# Anonymization and De-anonymization of Mobility Trajectories: Dissecting the Gaps between Theory and Practice

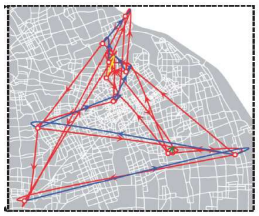
**Huandong Wang**

Tsinghua University

Aug. 17, 2019

# Increasing Concern on Privacy/Security

- **Anonymized user trajectories are increasingly collected by ISPs**
  - High research and business value
- **Growing privacy concern**
  - ISPs are motivated to monetize or share user trajectory data
- **De-anonymization attack**
  - How likely users can be de-anonymized in the shared ISP trajectory dataset?



Now Those Privacy Rules Are Gone, This Is How ISPs Will Actually Sell Your Personal Data

Thomas Fox-Brewster, FORBES STAFF

*I cover crime, privacy and security in digital and physical forms.* FULL BIO


# De-anonymization Attack: Theory and Practice

- **Appalling Theoretical Privacy Bound**
  - 4 location points uniquely re-identify 95% users [Scientific Report 2013]

**Is this true in practice?**

- **Practical Challenge: Lack of large real-world *ground-truth* datasets**
  - Small datasets
    - ✓1717 users in [WWW 2016]
  - Synthetized datasets
    - ✓Parts of the same dataset [TON 2011]

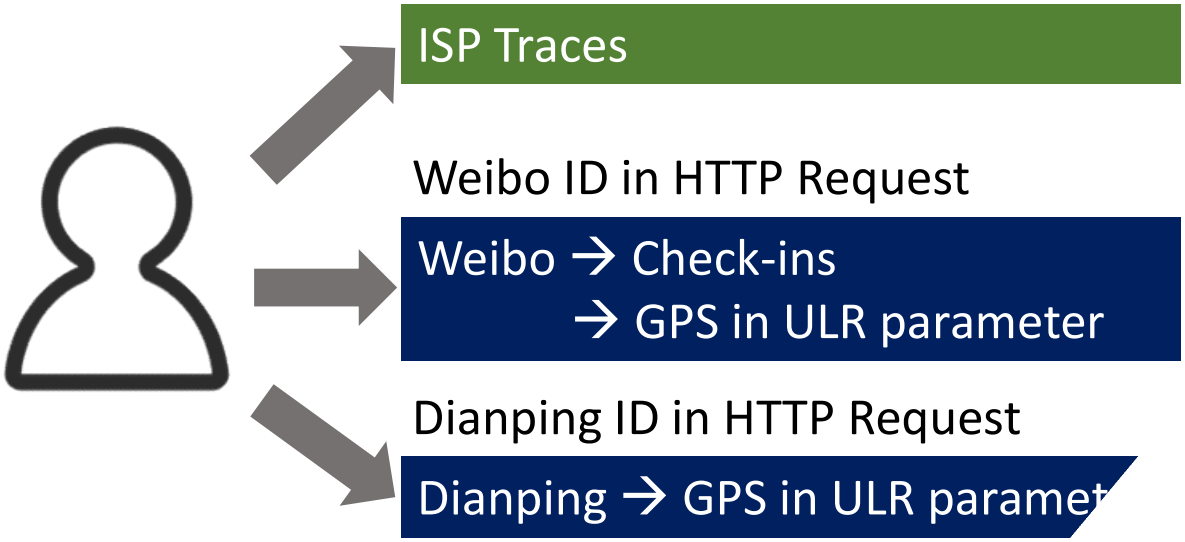
# Our Approach: Collect Three Real-world Ground-truth Datasets

**Ground-Truth: Traces from the same set of users**

| Dataset | Total# Users | Total# Records |
|---|---|---|
| ISP | 2,161,500 | 134,033,750 |
| Weibo App-level | 56,683 | 239,289 |
| Weibo Check-in (Historical) | 10,750 | 141,131 |
| Weibo Check-in (One-week) | 506 | 873 |
| Dianping App-level | 45,790 | 107,543 |


