Unintended Consequences of Disclosing Location Data

Cyrus Shahabi, Ph.D.
Professor of Computer Science, Electrical Engineering & Spatial Sciences
Chair, Department of Computer Science
Director, Integrated Media Systems Center (IMSC)
Director, Informatics Program
Viterbi School of Engineering
University of Southern California
Los Angeles, CA 900890781
shahabi@usc.edu

Outline

Motivation: Location-embedded social structure

Prior Work: Inferring Social Behaviors

Current Efforts: Protecting against social inferences

- But allow location disclosure

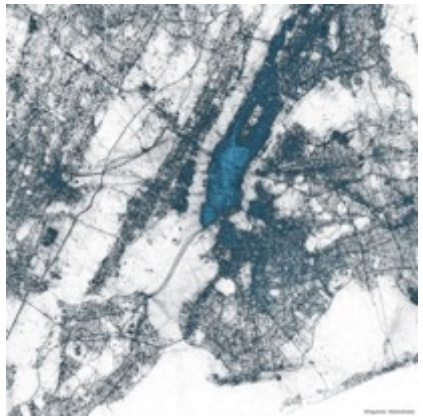
Open Problem: Protecting against location disclosure

- But allow social inferences

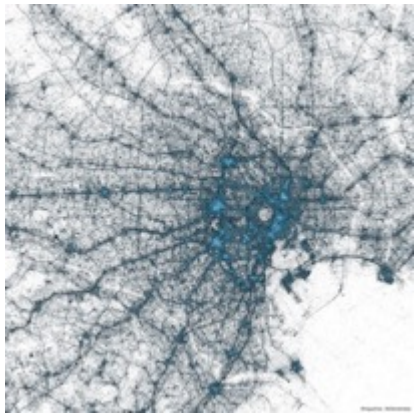
Location-Enriched Datasets

- Popularity of Location-Based Services

Twitter: 10M+ geo-tagged tweets/day mashable.com
Foursquare: 5M check-ins/day venturebeat.com/2015/08/09/



New York City



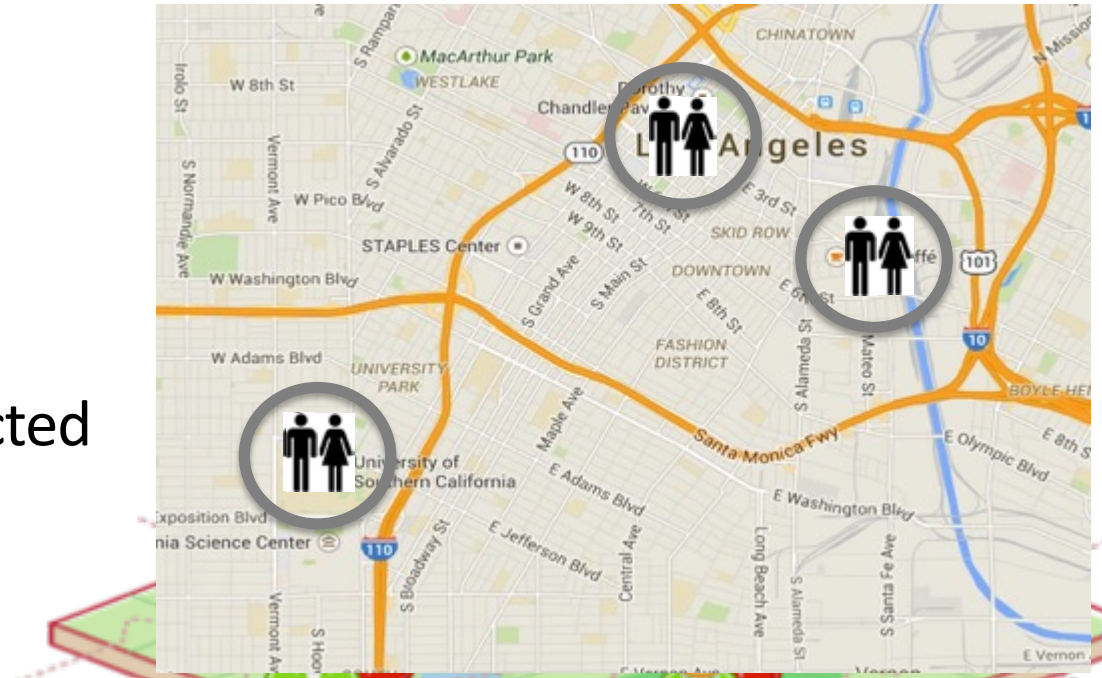
Tokyo

Europe

Geo-Tagged Tweets on Map
by Twitter mashable.com

Social Relationship Inference from Location Data

- Reachability [VLDB'12]
  - *u* is reachable to *v* in time period *T*
  - if there is a contact path
- Social Strength [SIGMOD'13]
  - *u* and *v* are socially connected
  - how often they meet and where
- Spatial Influence [ICDE'16]
  - *u* influences *v*
  - if *v* follows *u*



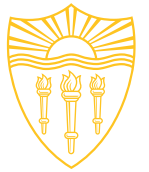
followship: *u* visits a place, and *v* visits the same place shortly after.

Applications

- Social Network
  - Marketing
  - Friendship suggestions
  - Social and cultural studies

- Geo-social Network
  - Criminology
    - identify the new or unknown members of a criminal gang or a terrorist cell
  - Epidemiology
    - spread of diseases through human contacts
  - Policy
    - induce local influence in electing a tribal representative

Outline



Motivation: Location-embedded social structure

Prior Work: Inferring Social Behaviors

Current Efforts: Protecting against social inferences

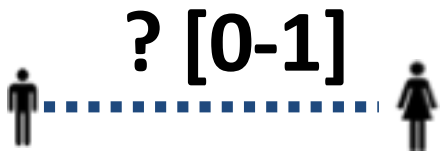
- But allow location disclosure

Open Problem: Protecting against location disclosure

- But allow social inferences

Real-World Social Strength - Intuition

Co-occurrence



From Real-World ***Co-occurrences*** to ***Social Strength***

Inferring friendship network structure by using mobile phone data (PNAS'09)

N. Eagle, A. Pentland, D. Lazer

- Study traces of 94 subjects using mobile phones
  - Subjects also reported their data: proximity and friendships
  - Analyzes proximity and friendships (inferred from recorded data) vs. ones that were self-reported by users

  - Conc-1: Two data sources is overlapping but distinct
  - Conc-2: Accurately infer 95% of friendships based on the observational data alone, where friend dyads demonstrate distinctive temporal and spatial patterns in their physical proximity and calling patterns.

Inferring social ties from geographic coincidences (in PNAS'10)

David J. Crandall, Lars Backstromb, Dan Cosleyc, Siddharth Surib, Daniel Huttenlocher, and Jon Kleinberg

- Probabilistic Model
  - Infer the probability of two people being friends given their co-occurrences in space and time
  - Does not consider the frequency of co-visit
  - Simplifies the social network: one connection for each person

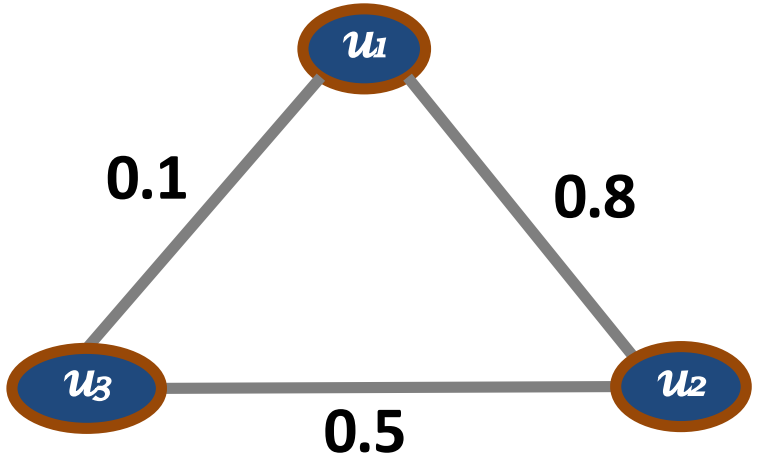
Bridging the Gap between Physical Location and Online Social Network (Ubicomp '10)

J. Cranshaw, E. Toch, J. Hong, A. Kittur, N. Sadeh

- Introduces a novel set of location based features for analyzing the social context of a geographical region
- **Location Entropy**: analyzes the context of the social interactions at that location: crowdedness and diversity
- **Regularity (Schedule_Entropy)**: High value reflects irregular movements, which produce high chance of making new friends
- Establishes a model of friendship in an online social network based on contextual features of co-locations

Example

	USC	SM Pier	Kodak Theater
			
(u_1, u_2)	4	3	2
(u_2, u_3)	2	2	1
(u_1, u_3)		5	



Problem Definition

Social strength is a quantitative measure that tells how socially close two people are.

