Trường ĐH Bách Khoa Tp.HCM
Khoa Khoa Học và Kỹ Thuật Máy Tính

PPDM: Geospatial privacy-preserving data mining of social media

GV: PGS.TS ĐẶNG TRẦN KHÁNH

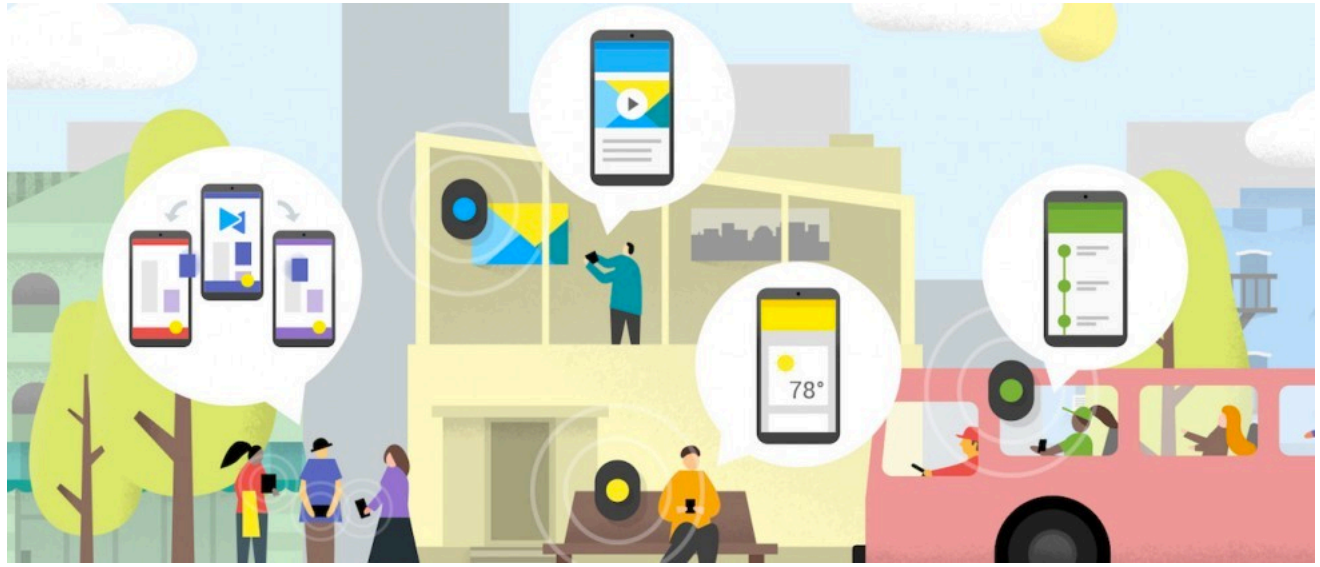
Chu Xuân Tình -1870583
Nguyễn Đức Huy -1870567

Outline

- Introduction
- Differentially Tree Spatial Decomposition
- DBSCAN
- Conclusion

Introduction

- What is Location Privacy
- Basic Techniques
  - Private Information Retrieval
  - Probabilistic Approach
    - Stationary
    - Temporal



What is Location Privacy

- Location Based Services(LBS)
  - Google, zalo, facebook, Instagram, Twitter, Yelp
  - Restaurant check-in, finding the nearest gas station, navigation, tourist city guide, ...
- Location Sharing
  - Find Friends, Find my iphone, ...
- Risks?
  - Give your location data for the service
  - Give your location traces to Google, Apple or other Service providers
  - Enable malicious apps to know your locations
  - Locations may be leaked to other attackers through network

Features of Location Privacy

- Vs Standard Differential Privacy
  - Differential Privacy: the outputs are similar whether a user opts in or out
  - For Location base services, only one user
- Data Type
  - Standard Differential Privacy: tuples in Database
  - Location Privacy: place, user location
- Location data is only two-dimensional
  - Or at most three-dimensional

Techniques

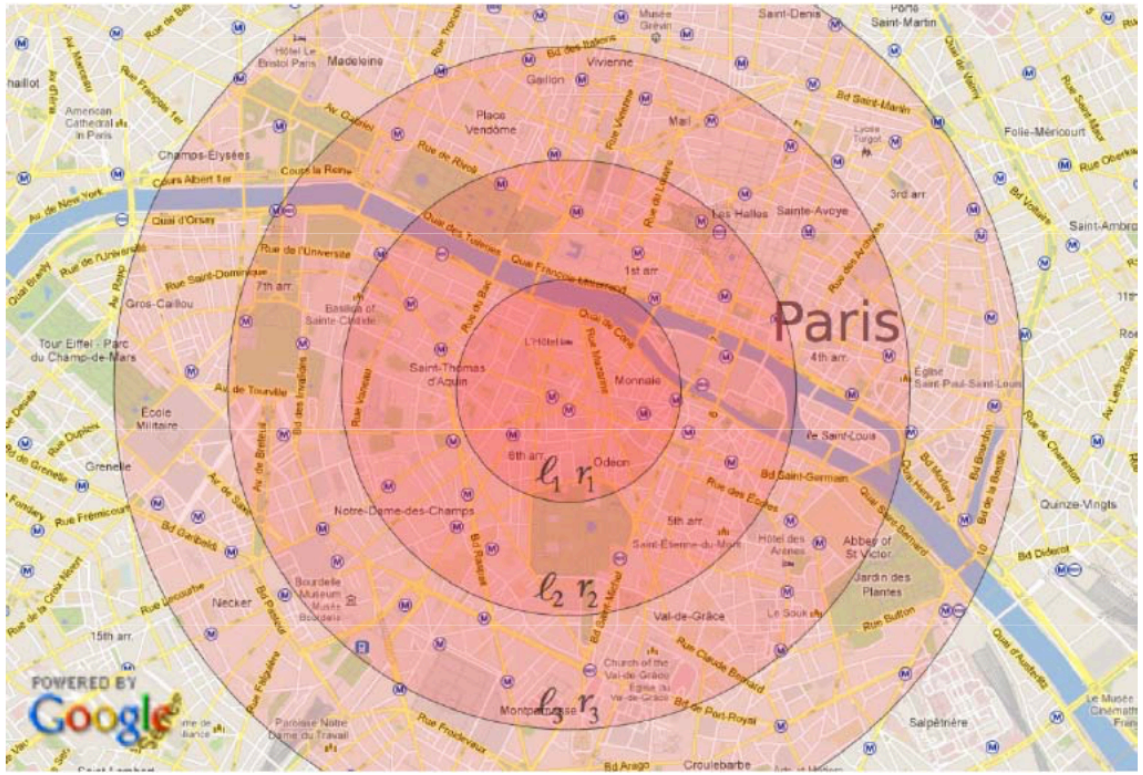
- Encryption-based Techniques
  - ✓ Private Information Retrieval Techniques
- Probabilistic Techniques
  - ✓ Location obfuscation, location cloaking
  - ✓ Location generalization

Probabilistic Techniques

- We develop a mechanism to achieve **geo-indistinguishability** by **perturbating** the user's location x.
- The inspiration comes from one of the most popular approaches for **differential privacy**, **namely the Laplacian noise**.
- We adopt a specific planar version of the Laplace distribution, allowing to draw points in a *geo-indistinguishable* way

Probabilistic Techniques

- Spatial Cloaking/Location Generalization
  - – Instead of sending the exact location to the service providers, a user can send a “general area”.



