Anonymization of trajectory data

Josep Domingo-Ferrer and **Rolando Trujillo-Rasua**

Universitat Rovira i Virgili, Tarragona, Catalonia

http://unescoprivacychair.urv.cat
2011 WS on Stat. Conf., Tarragona, Oct. 26, 2011

October 17, 2011

Spatio-temporal data

- Spatio-temporal data (trajectories)
  may be collected by several technologies like GPS, RFID, GSM, etc.
- Sharing, publishing and mining such data is beneficial for most applications dealing with moving objects (intelligent transportation, traffic monitoring, planning, etc.).
- However, releasing trajectories may disclose sensitive information on individuals.

Privacy preservation of trajectories

- Privacy preservation of trajectories means that no sensitive location should be linkable to a specific individual.
- Since defining quasi-identifiers in spatio-temporal data is a difficult task, conventional microdata anonymization does not fit spatio-temporal data due to high information loss.
- We present two methods for privacy-preserving trajectory publication based on microaggregation.
- Contrary to previous methods, our proposals do not ignore temporal information and release synthetic trajectories that preserve the locations covered by the original trajectories.

The SwapLocations method

The SwapLocations method works over a cluster of trajectories. The idea is that each location must be randomly swapped within a set of other $k-1$ locations selected from $k-1$ different trajectories. If some location could not be swapped, it is automatically removed from the anonymized dataset.

The SwapLocations method achieves trajectory k-anonymity

Every location $\ell \in T$ in an original trajectory T has the same probability to be a location of a set C of k different anonymized trajectories including the anonymized version of T. Therefore, any subset $S \preceq T$ has the same probability to be a subset of any trajectory in C. Considering that $S \preceq T$ is the adversary's knowledge, all trajectories in C are indistinguishable for the adversary and thus, $Pr_{T^\star}[T|S] = \frac{1}{k}$.

The ReachLocations method

- The ReachLocations method takes reachability constraints into account. From a given location, only those locations at a distance below a threshold, following a path in an underlying graph (e.g. urban pattern or road network), are considered to be directly reachable.
- Aimed at providing high utility, the ReachLocations method does not microaggregate trajectories. Instead, every location should be swapped within other $k-1$ different locations of different trajectories selected from the whole dataset.

Privacy guarantee of the ReachLocations method

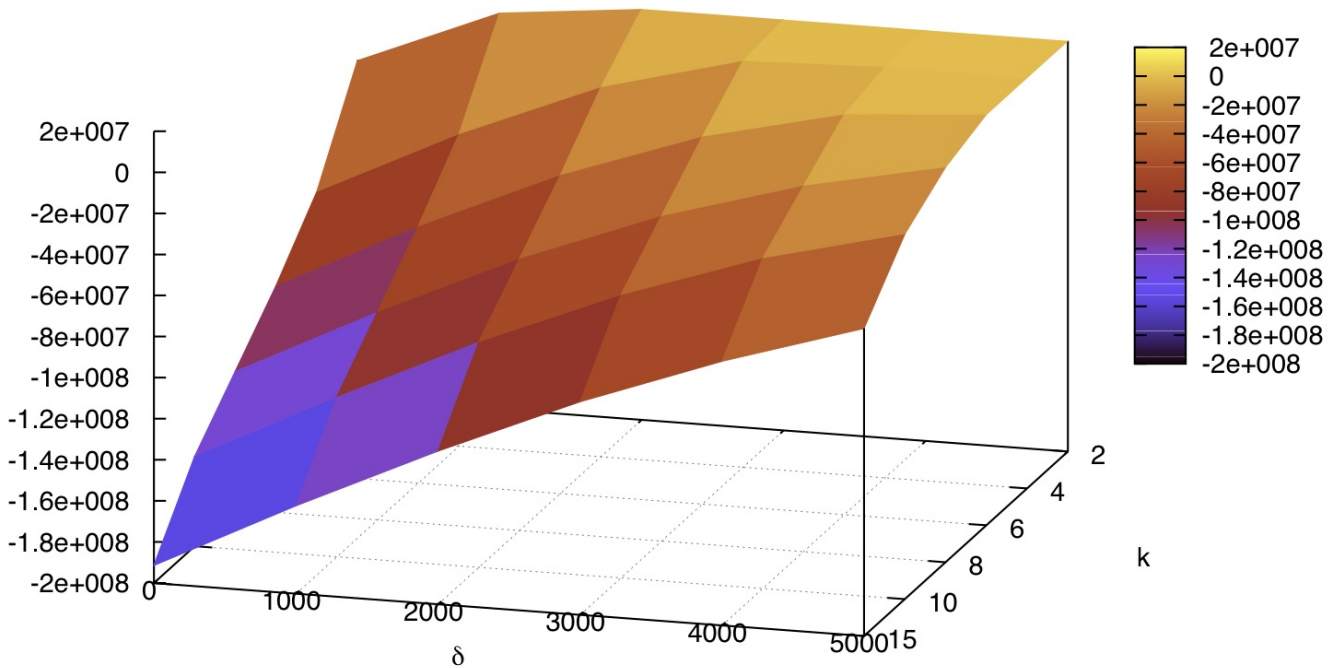
The ReachLocations method achieves location k-diversity