Input: Users: $U = (u_1, u_2, \ldots, u_M)$

Locations : $L = (l_1, l_2, \ldots, l_N)$

Spatiotemporal records $< user\_id, location, time >: < u, l, t >$

Output: a weighted social graph where the weights of the edges define social strengths.

Challenges

1. What features of co-occurrences matter?
   - Richness?
   - Frequency?
   - Coincidences?
2. Location
   - Popularity?
   - Semantics?
3. Quantify friendships
   - Social Strength in between [0,1]

Baseline Solution - Richness

Counting the number of unique locations

Co-occurrence Vectors	Richness
C_{12}= (10, 1, 0, 0, 9)	**3**
C_{23}= (2, 3, 2, 2, 3)	**5**
C_{13}= (10, 0, 0, 0, 10)	**2**

✘ Ignore multiple co-occurrences @ same places

Baseline Solution - Frequency

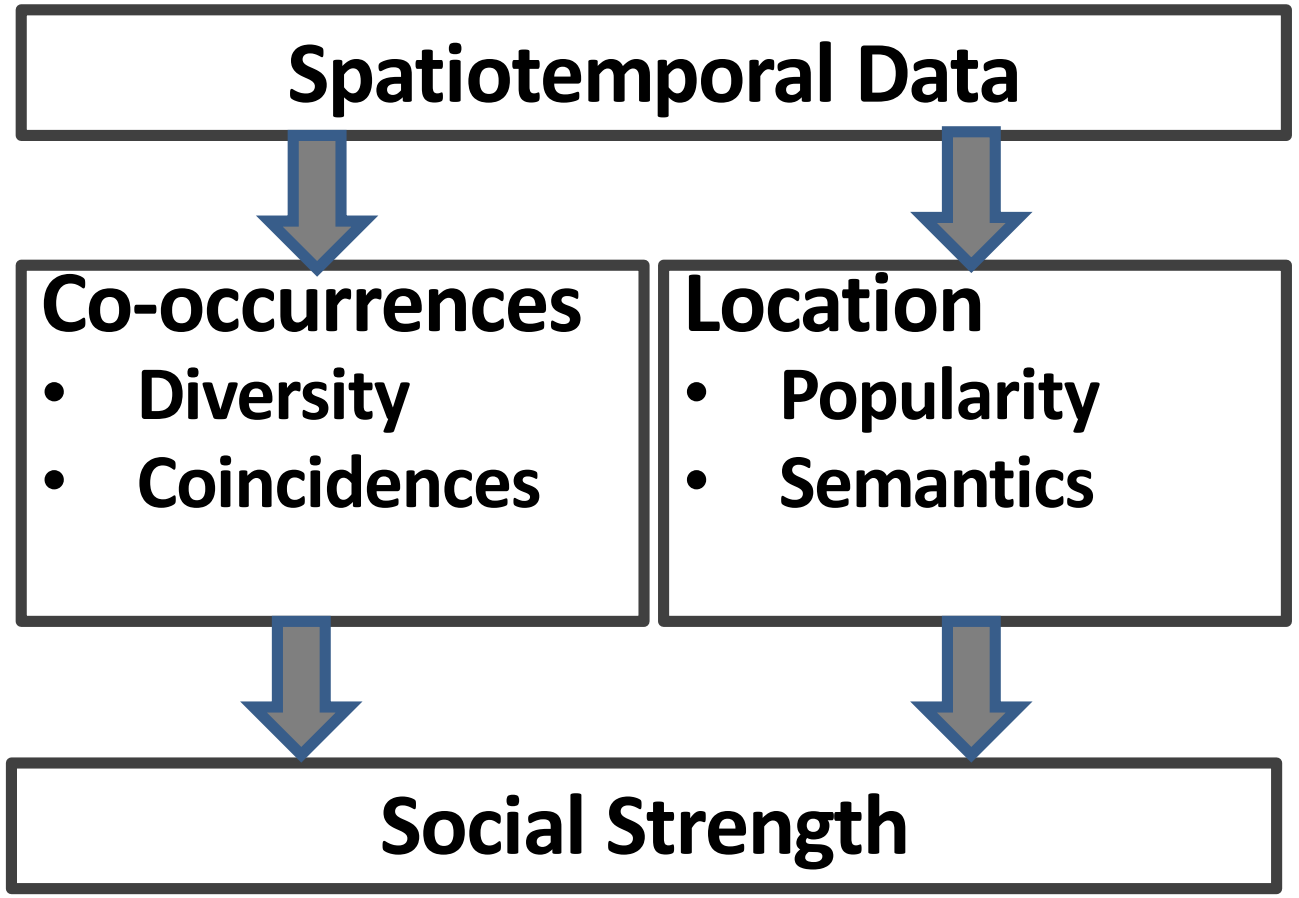
Counting the number of co-occurrences

Co-occurrence vectors	Frequency
C_{13}= (10, 1, 0, 0, 9)	**20**
C_{23}= (2, 3, 2, 2, 3)	**13**
C_{31}= (10, 0, 0, 0, 10)	**20**

✓ Captures local frequency

✗ Cannot capture the diversity of co-occurrences

EBM Model [SIGMOD'13]



Shannon Entropy

$$H_{ij}^{S} = -\sum_{l} P_{ij}^{l} \log P_{ij}^{l}$$

- If we select a random location, how predictable is whether i and j co-occurred there?
- More diverse places they co-occurred → Low predictability → High entropy

Co-occurrence vectors	H_{ij}^{s}
C_{12}= (10, 1, 0, 0, 9)	**0.86**
C_{23}= (2, 3, 2, 2, 3)	**1.59**
C_{13}= (10, 0, 0, 0, 10)	**0.69**

✓ The more locations, the higher entropy.

✓ The more diverse, the higher entropy.

✘ No control on diversity vs. frequency, e.g., may put too much weight on outliers (coincidences)

