New Directions in Anonymization: The Permutation Paradigm, Verifiability, Transparency and Co-Utility

Josep Domingo-Ferrer
(joint work with Krishnamurty Muralidhar and Jordi Soria-Comas)

Universitat Rovira i Virgili, Tarragona, Catalonia

UNESCO Chair in Data Priv@cy

Maynooth, September 28, 2015

1. Introduction
2. Permutation model of microdata masking
3. A new subject-verifiable privacy model: $(\mathbf{d}, \mathbf{v}, f)$-permuted privacy
4. Maximum-knowledge intruder
5. Record linkage by the intruder
6. Verifiability of record linkage
7. Evaluation of synthetic data vs masking
8. Benefits of linkage verification to protector and subjects
9. Anonymization transparency towards the user
10. Alternatives to centralized anonymization: collaborative anonymization and co-utility
11. Conclusions

Introduction

- In anonymization of microdata, the data protector:
  - either makes restrictive assumption on the intruder's background knowledge (*e.g.* *k*-anonymity) $\Longrightarrow$ risky!!
  - or makes no assumptions at all (*e.g.* differential privacy) $\Longrightarrow$ utility damaging!!

UNIVERSITAT ROVIRA I VIRGILI

Introduction

- A further complication in microdata anonymization is the diversity of principles inspiring anonymization methods.
- This diversity makes it difficult:
    - To select the best method;
    - To select the best method parameters to achieve an optimum trade-off between utility preservation and disclosure protection.

Challenges/desiderata for big data anonymization

- *Linkability*. Linking data on the same individuals coming from several sources should remain feasible to some extent on anonymized data.
- *Composability*. The privacy guarantees given by a privacy model for several separate data sets should hold to some extent when the data sets are merged.
- *Computational cost*. SDC methods used to reach a certain privacy model should be scalable to large data volumes.

Recommendation: tunable and verifiable anonymization

- Privacy-first anonymization (based on enforcing a privacy model, like k-anonymity, t-closeness or ε-differential privacy) often leads to poor data utility/linkability.
- Utility-first anonymization (iteratively changing parameters until empirical disclosure risk is low enough, as usual in official statistics) is slow and lacks formal privacy guarantees.
- Verifiable anonymization (based on the permutation model) allows *exactly tuning anonymization to achieve the desired linkability while offering formal privacy guarantees to the data administrator and the subjects.*

Permutation model: reverse mapping

Require: Original attribute $X = \{x_1, x_2, \cdots, x_n\}$
Require: Anonymized attribute $Y = \{y_1, y_2, \cdots, y_n\}$
  for $i = 1$ to n **do**
    Compute $j = \text{Rank}(y_i)$
    Set $z_i = x_{(j)}$ (where $x_{(j)}$ is the value of X of rank j)
  end for
  return $Z = \{z_1, z_2, \cdots, z_n\}$

Note. If there are several attributes in an original data set $\mathbf{X}$ and anonymized data set $\mathbf{Y}$, the above procedure is repeated for each attribute.

Permutation model: permutation plus residual noise

- A reverse-mapped attribute Z is a permutation of the corresponding original attribute X.
- The rank order of Z is the same as the rank order of Y.
- Therefore, any microdata anonymization technique is functionally equivalent to
  - **Permutation**. Each attribute of the original dataset $\mathbf{X}$ is permuted to obtain $\mathbf{Z}$.
  - **Residual noise addition**. Noise is added to each value of $\mathbf{Z}$ to obtain the anonymized data set $\mathbf{Y}$ (the noise is residual, because the ranks of $\mathbf{Z}$ and $\mathbf{Y}$ must stay the same).

