Privacy-Preserving Online Task Assignment in Spatial Crowdsourcing with Untrusted Server

Hien To[1], **Cyrus Shahabi**[2], Li Xiong[3]

[1] Amazon Mechanical Turk

[2] University of Southern California

[3] Emory University

USC Viterbi School of Engineering
Integrated Media Systems Center

[1] This work has been done while the author was a PhD student at University of Southern California

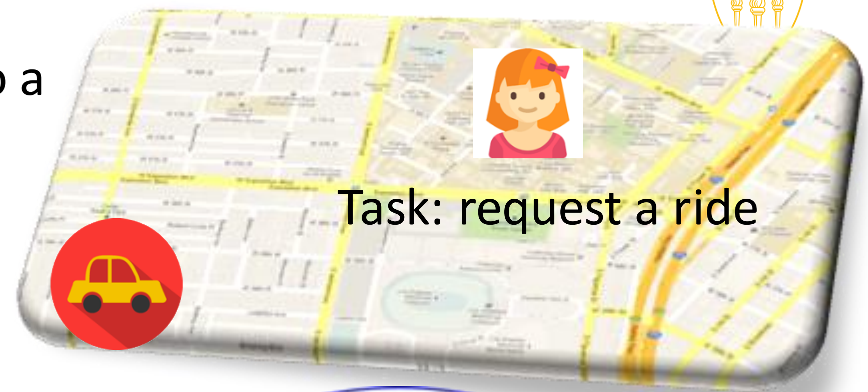
Outlines

- Introduction & Motivation
- Related work
- Background
- Proposed Approach
- Evaluation
- Conclusions

Spatial Crowdsourcing (SC)

Crowdsourcing: outsourcing a set of tasks to a set of workers amazon mechanical turk™ Artificial Artificial Intelligence

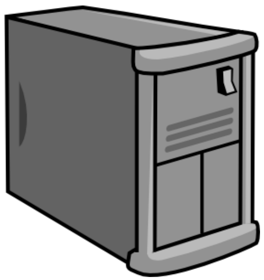
Spatial crowdsourcing (SC): requires workers to ***physically*** travel to task's location



Ubiquity of mobile users	**6.5 billion** mobile subscriptions, 93.5% of the world population [*]
Technology advances on mobiles	Smartphone's **sensors**. e.g., video cameras
Network bandwidth improvements	From 2.5G (384Kbps) to 3G (14.7Mbps) and recently **4G** (100 Mbps)

Task Assignment in SC

Requesters (e.g., request a ride)



Server (e.g., Uber)

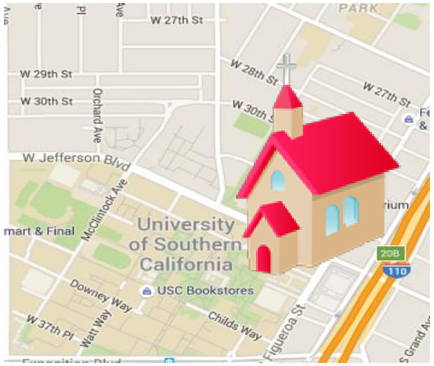
Workers (e.g., drivers)

Server chooses best workers for a task based on task-worker proximity *e.g., [Kazemi'12, Pournajaf '14, To'17]*

Server knows locations of workers and tasks ☹

Risks of Location Leaks

Location leaks sensitive information, e.g., religious view, health status



Attacks based on locations:

PRIVACY ROAD KILL 4/26/16 2:40 PM

If you use Waze, hackers can stalk you

'God View': Uber Allegedly Stalked Users

"Uber treated guests to Creepy Stalker View, showing them the whereabouts and movements of 30 Uber users in New York in real time."

Forbes

Related work

Location Privacy

Anonymity based (e.g., cloaking)

- Pseudonymity *[Pfitzmann et al. 2010]*
- K-anonymity/Cloaking *[Sweeney'02]*

Encryption-Based

- Private information retrieval *[Ghinita et al. SIGMOD 2008]*
- Space transformation *[Khoshgozaran & Shahabi SSTD 2007]*

Perturbation (e.g., differential privacy)

- Geo-indistinguishability *[Andrés et al CCS 2013]*
- δ-location set-based differential privacy *[Xiao & Xiong CCS 2015]*

Apple and Google adapted **differential privacy** to discover usage patterns from a large number of users

- Google Chrome web browser [1]
- Apple QuickType/Emoji [2] suggestions.

[1] Erlingsson et. al. *Rappor: Randomized aggregatable privacy-preserving ordinal response*. *ACM SIGSAC 2014.*
[2] Learning with Privacy at Scale.
https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html

Related work

Privacy-Preserving Task Assignment

Papers	*Privacy Techniques*			*Protection*		*Trusted Server*	
	Cloak	*Encrypt*	*Perturb*	*Worker*	*Task*	*Yes*	*No*
[Pournajaf et al. 2014]	x			x		x	
[Sun et al. 2017]	x			x		x	
[Pham et al. 2017]	x			x	x	x	
[Hu et al. 2015]	x			x		x	
[Shen et al. 2016]		x		x			x
[Liu et al. 2017]		x		x	x		x
[To et al. 2014]			x	x	☹	x	☹
[Gong et al. 2015]			x	x	☹	x	☹
[Zhang et al. 2015]			x	x	☹	x	☹
[To et al. 2016]			x	x	☹	x	☹

Existing work that use perturbation technique protect worker location only and assume trusted server ☹

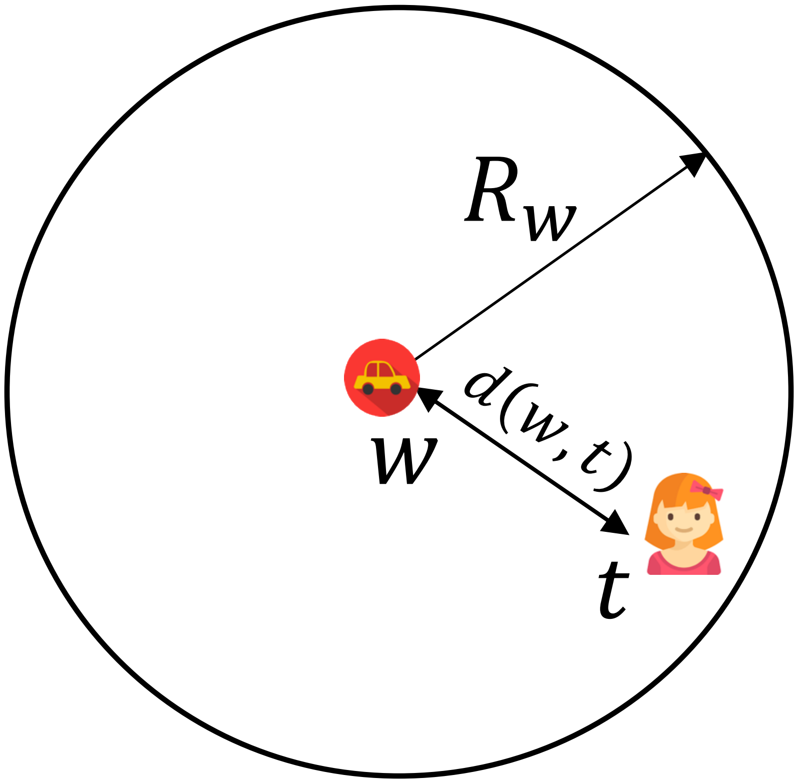
USC Viterbi School of Engineering Integrated Media Systems Center

