, the service quality of $\mathcal{K}$ is defined as follow:

$$QL(\mathcal{K}, \pi, d_Q) = \boxed{\sum_{x,z\in\mathcal{X}} \pi(x)\,\mathcal{K}(x)(z) d_Q(x,z).}$$

Expected distance in term of d_Q between the input and output of $\mathcal{K}$

- A large $QL(\mathcal{K}, \pi, d_Q)$ indicates poor quality
- The minimized $QL(\mathcal{K}, \pi, d_Q)$ indicates the optimal quality

Geo-Indistinguishability

- Achieving gi with Optimal Utility [CCS 2014]
  - Problem Definition
    - Given a prior π, a privacy metric $d_{\mathcal{X}}$, a privacy parameter ε and a quality metric d_Q, compute a mechanism $\mathcal{K}$ such that:
      - $\mathcal{K}$ is $\varepsilon d_{\mathcal{X}}$-private
      - For all $\varepsilon d_{\mathcal{X}}$-private mechanism $\mathcal{K}'$, $QL(\mathcal{K}, \pi, d_Q) \le QL(\mathcal{K}', \pi, d_Q)$

Geo-Indistinguishability

- Achieving gi with Optimal Utility [CCS 2014]
  - Solution
    - *Minimize*: $\sum_{x,z\in\mathcal{X}} \pi(x)\mathcal{K}(x)(z) d_Q(x,z)$
    - *Sub to*:

$$\begin{array}{ll} \mathcal{K}(x)(z) \le e^{\varepsilon d(x,x')}\mathcal{K}(x')(z) & x, x', z \in \mathcal{X} \\ \sum_{z\in\mathcal{X}} \mathcal{K}(x)(z) = 1 & x \in \mathcal{X} \\ \mathcal{K}(x)(z) \ge 0 & x, z \in \mathcal{X} \end{array}$$

Given $\pi(x)$ and $d_Q(x,z)$, solve the above LP for $\mathcal{K}(x)(z)$

Hierarchical Location Publishing

- Motivation
  - Location Datasets Are Valuable
    - Travel Pattern Mining
    - Traffic Analysis
  - Release The Original Datasets? No!
    - Re-identifying of Users and Their Sensitive Information

User	Location
u_1	$< x_{11}, y_{11} >, < x_{12}, y_{12} >, \dots$
u_2	$< x_{21}, y_{21} >, < x_{22}, y_{22} >, \dots$
$u_{\lvert U \rvert}$	$< x_{\lvert U \rvert 1}, y_{11} >, < x_{\lvert U \rvert 2}, y_{12} >, \dots$

Bob has visited locations including $< x_{21}, y_{21} >$ and $< x_{22}, y_{22} >$

u_2 is very likely to be Bob!

Hierarchical Location Publishing

- Location Dataset Representation
  - In a location dataset D, the information of a user u is presented by a profile
    - $P_u =< T(u), W(u) >$
    - $T(u)$ is the set of all locations in D
    - $W(u)$ is the weight vector representing the frequency distribution on $T(u)$

$U = \{u_1, u_2, u_3\}$
$T(u) = \{l_1, l_2, l_3, l_4\}$

$W(u_1) = \{3,4,0,0\}$
$W(u_1) = \{4,4,1,0\}$
$W(u_1) = \{0,0,4,3\}$

User\Location	l_1	l_2	l_3	l_4
u_1	3	4	0	0
u_2	4	4	1	0
u_3	0	0	4	3

Hierarchical Location Publishing

- Private Location Release
  - "Problem Definition": Private Location Release aims to publish all users' profiles by masking exact locations and weights under the notion of differential privacy