- **ISP Dataset**
  - Shanghai, 4/19-4/26, 2016 (victim dataset)
  - 2 million users
  - Access logs to cellular tower → Location traces
- **Weibo Dataset:** One of the largest social networks in China (external information)
- **Dianping Dataset:** "Chinese Yelp" (external information)

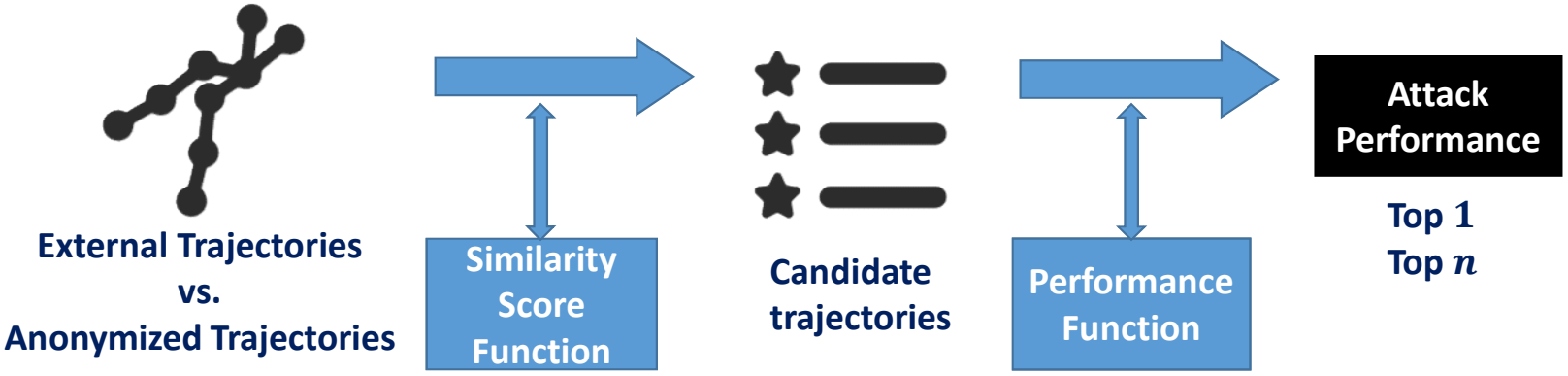
# How to Obtain the Ground-Truth?