Outlines

- Introduction & Motivation
- Related work
- Background
- Proposed Approach
- Evaluation
- Conclusions

Notations

Notation	Description
w, t	Actual locations of a worker, a task
w', t'	Perturbed locations
R_w	Reachable distance of worker w
$d(w, t)$	Euclidean distance between w and t



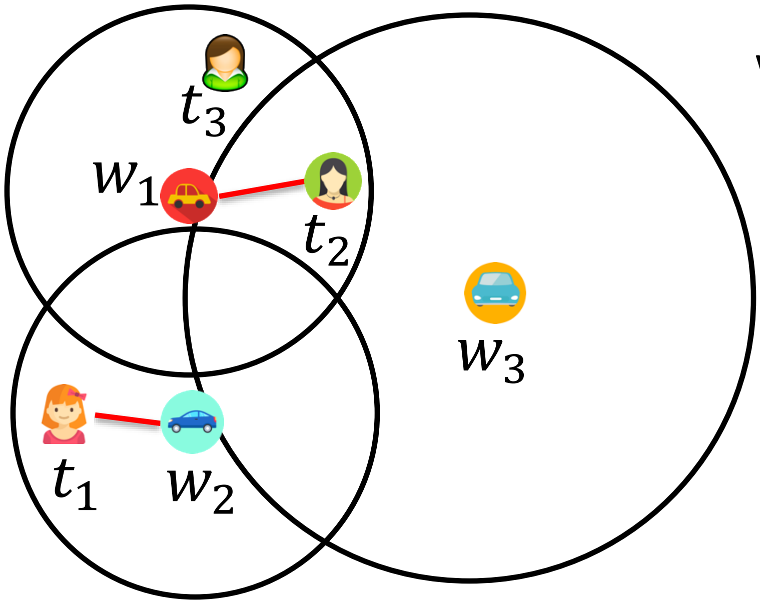
Task t is ***reachable*** from worker w if $d(w, t) \leq R_w$

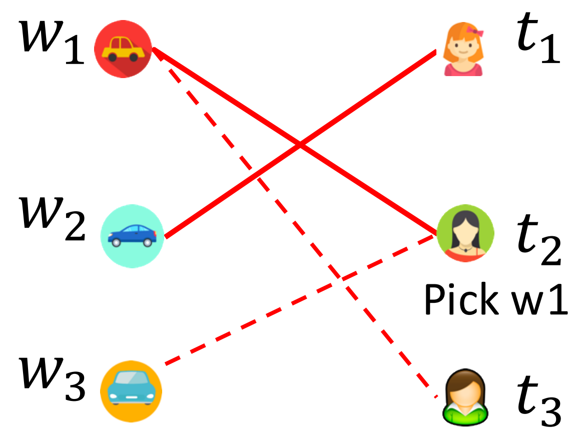
d can be non-Euclidean & R_w can be complex shapes like polygon

Background

Online Task Assignment

Worker set is known, each task arrives one-by-one





Assign as many tasks as possible to workers

Ranking algorithm[*] is optimal, competitive ratio 0.63

- Permutes workers and assigns a **random rank** to them
- Each task is matched to a reachable worker of the highest rank

USC Viterbi School of Engineering *Integrated Media Systems Center*

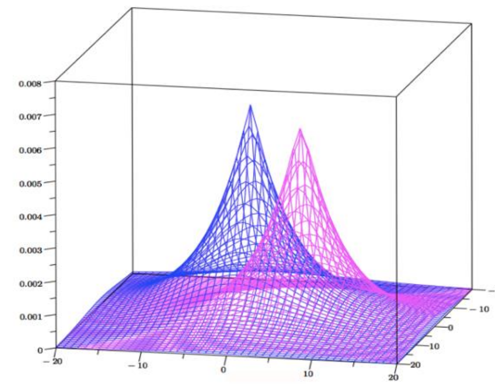
[*] Karp et al. *An optimal algorithm for on-line bipartite matching*, STC'90

(ϵ, r) Geo-indistinguishability[*]

<u>The goal:</u> An adversary cannot distinguish locations which are at most r distance away

<u>Approach:</u> Any two locations at distance at most r produce "similar" observations (bounded by ϵ),

- r is the radius of concern within which privacy is guaranteed (This means that an adversary cannot distinguish locations which are at most r distance away)
- The smaller eps is, the stronger privacy (as it gets harder for the attacker to detect the user's location among the points within this circle).



More formally:

Mechanism A satisfies (ϵ, r)-Geo-I iff for all x, y such that $d(x, y) \leq r$:

$$d_p\big(A(x), A(y)\big) \leq \epsilon d(x, y) \leq \epsilon r$$

- $d(x, y)$: Euclidean distance between x, y
- $d_p(,)$: multiplicative distance between two distributions

[*] Andrés et al. *Geo-indistinguishability: differential privacy for location-based systems*, CCS'13

(ϵ, r) Geo-indistinguishability[*]

it is sufficient to achieve (ϵ, r)-Geo-I by generating random point z (from actual point x ∈ X) according to planar Laplace distribution.

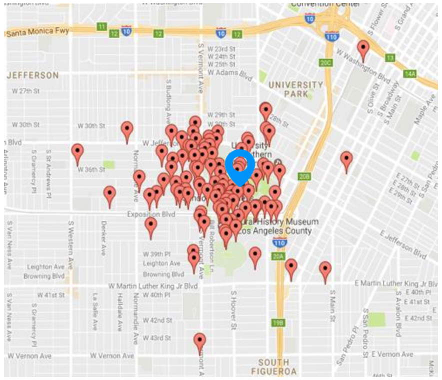
r (in meters) is the radius within which privacy is guaranteed

ϵ tunes how much privacy, smaller ϵ means higher privacy

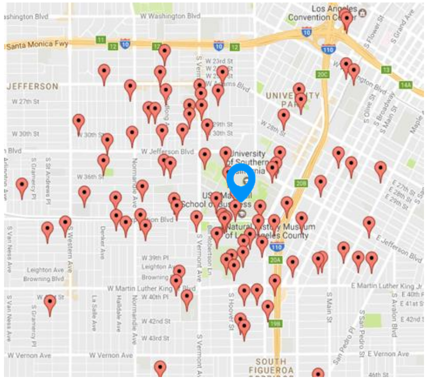
achieve privacy by injecting planar Laplace noise

