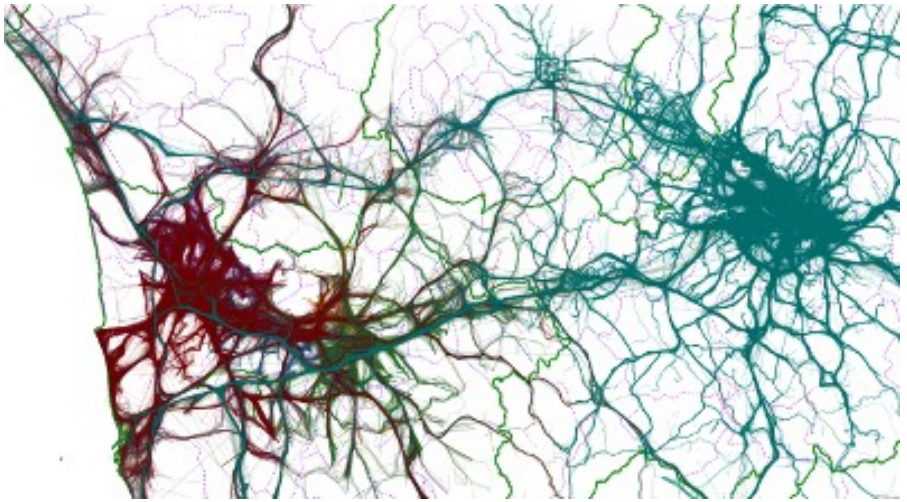
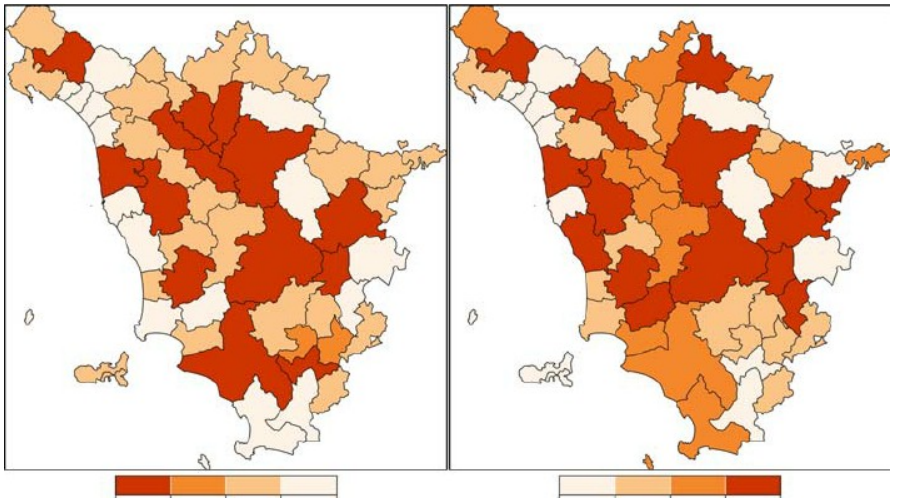
A Data Mining Approach to Assess Privacy Risk in Human Mobility Data