Rényi Entropy

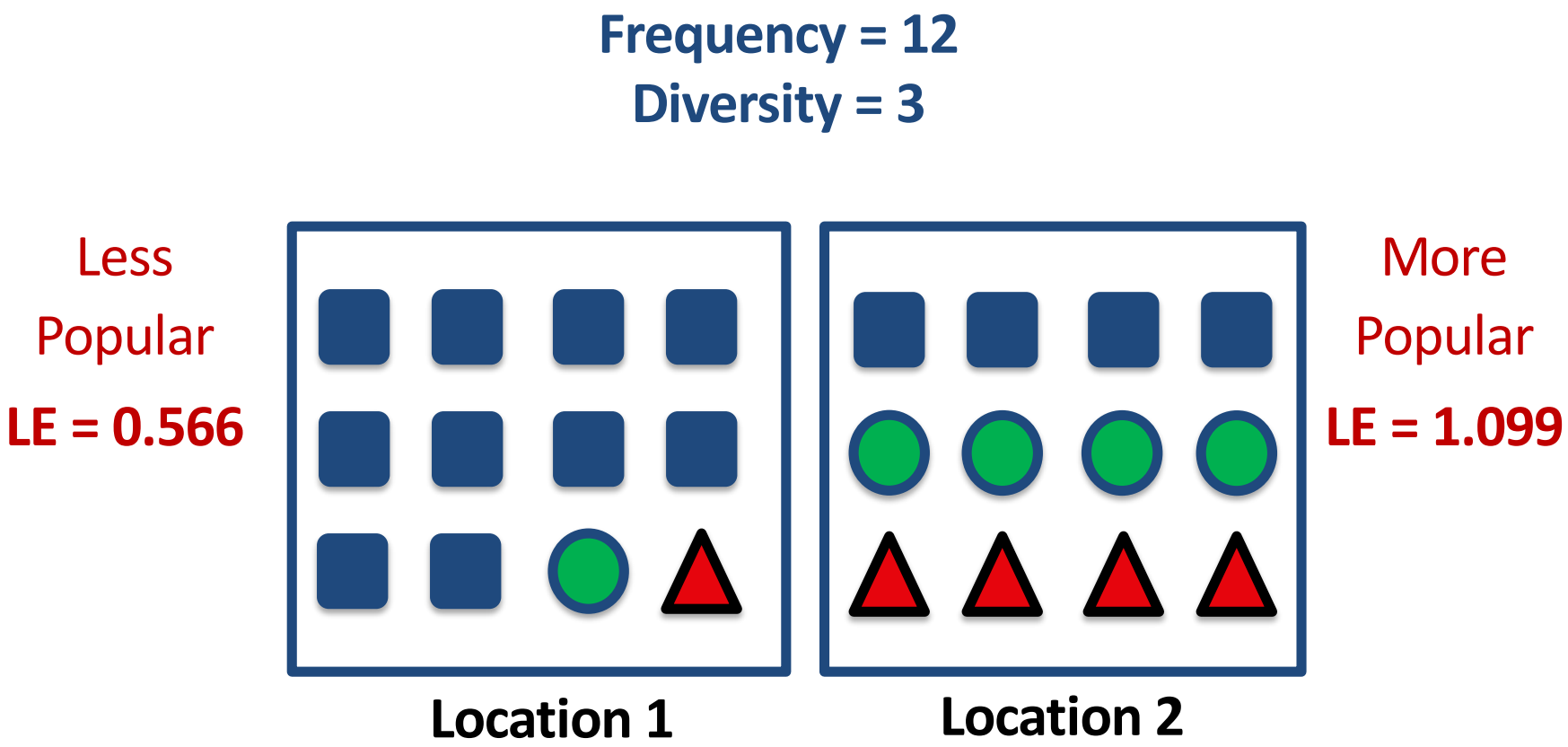
We want to control the impact of diversity vs. frequency

$$H_{ij}^{R} = \left(-\log \sum_{l} \left(P_{ij}^{l} \right)^{q} \right) / (q-1)$$

Order of diversity

- $q > 1$ – Renyi entropy more favorably considers high local frequencies. (less diversity)
  - ✓ Captures the diversity of co-occurrences.
- $q < 1$ – in *opposite*, it gives more weight to low local frequencies.
  - ✓ Limits impact of coincidences (outliers).
- $q = 1$ – Renyi entropy is *undefined*, but its limit exists and becomes ***Shannon*** entropy, where it is unbiased.
  - ✗ Still considers all locations equally important. We need to consider:
    - Location popularity
- q = 0 the entropy is *insensitive* to local frequencies ⇔ giving pure number of unique locations – ***richness***.

Location Entropy for Location Popularity



Location Entropy (LE)

$$H_l = -\sum_{u, P_{u,l} \neq 0} P_{u,l} \log P_{u,l}$$

- LE indicates the popularity of a location *Cranshaw, J., et al., (2010). Bridging the gap between physical locations and online social networks. UBICOMP, 119-128.*
- The more popular, the higher entropy, and vice versa
- LE captures how diverse the visitors of a *location* are
  - E.g., your home is not diverse as only 2-4 users visited there; Eifel tower is the opposite
- Pick a random visit v at location l; high entropy means:
  - less predictable who made v
  - *The location has more diverse set of visitors*

The Entropy Based Model (EBM)

- Renyi Entropy

$$H_{ij}^{R} = \left(-\log \sum_{l} \left(P_{ij}^{l} \right)^{q} \right) / (q-1)$$

(How often i and j meet in how diverse of locations)

- Location Entropy

$$H_l = -\sum_{u, P_{u,l} \neq 0} P_{u,l} \log P_{u,l}$$

(How popular a location is)

- Weighted Frequency

$$F_{ij} = \sum_{l} c_{ij,l} \times \exp(-H_l)$$

(More weights to meetings in unpopular locations)

- Social Strength

$$s_{ij} = \alpha . \exp\left(H_{ij}^{R}\right) + \beta . \sum c_{ij}^{l} \times \exp(-H^{l}) + \gamma$$

Social Strength (EBM model)

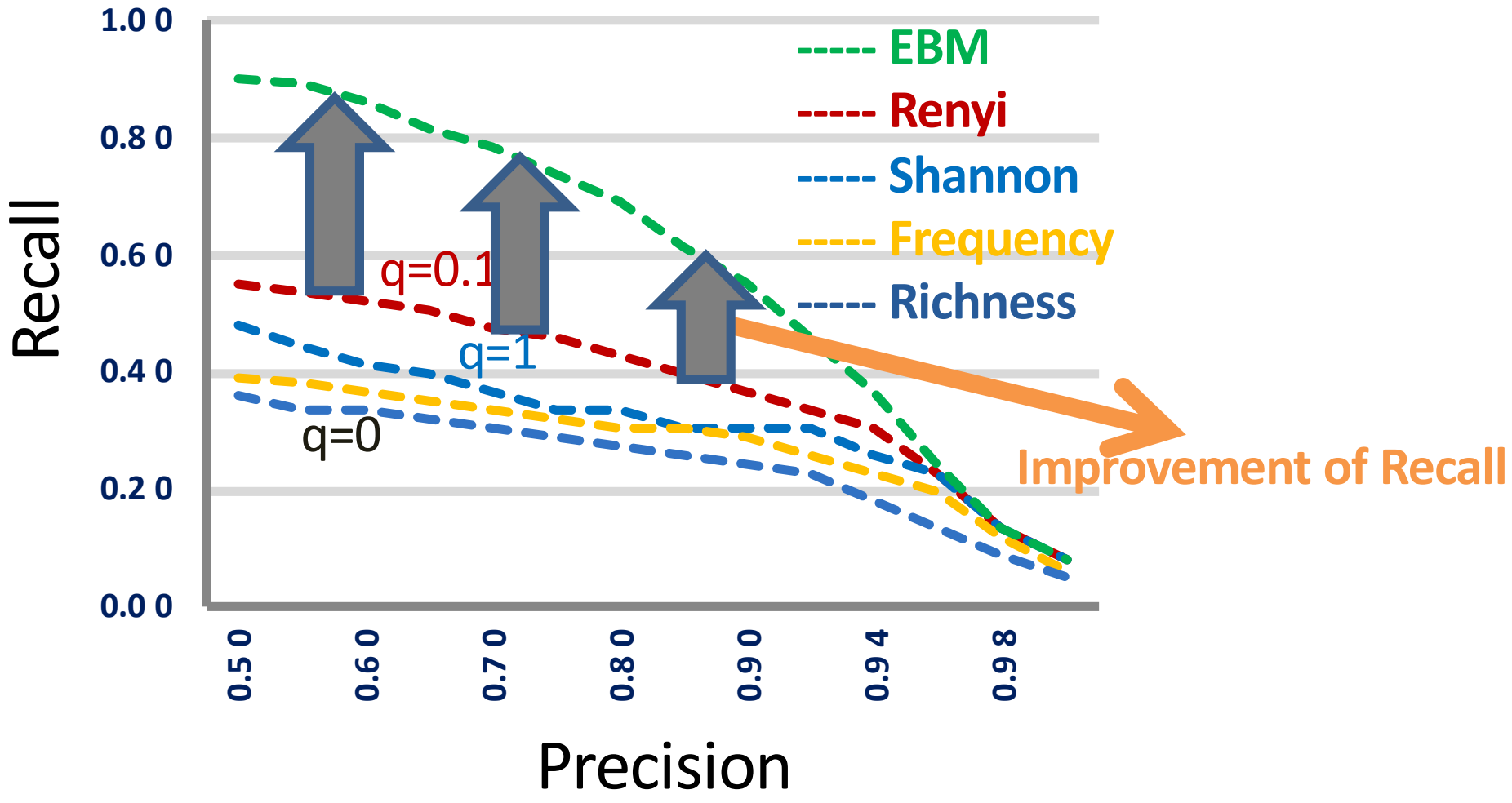
$$s_{ij} = \alpha . \exp\left(H_{ij}^{R}\right) + \beta . \sum c_{ij}^{l} \times \exp(-H^{l}) + \gamma$$

where parameter α, β and *ϒ can be learned from training data.*

Have addressed all the challenges mentioned earlier.