True locations

Perturbed locations



$\epsilon = \log(6)$
r=1 km

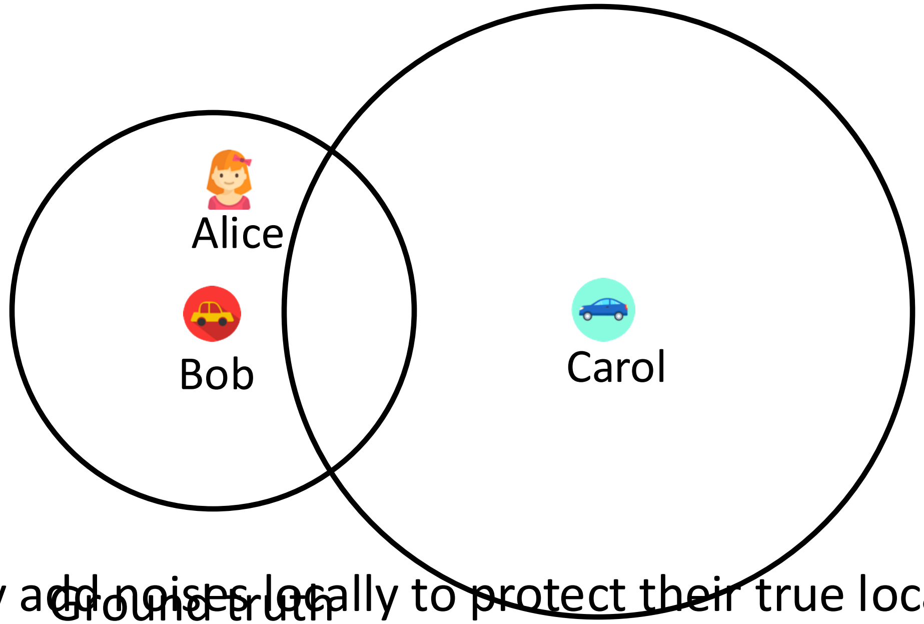


Better privacy: $\epsilon = \log(2)$
r=1 km

[*] Andrés et al. *Geo-indistinguishability: differential privacy for location-based systems*, CCS'13

Challenges with Perturbed Locations

Reachable worker-task pair is observed as unreachable, and vice versa



They add noises locally to protect their true locations

Ground truth

Alice is not assigned to Bob (not reachable) ☹

Alice's location is disclosed to Carol *unnecessarily* ☹

Outlines

- Introduction & Motivation
- Related work
- Background
- Proposed Approach
- Evaluation
- Conclusions

Three-Phase Framework

Alice requests for a ride

Finds ***candidate drivers*** *for* Alice: Bob Carol Dave

Server does not know *anyone's* location (works in perturbed space for both riders and drivers)

Sends perturbed locations of drivers to Alice

System Overhead

Finds the **most likely reachable driver:** Bob

Alice does not know *any driver's* location (works in perturbed space for drivers but knows her own location)

Reveals her location to Bob

Location Disclosure

Bob checks if Alice is reachable

Reachable → accepts (happy case)

Not reachable → rejects

Repeat until either task is assigned or no candidate worker left

Baseline Approach

System Overhead: size of the worker candidate set, captures communication and computational overhead
Location Disclosure (false hit): privacy leak occurs when Alice estimates an unreachable worker as reachable & reveals her location
Utility: number of assigned tasks
Worker Travel Cost: captures travel cost or assignment quality

"Oblivious" algorithm

- assumes perturbed locations as actual ones
- Direct adaptation of Ranking algorithm[*] to our framework
  - Consider both **random rank** and **distance-based rank**

Core idea:

- to use underlying distributions of noisy locations to estimate real locations

[*] Karp et al. *An optimal algorithm for on-line bipartite matching*, STC'90

Worker-Task Reachability

Compute the **reachability probability** of a worker-task pair given their observed distance

: $\Pr(d(w,t) \le R_w \mid d(w', t'))$

: $\Pr(d(w,t) \le R_w \mid d(w', t))$

I. Analytical approach, based on estimating the reachability probability

- Derive PDF of $d(w,t)$, given w', t'
  Subsequently, the reachability probability can be computed efficiently
- Planar Laplace distribution is difficult to analyze so we approximate it by bivariate normal distribution (BND)

II. Empirical approach, based on synthetic or historical data

Bivariate Normal Distribution (BND)

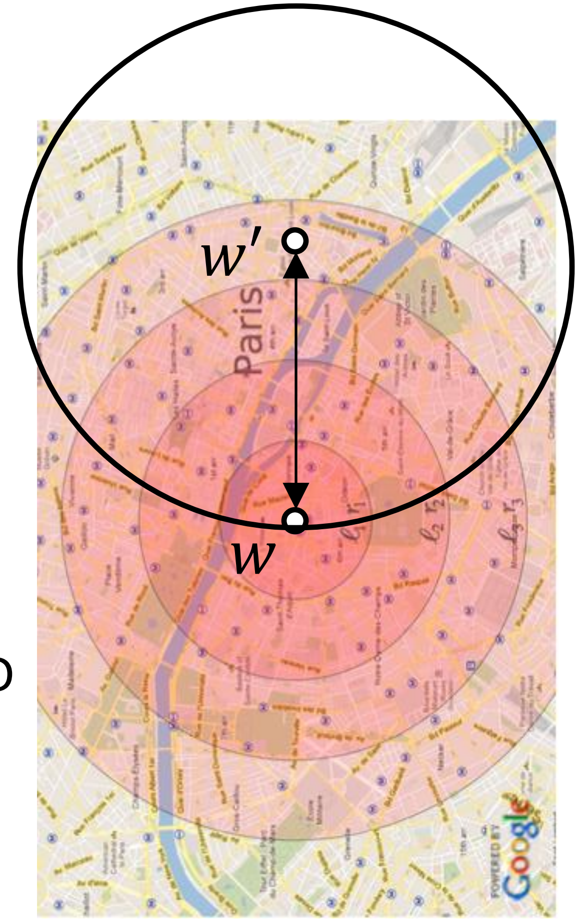
(ϵ, r)-Geo-Indistinguishability uses planar Laplace distribution (PLD) to inject noise

- PLD is difficult to analyze

Approximate PLD by a circular BND with same mean (w_x, w_y) & covariance matrix $\begin{bmatrix} \frac{2r^2}{\epsilon^2} & \\ & \frac{2r^2}{\epsilon^2} \end{bmatrix}$

- BND is made up of two random variables x and y; both normally distributed
- PLD is symmetric to its center → approximated BND should be symmetric to the same center

w' is known → w follows circular BND centering at w': circular $BND(w', \Sigma)$



[*] Andrés et al. *Geo-indistinguishability: differential privacy for location-based systems*, CCS'13

