Privacy risk assessment for mobility data

What’s the meaning of privacy?

Privacy has many connotations:

- A series of principles
  Right to be let alone, to limit access to information etc.

- Data privacy
  Regulated by national and international laws. Protect an individual's privacy and their personally identifiable information

Why privacy for mobility data is a concern?

- Mobility is a *sensitive* type of information
  Depending on the location visited, one could infer religious preferences, daily habits, health problems.

- Mobility data is *abundant* and readily available
  Location based services, social media access etc.

K-anonymity

- Hide individuals amongst k-1 others
  - Generalization
  - Suppression
- Privacy vs Utility Tradeoff
- NP-Hard
- Vulnerabilities: l-diversity and t-closeness

Key Attribute | **Quasi-Identifier** | **Sensive Attribute**

Name	Birthday	Sex	ZIP	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

Structure of an attack

- Individual record

UserId	Age	Weight	Heart rate	Pressure	Disease
u_1	>40	95 kg	110 bpm	150	Arrhythmia

- Assumed adversary knowledge

Age	Weight	Heart rate	Pressure
>40	95 kg	110 bpm	150

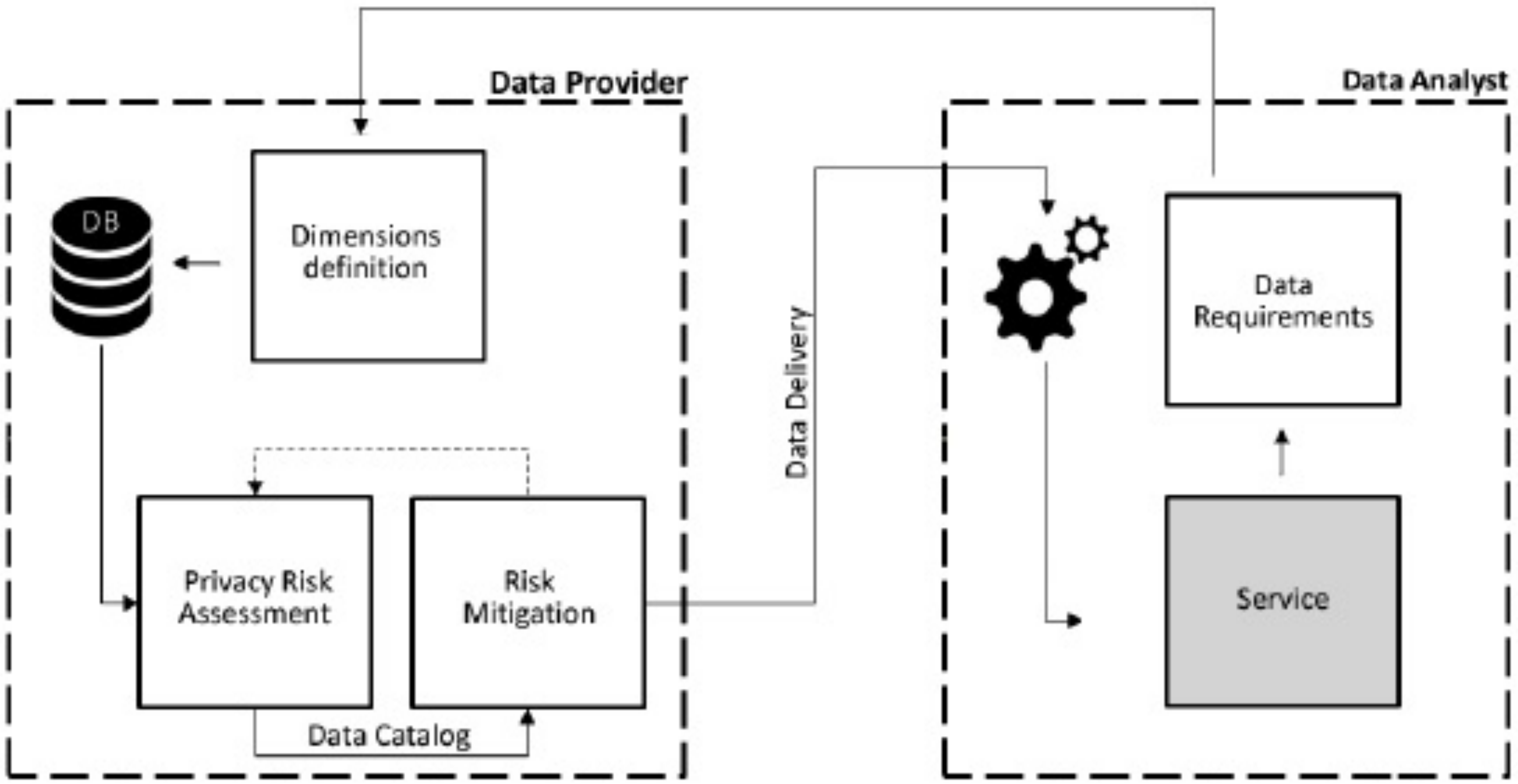
Removing ids may not be enough

For mobility data

- Trajectory example
  - Day $T = \langle (l_1, t_1), (l_2, t_2), (l_3, t_3) \rangle$
  - Month $T = \langle (l_1, t_1), (l_2, t_2), (l_3, t_3), \ldots, (l_n, t_n) \rangle$

- Worst case scenario approach
  - We assume that the adversary knows everything

PRUDEnce privacy framework



Pellungrini et al., A Data Mining Approach to Assess Privacy Risk in Human Mobility Data, ACM TIST 2018

Pratesi et al., PRUDEnce: a System for Assessing Privacy Risk vs Utility in Data Sharing Ecosystems, Transactions on Data Privacy 2018.