- ✓ Eliminate the impact of coincidences.
- ✓ Take into account the impact of locations.
- ✓ Data Sparseness.

Comparison of Various Social Strength Measures



Privacy Twist

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)

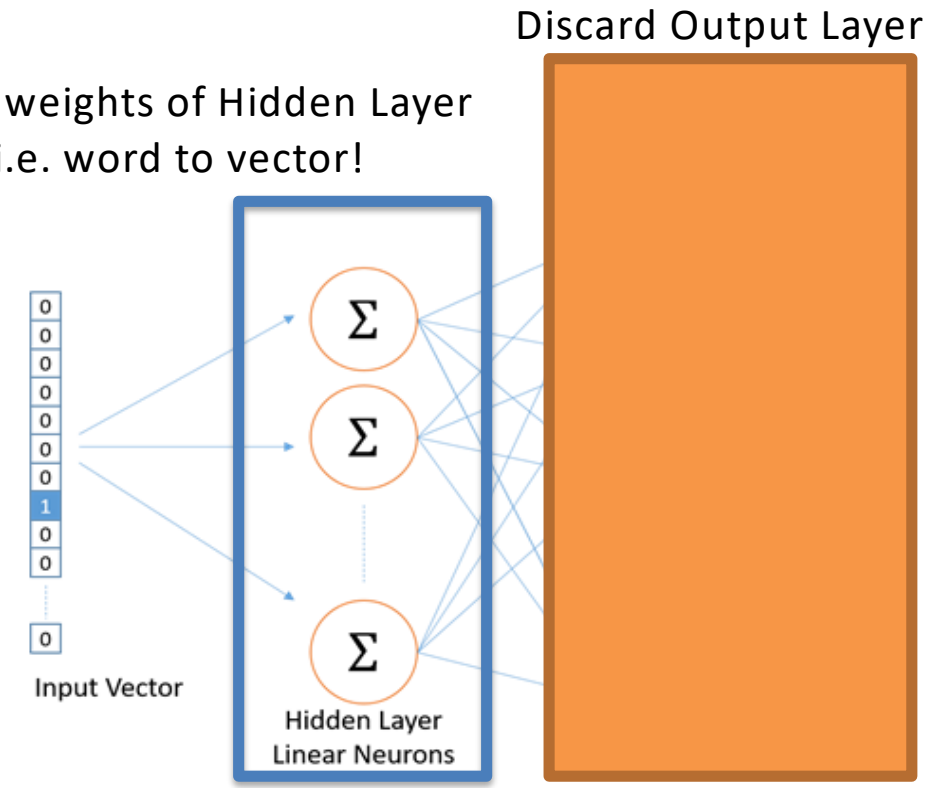
A glance at the Skip-Gram Model

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.

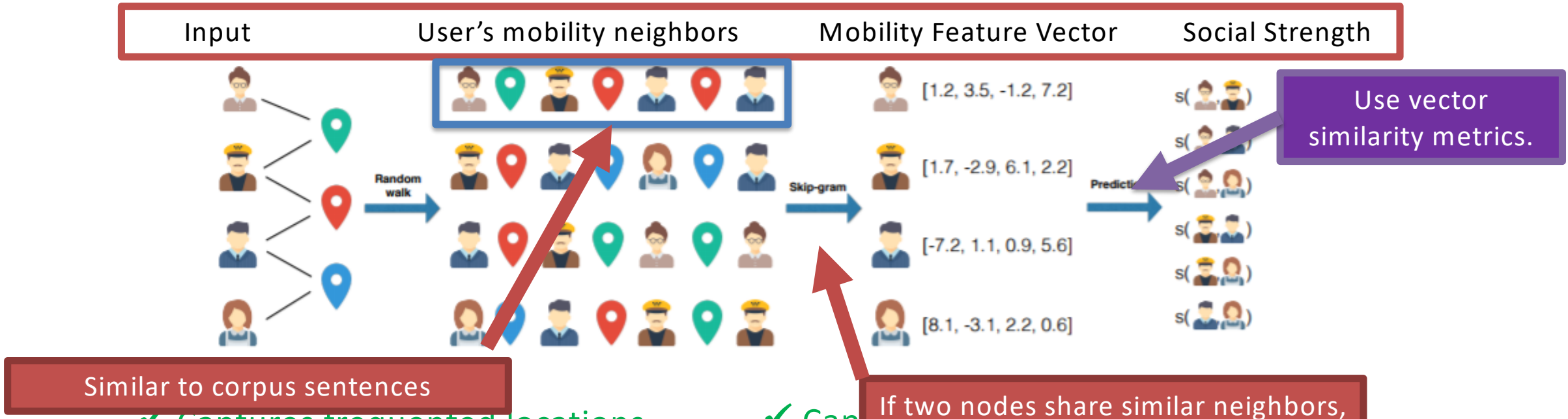
Corpus training (NN)

The quick brown fox jumps over the lazy dog.

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



walk2friends: Extending to locations based networks.



✓ Captures frequented locations.

✓ Cap

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

✗ 3-5% worse on sparse datasets.

Outline

Motivation: Location-embedded social structure

Prior Work: Inferring Social Behaviors

Current Efforts: Protecting against social inferences

- But allow location disclosure

Open Problem: Protecting against location disclosure

- But allow social inferences

Co-Location Privacy Risks

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

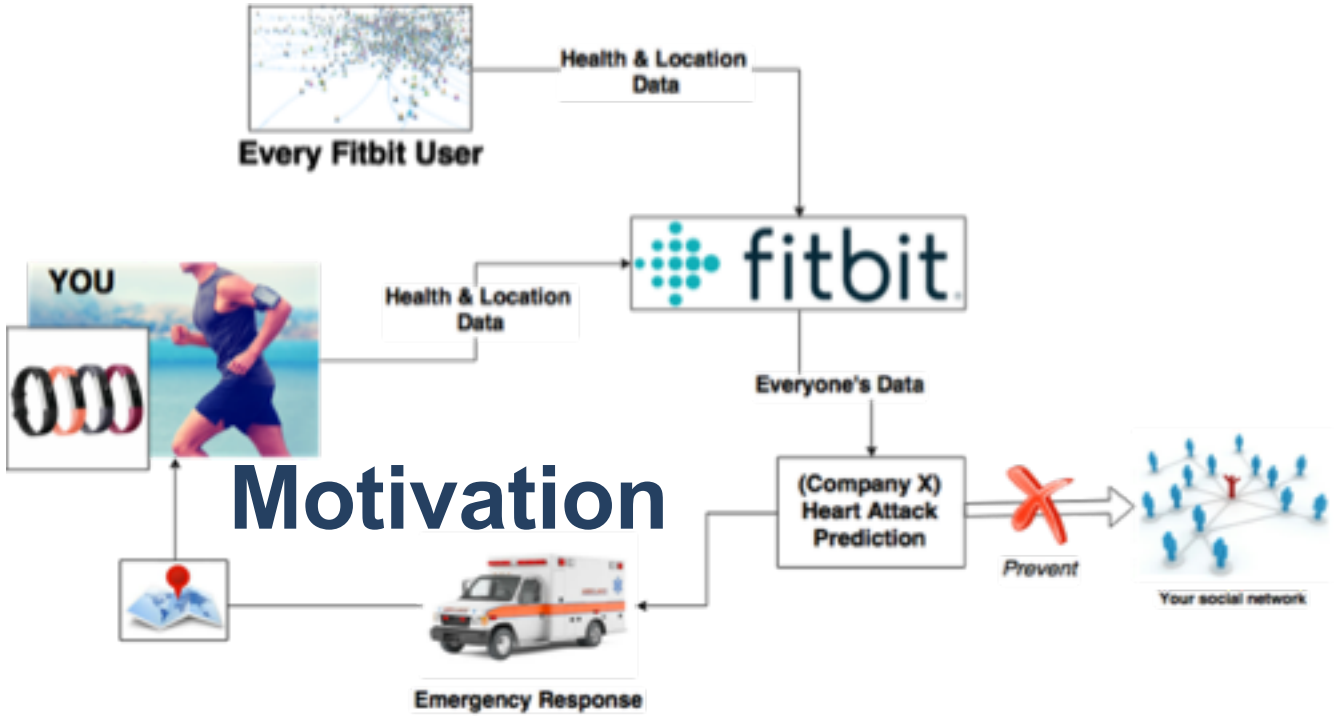
2. NSA's Co-Traveler program (exposed 2013):
   Identifies unknown associates of a known target.

