Protecting Against Maximum-Knowledge Adversaries in Microdata Release: Analysis of Masking and Synthetic Data Using the Permutation Model

Josep Domingo-Ferrer and Krishnamurty Muralidhar

Universitat Rovira i Virgili, Tarragona, Catalonia

Luxemburg, March 24, 2015

Introduction

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

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.

Plan of this presentation

1. Definition of a maximum-knowledge adversary inspired on cryptology.
2. Permutation model of any microdata anonymization method.
3. Record linkage by the adversary.
4. Verifiability of record linkage by the adversary.
5. Evaluation of synthetic data vs masking techniques.
6. The data administrator: data set-level and record-level privacy measures.

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 adversary

- 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 adversary:
  - 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 adversarial model

- Our adversary is stronger than the one considered in differential privacy.
- Our adversary 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 adversary.
- In anonymization, there is one (or few) adversary(ies) and everyone else is deemed a user.

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 $Z = \{z_1, z_2, \cdots, z_n\}$

Notes. If there are several attributes in an original data set **X** and anonymized data set **Y**, the above procedure is repeated for each attribute. Our adversary can run reverse mapping.

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 **X** is permuted to obtain **Z**.
  - **Residual noise addition**. Noise is added to each value of **Z** to obtain the anonymized data set **Y** (the noise is residual, because the ranks of **Z** and **Y** must stay the same).

Record linkage by the adversary: search procedure

Require: $\mathbf{x} = \{x_1, x_i, \cdots, x_m\}$ (target record from the original data set $\mathbf{X}$, say i-th record in $\mathbf{X}$)

Require: Anonymized data set $\mathbf{Y}$

Evaluate the rank of $\mathbf{x}$ in $\mathbf{X}$ w.r.t. each attribute

$d = 0$

while no record in $\mathbf{Y}$ is found whose rank is within $\pm d$ of $\mathbf{x}$ across all attributes **do**

$d = d + 1$

end while

return match (permutation) distance d and found record(s) in $\mathbf{Y}$

Comments on the adversary's search procedure

- The search procedure can be run by the maximum-knowledge adversary, but also by the data administrator.
- Match distance $d \leq n$ (no. of records).
- If a single record in $\mathbf{Y}$ is found, say $\mathbf{y}$, the adversary concludes that the correct linkage of $\mathbf{x} \in \mathbf{X}$ is $\mathbf{y}$.
- When multiple records in $\mathbf{Y}$ are found, the adversary is unsure about the linkage, but repeating the search for every record in $\mathbf{X}$ may allow to refine the linkage.
- E.g. if for $\mathbf{x}_{i_1} \in \mathbf{X}$ both $\mathbf{y}_{j_1}, \mathbf{y}_{j_2} \in \mathbf{Y}$ are found, but for, say, $\mathbf{x}_{i_2} \in \mathbf{X}$, $\mathbf{y}_{j_2} \in \mathbf{Y}$ is again found whereas $\mathbf{y}_{j_1}$ is found for no other record in $\mathbf{X}$, then the adversary concludes that $\mathbf{x}_{i_1}$ is linked to $\mathbf{y}_{j_1}$.

Verifiability of record linkage

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

Verification procedure by the adversary

Require: $\mathbf{X}$ original data set with n records and m attributes $X_1, \cdots, X_m$; results from the search procedure above for all records in $\mathbf{X}$; N large integer ($n \gg N \leq n^m$)

for $i = 1$ to N **do**

  for $j = 1$ to m **do**

    Set t_{ij} to a value randomly selected from X_j

  end for

  Add $\mathbf{t}_i = \{t_{i1}, \cdots, t_{im}\}$ as the i-th record of a random data set $\mathbf{T}$

end for

Use the search procedure taking as targets all records in $\mathbf{T}$

return Return similarity measure between distributions of the match distance for linkages from $\mathbf{X}$ and $\mathbf{T}$