Probabilistic Techniques

❖ Location Obfuscation

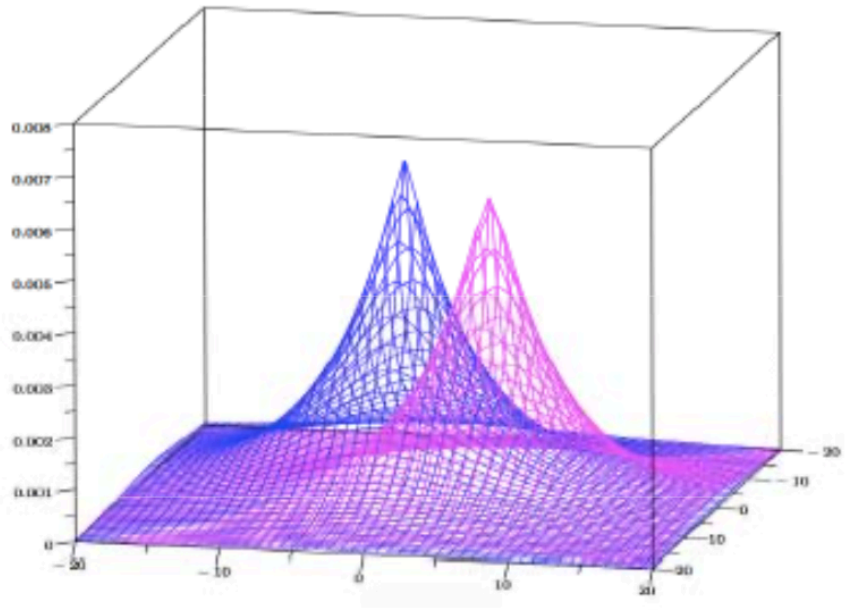
- Instead of sending the exact location to the service providers, a user can send a "noisy" location.
- Essentially, similar to spatial cloaking.
  - With the "general area", a point can be randomly chosen to represent the "noisy" location.
  - The posterior probability of the "noisy" location will be the same as the "general area". Can you prove it?

Probabilistic Techniques

- Privacy Guarantee
  - Uniform distribution in a circle
  - Uniform distribution in a polygon
  - Laplace distribution
  - Other distributions: 2D Gaussian distribution
- The trade-off between utility and privacy
  - What is the expected distance between the noisy location and the real location?
  - How much extra information does the noisy location give to attackers?
  - Can you derive the above distance function and the privacy function?

Geo-indistinguishability

- Geo-indistinguishability
  - ✓ A “differentially private” cloaking method
  - ✓ Based on the 2D Laplace distribution
  - ✓ Randomly draw a point from the distribution



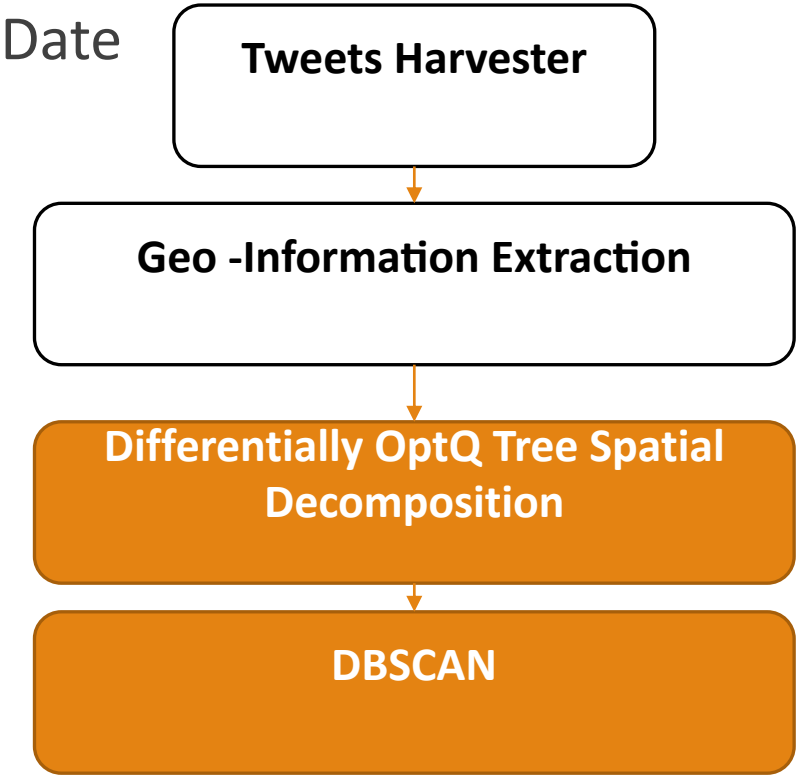
Geo-indistinguishability

- Definition
  - ✓ Pr(z|x)<=e^{epsilon}*Pr(z|x')
  - ✓ Where x and x' are any two locations in a circle with a radius r, z is the noisy location
- Features
  - ✓ Location data: x and x' are two points on a map
  - ✓ Neighboring databases: any points in the circle
  - ✓ Protection: indistinguishability in the circle

Example: Traffic information alerting.

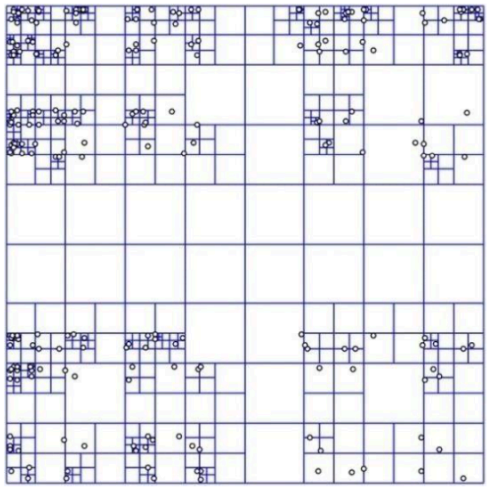
Input: UserId| PointID| Longitude|Latitude|Date

Output: Traffic density

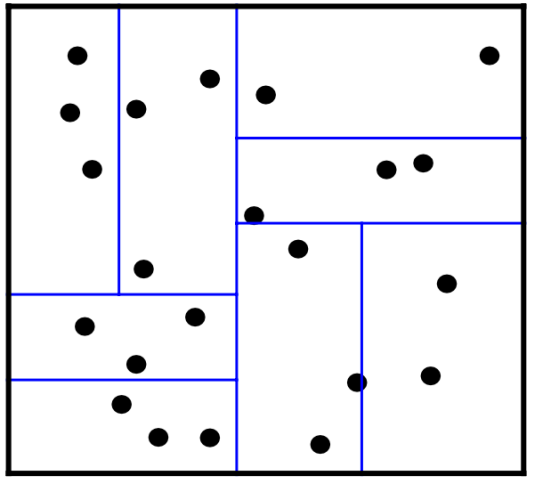


Partitioning strategy

- A more representative example of data-independent tree partitioning for two-dimensional data is **quad-tree**
- Build: partitioning with differential privacy
- Release: a private description of data distribution (in the form of bounding boxes and noisy counts)