[Source: Washington Post]

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.

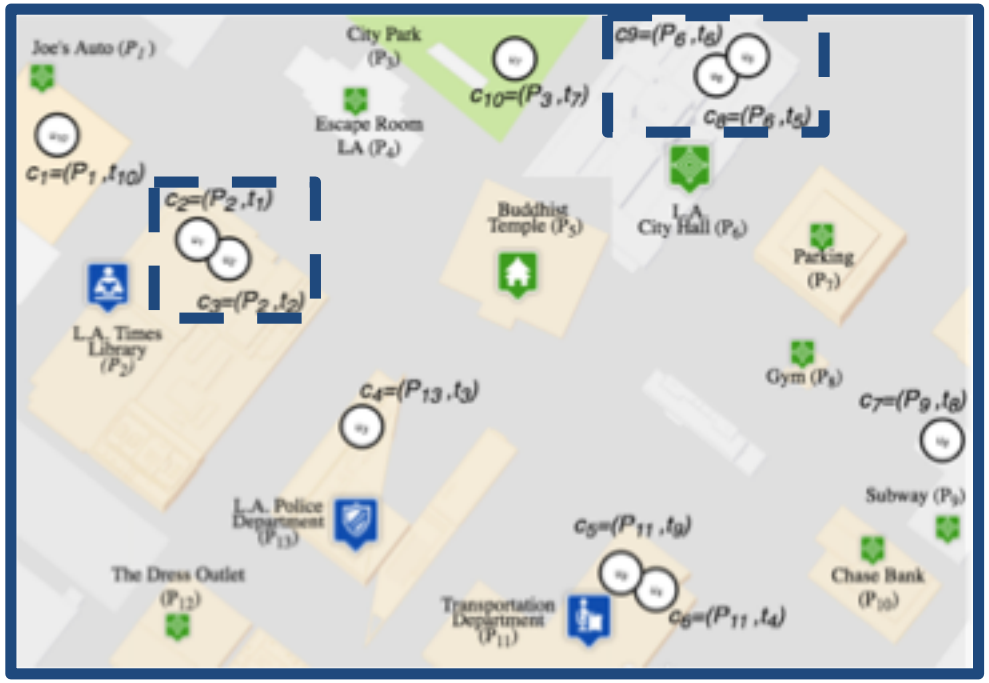


Enable LBS to provide recommendation, advertisement, and other services.

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,

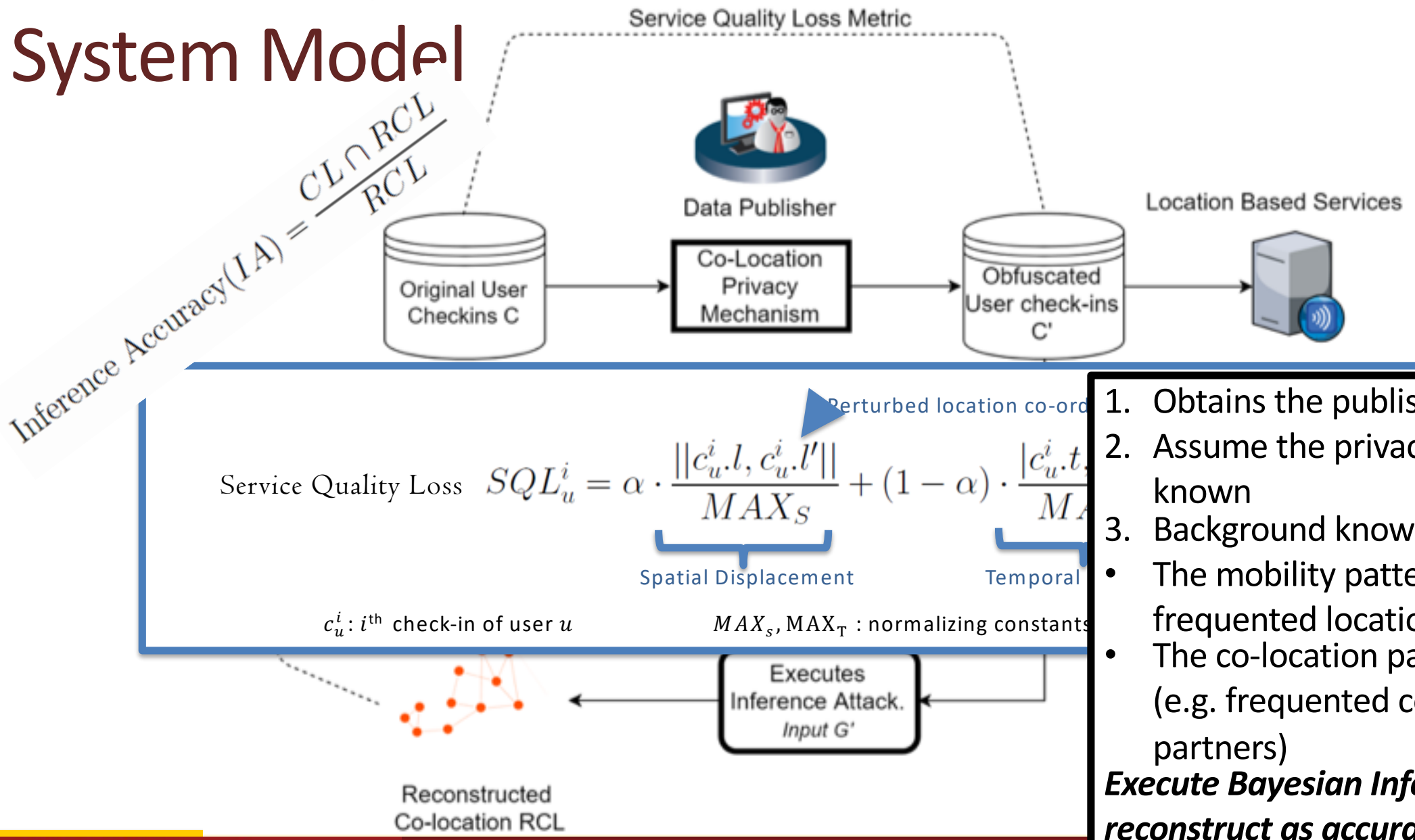
- Assume buildings are points

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

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

Δ_s and Δ_t are application specific.

System Model



$$\text{Inference Accuracy}(IA) = \frac{CL \cap RCL}{RCL}$$

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.