Risk definition

- Background knowledge $B = B_1, B_2, \ldots, B_k$
- Background knowledge instance $b \in B_k$
- Probability of re-identification $PR_D(d = u | b) = \frac{1}{|M(D, b)|}$
- Privacy risk $Risk(u, D) = max(PR_D(d = u | t))$

$$M(D, b) = \{ d \in D \,|\, matching(d, B) = True \}$$

Matching

$$M(D,b)=\{d\in D \mid matching(d,B)=True\}$$

- Location attack

$$matching(d,B)=\begin{cases} true & b\subseteq L(T_u) \\ false & \text{otherwise} \end{cases}$$

- Frequency attack

$$matching(d,B)=\begin{cases} true & \forall (l_i,wi)\in b, \exists (l_i^d,w_i^d)\in W \mid l_i=l_i^d \wedge w_i\leq w_i^d \\ false & \text{otherwise} \end{cases}$$

Defining attacks

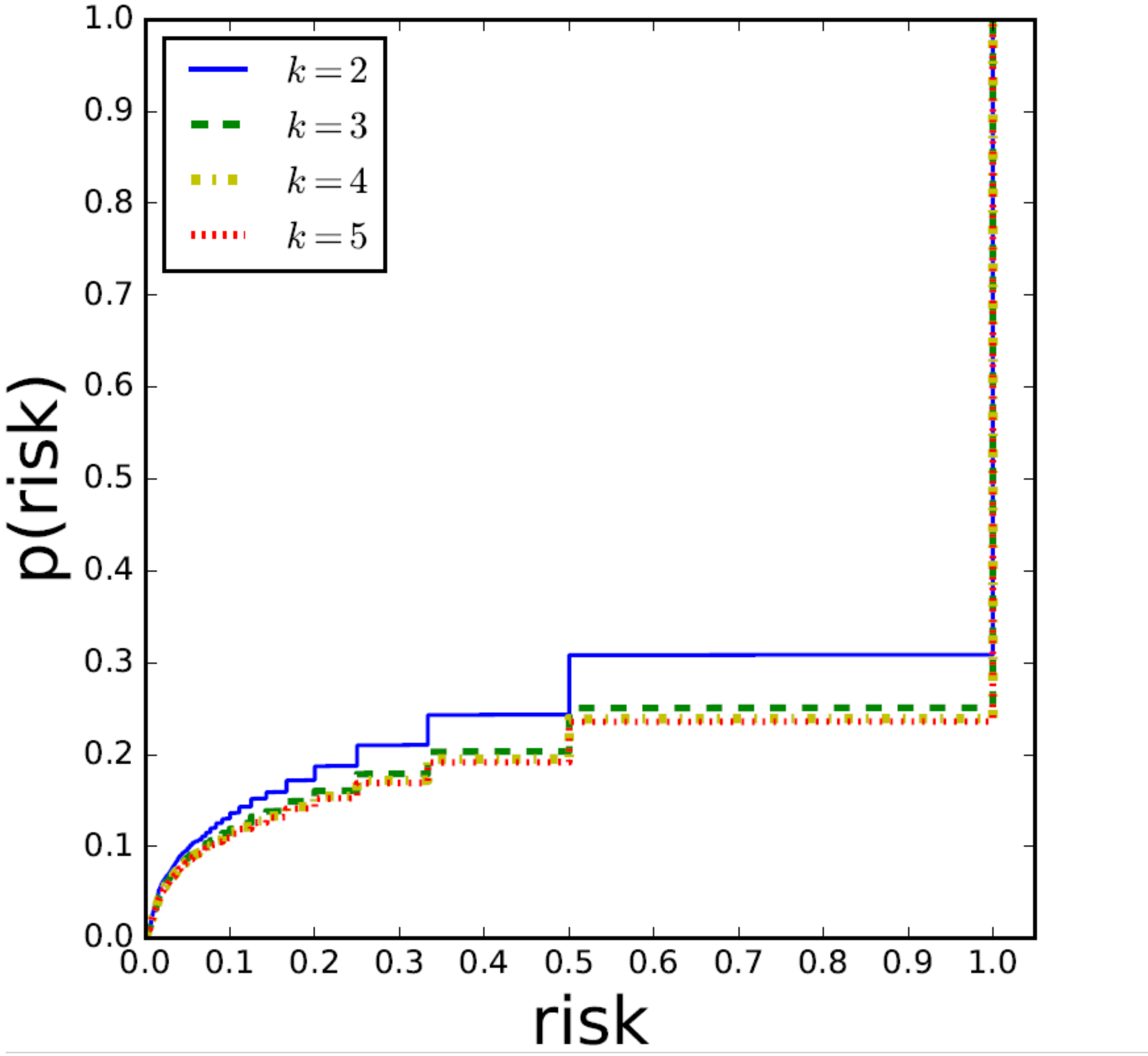
- Trajectories
  - Location
  - Sequence
  - Location + time

- Derived structures: frequency and probability vectors
  - Unique locations
  - Frequency
  - Probability
  - Proportion
  - Home and work

Pellungrini et al., Analyzing Privacy Risk in Human Mobility Data, STAF Workshops 2018

An example of real results

- Location attack performed on real gps data from the city of Florence



Computational complexity

- For each individual compute all possible instances of background knowledge
  - For each instance, scan the data
    - Determine match between instance and individuals in the data

- Complexity: $O(\binom{len}{k} N * matching)$

Further extensions

- New attacks
- Anonymization techniques
- Dataset matching algorithms

Coming soon…