A new subject-verifiable privacy model: $(\mathbf{d}, \mathbf{v}, f)$-permuted privacy I

Given a vector $\mathbf{d} = (d^1, \ldots, d^m)$ of non-negative integers, a vector $\mathbf{v} = (v^1, \ldots, v^m)$ of non-negative real numbers, an original data set $\mathbf{X}$ and an anonymized data set $\mathbf{Y}$ both with m attributes, and a record-level mapping $f : \mathbf{X} \longrightarrow \mathbf{Y}$, we say $\mathbf{Y}$ satisfies $(\mathbf{d}, \mathbf{v}, f)$-permuted privacy *with respect to original record* $\mathbf{x} = (x^1, \ldots, x^m) \in \mathbf{X}$ if y_*^j being the value of the j-attribute Y^j in the anonymized data set closest to x^j for $j = 1, \ldots, m$,

1. The anonymized record $f(\mathbf{x}) = (y^1, \ldots, y^m)$ satisfies

$$|\mathrm{Rank}(y^j) - \mathrm{Rank}(y_*^j)| \geq d^j \quad (j = 1, 2, \ldots, m)$$

(d^j is called the *permutation distance* for the j-th attribute);

A new subject-verifiable privacy model: $(\mathbf{d}, \mathbf{v}, f)$-permuted privacy II

2. If $S^j(d_j)$ is the set of values of the sorted Y^j whose rank differs no more than d_j from the rank of y_*^j, then the diversity of $S^j(d_j)$ is greater than v^j according to a given diversity criterion.

Explanations on the definition

- If anonymization is just a permutation, then $y_*^j = x^j$.
- For each original record $\mathbf{x}$, the data protector can take as $f(\mathbf{x})$ the anonymized record derived from $\mathbf{x}$.
- The subject can take as a possible approximation for $f(\mathbf{x})$ the record in $\mathbf{Y}$ whose attribute values have the smallest rank difference with $(y_*^1, \ldots, y_*^m)$.
- Diversity criteria for $S^j(d_j)$ may be the variance, one of the l-diversity criteria, or the t-closeness criterion.
- If $(\mathbf{d}, \mathbf{v}, f)$-permuted privacy holds w.r.t. all records in $\mathbf{X}$, then we say it holds for the dataset $\mathbf{X}$.

Computing the vector $\mathbf{d}$ of permutation distances I

Require: $\mathbf{x} = (x^1, \ldots, x^m)$ {Original record containing m attribute values}
Require: $\mathbf{Y} = \{(y_i^1, \ldots, y_i^m) : i = 1, \ldots, n\}$ {Anonymized data set containing n records with m attributes $Y^1, \ldots, Y^m$}
Require: $f : \mathbf{X} \longrightarrow \mathbf{Y}$

```
for j = 1 to m do
    Let y_*^j be the value of Y^j closest to x^j
    Sort Y by Y^j
    Let Rank(y_*^j) be the rank (record no.) of y_*^j in the sorted Y
    for i = 1 to n do
        Let Rank(y_i^j) be the rank of y_i^j in the sorted Y
    end for
end for
Let f(x) = (y_p^1, ..., y_p^m)
```

Computing the vector $\mathbf{d}$ of permutation distances II

for $j = 1$ to m **do**
  $d^j = |\mathrm{Rank}(y^j_p) - \mathrm{Rank}(y^j_*)|$
end for
return $\mathbf{d} = (d^1, \ldots, d^m)$

Verifiability of $(\mathbf{d}, \mathbf{v}, f)$-permuted privacy

Given $\mathbf{Y}$, not only the data protector, but also the subject can verify $(\mathbf{d}, \mathbf{v}, f)$-permuted privacy for her original record $\mathbf{x}$ because:

- Using $\mathbf{Y}$ and $\mathbf{x}$, the subject can compute $(y_*^1, \ldots, y_*^m)$ and $\mathbf{d} = (d^1, \ldots, d^m)$ with the above algorithm.
- Then the subject can check the diversity condition $\mathbf{v}$ on $\mathbf{Y}$.

Hence, the subject can make sure the values in her record $\mathbf{x}$ have been sufficiently protected in $\mathbf{Y}$ (enough permutation and enough diversity).