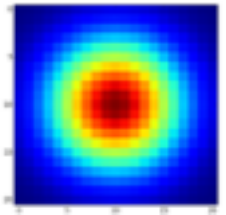
Co-Location Privacy Mechanism 1: Gaussian Perturbation (Naïve)

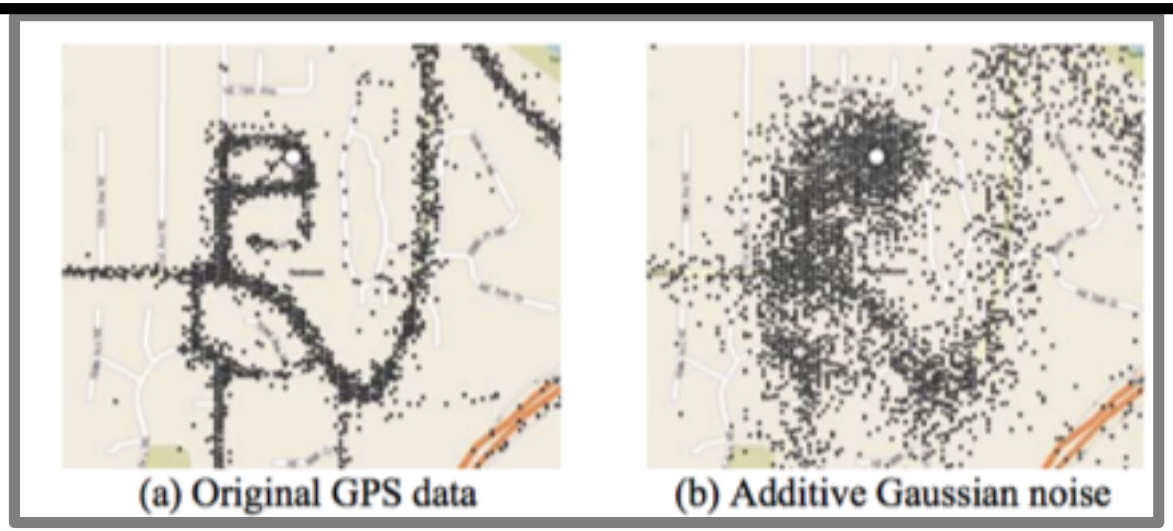
Most popular methods in statistical data privacy.

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



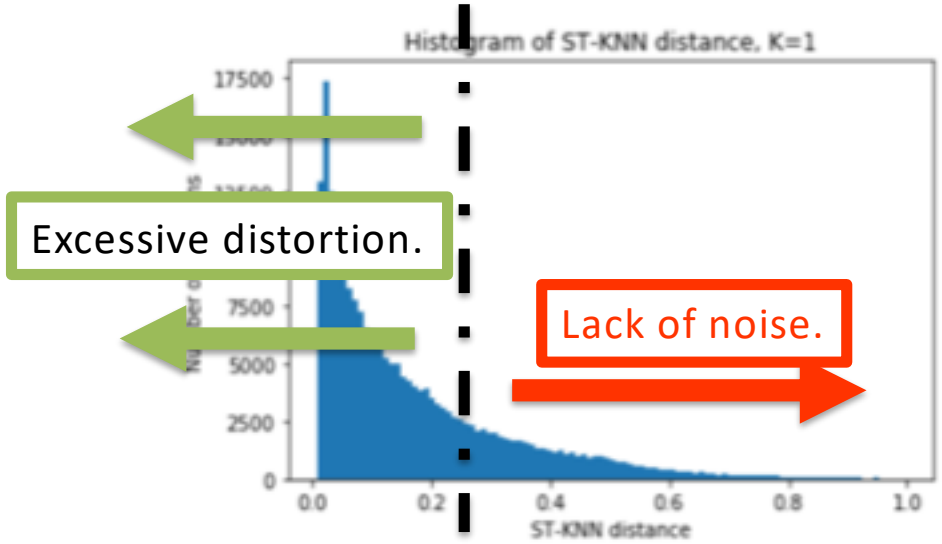


Krumm, [PerCom'07]

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.



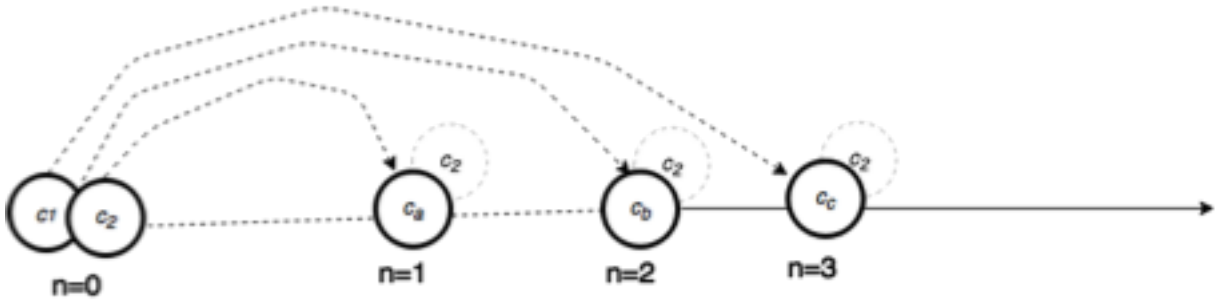
$ST_{dist} = 0.1 = 100m + 40\ min$

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 spatio-temporal 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.



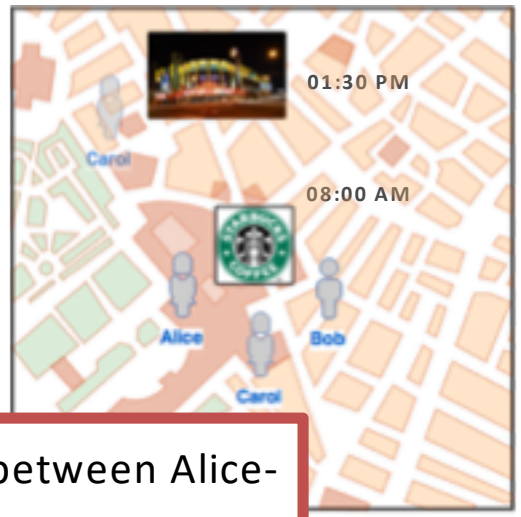
Move c2 to any of 'b=4' positions at random

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

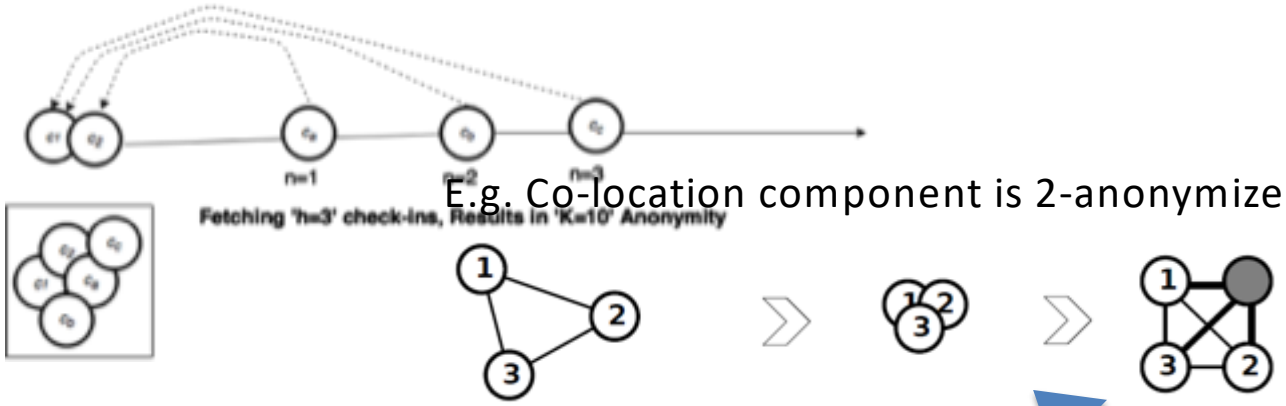
Co-Location Privacy Mechanism 3: Co-Location K-Anonymity

Definition: A co-location is k-anonymous if it is spatio-temporally 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.