**Ethical approval obtained from We**

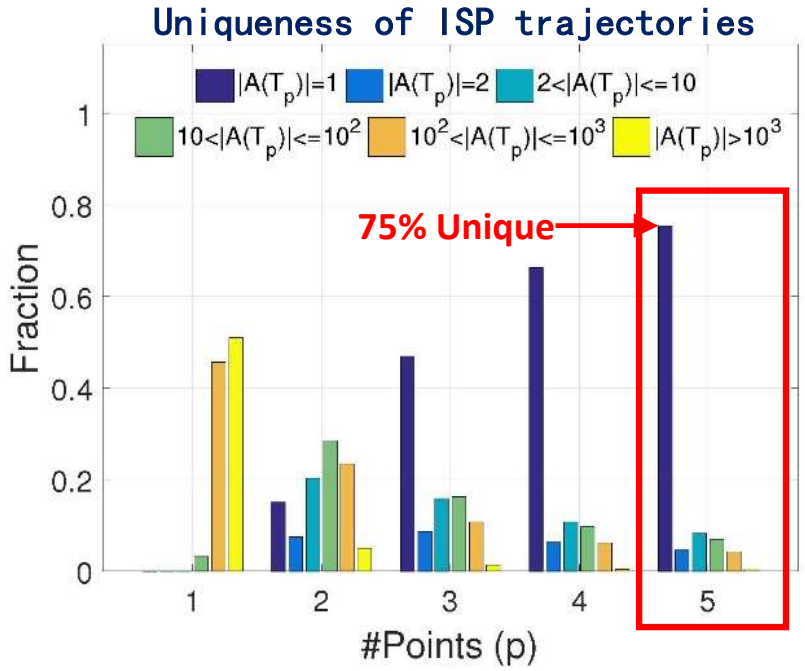
# De-anonymization Attack: Threat Model

- **Anonymized Trajectory Data Published by ISP**
  - Anonymization: Replace user identity with the pseudonym
  - Obfuscation: Perturbation, Location hiding
- **Adversary**
  - Match the anonymized traces (e.g., ISP traces) and external traces (e.g., Weibo/Dianping traces)
  - Social network has PII → real-world identifier



# De-anonymization: Theoretical Bound based on Uniqueness

- Number of points sufficient to uniquely identify a trajectory
- $T_p$: Randomly sampled *p* points
- $A(T_p)$: find all trajectories containing the *p* points of $T_p$
- **Uniqueness**: $|A(T_p)| = 1$?




**5 points are sufficient to uniquely identify 75% trajectories!**

**High potential risk of trajectories to be de-anonymized!**

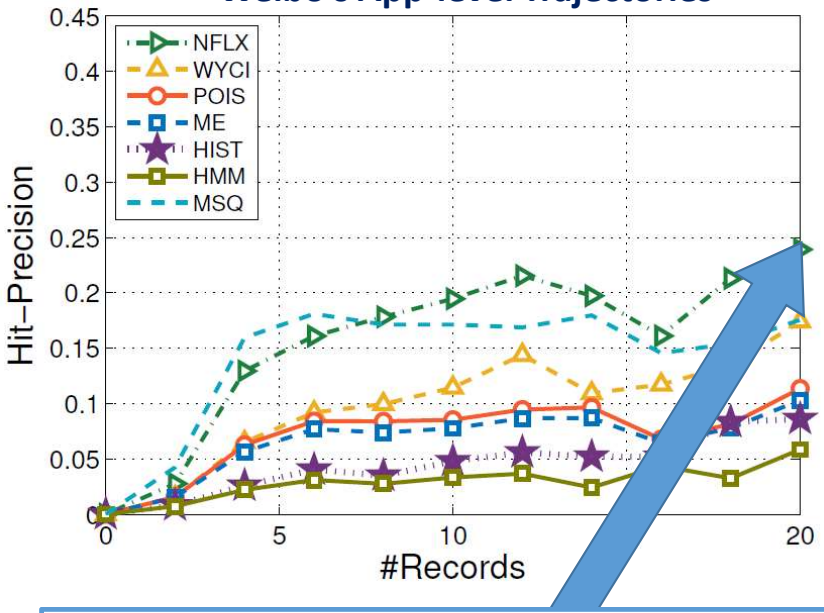
# De-anonymization Attack: Actual Performance

**Implement 7 state-of-the-art algorithms**

- **"Encountering" event**
  - **POIS** [WWW 2016]
  - **ME** [AIHC 2016]
- **Individual user's mobility patterns**
  - **HMM** [IEEE SP 2011]
  - **WYCI** [WOSN 2014]
  - **HIST** [TIFS 2016]
- **Tolerating temporal/spatial mismatches**
  - **NFLX** [IEEE SP 2008]
  - **MSQ** [TON 2013]