Adversarial model: crypto attacks adapted to anonymization

- Ciphertext-only. Adversary has access only to ciphertext (*i.e.* anonymized data set).
- Known-plaintext. Adversary has access to pairs plaintext/ciphertext (*i.e.* pairs original and anonymized records).
- Chosen-plaintext. Adversary can choose a plaintext (original record) and get the corresponding ciphertext (anonymized record).
- Chosen-ciphertext. Adversary can choose a ciphertext (anonymized record) and get the corresponding plaintext (original record).

Maximum-knowledge intruder

- In a non-interactive setting (microdata set anonymization), known-plaintext is the strongest possible attack.
- We take the worst known-plaintext case and we assume that the intruder:
  - Knows the entire original data set **X** and the entire masked data set **Y**;
  - Wants to find the mapping between records in **X** and records in **Y**.

Comments on the intruder model

- Our intruder is stronger than the one considered in differential privacy.
- Our intruder is purely malicious and has nothing to gain from the released data (unlike a normal user).
- In cryptography, there is one (or few) legitimate receiver(s) and everyone else is deemed an intruder.
- In anonymization, there is one (or few) intruder(ies) and everyone else is deemed a user.

Record linkage by the intruder: search procedure

- Our powerful intruder can do reverse mapping and obtain the permuted dataset $\mathbf{Z}$ from the anonymized dataset $\mathbf{Y}$.
- Then he can link any original record $\mathbf{x} \in \mathbf{X}$ to (at least) one record $f(\mathbf{x}) = \mathbf{z}_p = (z_p^1, \ldots, z_p^m)$ computed the way he prefers, for example as:

Set $d = 0$
while $\nexists (z_p^1, \ldots, z_p^m) \in \mathbf{Z}$ such that $\forall j = 1, \ldots, m$, $|\mathrm{Rank}(z_p^j) - \mathrm{Rank}(x^j)| \leq d$ holds **do**
  $d = d + 1$
end while
return $f(\mathbf{x}) = (z_p^1, \ldots, z_p^m)$

Verifiability of record linkage

- Data protectors often dismiss record linkages by the intruder with the argument that the intruder cannot verify their correctness (plausible deniability).
- However, we show that our maximum-knowledge intruder can demonstrate that a linkage did not occur by chance alone.

Verification procedure by the intruder

1. Generate a large random set **T** of values by drawing from the original data **X**.
2. Determine the permutation distances at which matches occur between records in **T** and records in **Z**.
3. If the distribution of the permutation distances for matches between **T** and **Z** overlaps with the distribution of permutation distances for matches between **X** and **Z**, then the intruder's matches are plausibly random and he cannot claim them.

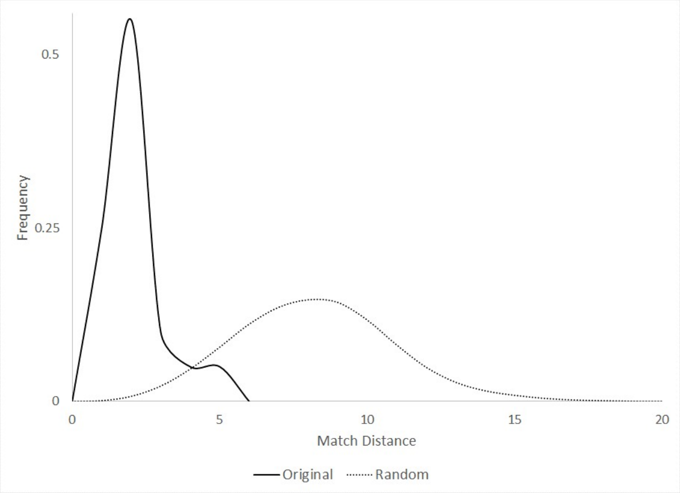
Evaluation of methods: deterministic masking

- Any deterministic masking method allows our maximum-knowledge intruder to exactly reconstruct the anonymization process from $\mathbf{X}$.
- Hence, it allows the intruder to determine the correct linkage between records of $\mathbf{X}$ and records of $\mathbf{Y}$.
- Deterministic methods include rounding, generalization, microaggregation, etc.