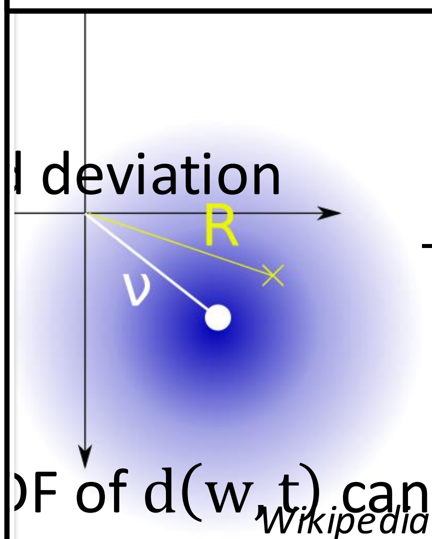
derives PDF of $d(w,t)$

Given true location of Alice t and perturbed location of Bob w'

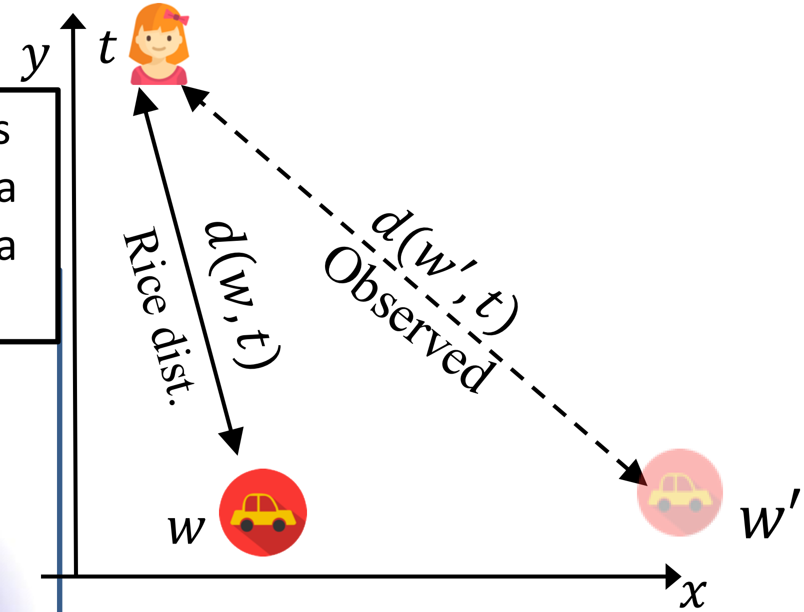
estimates PDF of $d(w,t)$

In the 2D plane, pick a fixed point at distance v from the origin. Generate a distribution of 2D points centered around that point, where the x and y coordinates are chosen independently from a gaussian distribution with standard deviation σ (blue region). If R is the distance from these points to the origin, then R has a Rice distribution.

Rice distribution is the magnitude of a circular BND with a non-zero mean



l deviation



DF of d(w,t) can be found in the paper

[*] Stüber. *Principles of mobile communication*, volume 2. Springer, 2001

Proposed Approach

Probability-based Solution

The key idea is to use the probabilistic model (either the analytical or the empirical approach), for quantifying reachability between a worker and a task.

finds *candidate drivers* N_j based on *reachability threshold* α

$$N_j = \{w_i : \Pr(reachability(w'_i, t'_j)) \geq \alpha\}$$

The smaller α, the higher the overhead, but less chance of missing a reachable worker

reveals her location to highly *likely reachable drivers*

$$Rank_{w_i} = \Pr(reachability(w'_i, t_j))$$

Heuristic:

can reduces disclosure of her location based on *reachability threshold* β $(\beta > \alpha)$

e.g., if $Rank_{w_i} < \beta$, cancel this task

Outlines

- Introduction & Motivation
- Related work
- Background
- Proposed Approach
- Evaluation
- Conclusions

Experimental Evaluation

- GPS-equipped taxis dataset [1]
  - Workers' locations are the most recent drop-off locations
  - Tasks' locations at the pick-up locations
  - 500 tasks and 500 workers were randomly sampled

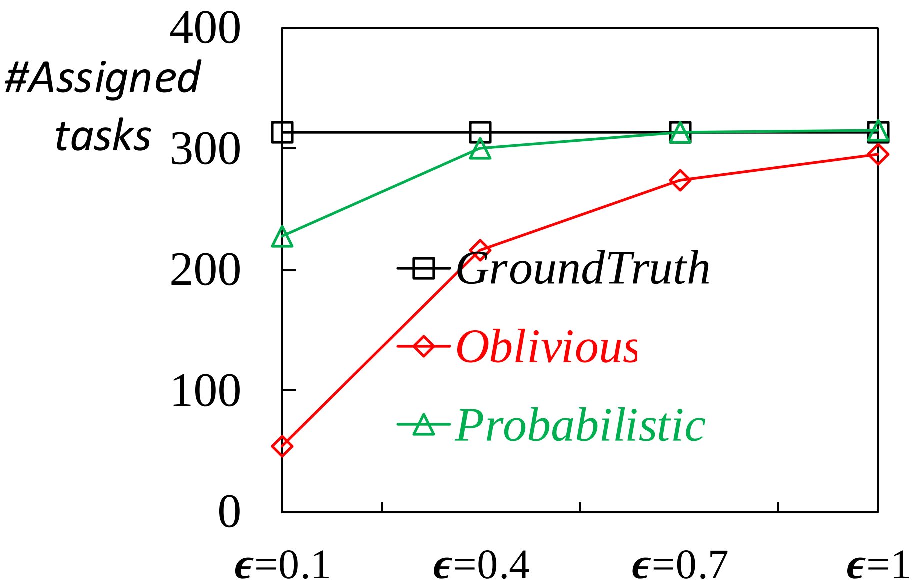
	#Passengers	#Drivers	Area
T-Drive	100,000+	9,019	Beijing City

- Performance metrics
  - **Utility**: number of assigned tasks
  - **Worker Travel Cost**: captures travel cost or assignment quality
  - **System Overhead**: size of the worker candidate set, captures communication and computational overhead
  - **Location Disclosure** (false hit): privacy leak occurs when requester estimates an unreachable worker as reachable

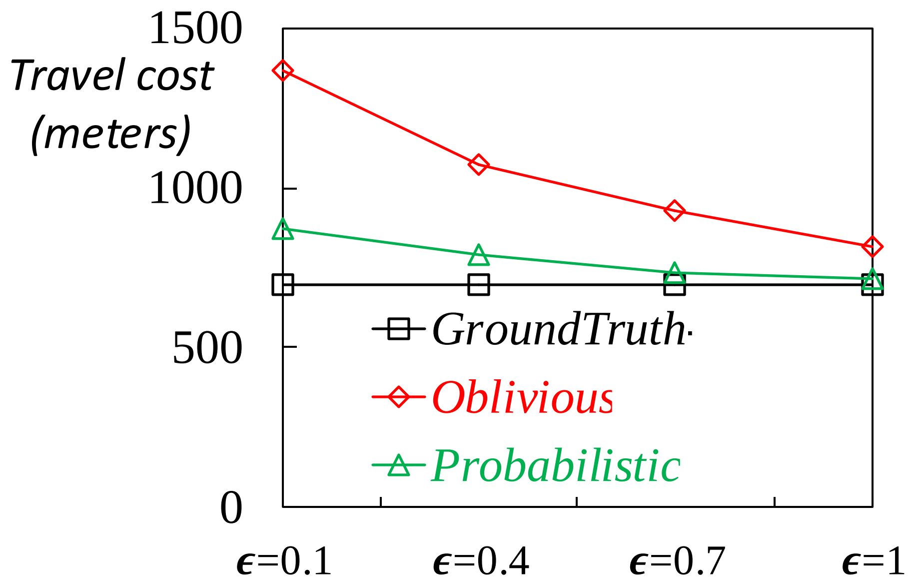
[1] Yuan et al. *T-drive: driving directions based on taxi trajectories*. SIGSPATIAL 2010