quadtree



kd-tree

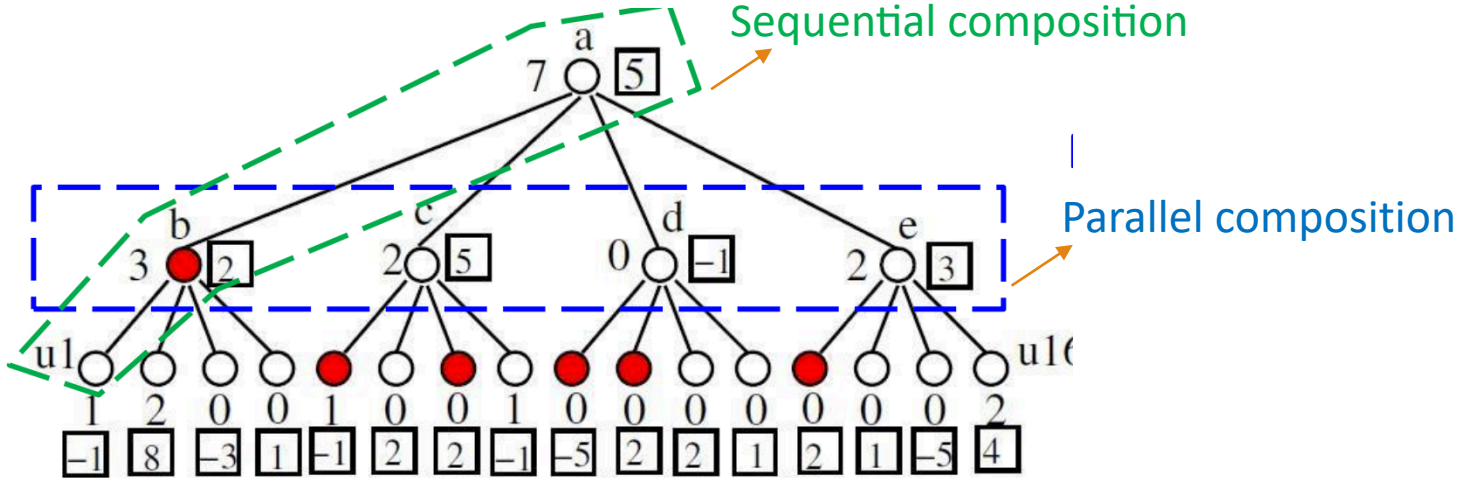
Building a Private quad-tree

- Process to build a private **quad-tree**
  - Input: maximum height h, minimum leaf size L, data set
  - Choose dimension to split (*k*-dimensional dataset)
  - Get (private) median in this dimension (*m* as the pivot to divide the dataset in the dimension)
  - Create child nodes and add noise to the counts
  - Recurse until:
    - Max height is reached
    - Noisy count of this node less than L
    - Budget along the root-leaf path has used up
- The entire PSD satisfies DP by the composition property

Building a Private quad-tree

Building PSDs – privacy budget allocation

- ❖ Budget is split between medians and counts at each node – Tradeoff accuracy of division with accuracy of counts
- ❖ Budget is split across levels of the tree
- – Privacy budget used along any root-leaf path should total
- – Optimal budget allocation
- – Post processing with consistency check

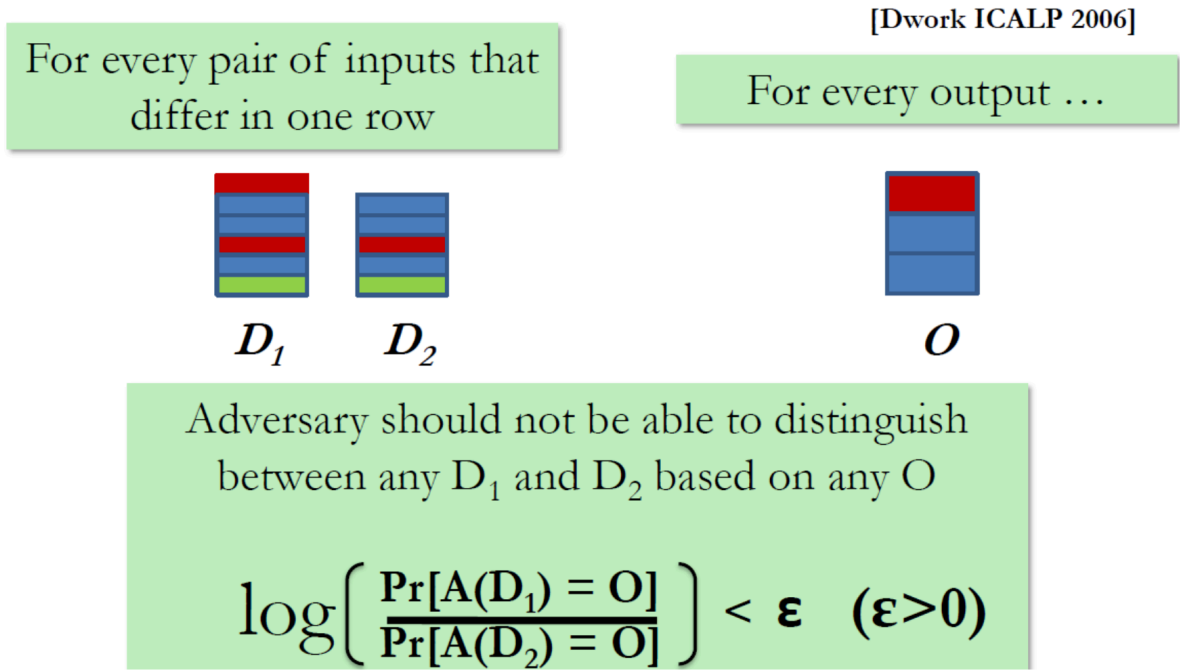


Building a Private quad-tree

- **Focus on the construction of quad-tree and the process of combining with DP.** The algorithm that uses DP based on quad-tree is called quad-tree -standard
- The privacy budget of quad-tree -standard is divided into two parts:
  - First is to determine the median, because if the differential process is not used to protect the segmentation process, the segmentation line may leak the true value of the median.
  - Second, privacy budget is used to add Laplace noise to each level of the quad-tree

Differential Privacy

- Differential privacy was introduced by Dwork et al. (2006). It ensures that useful information can be inquired and mined from a statistical database comprised of individually identifying information, while protecting a given individual's privacy.



