**Need to review: 2008, 2014, 2004, 2003**

**Monreale, A. & Pellungrini, R. (2023). A Survey on Privacy in Human Mobility. *Transactions on Data Privacy*, 16, pp. 51-82.**

*Summary*

* Describes the privacy attacks and privacy models considered for mobility data – privacy preserving trajectory data publishing
* Present frameworks for privacy risk assessment
* Discuss three types of privacy preserving strategies
  + Anonymization
  + Differential privacy
  + Generative models

*Conclusions*

* Many privacy attacks studied in the literature assume a worst-case scenario, *e.g.,* the adversary has knowledge that in practice would be very difficult to collect – protecting against the worst case leads to lower data utility
* **There should be a focus on realistic adversarial attacks**

*Relevant Details*

* Private trajectory data should still serve as a proxy for human behavior – destroying this makes the data useless
* **PRIVACY ATTACKS**
  + **Record Linkage Attacks**
    - **Background knowledge attacks**
      * Adversary has some knowledge concerning the target individual
        + Generally some data related to the individual collected through observation or separate data sources
      * Adversary’s goal is to match the data in their possession to the anonymized data corresponding to the individual of interest
      * Background knowledge assumes the data of the adversary is present in the anonymized data, and the attack can be simulated using sub-sampling
    - **External or Separate Data base attacks**
      * Adversary attempts to cross-reference information for the same individual stored in different databases
      * Examples include background knowledge extracted from social graphs or performing distance based record linkage across data bases which is commonly used for database integration
  + **Attribute Linkage Attacks**
    - Also known as homogeneity attacks, where background knowledge is linked on quasi-identifiers but homogeneity of sensitive attributes enables a privacy breach
    - Can be used to infer sensitive locations or sensitive attributes associated with the trajectory
  + **Probabilistic or Inference Attacks**
    - Adversary tries to increase knowledge by accessing the target data base
    - Accessing the data base should not reveal too much additional information than what is already known by the adversary
    - Can be viewed as a generalization of the attribute linkage attack since it measures the increase in knowledge
      * Examples:
        + the number of additional locations acquired after the attack
        + similarity attack – if there are three individuals with same trajectories and different stomach diseases, adversary can still infer the individual has a stomach related problem
        + Semantic attack: infer information based on the frequency of visiting regions with a lot of POIs, e.g., health status based on visits to hospital/doctors
        + Activity attack: adversary estimates the probability of an activity trajectory
* **PRIVACY MODELS**
  + K-anonymity is used to counter record linkage attacks
    - **One of the most extensively applied techniques to trajectory anonymity**
    - Assumes that the attributes in a data set are divided into sensitive and quasi identifiers
      * Sensitive attributes need to be protected
      * Quasi-identifiers make be linked to external information for a linking attack
  + L-diversity provides (some) protection against attribute linking attacks
    - Every equivalence class should contain at least L well represented values for a sensitive attribute
  + T-closeness provides additional protection against attribute linkage attacks
    - The distance between the distributions of a sensitive attribute in an equivalence class and the entire data set must be bounded by a threshold t
  + **For trajectory data, splitting the data into quasi-identifiers and sensitive attributes does not work – locations can be both QIs and sensitive**
    - **There are many k-anonymity based approaches, we can refer the reader here**
  + **Approaches Based on Sensitive Information**
    - Approaches are generally designed to counter two situations:
      * Protecting against inferring sensitive information associated with trajectories
      * Protecting against inferring sensitive location data
    - ***Need to choose to protect against reasonable inference, i.e., what is likely to be inferred by an adversary from this data set***

**Skinner, C.J. & Elliot, M.J. (2002). A Measure of Disclosure Risk for Microdata. *J.R. Statist. Soc. B*, 64, part 4, pp. 855-867.**

*Summary*

* Propose a measure of disclosure risk based on the probability that a unique match between a microdata record and a population unit is correct

*Conclusions*

*Relevant Details*

* Population Uniqueness
  + = the proportion of units in the population which are population unique
  + If an adversary can link an population unique unit to the microdata sample they would know that the match was correct
  + This is the probability that a unit is population unique when the intruder draws the target unit at random from the population
* Population and Sample Uniqueness
  + The probability that a sample unique unit is population unique when an adversary draws the unit at random from the sample unique units
* **Proposed measure**
  + Interpreted as the probability that a unique match between the population and sample will be correct
  + Denominator: the number of possible chosen units (from population) for which a unique match will exist
  + Numerator: the number of these units for which the match is correct (for attribute combination j)
* **This relates to both singling out and linkability. Singling out is concerned with whether the record is sample unique. If we have a sample unique record then we could still have a many-to-one linkage if there are multiple matching population records – however we won’t know this for sure – and, it is impossible to enumerate all combinations of identifying attributes – it is likely that with a few locations everyone is population unique**