Utility and Travel Cost

GroundTruth	Has access to exact locations (distance-based rank)
Oblivious	Assumes perturbed locations as actual ones (distance-based rank)
Probabilistic	Estimates worker-task reachability (probability-based rank)



Probabilistic obtains much **higher utility** than *Oblivious* (by 300%)



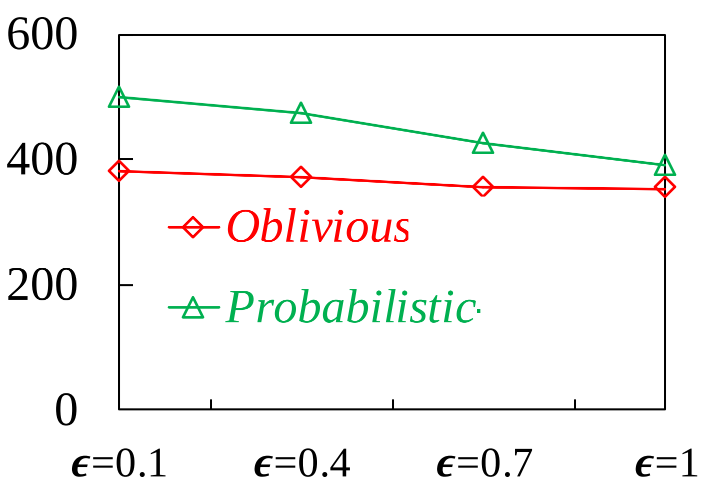
Probabilistic obtains significantly **lower travel cost** than *Oblivious* (by 30%)

USC Viterbi
School of Engineering
Integrated Media Systems Center

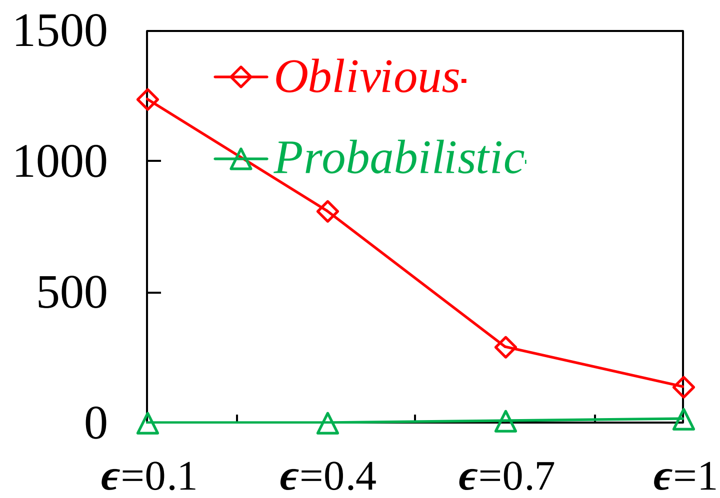
System Overhead and Privacy Leak

Oblivious	Assumes perturbed locations as actual ones (distance-based rank)
Probabilistic	Estimates worker-task reachability (probability-based rank)

#workers (overhead)



#false hits (disclosure)



Although the overhead of *Probabilistic* is slightly higher than *Oblivious's*, *Probabilistic* has **much smaller false hits**

Average **#false hits** before a task can be assigned: 23 workers vs 1.05 workers

Conclusions and Future Work

- Protected locations of both workers and tasks
  - Introduced privacy-aware framework with untrusted server
  - Proposed models for quantifying worker-task pair reachability
  - Proposed algorithms, heuristics for effective online tasking
- Confirmed the cost of privacy is practical
  - Low cost and low overhead without compromising utility
- Future directions
  - Consider malicious adversaries: requesters send fake tasks to estimate workers' locations, server colludes with workers (driverless cars)
  - Consider protection for dynamic workers and task: workers' traces and task locations of individual requesters can follow a specific pattern
  - Consider tasks that may require redundant assignment: taking pictures of a particular location, reporting how crowded a restaurant is

Two Sides of the Coin

Protecting against social inferences
**** But allow for LBS***

Privacy Twist

Inferring Social Relationships

- Privacy attack

walk2friends: Inferring Social Links from Mobility Profiles
[CCS, Nov '17] Backes M, Humbert M, Pang J, Zhang Y.

walk2friends: Inferring Social Links from Mobility Profiles [CCS, Nov ‘17] Backes M, Humbert M, Pang J, Zhang Y.

- Can we do better in very dense datasets ?

- Feature learning method – Unsupervised
  - As opposed to EBM’s supervised linear regression.
  - Claims to exploit followship in addition to EBM’s co-occurrence

- Inspired by Deep Learning in NLP – word2vec
  - Skip-gram Model
    (Tomas Mikolov et. al., at Google Research, 2013)

A glance at the Skip-Gram Model

Goal: Given a specific word in a sentence, tell us the probability for every word in our vocabulary of being the "nearby word" to the one we chose.

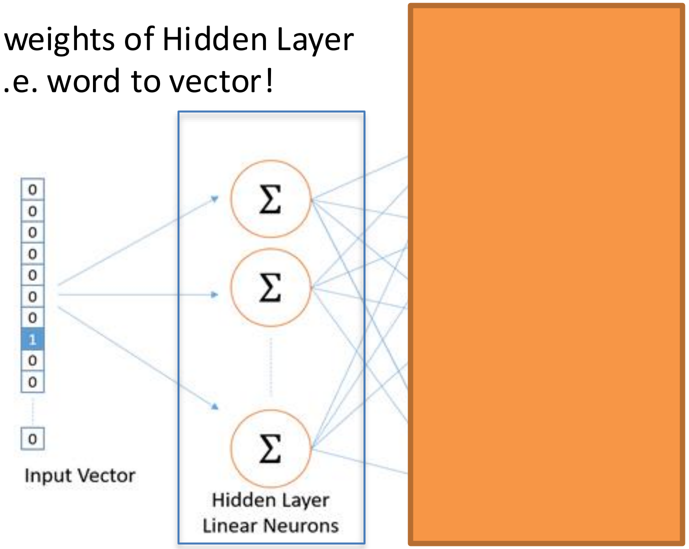
Corpus training (NN)

The quick brown **fox** jumps over the lazy dog.