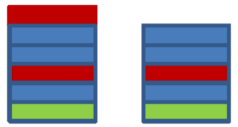
Differential Privacy

- We will introduce the definition of DP and partitioning in detail.
- Now that DP is defined on neighbor datasets, and it is necessary for us to show the definition of neighboring datasets
- Definition 1: Neighboring dataset

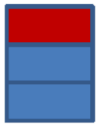
[Dwork ICALP 2006]

For every pair of inputs that differ in one row

For every output …



D_1 D_2



O

Adversary should not be able to distinguish between any D_1 and D_2 based on any O

$$\log\left(\frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]}\right) < \varepsilon \quad (\varepsilon>0)$$

Differential Privacy

Definition 2: ε-DP

- A randomized algorithm B gives ε-DP, for any pair of neighboring datasets $D1$ and $D2$, and for every set of outcomes O (O ∈ Range(B)), B satisfies

$$\Pr[B(D_1) = O] \leq e^{\varepsilon} \cdot \Pr[B(D_2) = O] \qquad (1)$$

- We call ε in formula (1) as privacy budget which represents the level of the privacy protection level. The smaller the ε, the higher the protection level.
- Achieving DP generally adopts two mechanisms: the Laplace mechanism and the exponential mechanism. These mechanisms contain the definition of sensitivity

Differential Privacy

Definition 3: Global sensitivity

- For a function F: D→R^d, for any neighboring datasets, the global sensitivity ΔF of function *F* is defined as

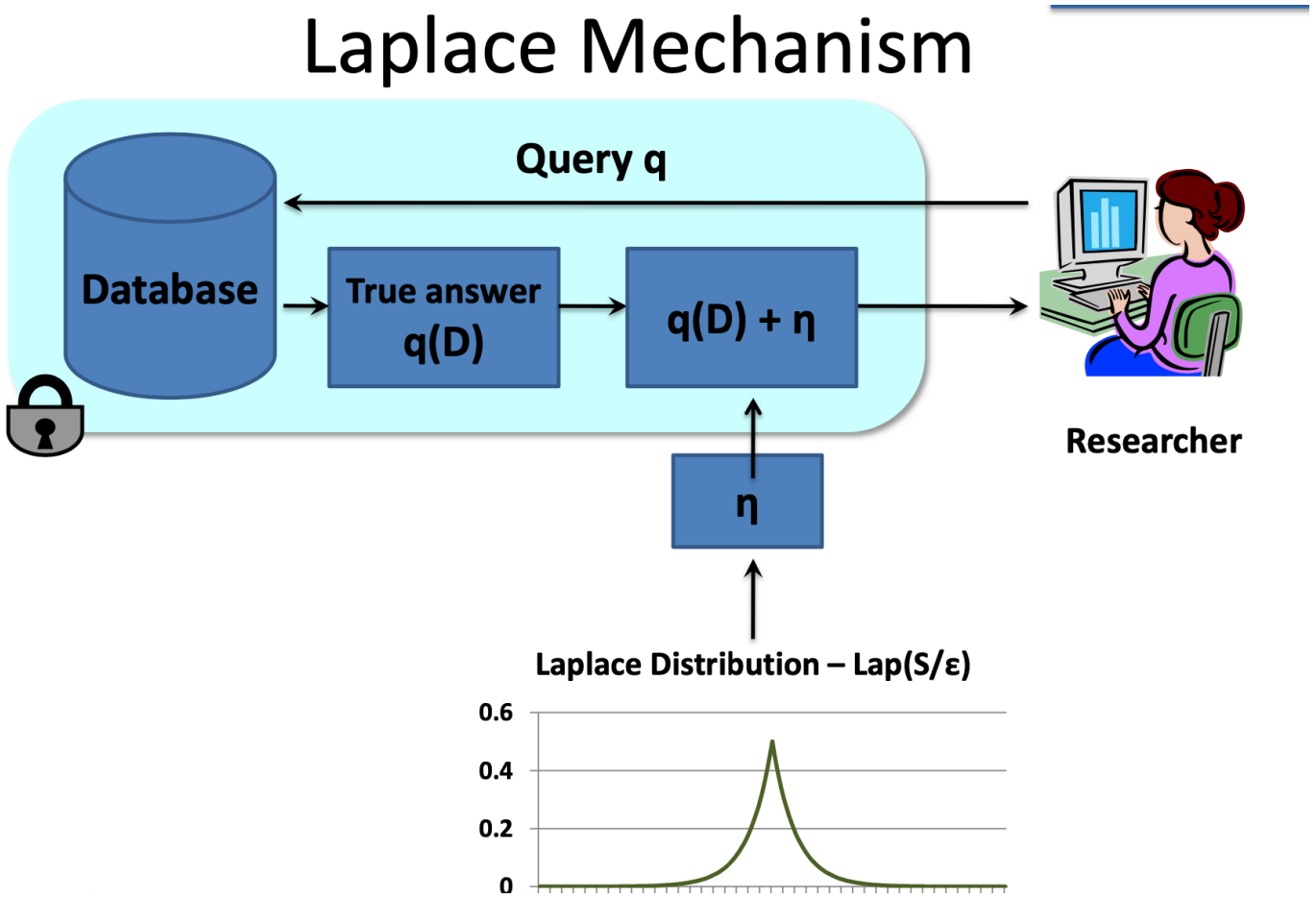
$$\Delta F = \max_{D_1\ D_2} ||F(D_1) - F(D_2)||_1 \qquad (2)$$

where *D* refers to the dataset, *d*-dimensional vector, *d* is a positive integer,

$||F(D1)-F(D2)||1$ represents the first-order distance between $F(D1)$ and $F(D2)$

- Global sensitivity represents the change in the output of the algorithm when changing any record in the dataset.

Differential Privacy



Differential Privacy

Laplace Mechanism

❖There are two common mechanisms to complete the DP model, namely, the Laplace mechanism and the exponential mechanism. They mainly add noise to statistical data. There is also a significant difference between them.

❖ The Laplace mechanism mainly adds noise to the numeric query result to complete. The exponential mechanism has its own scoring function and finally publish the data according to the level of the score. We give their definitions below respectively.

❖Why use Laplace ?

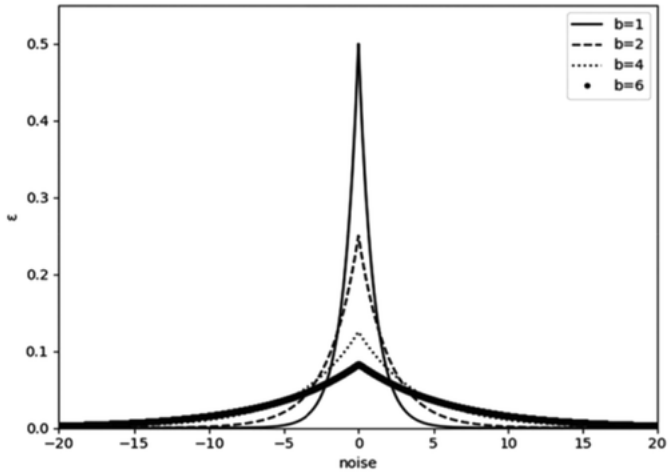
- ➢ When the possible locations of the users are modeled by a continuous region, the Laplacian was formally proved to provide the minimal noise required to satisfy geo-indistinguishability on this region
- ➢The generation of Laplace noise can be done efficiently and at low-cost using an analytic expression, so the mechanism can be implemented easily even in a computationally limited device such as a smart phone