Roberto Pellungrini

Anna Monreale
Luca Pappalardo
Francesca Pratesi

Privacy and Mobility

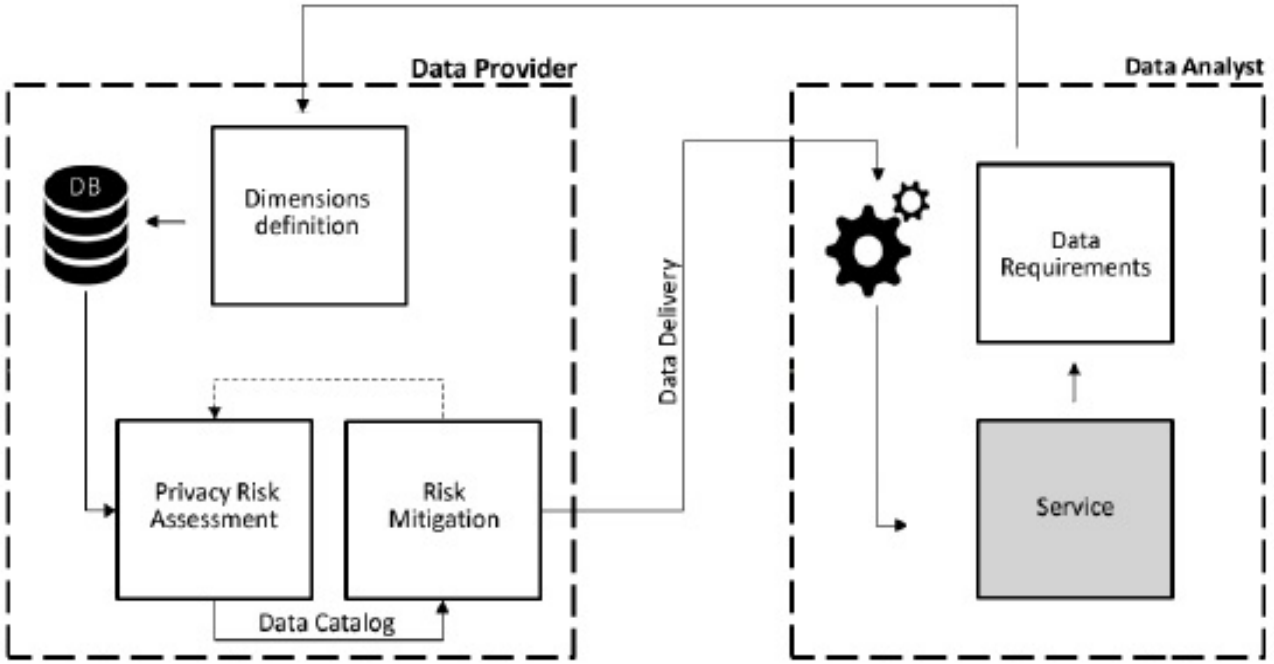




- Mobility data are rich and useful, but they bare a great risk for the privacy of the individuals involved.

- EU’s 2016 Regulation concerning privacy requires data holders to assess the privacy risk when managing data, according to privacy by desing principles.

Privacy Framework



- Trajectory $T = \langle (l_1, t_1), \ldots, (l_i, t_i) \rangle$
- Frequency Vector $W = \langle (l_1, w_1), \ldots, (l_i, w_i) \rangle$
- Probability Vector $P = \langle (l_1, p_1), \ldots, (l_i, p_i) \rangle$
- Mobility Dataset $D = \{ S_1, \ldots, S_n \}$

Risk definition

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

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

Attacks

Trajectories:

- Location
- Location Sequence
- Visit

Vectors:

- Frequent Location
- Frequent Location Sequence
- Frequency
- Probability
- Proportion
- Home&Work

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

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

Basic Structure

- For each individual calculate all possible instances of background knowledge
  - For each instance cycle through each individual
    - Find a match between the instance and the individual