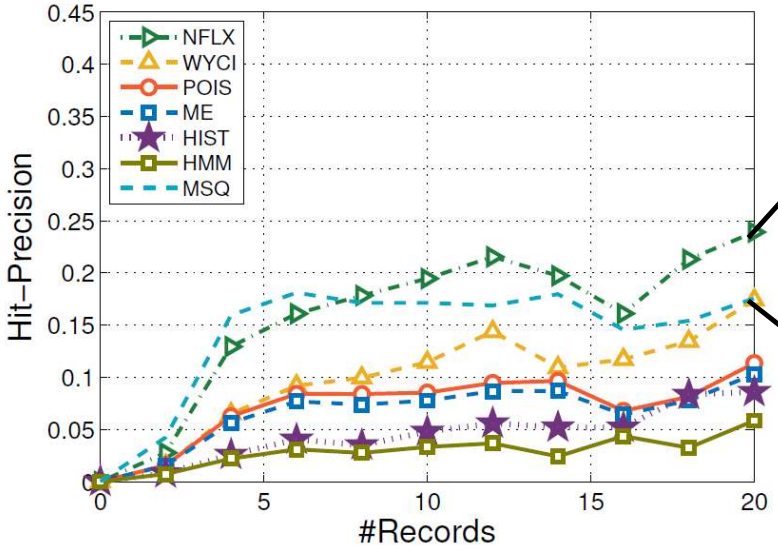
**Hit-precision** $h(x) = \begin{cases} \frac{k-(x-1)}{k}, & \text{if } k \geq x \geq 1, \\ 0, & \text{if } x > k. \end{cases}$

**Actual Performance Based on Weibo's App-level Trajectories**




**Maximum hit-precision is only 25% !**
**Far from the privacy bound !**

# Reasons Behind Underperformance




**Algorithms with best performance**

**NFLX** [IEEE SP 2008]

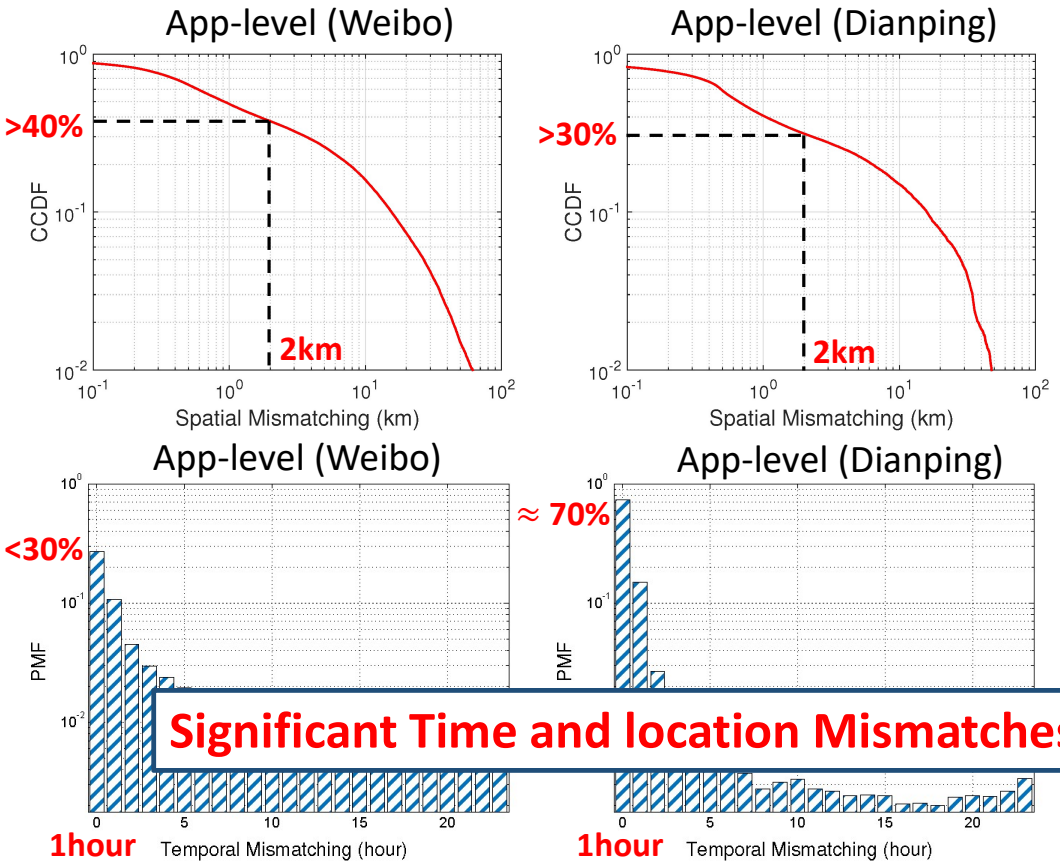
- Similarity function
  - Minimum time gap between users' visits to the same location
- **Tolerate temporal mismatches**