**Domingo-Ferrer, J. & Torra, V. (2003). Disclosure Risk Assessment in Statistical Microdata Protection via Advanced Record Linkage. *Statistics and Computing,* 13, pp. 343-354.**

*Summary*

* One form of empirical disclosure risk assessment examines record linkage
* This paper reviews record linkage and shows that it is still possible even without shared variables between data sets

*Conclusions*

*Relevant Details*

* **Authors assume two data sets corresponding to the same n individuals**
* Traditional record linkage techniques assume a set of common variables between data sets
* The problem is that true matches often contain errors between the shared variables (e.g., measurement error)
* Four linkage situations exist:
  + Same variables, same terms
  + Same variables, different terms
  + Different variables, same terms
  + different variables, different terms
* **How is record linkage performed?**
  + Benchmark methods
    - Random permutation
    - Ranking based on a single dimensions (principal component, sum of z-scores)
  + Commonly Used Methods
    - Probabilistic record linkage
    - Distance based record linkage
* **How do we justify our approach? Can we say that no matter the record linkage method performed, having at least *k* individuals in an equivalence class means an external record is tied to at least *k* records in the anonymized data?**
* Group-level vs. record-level reidentification
  + The proposed approach based on clustering is meant for group-level reidentification: a mapping between groups of records between data sets
  + If a group contains a single record the approach may produce a many-to-one or a few-to-one linkage
  + Outliers are most likely to result in one-to-one linkage
* **Empirical Results**
  + 29/90 correct group-level reidentifications, 24 of which are record level (a single record in one data set) – 26.6% of record-level re-identifications are correct
  + Benchmark one dimensional rankings correctly re-identified 5.5%
  + Permutation approach was negligible

**Domingo-Ferrer, J. & Torra, V. (2004). Disclosure risk assessment in statistical data protection. *Journal of Computational and Applied Mathematics*, 164-165, pp. 285-293.**

*Summary*

* Propose a unified approach to measuring disclosure risk for micro data using record linkage
* Shows that the most widely used rule for disclosure risk assessment in tabular data is flawed

*Relevant Details*

* 2 data types: Micro data sets (individual respondent records) and tabular data
* **An accurate utility assessment must take into account the specific data uses the user is interested in**
* **Uniqueness:** Some disclosure risk measurement for micro data relates to *nonperturbative* methods based on sampling: they measure the probability that a sample unique is a population unique. If sample is large relative to the population, a unique value in the sample is likely unique in the population and identification can occur
* **Record linkage:** used to assess disclosure risk for perturbative methods
  + Distance based record linkage
  + Probabilistic record linkage

**Winkler, William (2014). Matching and Record Linkage. *WIREs Comput Stat,* 6, pp. 313-325. Doi: 10.1002/wics.1317.**

*Summary*

* Provides background on the FS model of record linkage (probabilistic model)

*Relevant Details*

* Appropriate standardization and pre-processing of data has a larger effect on record linkage success than improving models

**Skinner, Chris (2008). Assessing disclosure risk for record linkage. *Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference.***

*Summary*

* Discusses estimating identification risk as the probability that a match between a microdata file and an external file is correct.
* **Matching takes place on ‘key’ variables – should we use this terminology? We need specific notation for the key variables.**

*Conclusions*

* Group and record level re-identification is possible without any shared variables between data sets

*Relevant Details*

* **Record linkage can be successfully performed even with masked data – risk is the proportion of matches which are correct, but this often uses the original (unmasked) data compared to the masked data – this is highly conservative and ignores the protection of sampling**
* **Also focuses on the probabilistic FS method for record linkage**
* Other things to note are that if, *e.g.*, only 5% of matches are correct but the adversary knows which 5% that is, then they can claim matches with 100% confidence (too high) – or, if the probability of a correct match varies across data subgroups then probabilities in some subgroups may be deemed too high. Overall, it is unrealistic that the adversary will know the proportion of correct matches since it requires knowing the true identities of the records in the microdata)