Evaluation of methods: additive noise

- We take as $\mathbf{X}$ a simulated data set with $n = 40$ records and $m = 4$ attributes X_1, X_2, X_3, X_4.
- We anonymize as $y_{ij} = x_{ij} + e_{ij}$, for $i = 1, \cdots, n$, $j = 1, \cdots, m$, with $e_{ij} \sim N(0, 0.01 \times \sigma_j^2)$, where σ_j is the variance of attribute X_j.
- The distributions of the match distance for linkages from $\mathbf{X}$ (original) and $\mathbf{T}$ (random) turn out to be quite different $\Longrightarrow$ linkages by the intruder are not plausibly deniable by the protector.
- The protector needs to increase the noise until both distributions are more similar/overlap more.

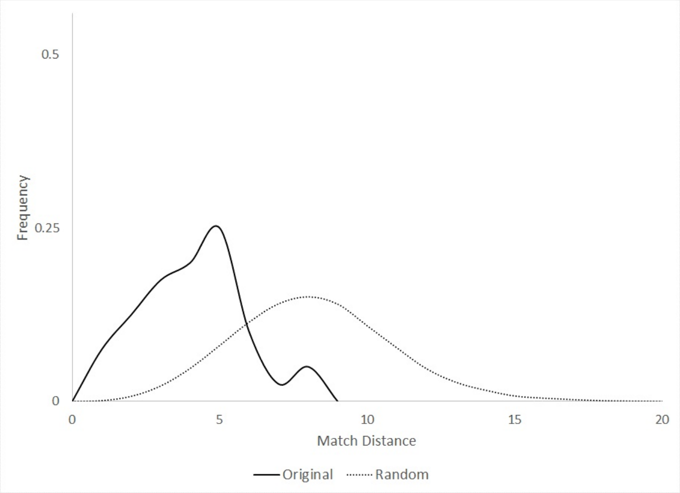
Evaluation of methods: additive noise (II)



Evaluation of methods: multiplicative noise

- We use the same **X** as for additive noise.
- We anonymize as $y_{ij} = x_{ij} \times e_{ij}$, for $i = 1, \cdots, n$, $j = 1, \cdots, m$, with $e_{ij} \sim \text{Uniform}(0.95, 1.05)$.
- The distributions of the match distance for linkages from **X** (original) and **T** (random) turn out to be different (but less different than for additive noise) $\Longrightarrow$ linkages by the intruder are still not plausibly deniable by the protector.
- The protector needs to increase the noise until both distributions are more similar/overlap more.

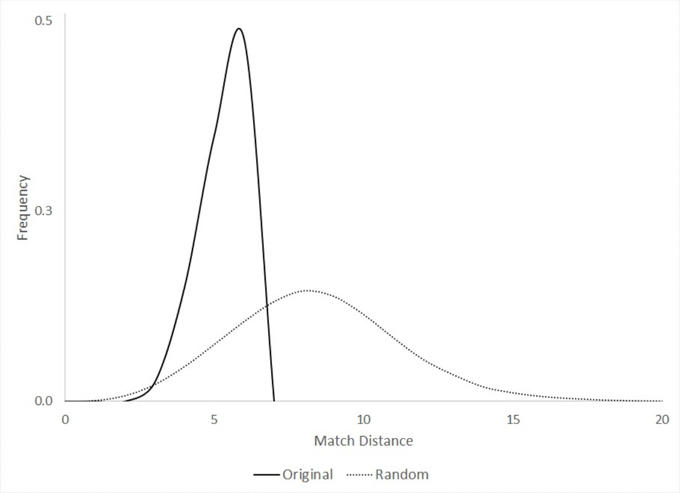
Evaluation of methods: multiplicative noise (II)



Evaluation of methods: rank swapping

- We swap with parameter 15%, that is, for each attribute, the values of records that are within a rank of 6 (15% of $n = 40$) are swapped randomly.
- The distributions of the match distance for linkages from **X** (original) and **T** (random) substantially overlap but are still quite different $\Longrightarrow$ linkages by the intruder are still not plausibly deniable by the protector.
- The protector possibly needs to increase the swapping parameter until both distributions are more similar.