**MSQ** [TON 2013]

- Similarity function
  - Square root of distance between trajectories
- **Tolerate spatial mismatches**

**Existing algorithms tolerating spatio-temporal mismatches have the best performance**

# Reasons Behind Underperformance: Large Spatio-Temporal Mismatches




Spatial mismatches of over 40% records $\geq$ 2km

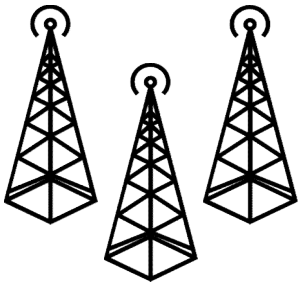
Temporal mismatches of over 30% records

**Significant Time and location Mismatches between Different Datasets!**

# Potential Reasons behind the Mismatches

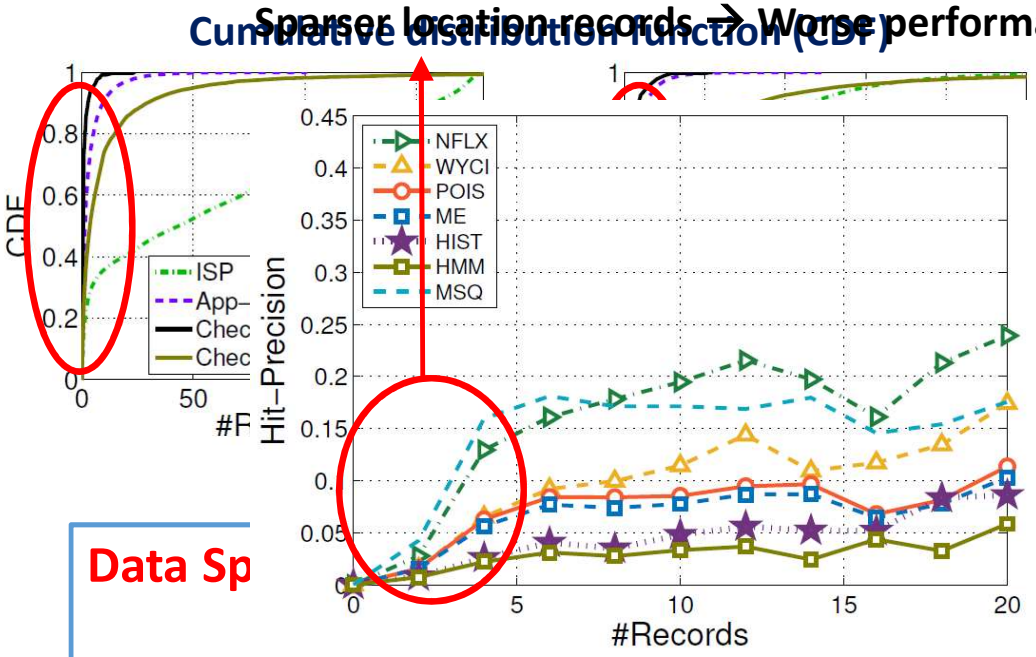
- **GPS errors**
  - GPS unreachable locations (Indoor, underground)
  - Lazy GPS updating mechanisms [UbiComp 2007]

- **Deployment of base stations**
  - Lower density → larger mismatches



- **User behavior**
  - 39.9% remote (fake) check-ins [ICWSM 2016]
  - Earn virtual rewords, compete with their friends