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

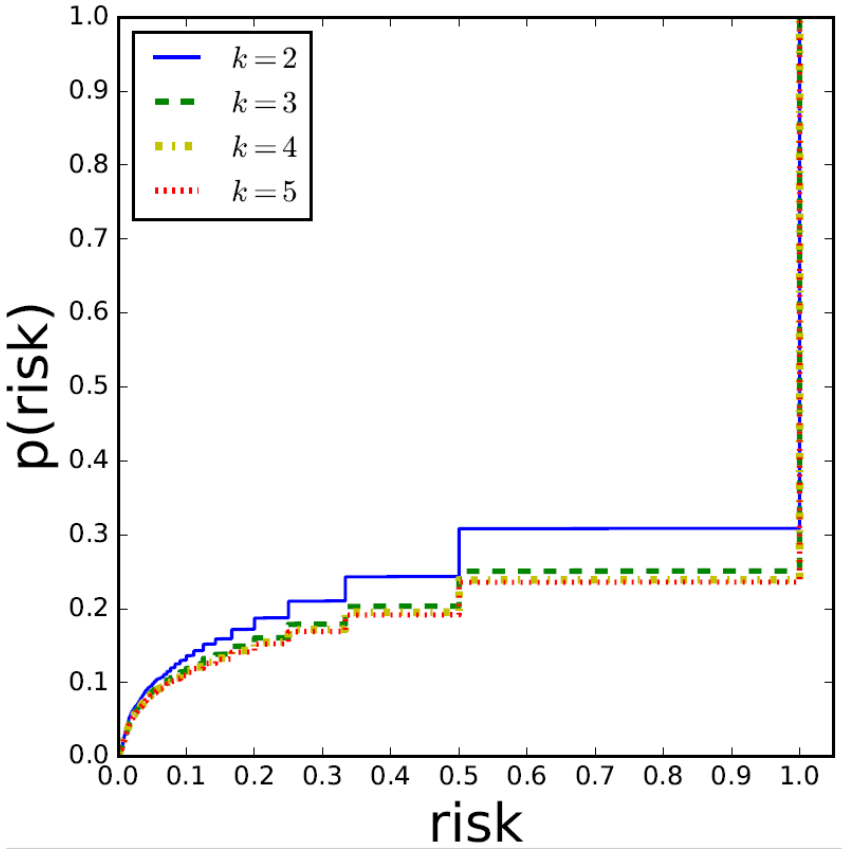
Dataset

- GPS data from Octo-Telematics
- May 2011, Tuscany

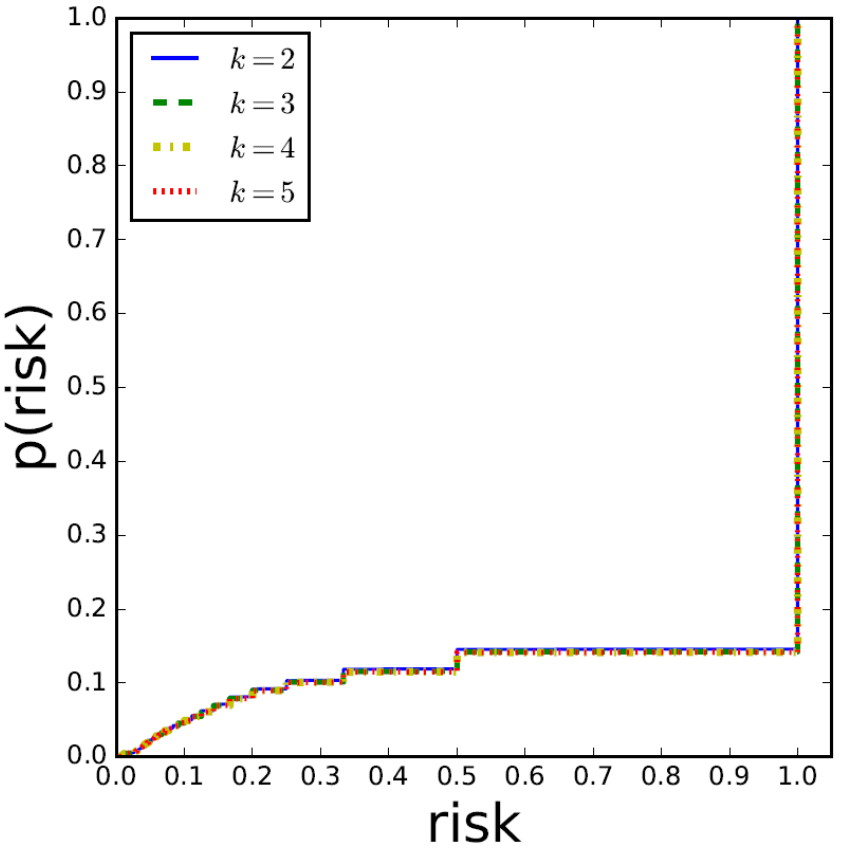
Projections:

- Florence: 9715 individuals
- Pisa: 2280 individuals

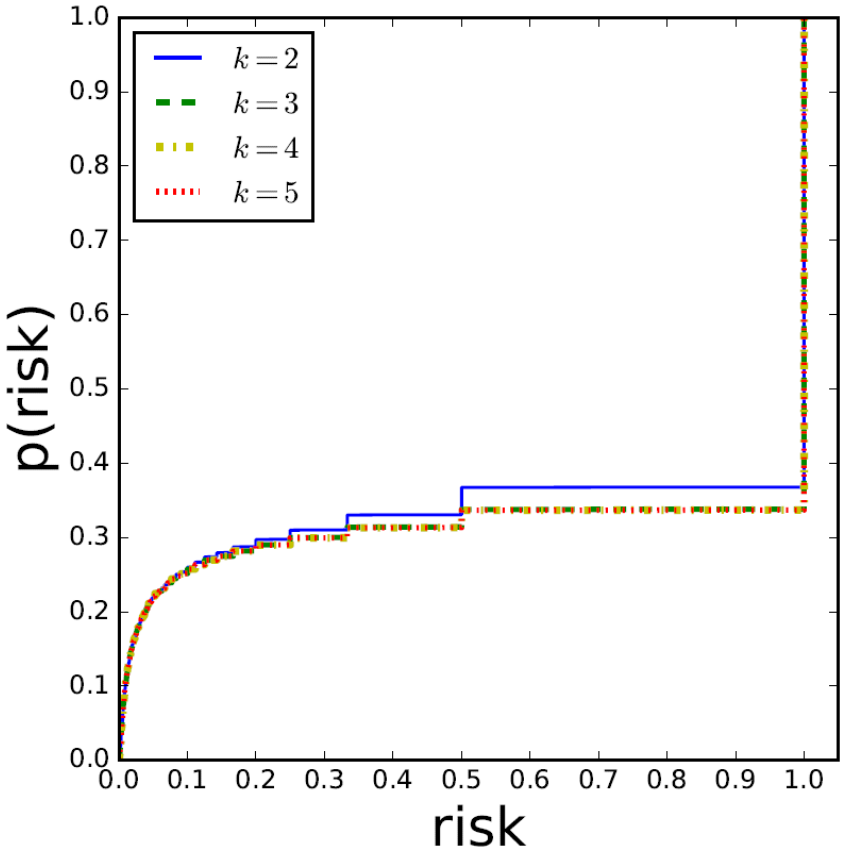
Florence Location attack



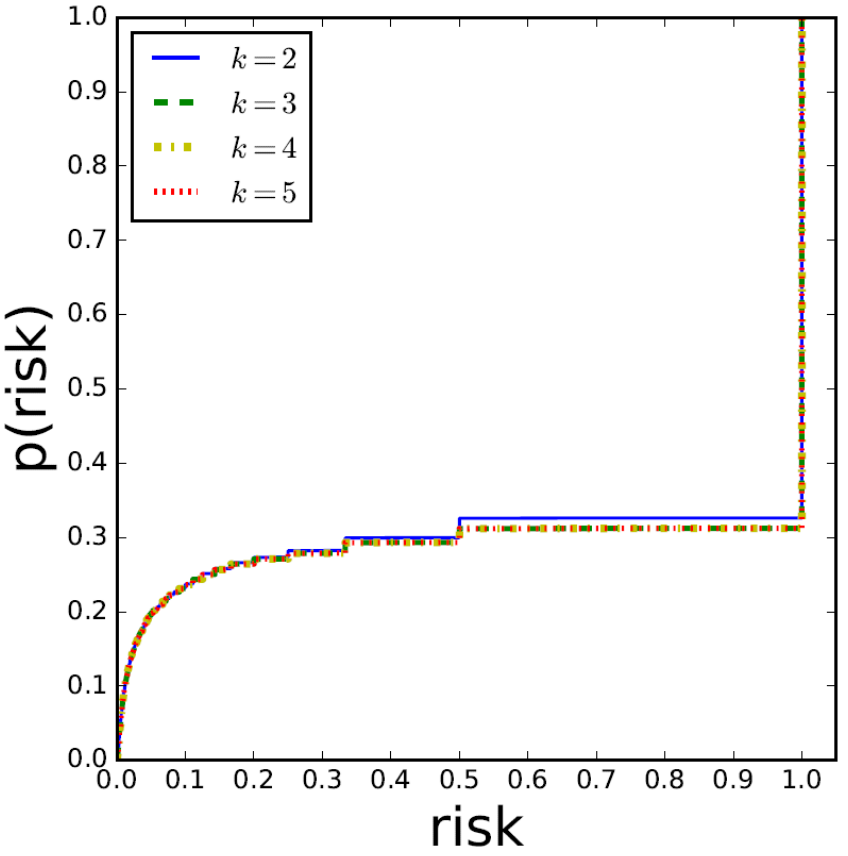
Florence Visit attack



Florence Probability Attack



Florence frequency attack



Pros

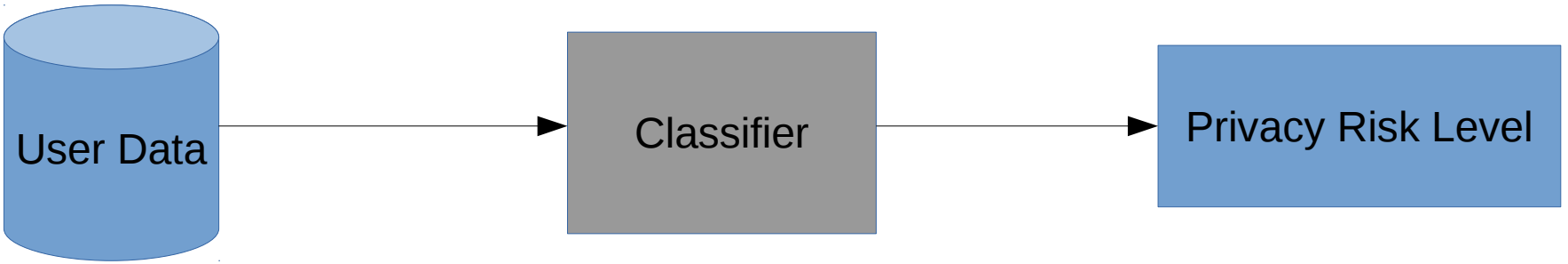
- Worst case re-identification scenario
- Easy to extend

Cons

- Computational Complexity
- Non flexible to different selections of users

A possible solution

- Using individual mobility features to predict the level of privacy risk



Mobility features

- Radius of Gyration
- Individual’s Entropy