User\Location	l_1	l_2	l_3	l_4
u_1	3	4	0	0
u_2	4	4	1	0
u_3	0	0	4	3

D

➡

User\Location	l_1	l_2	l_3	l_4
u_1	2	5	1	1
u_2	3	4	0	1
u_3	1	1	2	5

Released Noisy D

Hierarchical Location Publishing

- Private Location Release
  - Naïve Solution
    - Add randomized noise to $W(u)$ with standard differential privacy, and get $\widehat{W}(u)$
      - $\Delta = 1$
      - Add $Lap(\frac{1}{\varepsilon})$ on each dimension of $W(u)$
    - Release noisy profile $\widehat{P_u} =< T(u), \widehat{W}(u) >$
  - Disadvantages
    - $|T(u)|$ is large and $W(u)$ is a sparse vector
    - $\widehat{W}(u)$ will contain a large amount of noise
      - $0 \rightarrow positive\ number$

Hierarchical Location Publishing

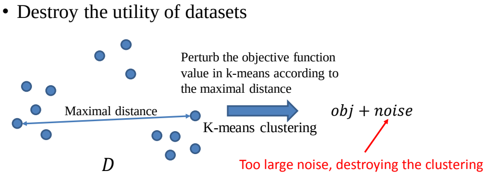
- *PriLocation* Algorithm [KAIS 2016]
  - [Step 1] Private Location Cluster
    - Group all locations into η clusters
      - Mask exact number of locations as well as the center of each cluster
      - From the cluster outputs, the adversary could not infer to which cluster a location exactly belongs
      - Aims to reduce the amount of noise added to profile

Hierarchical Location Publishing

- *PriLocation* Algorithm [KAIS 2016]
  - [Step 2] Cluster Weight Perturbation
    - Perturb the weight of each cluster with Laplace noise
      - Mask the weights of locations in a user's profile
      - Prevent the adversary from inferring how many locations a user has visited in a certain cluster
  - [Step 3] Private Location Selection
    - Select new locations for original ones
      - Aims to Mask a user's profile
      - Prevent the adversary from inferring the locations visited by a user

Hierarchical Location Publishing

- Sensitivity for Location Dataset
  - The standard sensitivity of a query is calibrated by the maximal distance between locations
    - Mask the true distance
    - Destroy the utility of datasets



Hierarchical Location Publishing

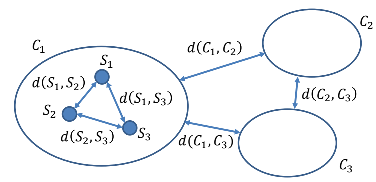
- Sensitivity for Location Dataset
  - Location datasets have inherent hierarchy
    - Different semantics on each level
    - For instance, $country \rightarrow city \rightarrow street$
    - Users may have different level of privacy requirement
      - Hide the street, hide the city or even hide the country

Hierarchical Location Publishing

- Hierarchical Sensitivity
  - For a given level L, the hierarchical sensitivity of L is $HS_L = \max_{t_i, t_j \in L} d(t_i, t_j)$, where $d(t_i, t_j)$ represents the distance between t_i and t_j.
  - Privacy on city level
    - Sensitivity is measured by the maximal distance between cities
    - Hide the city rather than the country for a user

Hierarchical Location Publishing

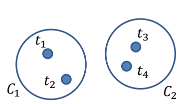
- Hierarchical Sensitivity
  - Sensitivity on different levels
    - $HS_{city} = \max\{d(C_1, C_2), d(C_2, C_3), d(C_1, C_3)\}$
    - $HS_{street} = \max\{d(S_1, S_2), d(S_2, S_3), d(S_1, S_3)\}$



Hierarchical Location Publishing

- Private Location Cluster
  - Create location clusters
    - $< T(u), W(u) > \rightarrow < T_C(u), W_C(u) >$
  - Private clustering algorithm based on k-means
  - Distance measure for locations

$$d(t_i, t_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$



$T(u) = \{t_1, t_2, t_3, t_4\}$
$W(u) = < 3,4,0,0 >$

$T_C(u) = \{c_1, c_2\}$
$W_C(u) = < 7,0 >$

Reduce the number of "0"