→ (fox, quick)
(fox, brown)
(fox, jumps)
(fox, over)

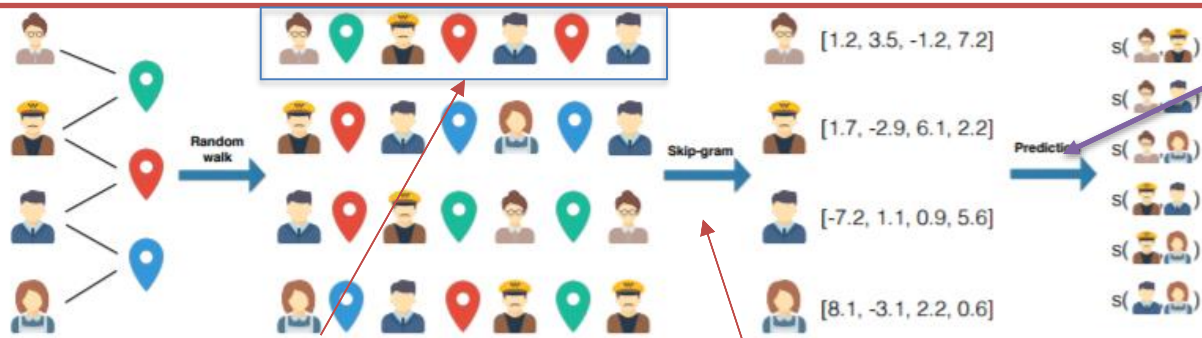
Discard Output Layer!

Keep weights of Hidden Layer
i.e. word to vector!



walk2friends: Extending to locations based networks.

Input | User's mobility neighbors | Mobility Feature Vector | Social Strength



Use vector similarity metrics.

Similar to corpus sentences

If two nodes share similar neighbors, then their vectors will be similar.

✓ Captures frequented locations.

✓ Cap

✓ Performs~10-15% percent better than EBM on relatively dense datasets.

✗ 3-5% worse on sparse datasets.

How to protect against social inference attack?

Co-Location Privacy Risks

1. NSA PRISM (began 2007):
   Mass surveillance of location data from Google, FB, Microsoft.

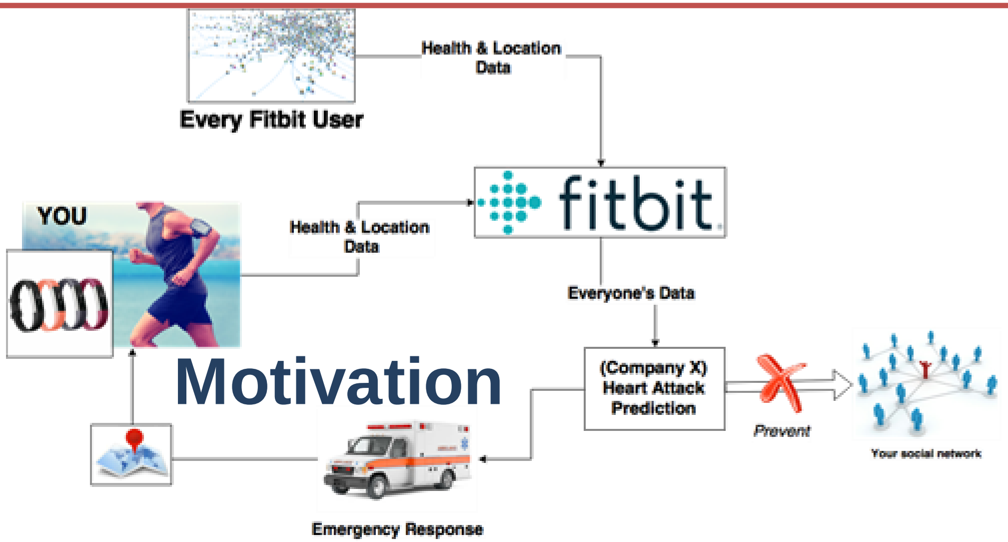
[Source: Washington Post]

2. NSA's Co-Traveler program (exposed 2013):
   Identifies unknown associates of a known target.

3. Domestic prosecution facilitated by co-location information as evidence of wrongdoing. [United States v. Jones, 132 S.Ct. 945 (2012)]

Motivation

Location Data is necessary for service but social connectivity is sensitive.

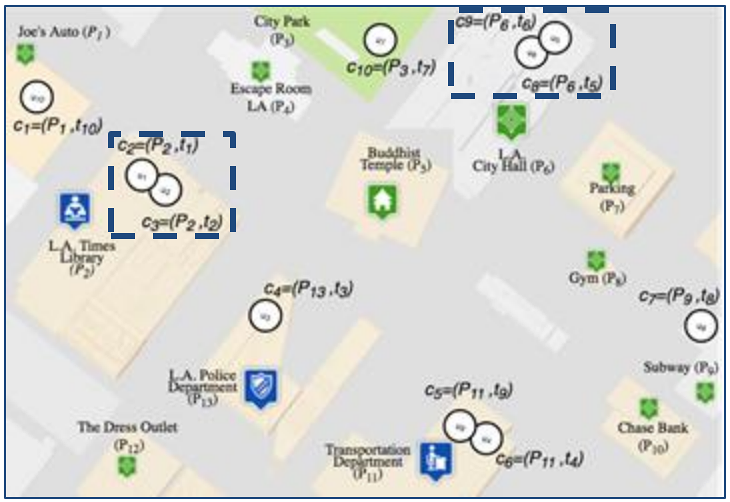


Enable LBS to provide recommendation, advertisement, and other services.

Target Co-locations

The building blocks for social inference techniques.

Co-Location: Two people at *roughly* the same geographic locale at roughly the same time.



We quantify 'roughly' based on parameters Δ_s and Δ_t .