# Reasons Behind Underperformance: Data Sparsity



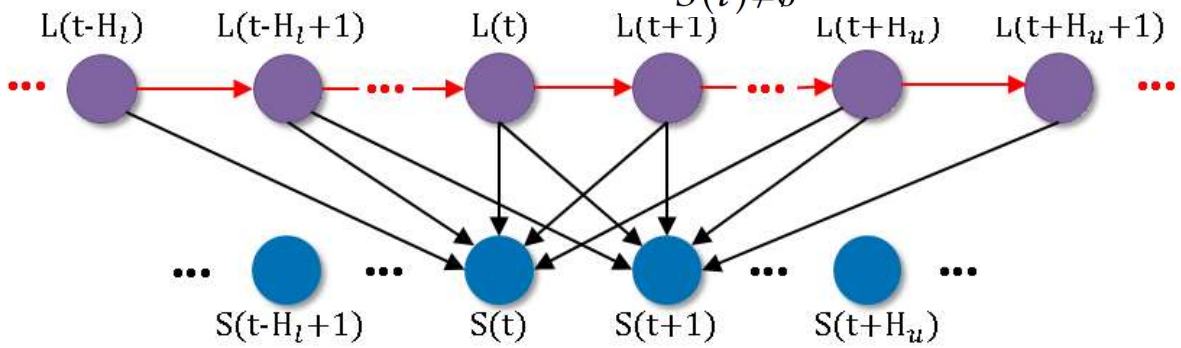

The vast majority of users have sparse location records!

Data Sp ! ing!

# Can we bridge this gap?

# Our De-anonymization Method

$$D_{\mathrm{GM}}(\boldsymbol{S},\boldsymbol{L}) = \log p(\boldsymbol{S}|\boldsymbol{L}) = \prod_{S(t)\neq\emptyset} p(S(t)|\boldsymbol{L}).$$




■**1) Modelling Spatio-Temporal Mismatches**: Gaussian Mixture Model (GMM)

$$P(S(t)|L) = \sum_{p=-H_l}^{H_u} \pi(p) \cdot \mathcal{N}(S(t)|L(t-p), \sigma^2(p))$$

➢Parameters chosen by empirical values or estimated by EM algorithm

■**2) Modelling Users' Mobility Pattern**: Markov Model

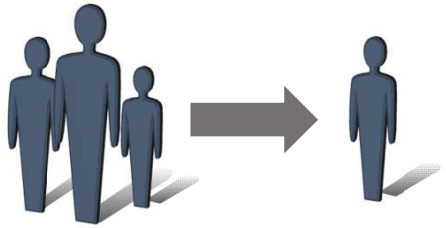
➢Solving the **data sparsity** issue: rare "encountering" event

➢Missing locations are estimated by Markov Model

# Our De-anonymization Method

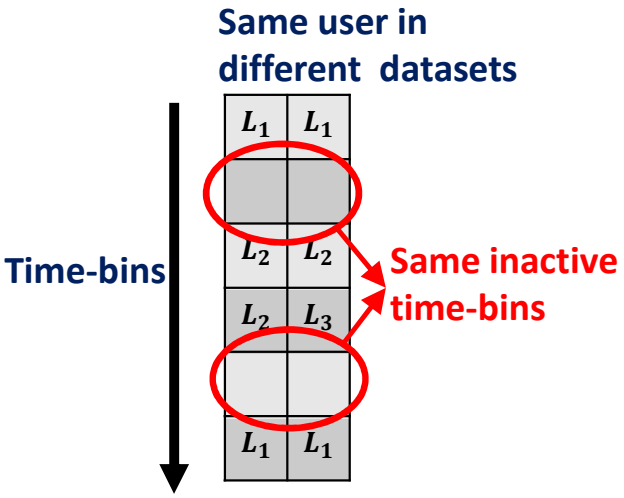
## ■3) Use Location Context

- ➢Solve the **data sparsity** issue
- ➢Use aggregated user behavior at locations
- ➢To infer individual user behavior (location transition probability)