Laplace Mechanism

- Given a function $f:D\rightarrow R^d$ over a dataset *D*, a privacy budget ε, and the global sensitivity ΔF, *F* satisfies Laplace mechanism when

$$A(D) = F(D) + \mathrm{Lap}\left(\frac{\Delta F}{\varepsilon}\right) \tag{3}$$

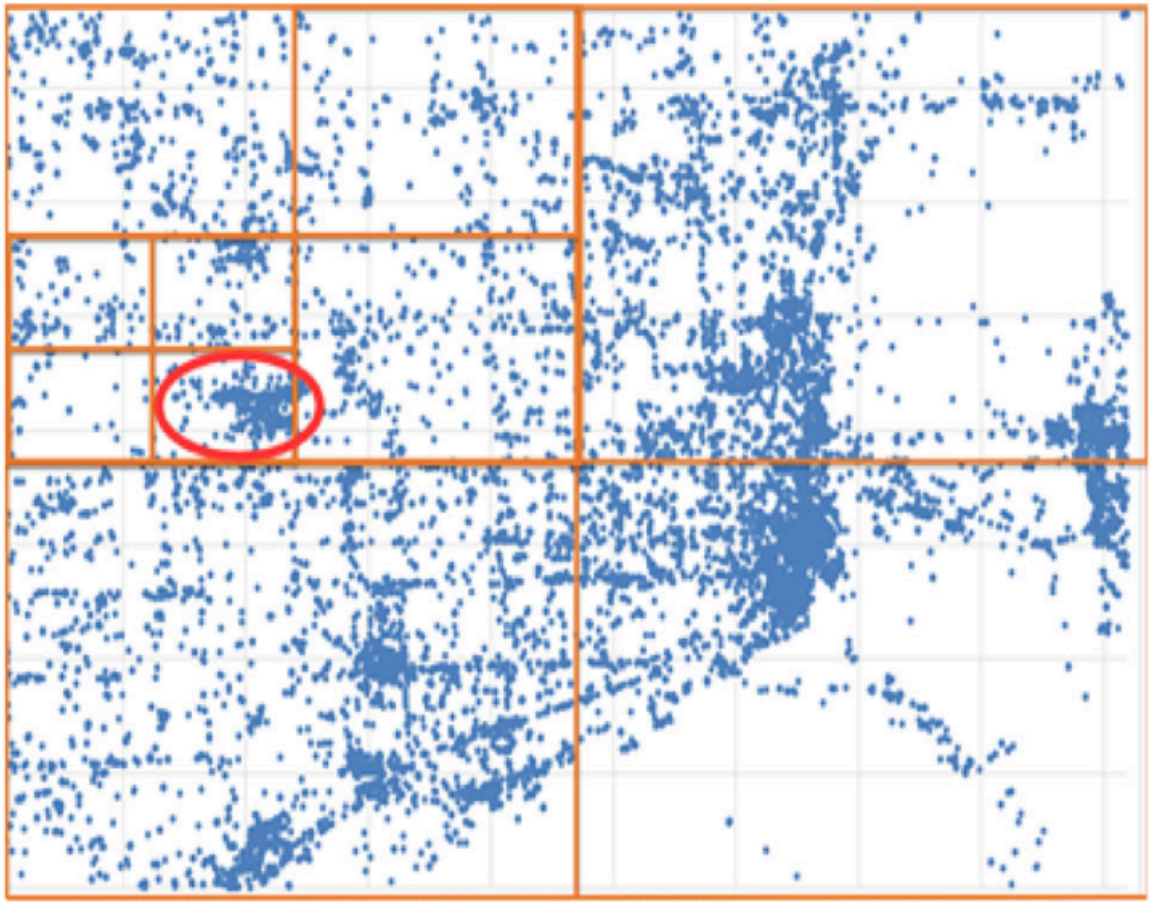
where Lap(λ) means the position parameter of Laplace distribution is 0, and the scale parameter is λ, same as $\Delta F/\varepsilon$.



Laplace distribution function with different value *b*.

Differential privacy-based spatial decomposition

- Differential privacy spatial decomposition can be divided into adding noise to counts and index structures satisfying differential privacy.
- A **quad-tree**- based spatial decomposition was adopted here to create sets of locations that group points within a certain area from the leaf of the quad-tree.
- Perturb the count of the sub-regions to protect the differential privacy of the count query outputs



Differential privacy-based spatial decomposition

Algorithm 1. Optimal quad-tree spatial decomposition with differential privacy (OptQ-SDDP)

Variables: $P = \{\}$; Sp $= \{\}$; $H = 8$; $T = 3L$

OptQ -SDDP (S, R, T)

0: Obtain ε_i according to geometric privacy budget strategy

1: CountWithNoise $= |S| + \text{Lap}(\Delta f/\varepsilon_i)$;

2: **if** $h > 8$ then

3: $P = P \cup \{R\}$; Sp $=$ Sp$\cup\{S\}$;

4: return

5: **else if** CountWithNoise $< L$ then

6: $P = P \cup \{R\}$; Sp $=$ Sp$\cup\{S\}$;

7: return

8: **else**

9: Split spatial region R into 4 equal quadrants

10: OptQ -SDDP (S{q1}; Rn{q1}; T);

11: OptQ -SDDP (S{q2}; Rn{q2}; T);

12: OptQ -SDDP (S{q3}; Rn{q3}; T);

13: OptQ -SDDP (S{q4}; Rn{q4}; T);

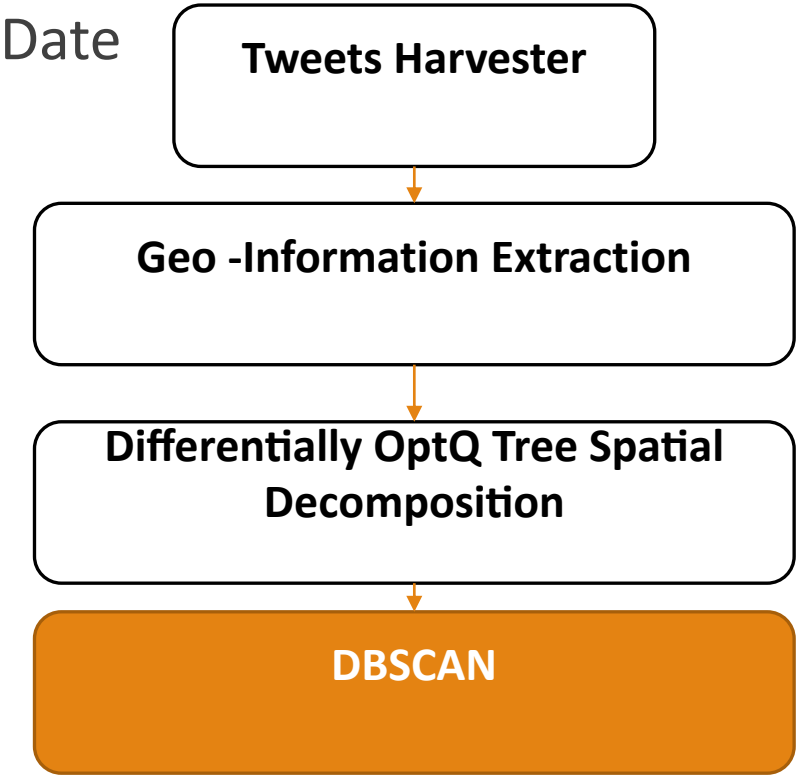
14: **end if**

15: **return**

Example: Traffic information alerting.