In running example**,**

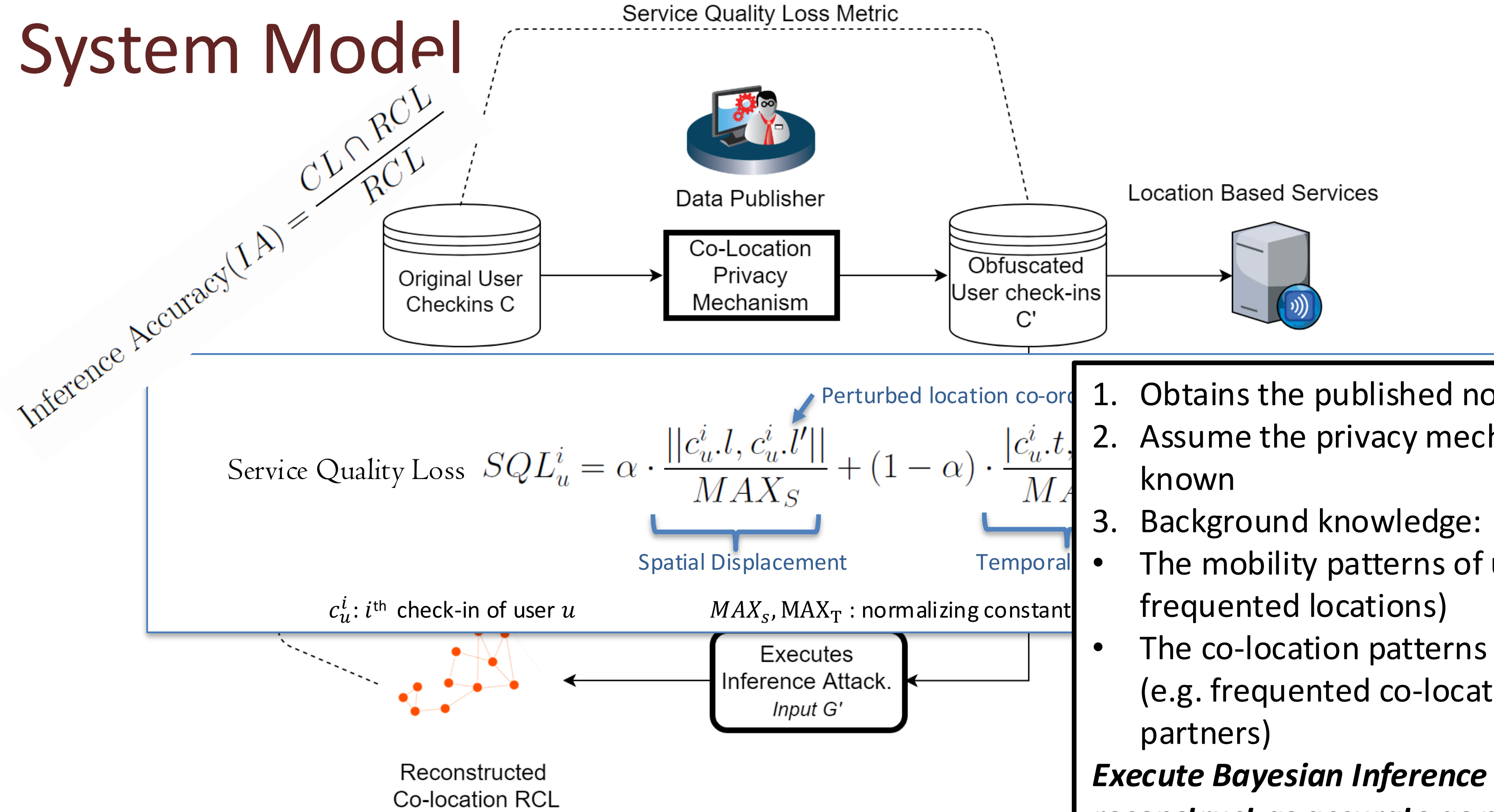
- Assume buildings are points

$\Delta_s = SameBuilding, \Delta_t = 1t$

Co-Locations: $(u_1, u_2), (u_5, u_6)$

Δ_s and Δ_t are application specific.

System Model



1. Obtains the published noisy data
2. Assume the privacy mechanism is known
3. Background knowledge:

- The mobility patterns of users. (e.g. frequented locations)
- The co-location patterns of users. (e.g. frequented co-locating partners)

Execute Bayesian Inference to reconstruct as accurate as possible representation of the original co-locations.

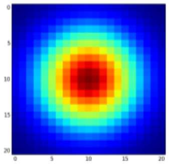
Co-Location Privacy Mechanism 1: Gaussian Perturbation (Naïve)

Most popular methods in statistical data privacy.

Simplest method in Location Privacy and a mechanism of noise for advanced methods like *probabilistic differential privacy*.

Method:

1. For every co-location, it is enough to perturb one check-in.
2. Translate both coordinates with 2d-gaussian noise.
3. Translate timestamp with 1d-gaussian noise



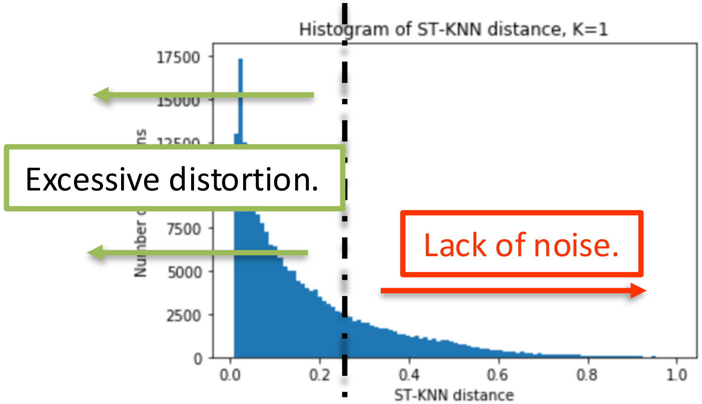
(a) Original GPS data (b) Additive Gaussian noise

Krumm, [PerCom'07]

Shortcomings of Gaussian Perturbation

1. Skewed nature of the distribution of the closest neighbor:
   large number of users have NN very close, while some have their NN very far.

2. Any fixed magnitude of noise will lead to either:
   - Low Privacy: Under-protected in sparse areas, or
   - Low Utility: Over-protected In dense areas inhibiting quality of LBSs.



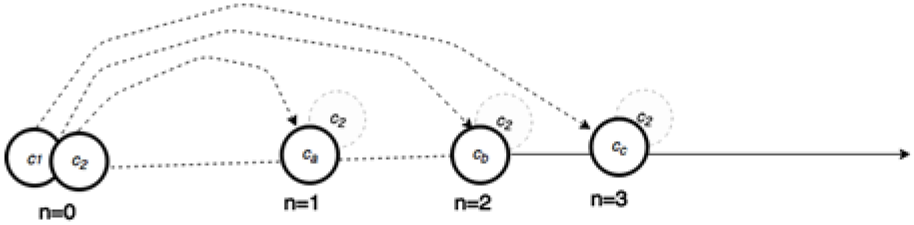
$ST_{dist} = 0.1 = 100m + \ 40\,min$

On X-Axis, 0.01 is the first 1% percent of **co-locations** (i.e. the 1st percentile) with the smallest STdist to their nearest neighbor.

Co-Location Privacy Mechanism 2: Adaptive Perturbation

Use the presence of spatiotemporal nearest neighbors as an estimate for density.

Method: 1. For every co-location pair, pick one check-in at random;
2. Chose p uniformly over the set of
(i) the b nearest neighbors,
(ii) together with the current location.
3. Move to p.



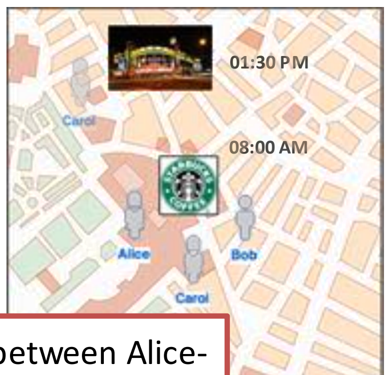
Move c2 to any of 'b=4' positions at random

$^*ST_{dist}(c, c')$ = sum of normalized spatial and temporal distances

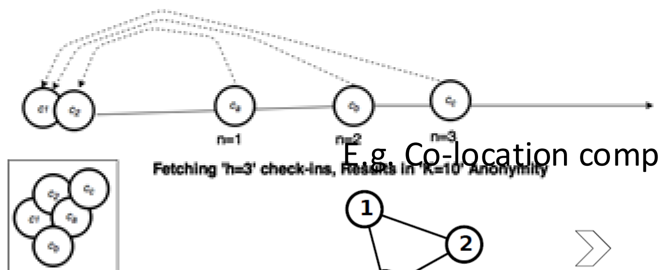
Co-Location Privacy Mechanism 3: Co-Location K-Anonymity

Definition: A co-location is k-anonymous if it is spatiotemporally indistinguishable to $k-1$ other co-locations.