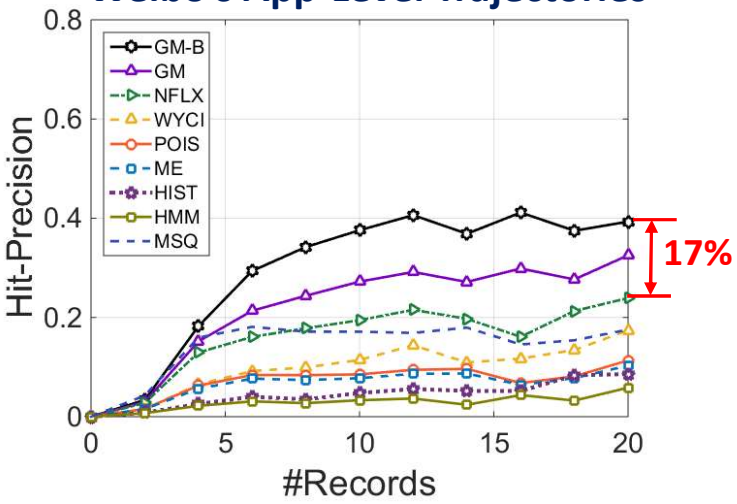
## ■4) Use Time Context

- ➢"Whether the user is active" is helpful
- ➢Modelling user inactive period (previously ignored feature)

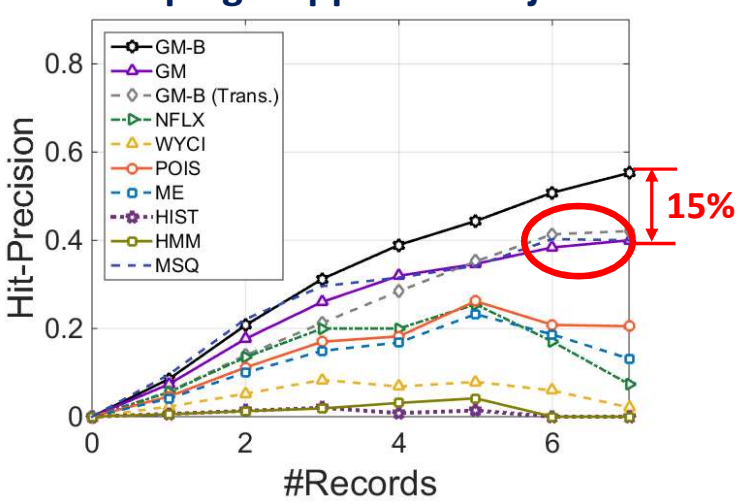


# Our De-anonymization Method: Performance Evaluation

**Weibo's App-Level Trajectories**




**Dianping's App-Level Trajectories**




- 7 state-of-the-art algorithms
- Our proposed algorithm: **GM-B**, **GM**
- Transferred parameters: GM-B (Trans.)

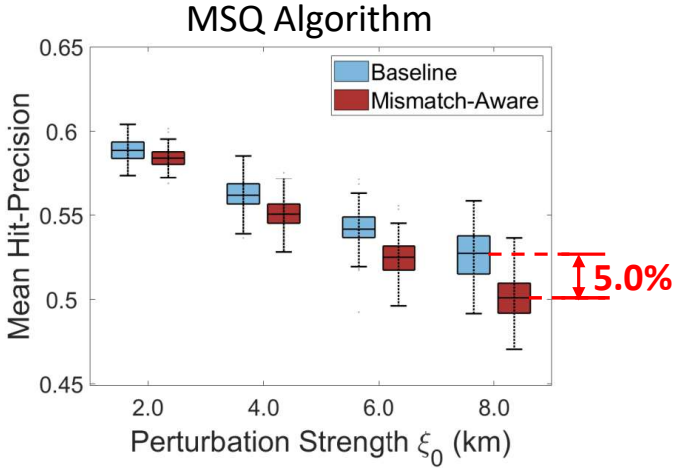
**Our proposed algorithms outperform baselines by over 17%**