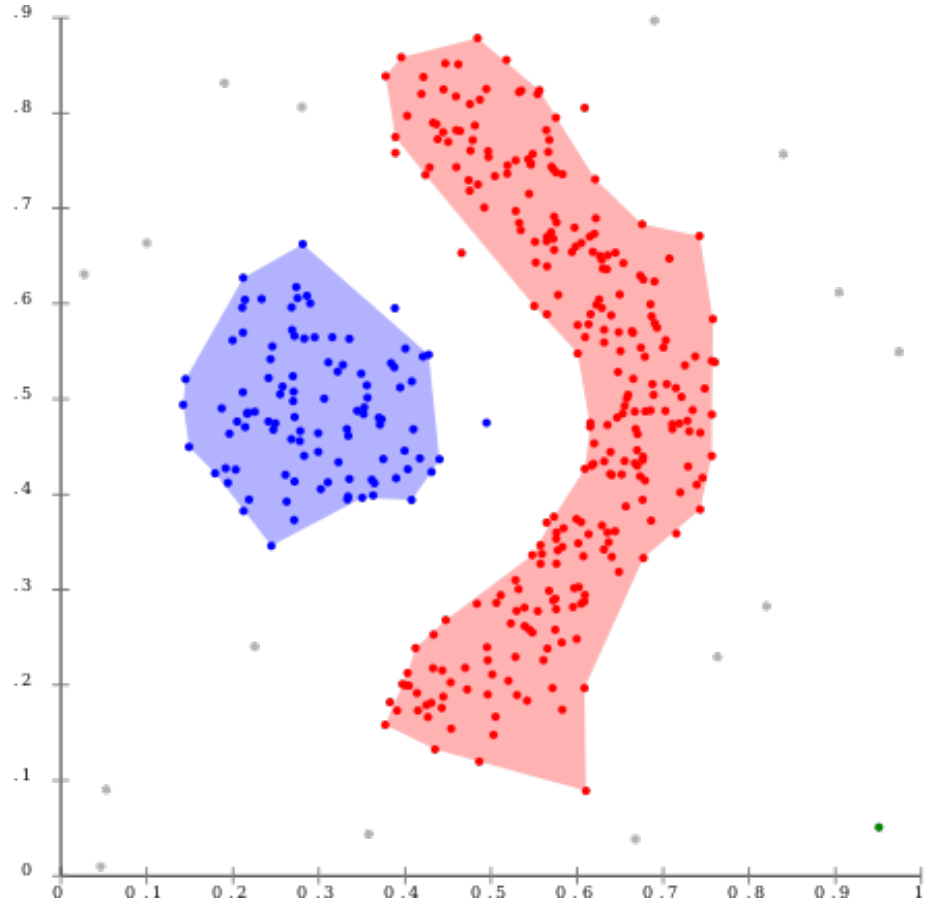
Input: UserId| PointID| Longitude|Latitude|Date

Output: Traffic density



DBSCAN Algorithm

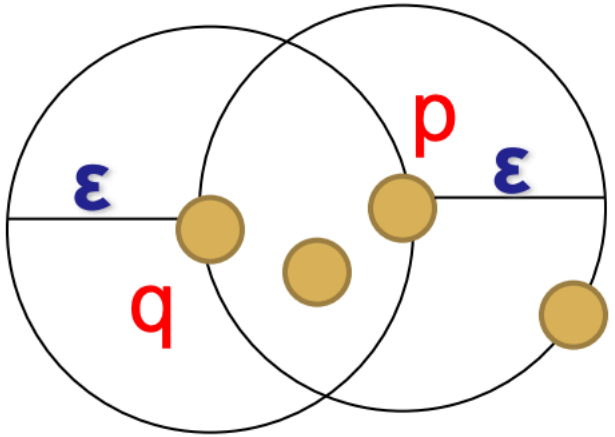
- Density based clustering



DBSCAN Algorithm

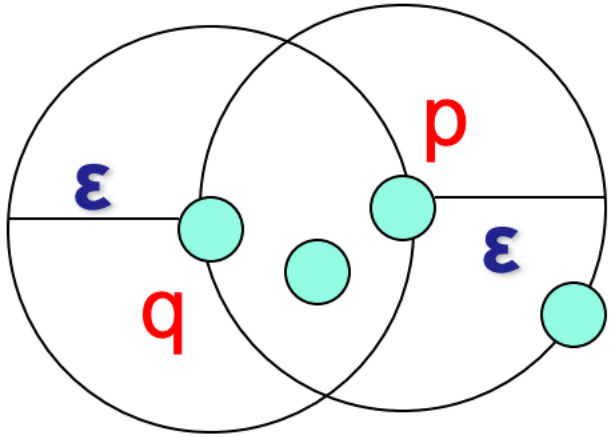
- Parameters:
  - ε: the radius of a neighborhood with respect to some point, called ε-neighborhood.
  - MinPts: the min number of points that are required in ε-neighborhood of a point.

=> A point *p* is a *core point* if at least minPts points are within distance *ε* of it.



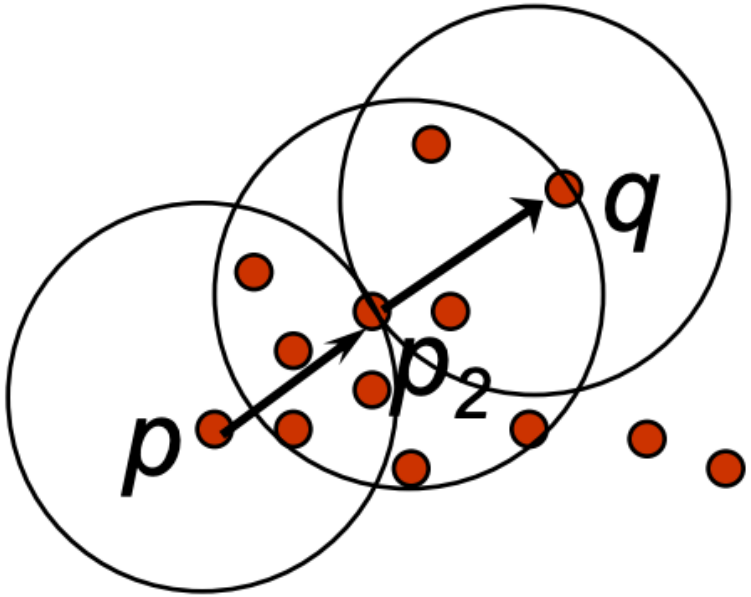
DBSCAN Algorithm

- A point *q* is *directly reachable* from *p* if point *q* is within distance *ε* from core point *p*. Points are only said to be directly reachable from core points.



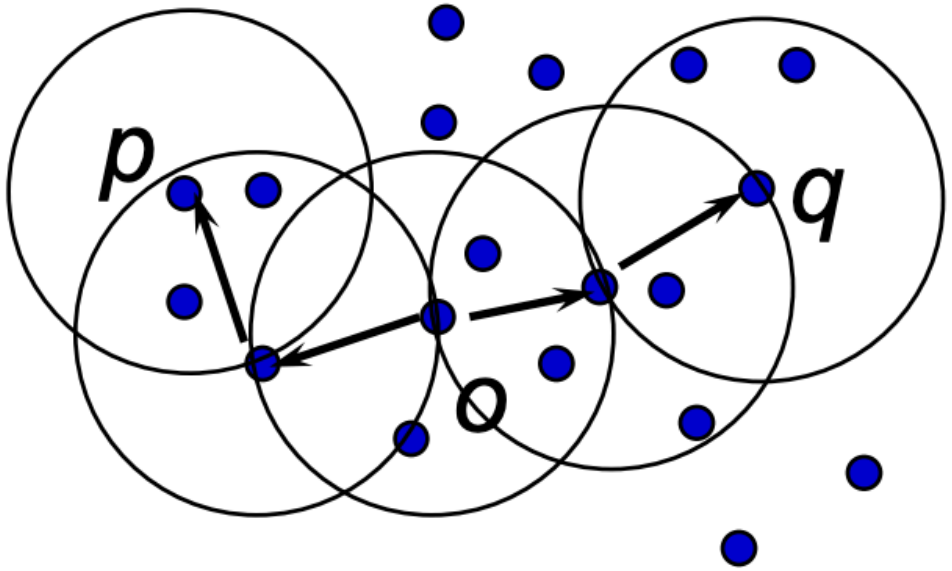
DBSCAN Algorithm

- A point q is *reachable* from p if there is a path p_1, ..., p_n with $p_1 = p$ and $p_n = q$, where each p_{i+1} is directly reachable from p_i. Note that this implies that the initial point and all points on the path must be core points, with the possible exception of q.