- Max Distance
- Sum Distances
- # Distinct locations

- Location Entropy
- Location Density
- Number of Visits
- Average # Daily Visits

Training Set

- Simulation of an attack
- Calculation of mobility features

UserId	Entropy	Daily Visits	Visit Num.	Radius Gyr.	Risk
u_1	0.9	9	280	600	1.0
u_2	1	13	400	750	1.0
u_3	0.12	2	58	180	0.15
u_4	0.09	2	61	120	0.075
u_5	0.22	4	120	280	0.25

Classification Problem

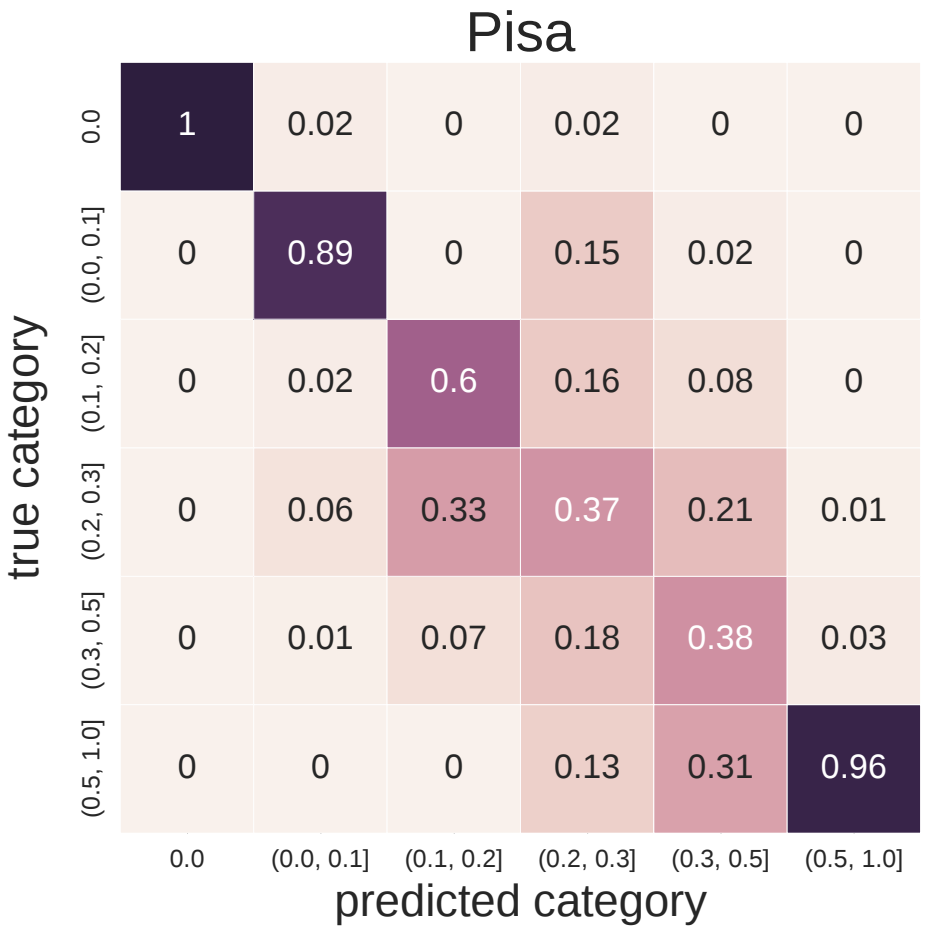
- Risk discretization in labels
- [0], (0,0.1], (0.1,0.2], (0.2, 0.3], (0.3,0.5], (0.5,1]
- Random Forest Classifier
- K-fold cross-validation