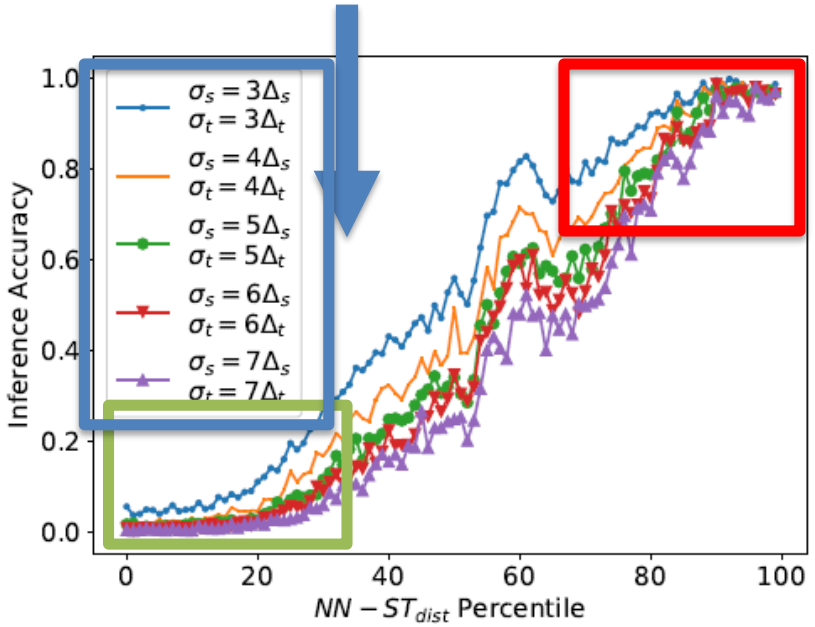
E.g. Co-location component is 2-anonymized

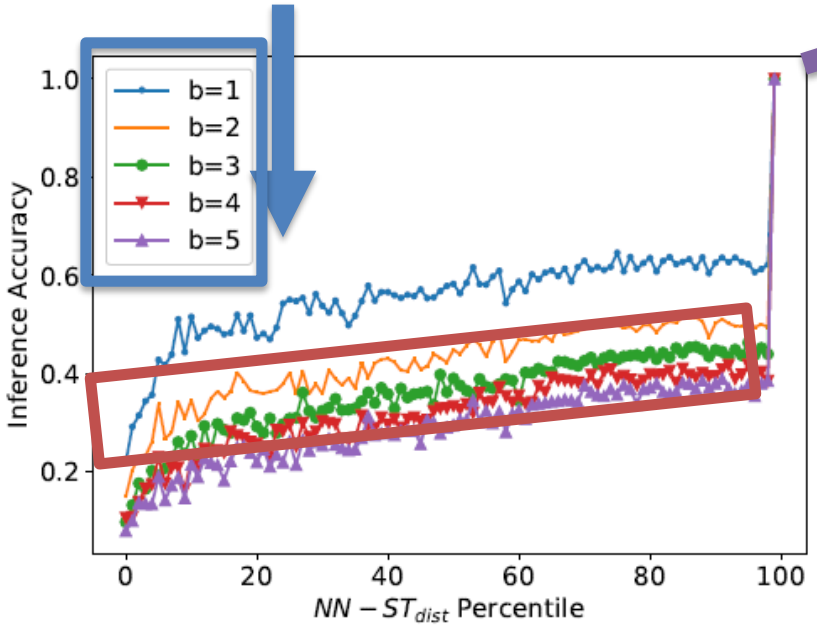
On seeing any co-location the adversary can only tell it's truthfulness with a certainty of $1/2$ $(i.e. 1/k)$.

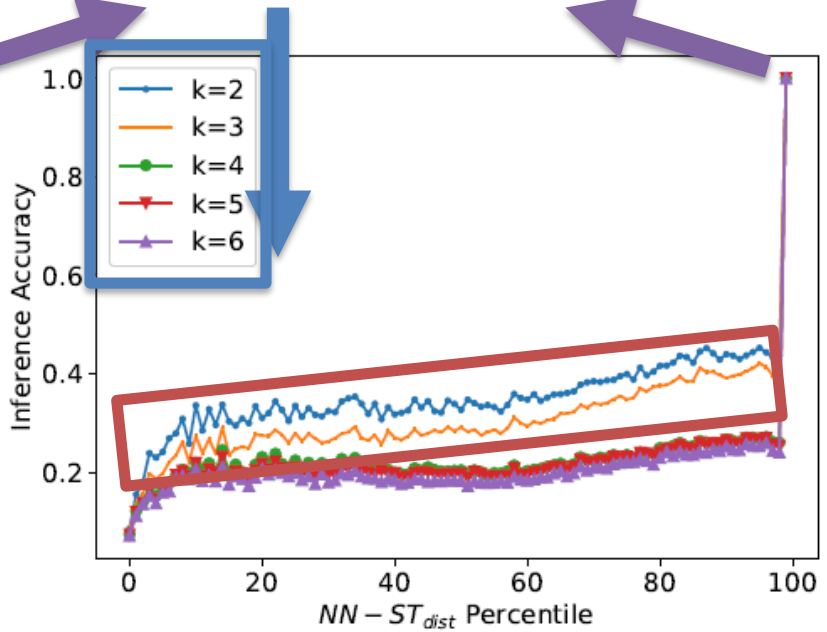
Attack Accuracy on Privacy Mecha

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

Increasing level of distortion.