Every location $\ell \in T$ in an original trajectory T is randomly permuted with other $k-1$ different locations of different trajectories. This means that any location in the anonymized dataset is indistinguishable from a set C of $k-1$ other locations. Considering that ℓ and the locations in C are pairwise different and belong to different trajectories, $Pr_\ell[T|S] = \frac{1}{k}$.

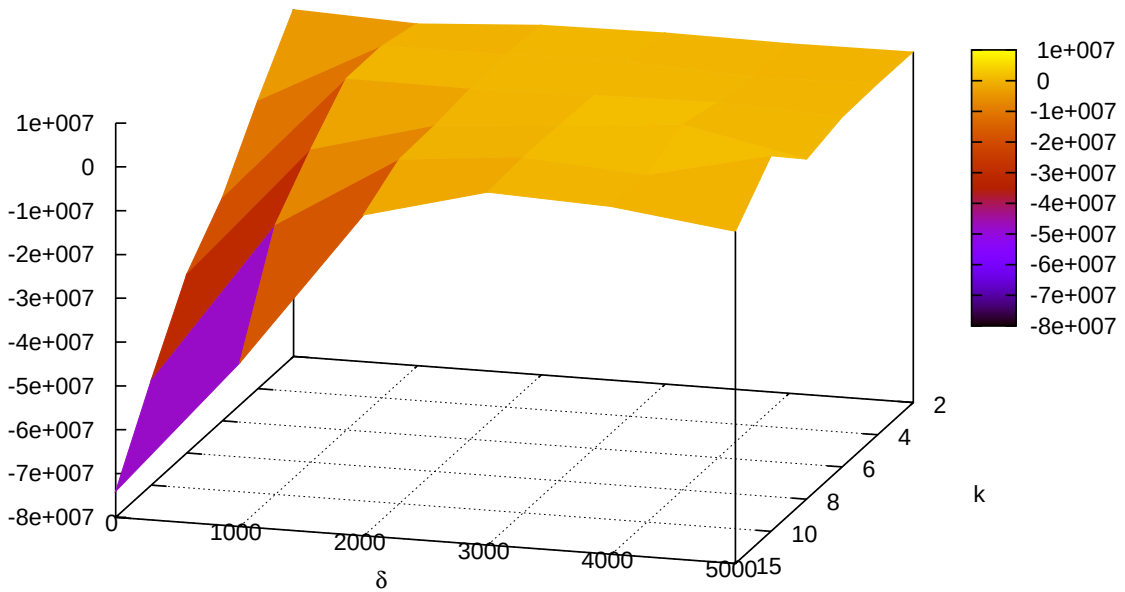
Datasets of trajectories

- We used synthetic data to compare SwapLocations against the (k, δ)-anonymity method and next we used real-life data.
- The synthetic data were generated with Brinkhoff's generator: 1,000 synthetic trajectories that visited 45,505 locations in the German city of Oldenburg.
- The real-life dataset consists of cab mobility traces that were collected in the city of San Francisco.
- Our results indicate that we discard significantly fewer trajectories than (k, δ)-anonymity, due to our being able to consider
  also trajectories that do not have the same time span.

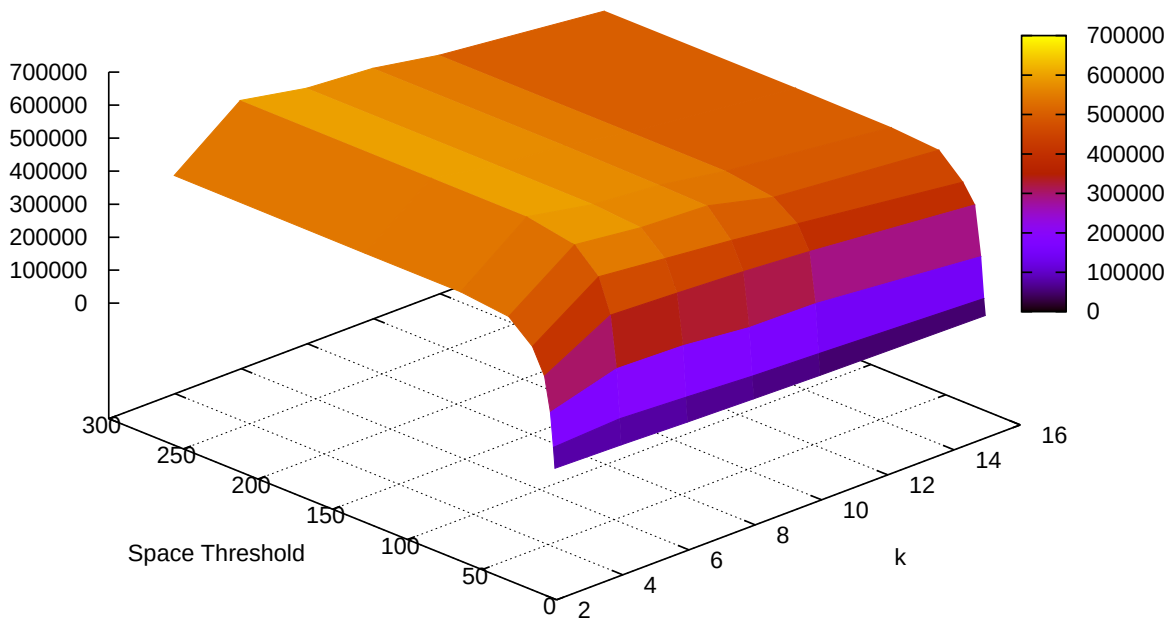
Experiments on synthetic data: space distortion of SwapLocations minus space distortion of (k, δ)-anonymity



Experiments on synthetic data: space distortion of ReachLocations minus space distortion of (k, δ)-anonymity



Experiments on real data: Total space distortion of the SwapLocations considering different space thresholds and cluster sizes.



Conclusions

- We proposed a distance measure for trajectories which:
  - considers both spatial and temporal aspects of trajectories;
  - is computable in polynomial time;
  - can cluster trajectories not defined over the same time span.
- We proposed two method for trajectory anonymization such that:
  - Places and times in the anonymized trajectories are true original places and times with full accuracy;
  - Trajectories with partial or no time overlap can be anonymized together;
  - The first method satisfies trajectory k-anonymity;
  - The second method satisfies location k-diversity.
- Experimental results on synthetic and real-life trajectory data show the viability of our proposals.