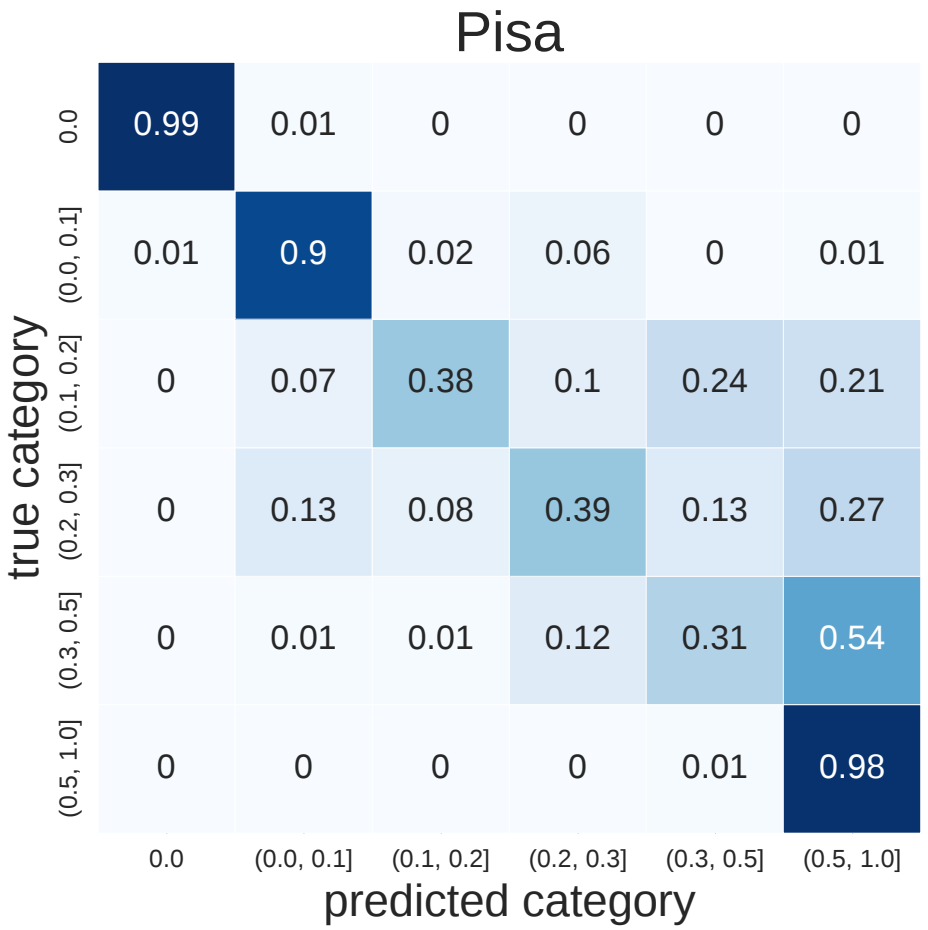
UserId	Entropy	Daily Visits	Visit Num.	Radius Gyr.	Risk	RiskLabel
u_1	0.9	9	280	600	1.0	(0.5,1]
u_2	1	13	400	750	1.0	(0.5,1]
u_3	0.12	2	58	180	0.15	(0.1,0.2]
u_4	0.09	2	61	120	0.075	(0,0.1]
u_5	0.22	4	120	280	0.25	(0.2,0.3]

Execution Times

variable ($\sum_2^5 k$)	**Florence**		**Pisa**	
	simulation	**classifier**	**simulation**	**classifier**
HomeWork	149s (2.5m)	7s	5s	3s
Frequency	645s (10m)	22s	20s	10s
Frequent Location Sequence	846s (14m)	22s	23s	10s
Proportion	900s (15m)	24s	30s	10s
Frequent Location	997s (10m)	22s	30s	10s
Probability	1,165s (20m)	22s	37s	10s
Visit	2,274s (38m)	16s	95s (1.5m)	9s
LocationSequence	$>$ 168h (1week)	22s	$>$ 168h (1week)	10s
Location	$>$ 168h (1week)	22s	$>$ 168h (1week)	10s
total	$>$ **2weeks**	**172s**	$>$ **2weeks**	**79s**