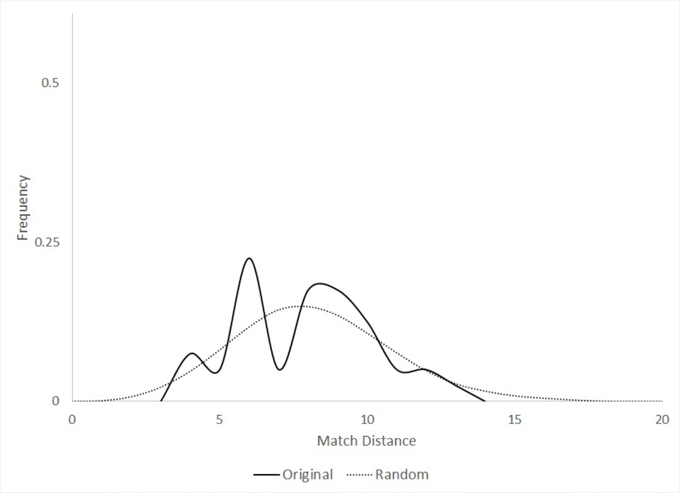
Evaluation of methods: rank swapping (II)



Evaluation of methods: synthetic data

- We generate a synthetic data $\mathbf{Y}$ by sampling from a multivariate normal distribution with mean vector the mean vector of $\mathbf{X}$ and covariance the covariance of $\mathbf{X}$.
- The distributions of the match distance for linkages from $\mathbf{X}$ (original) and $\mathbf{T}$ (random) are quite similar/overlapping $\Longrightarrow$ linkages by the intruder can be plausibly denied by the protector.

Evaluation of methods: synthetic data (II)



Anonymization tuning by the data protector

- The above record linkage verification can also be made by the data protector, who can use it to optimize the amount of permutation that anonymization should introduce.
- The distribution of the record-level permutation distance d for records in $\mathbf{X}$ depends only on the level of anonymization.
- In expectation, d for random records grows with the number N of records, the number m of attributes and is independent of the anonymization level used (the random data set $\mathbf{T}$ contains all possible permutations of original records or a random large subset of them).

Anonymization checking by the data subject

If anonymization involves only permutation without noise addition (swapping, shuffling, etc.), a data subject with access to just her own record in **X** can not only lead *d* for her record, but also verify whether *d* is safe.:

1. The subject generates **T** from the masked data **Z** (**Z** can be used instead of **X** to this end, because one is a permutation of the other).
2. The subject checks whether a match at distance *d* is plausible as a random match.
3. If yes, *d* is safe.

Anonymization transparency towards the user

- Privacy parameters are only explicit under the privacy-first approach (privacy model), but utility-first is more usual.
- Under utility-first, statistical agencies often withhold the parameters used for anonymization (variance of added noise, proximity of swapped values, etc.).
- This is problematic:
  - Users cannot properly evaluate utility.
  - Basing protection on parameter secrecy is a poor idea (violates Kerckhoff's principle).

Anonymization transparency towards the user (II)

- Applying Kerckhoff's principle to anonymization means that the user must be given all anonymization parameters except the random seed(s) (if any are used for pseudo-randomization).
- Transparency does not favor our maximum-knowledge intruder, who can compute record linkages and verify them without any information about the anonymization mechanism.
- Hence, transparency is neutral to intruder and subject and very good to the user.