# Can we utilize spatio-temporal mismatches to better protect users' privacy?

# Proposed Mismatch-Aware Location-Privacy Perserving Mechanism

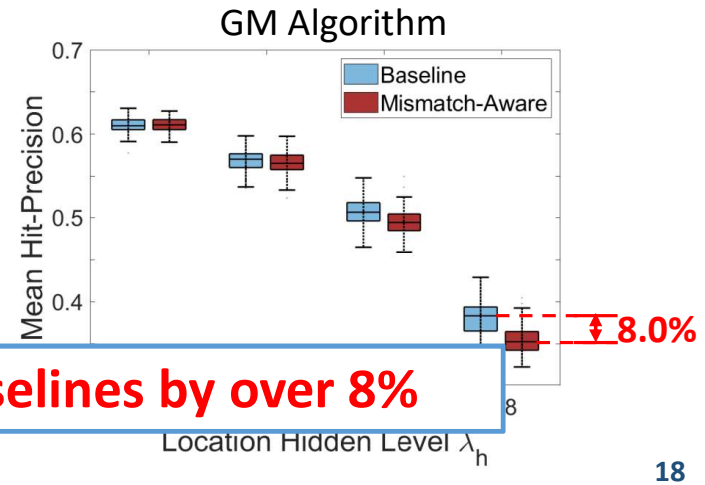
## Mismatch-Aware Perturbation Mechanism

- Idea: add larger noise to more important mobility records
  - ✓More Important: Larger contribution to the similarity score (less mismatches)
  - ✓Keep total energy unchanged




## Mismatch-Aware Location Hiding Mechanism

- Idea: eliminate more important mobility records with larger probability
  - ✓Keep total number of eliminated records unchanged.




**Mismatch-Aware LPPMs outperform baselines by over 8%**

# Summary

- **Large-scale Ground-truth Datasets**
  - ISP trajectories with over 2 million users
  - 2 different social networks, 2 different types of external information
- **Demonstrate the Gaps between Theory and Practice**
  - High theoretical bound
  - Low actual performance
- **Bridge the Gaps between Theory and Practice**
  - Considering spatio-temporal mismatches, data sparsity, location/time context
  - Utilize spatio-temporal mismatches in LPPMs
  - Improve the performance → confirm our observations

# Thanks you!

For Data Sample and Code, Please Contact

whd14@mails.tsinghua.edu.cn
liyong07@tsinghua.edu.cn

# Reference


**[Scientific Report 2013]** Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the crowd: The privacy bounds of human mobility,” Scientific reports, vol. 3, p. 1376, 2013.

**[WWW 2016]** C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, “Linking users across domains with location data: Theory and validation,” in Proc. WWW, 2016.

**[AIHC 2016]** A. Cecaj, M. Mamei, and F. Zambonelli, “Re-identification and information fusion between anonymized cdr and social network data,” Journal of Ambient Intelligence and Humanized Computing, vol. 7, no. 1, pp. 83–96, 2016.

**[WOSN 2014]** L. Rossi and M. Musolesi, “It’s the way you check-in: identifying users in location-based social networks,” in Proc. ACM WOSN, 2014.

**[TIFS 2016]** F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, “Where you are is who you are: User identification by matching statistics,” IEEE Transactions on Information Forensics and Security (TIFS), vol. 11, no. 2, pp. 358–372, 2016.

**[IEEE SP 2008]** A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” in Proc. IEEE SP, 2008.

**[IEEE SP 2011]** R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, “Quantifying location privacy,” in Proc. IEEE SP, 2011.

**[TON 2013]** C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao, “Privacy vulnerability of published anonymous mobility traces,” IEEE/ACM Transactions on Networking (TON), vol. 21, no. 3, pp. 720–733, 2013.

**[UbiComp 2007]** N. Banerjee, A. Rahmati, M. Corner, S. Rollins, and L. Zhong, “Users and batteries: interactions and adaptive energy management in mobile systems,” Proc. ACM UbiComp, 2007.

**[ICWSM 2016]** G. Wang, S. Y. Schoenebeck, H. Zheng, and B. Y. Zhao, ““will checkin for badges”: Understanding bias and misbehavior on location-based social networks.” in Proc. ICWSM, 2016.