Precision and Recall



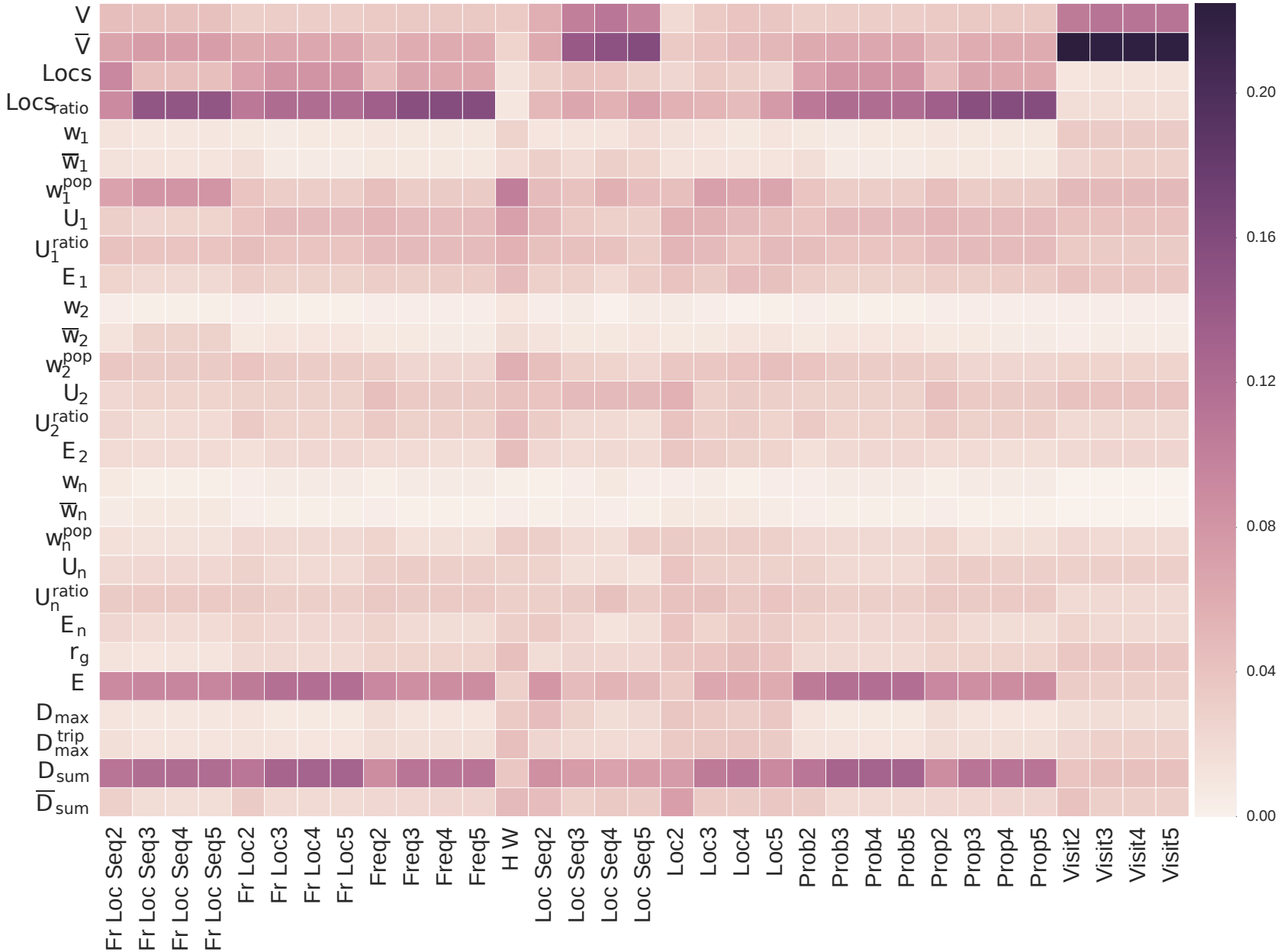


Transfer Learning

Visit

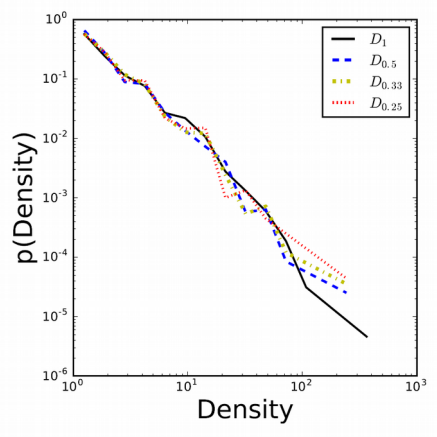
configuration		Florence		Pisa		FI → PI		PI → FI	
		acc.	**f1**	**acc.**	**f1**	**acc.**	**f1**	**acc.**	**f1**
locations with timestamps	$k = 2$	0.94	0.94	0.93	0.93	0.93	0.92	0.93	0.93
	$k = 3$	0.94	0.94	0.93	0.93	0.93	0.93	0.93	0.93
	$k = 4$	0.94	0.94	0.93	0.93	0.93	0.93	0.92	0.92
	$k = 5$	0.94	0.94	0.92	0.92	0.93	0.93	0.91	0.92
avg baseline		0.82	0.81	0.81	0.80				