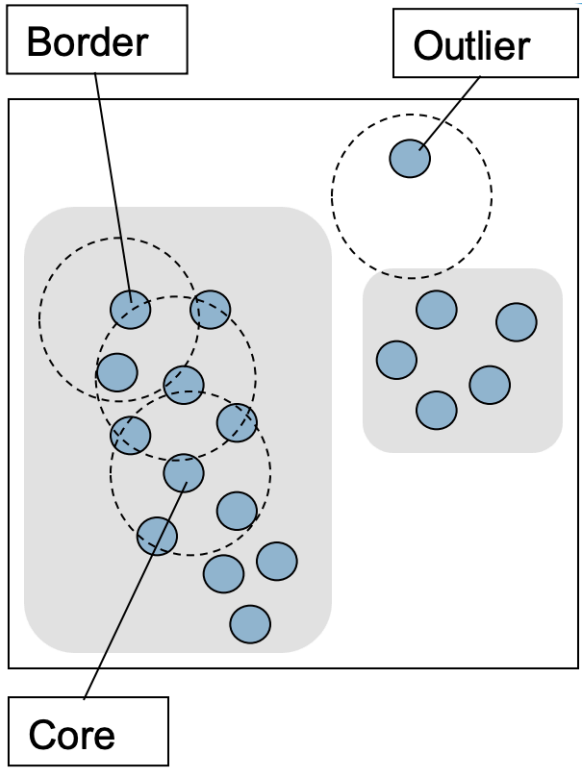
DBSCAN Algorithm

- Two points *p* and *q* are density-connected if there is a point *o* such that both *p* and *q* are reachable from *o*.