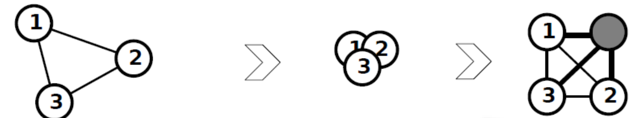
Method: For every co-location pair, Make each co-location k-anonymous by moving "h" closest check-ins to form a group.



The co-Location between Alice-Bob is now 3-anonymous.



E.g. Co-location component is 2-anonymized

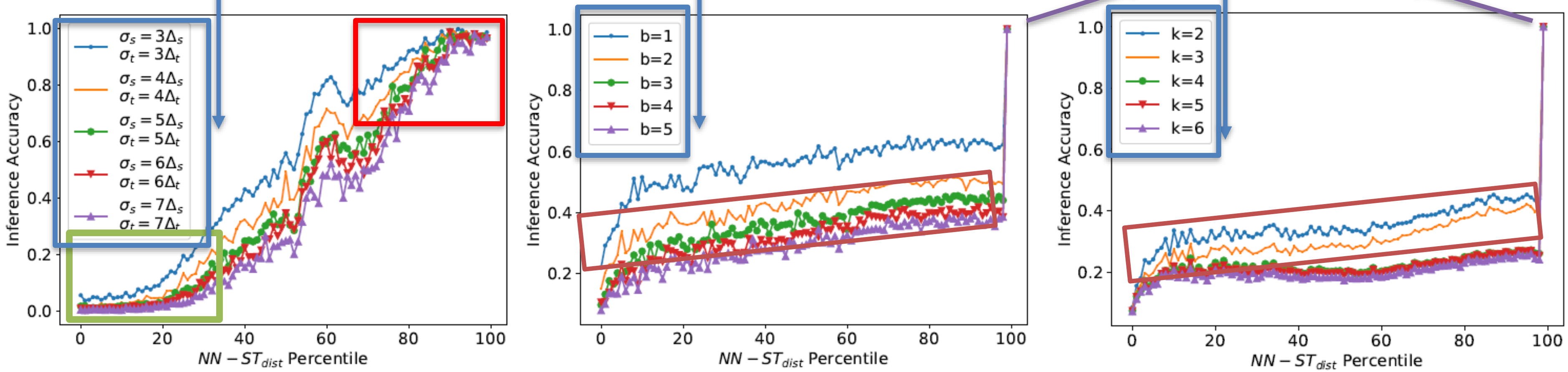


On seeing any co-location the adversary can only tell it's truthfulness with a certainty of $1/2$ ($i.e.\ 1/k$).

Attack Accuracy on Privacy Mecha

Ignoring a few hundred co-locations in extremely remote
or fair comparison.

Increasing level of distortion.



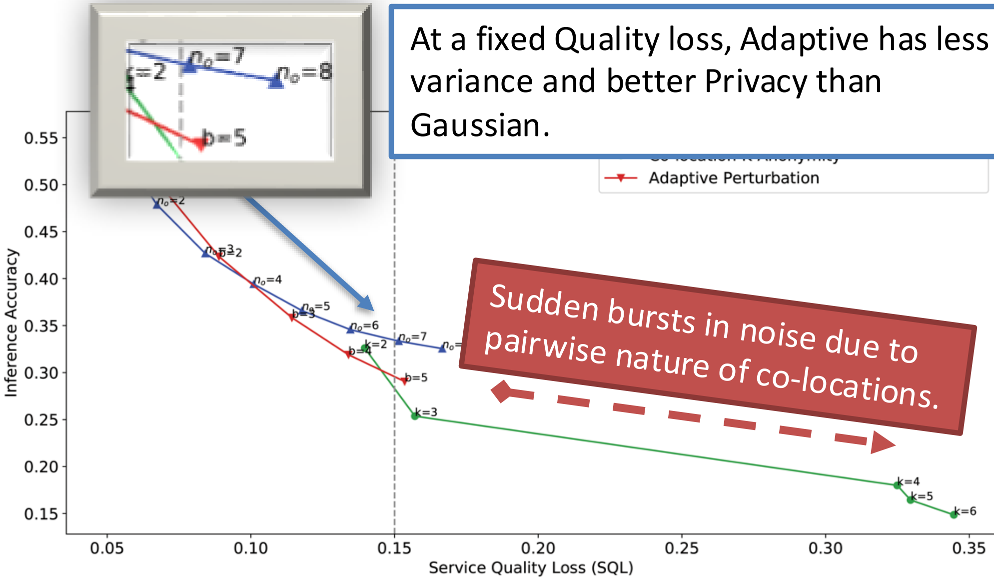
Dense → Sparse

Gaussian exposes a significant portion of the population to highly accurate inferences.

Adaptive and k-anonymity provide consistent protection (i.e. with low variance) against an adversary.

Analysis of Quality Loss and LBS Range Utility

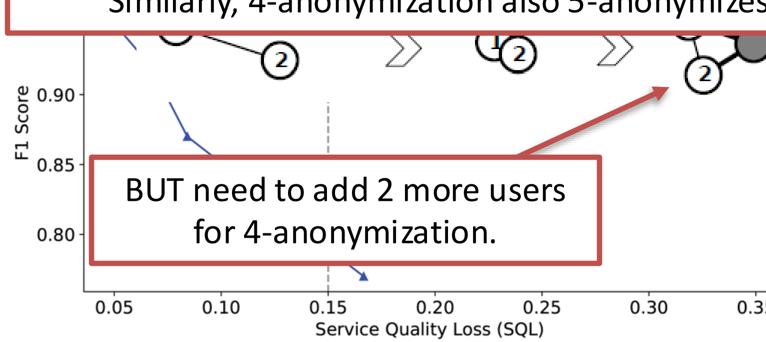
At a fixed Quality loss, Adaptive has less variance and better Privacy than Gaussian.



Sudden bursts in noise due to pairwise nature of co-locations.

At a given level of total displacement, Adaptive performs best in Range queries.

For a co-located pair 2-anonyimization also 3-anoymizes.
Similarly, 4-anonymization also 5-anonymizes



BUT need to add 2 more users for 4-anonymization.

- Adaptive outperforms Gaussian by achieving better privacy at a given *SQL*.
- Co-location k-anonymity offers limited flexibility in calibrating noise.
- Adaptive distorts to the NNs, hence is ideal for location-based advertising.

Thanks!