Hierarchical Location Publishing

- Private Location Cluster
  - Differentially private k-means
    - Initialize k clustering centers
    - Run in iterations
      - Assign each location to its closest cluster
      - Calculate the clustering objective function
      - Add noise to the clustering objective function value
      - If a smaller noise value is obtained, update the clustering
    - Output the clustering

Hierarchical Location Publishing

- Private Location Cluster
  - Objective function in each iteration of k-Means
    - Let c_l denote the center of cluster C_l, T is the set of location, and η is the number of cluster
    - Objective function g measures the total distance between the location and the cluster center it belongs to

$$g = \sum_{i=1}^{|T|} \sum_{l=1}^{\eta} \gamma_{il}\, d(t_i, c_l)$$

    - γ_{il} is an indicator that

$$\gamma_{il} = \begin{cases} 1 & t_i \in c_l \\ 0 & t_i \notin c_l \end{cases}$$

Hierarchical Location Publishing

- Private Location Cluster
  - Introduce Differential Privacy into k-Means
    - Laplace noise calibrated by hierarchical sensitivity HS of the objective function G and the privacy budget
    - Private location cluster consumes $\varepsilon/2$ privacy budget
    - Each iteration costs $\varepsilon/2p$ privacy budget
    - $\hat{G} = \sum_{i=1}^{m} \sum_{l=1}^{\eta} \gamma_{il}\, d(t_i, c_l) + Lap(\frac{2p \times HS}{\varepsilon})$
    - After p iterations, private location clustering outputs

$$\hat{C} = \{C_1, \dots, C_\eta\}$$

Hierarchical Location Publishing

- Cluster Weight Perturbation
  - After private location clustering, user u's weight in cluster C is denoted $W_c(u) = \sum_{t \in C} W_t(u)$
  - For each user u, Laplace noise is added to mask the counts of locations in each cluster

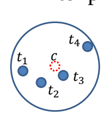
$$\widehat{W_c}(u) = W_c(u) + (Lap\left(\frac{4}{\varepsilon}\right))^\eta$$

  - Cluster weight perturbation consumes $\varepsilon/4$ privacy budget

Cluster Weight Perturbation

- Cluster Weight Perturbation
  - The added Laplace noise could either positive or negative
    - For positive noise, add locations close to the center of the cluster
    - For negative noise, delete locations with largest distance to the cluster center
  - After perturbation, we get each user u's profile as

$$\widetilde{P_C}(u) =< \widetilde{P_C}(u), \widehat{W_c}(u) >$$



$W(u) =< 2,3,4,1 > \to < 2,3,5,1 >$ $\quad noise\ 1\ added$
$W(u) =< 2,3,4,1 > \to < 2,3,4,0 >$ $\quad noise\ -1\ added$
$W(u) =< 2,3,4,1 > \to < 1,3,4,0 >$ $\quad noise\ -2\ added$

Hierarchical Location Publishing

- Private Location Selection
  - $\widetilde{P_C}(u) =< \widetilde{T_C}(u), \widehat{W_c}(u) >$ has the high probability to be re-identified since $\widetilde{P_C}(u)$ contains a major part of original locations
  - Private Location Selection replaces original locations with selected new locations

Replace t_1 with t_2

$\widehat{W_c}(u) =< 3,0,0,0 > \rightarrow < 0,3,0,0 >$

$\widehat{W_c}(u) =< 1,2,1,1 > \rightarrow < 0,3,1,1 >$

$\widehat{W_c}(u) =< 0,2,1,1 > \rightarrow < 0,2,1,1 >$

Hierarchical Location Publishing

- Private Location Selection
  - How to select a location to replace $t \in C_l$?
    - Uniformly selecting a location $t' \in C_l$ ← Poor utility
    - Selecting the most similar t' with t ← Poor privacy
  - Considerations on Selecting a New location to replace $t \in C_l$?
    - Retain utility of locations
    - Mask the similarity between locations