Alternatives to centralized anonymization: local anonymization

- If a subject can verify the level of anonymization provided by a centralized data protector and she is not satisfied, she may prefer local anonymization.
- Each subject anonymizes her own data before handling them to the data collector.
- Local anonymization requires subjects to anonymize their data without seeing the data of other subjects $\Longrightarrow$ overkill likely $\Longrightarrow$ more information loss than in centralized anonymization.

Universitat Rovira i Virgili

Alternatives to centralized anonymization: collaborative anonymization

- Seeks to empower each subject to anonymize her own data while preserving the utility as in the centralized paradigm.
- Subjects generate the anonymized data set in a distributed and collaborative manner.
- We seek two main properties:
    - Information loss must be equivalent to the information loss that would result from the centralized paradigm for the same privacy level.
- Neither the data collector nor subjects gain more knowledge about the confidential information of a specific subject than disclosed by the anonymized data set.

Subject's motivations

- A rational subject should only contribute if the benefit she gets from participating compensates her privacy loss.
- A *subject without any interest in the collected data* is better off by declining to contribute.
- A *subject without privacy concerns* can directly supply her data without any anonymization requirements.
- A subject who is interested in the collected data but has privacy concerns should prefer the collaborative approach:
  - It outperforms the centralized approach by offering also privacy versus the data collector.
  - It outperforms the local approach in that it yields less information loss.

Co-utility in collaborative anonymization

Co-Utility

The best strategy to attain one's goal is to help others in attaining theirs.

- Co-utility leads to protocols that work smoothly without external enforcing mechanisms.
- In microdata anonymization the privacy protection obtained by a subject affects the privacy protection that others get.
- When masking the identity of a subject within a group, none of the subjects in the group is interested in making any of the other subjects re-identifiable, because that makes her own data more easily re-identifiable.

Universitat Rovira i Virgili

More on co-utility

"CO-UTILITY" project (2014-2017), funded at URV by Templeton World Charity Foundation

`http://crises-deim.urv.cat/co-utility`

Co-utile collaborative k-anonymity

k-Anonymity

Each combination of quasi-identifier values in the data set must be shared by k, or more, records.

- The probability of correctly re-identifying a record in a k-anonymous data set is upper bounded by $1/k$.
- k-Anonymity usually assumes that an attribute is either a quasi-identifier or confidential but not both.
- Collaborative k-anonymity steps:
    - First share the QI so that groups can be generated.
    - Share confidential data at the group level.

Conclusions

- We have presented a new permutation model of anonymization.
- We have introduced a new privacy model, $(\mathbf{d}, \mathbf{v}, f)$-permuted privacy to capture that permutation is the essential principle of anonymization.
- Privacy in this model is verifiable by the subject.
- We have defined a maximum-knowledge intruder, and shown how he can verify the plausibility of record linkages.
- We have applied this to evaluate several anonymization methods.
- We have made the case for anonymization transparency towards the data user.
- We have explored alternatives to centralized anonymization, including collaborative anonymization, which can be sustained by the principle of co-utility.

Further details

- J. Soria-Comas and J. Domingo-Ferrer, "Big data privacy: challenges to privacy principles and data models", *Data Science and Engineering*, 1(1), 2015 (to appear).
- Josep Domingo-Ferrer and Krishnamurty Muralidhar, "New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users", Technical Report, Jan. 17, 2015. `http://arxiv.org/abs/1501.04186`
- Josep Domingo-Ferrer, Jordi Soria-Comas and Oana Ciobotaru, "Co-utility: self-enforcing protocols without coordination mechanisms", in *Proc. of the 2015 International Conference on Industrial Engineering and Operations Management-IEOM 2015*, pp. 1-7.
- Jordi Soria-Comas and Josep Domingo-Ferrer, "Co-utile collaborative anonymization of microdata", in *MDAI 2015-Modeling Decisions for Artificial Intelligence*, LNCS 9321, pp. 192-206, 2015.

Universitat Rovira i Virgili