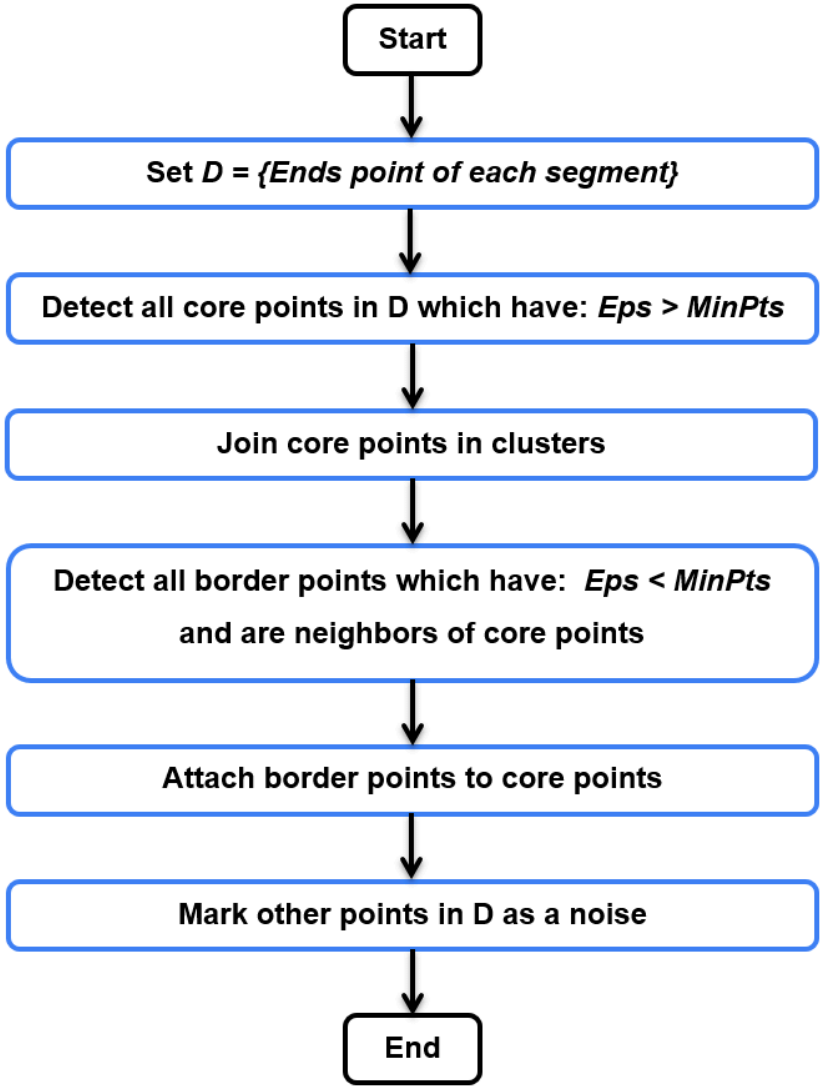


DBSCAN Algorithm

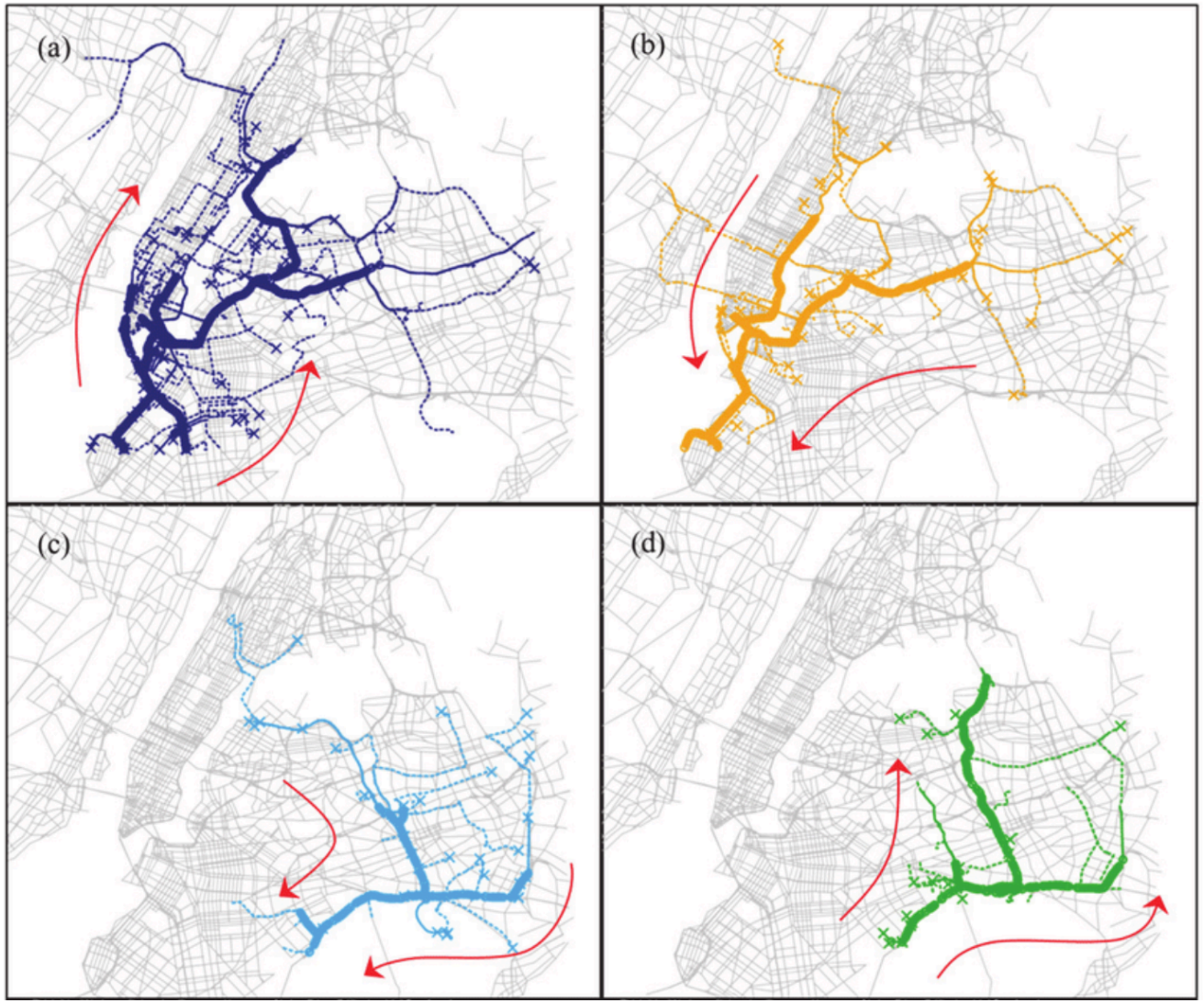
- A cluster then satisfies two properties:
  - All points within the cluster are mutually density-connected.
  - If a point is density-reachable from some point of the cluster, it is part of the cluster as well.



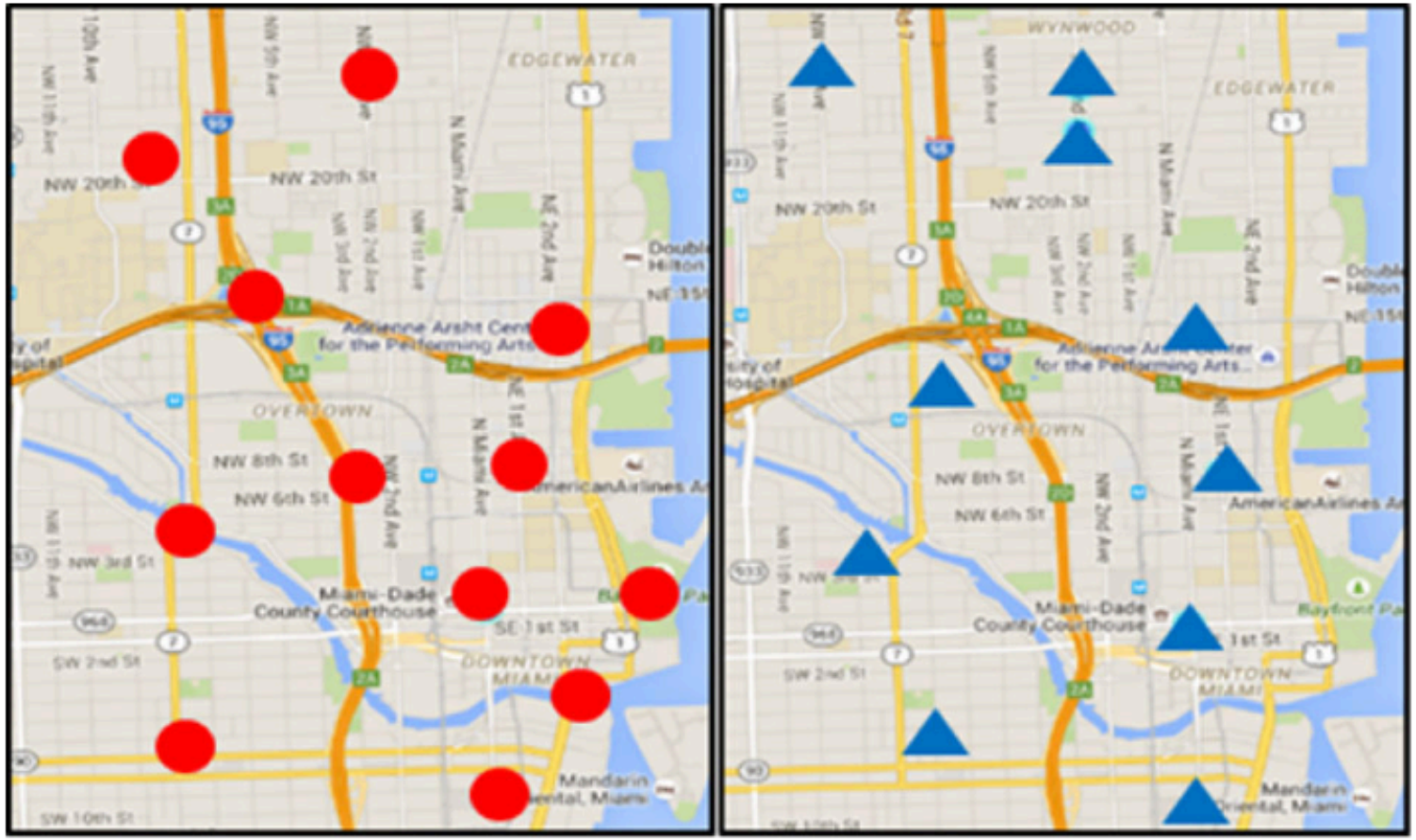
DBSCAN Algorithm



DBSCAN Algorithm



Result


EDGEWATER
NW 20th St
OVERTOWN
NW 8th St
NW 6th St
NW 3rd St
SW 2nd St
SE 1st St
DOWNTOWN MIAMI
Miami-Dade County Courthouse
WYNWOOD

Conclusion

- This method can not only protect a user's location privacy while efficiently ensuring the accuracy of the location-based service through differential privacy, but also protect the privacy of each individual user by adding noise to the statistical reports so that a user's tweets cannot significantly change the alert status.
- We explored adding differential privacy capabilities to Twitter data. Through the application of RDBC to cluster sub-regions split by differentially privacy optimal quad-tree spatial decomposition
- We showed that privacy and precision are trade-offs
- One key area of application of Twitter is real-time information on transport. Tweets about traffic conditions such as traffic congestion or traffic accidents provide near real-time traffic information that is useful for travelers and could allow them to take alternative routes.

References

[1] Shuo Wang1 & Richard O “Supporting geospatial privacy-preserving data mining of social media”

[2] Miguel Andre´s, Nicola´s Bordenabe “Geo-Indistinguishability: Differential Privacy for Location-Based Systems”

[3] Internet

THANK YOU!