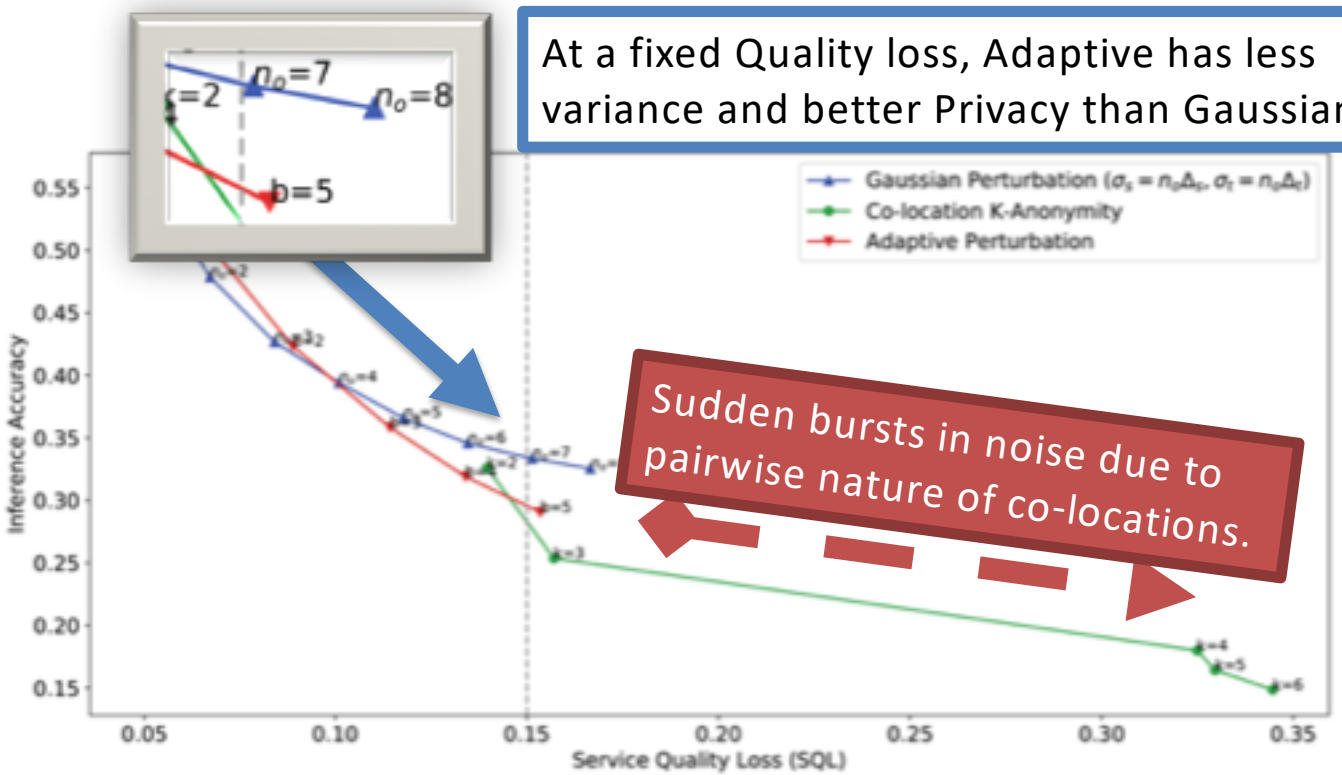


Dense → Sparse

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

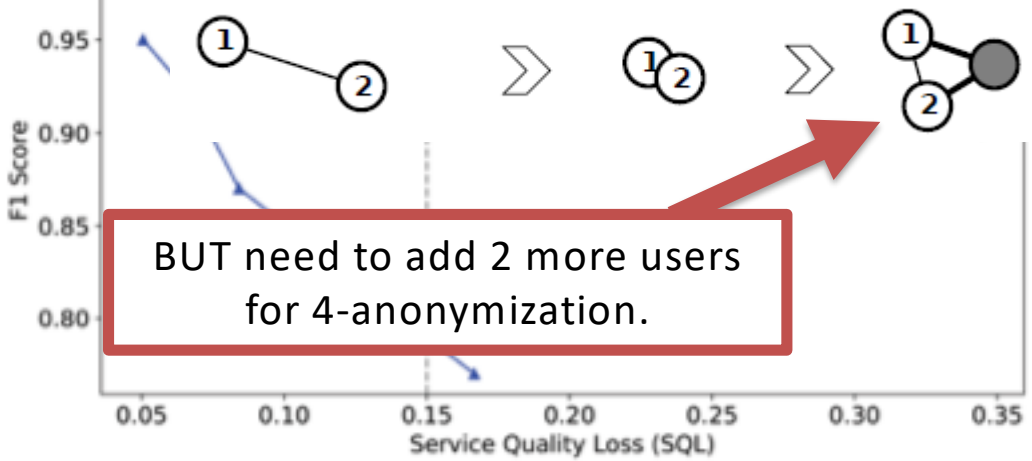


At a fixed Quality loss, Adaptive has less variance and better Privacy than Gaussian.

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.

Outline

Motivation: Location-embedded social structure

Prior Work: Inferring Social Behaviors

Current Efforts: Protecting against social inferences

- But allow location disclosure

Open Problem: Protecting against location disclosure

- But allow social inferences

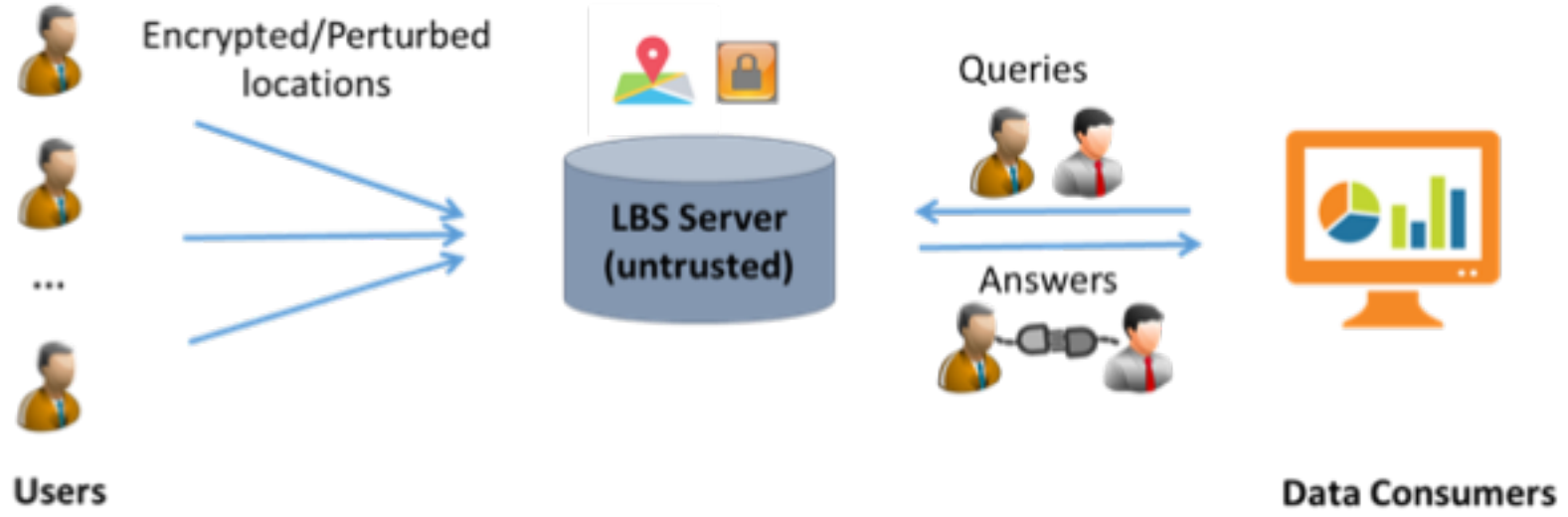
Two Sides of the Coin

Privacy-Preserving Social Inferecne

Criminology
- identify the new or unknown members of a criminal gang or a terrorist cell

Epidemiology
- spread of diseases through human contacts

Policy
- induce local influence in electing a tribal representative



Q&A

Thanks!

References

- [KDD'03] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '03).
- [PNAS'10] David J. Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. "Inferring social ties from geographic coincidences." PNAS 2010
- [KDD'10] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. 2010. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '10).
- [Ubi'10] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, Norman Sadeh."Bridging the Gap Between Physical Location and Online Social Networks." Ubicom 2010
- [VLDB'12] H. Shirani-Mehr, F. Banaei-Kashani, and C. Shahabi. "Efficient reachability query evaluation in large spatiotemporal contact datasets." Proc. VLDB Endow, 5(9), May 2012
- [SIGMOD'13] Pham, Huy, Cyrus Shahabi, and Yan Liu. "Ebm: an entropy-based model to infer social strength from spatiotemporal data." Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. ACM, 2013
- [Nature'13] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. "Unique in the Crowd: The privacy bounds of human mobility." Scientic Reports, 3, Mar.2013
- [ICDE-Bulletin'15] Cyrus Shahabi, Huy Pham: Inferring Real-World Relationships from Spatiotemporal Data. IEEE Data Eng. Bull. 38(2): 14-26 (2015)
- [CNS'15] B. Wang, M. Li, H. Wang, and H. Li. "Circular range search on encrypted spatial data." In IEEE CNS, 2015
- [CCS'15] Y. Xiao and L. Xiong. "Protecting locations with differential privacy under temporal correlations." In CCS, 2015
- [PerCom'07] Krumm, John. "Inference attacks on location tracks." *Pervasive computing* (2007): 127-143.
- [ACM TOPS'17] Argyros, George, et al. "Evaluating the Privacy Guarantees of Location Proximity Services." *ACM Transactions on Privacy and Security (TOPS)* 19.4 (2017): 12.
- [RSLB 2003] K. T. Eames and M. J. Keeling, "Contact tracing and disease control," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 270, no. 1533, pp. 2565–2571, 2003.

Reference (other)

- Chen et al., "Scalable influence maximization for prevalent viral marketing in large-scale social networks", ACM SIGKDD, 2010
- Goyal et al., "A data-based approach to social influence maximization", VLDB, 2011
- Kempe et al., “Maximizing the spread of influence through a social network”, ACM SIGKDD, 2003
- Leskovec et al., “Cost-effective outbreak detection in networks”, ACM SIGKDD, 2007
- Saito et al., "Prediction of information diffusion probabilities for independent cascade model" Knowledge-Based Intelligent Information and Engineering Systems, Springer, 2008.
- Goyal et al., “Learning influence probabilities in social networks” ACM WSDM, 2010
- Cho et al., “Friendship and mobility: user movement in location-based social networks”, ACM SIGKDD, 2011
- Zhang et al., “Understanding spatial homophily: the case of peer influence and social selection”, WWW, 2014
- Cranshaw et al., “Bridging the gap between physical location and online social networks”, ACM UbiComp, 2010
- Liben-Nowell et al., “The link-prediction problem for social networks”, Journal of the ASIST, 2007
- Pham et al., "Ebm: an entropy-based model to infer social strength from spatiotemporal data", ACM SIGMOD, 2013