Hierarchical Location Publishing

- Private Location Selection
  - Exponential Mechanism based Selection
    - For a location $t \in C_l$, the candidate set $I = \widetilde{T_C}(u)$
    - Score function is defined based on distance
      $q_i(I, t_j) = HS - d(t_i, t_j)$
    - The sensitivity for score function is measured by the maximal change in distance between t_i and t_j
      $\Delta q_i = HS$

Hierarchical Location Publishing

- Private Location Selection
  - Exponential Mechanism based Selection
    - Private location selection consumes $\varepsilon/4$ privacy budget
    - The probability arranged to each location t_j is

$$\Pr(t_j) = \frac{\exp(\frac{\varepsilon \times q_i(I,t_j)}{8 \times HS})}{\sum_{t_k \in I} \exp(\frac{\varepsilon \times q_i(I,t_k)}{8 \times HS})}$$

    - For each C_l, replace locations in C_l and output $\widehat{T_{C_l}}(u)$

Hierarchical Location Publishing

- Utility Analysis
  - $Distance\ Error$: the distance between P_u and $\widehat{P_u}$ which measures for user u (t is replaced by $\hat{t}$)

$$DE_u = \frac{\sum_{\hat{t} \in \widehat{T_C}(u)} d(t, \hat{t})}{HS \times |\widehat{T_C}(u)|}$$

  - For the entire dataset, Average $Distance\ Error$ is defined as

$$DE = \frac{1}{|U|} \sum_{u \in U} DE_u$$

Hierarchical Location Publishing

- Utility Analysis
  - For any user $u \in U$, for all $\delta > 0$, with probability at least $1 - \beta$, the distance error of the released dataset is less than α, where

$$\alpha = \max_{u \in U} \frac{\sum_{t_i \in \widehat{T_C}(u), t_j \in C_{t_i}} E[d(t_i, t_j)]}{HS \times |\widehat{T_C}(u)| \times \beta}.$$

Hierarchical Location Publishing

- Utility Analysis
  - Proof Sketch
    - According to Markov inequality we have

$$\Pr(DE_u > \alpha) \leq \frac{E[DE_u]}{\alpha}$$

    - That is to say

$$\Pr(DE_u \leq \alpha) \leq 1 - \frac{E[DE_u]}{\alpha}$$

    - Let $\beta = \frac{E[DE_u]}{\alpha}$, then $\alpha = \frac{E[DE_u]}{\beta}$
    - We can get $E[DE_u] = \frac{\sum_{t_i \in \widehat{T_C}(u), t_j \in C_{t_i}} E[d(t_i, t_j)]}{HS \times |\widehat{T_C}(u)|}$

Markov Inequality: $\Pr(X > \alpha) \leq \frac{E[X]}{\alpha}$

Hierarchical Location Publishing

- Utility Analysis
  - Proof Sketch
    - Thus for user u, the value of α should be

$$\alpha = \frac{\sum_{t_i \in \widehat{T_C}(u), t_j \in C_{t_i}} E[d(t_i, t_j)]}{HS \times |\widehat{T_C}(u)| \times \beta}$$

    - By traversing all the users

$$\alpha = \max_{u \in U} \frac{\sum_{t_i \in \widehat{T_C}(u), t_j \in C_{t_i}} E[d(t_i, t_j)]}{HS \times |\widehat{T_C}(u)| \times \beta}$$

Markov Inequality: $\Pr(X > \alpha) \leq \frac{E[X]}{\alpha}$

Hierarchical Location Publishing

- Privacy Analysis

Operations	Privacy Budget
[Step 1] Private Location Cluster	$\varepsilon/2$
[Step 2] Cluster Weight Perturbation	$\varepsilon/4$
[Step 3] Private Location Selection	$\varepsilon/4$

  – According to sequential composition theorem, $PriLocation$ algorithm satisfies ε-dp