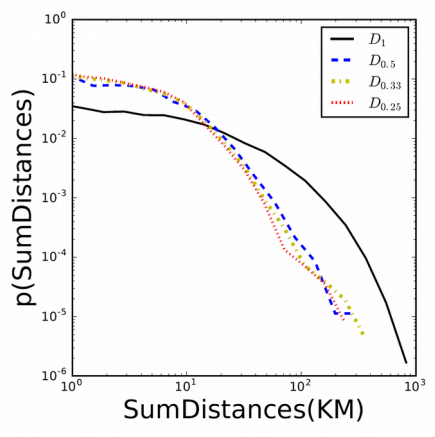
Feature Importance

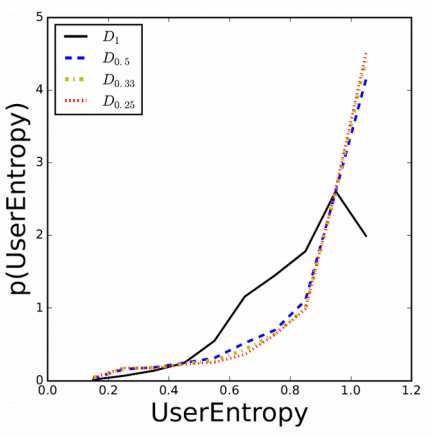


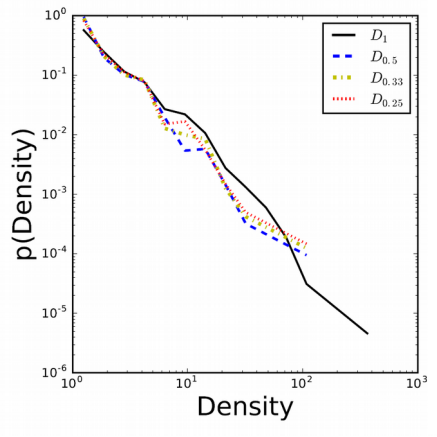
Feature distributions

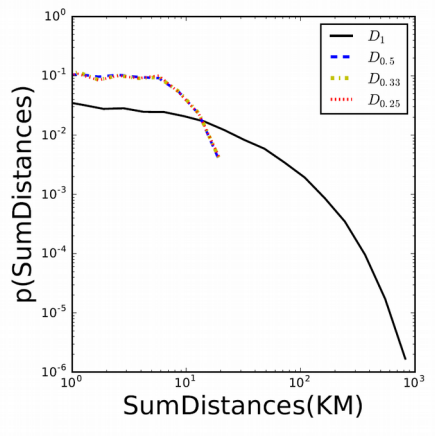
Florence location attack

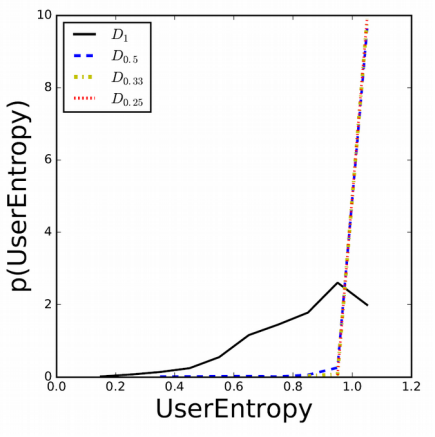












Florence visit attack

Future works

- Applying the data mining approach to
  - social networks
  - retail data
- Developing multidimensional attacks based on multiple data sources
  - Mobility data and retail data
  - Mobility data and social data