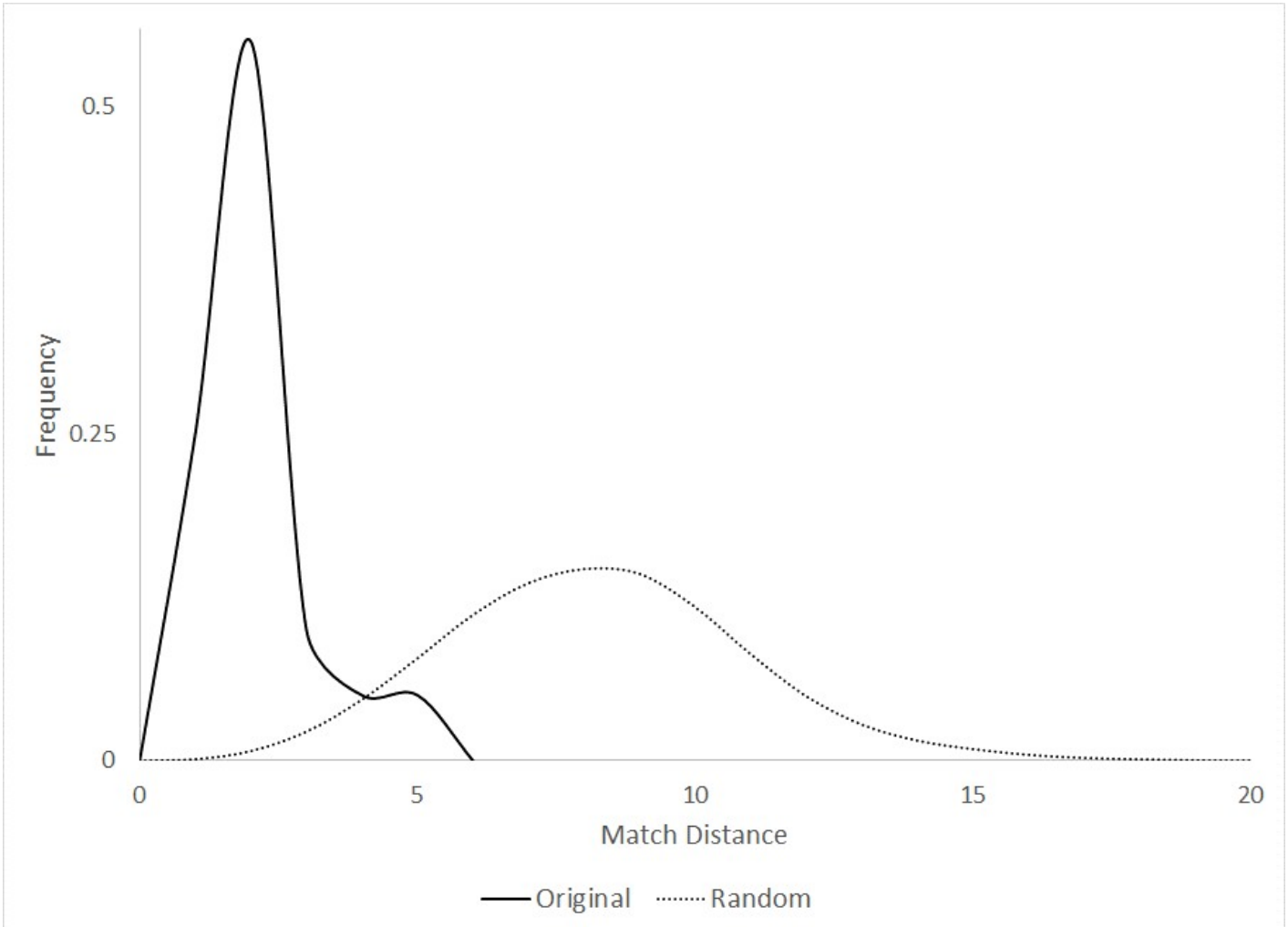
Evaluation of methods: deterministic masking

- Any deterministic masking method allows our maximum-knowledge adversary to exactly reconstruct the anonymization process from $\mathbf{X}$.
- Hence, it allows the adversary 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 adversary are not plausibly deniable by the administrator.
- The administrator needs to increase the noise until both distributions are more similar/overlap more.

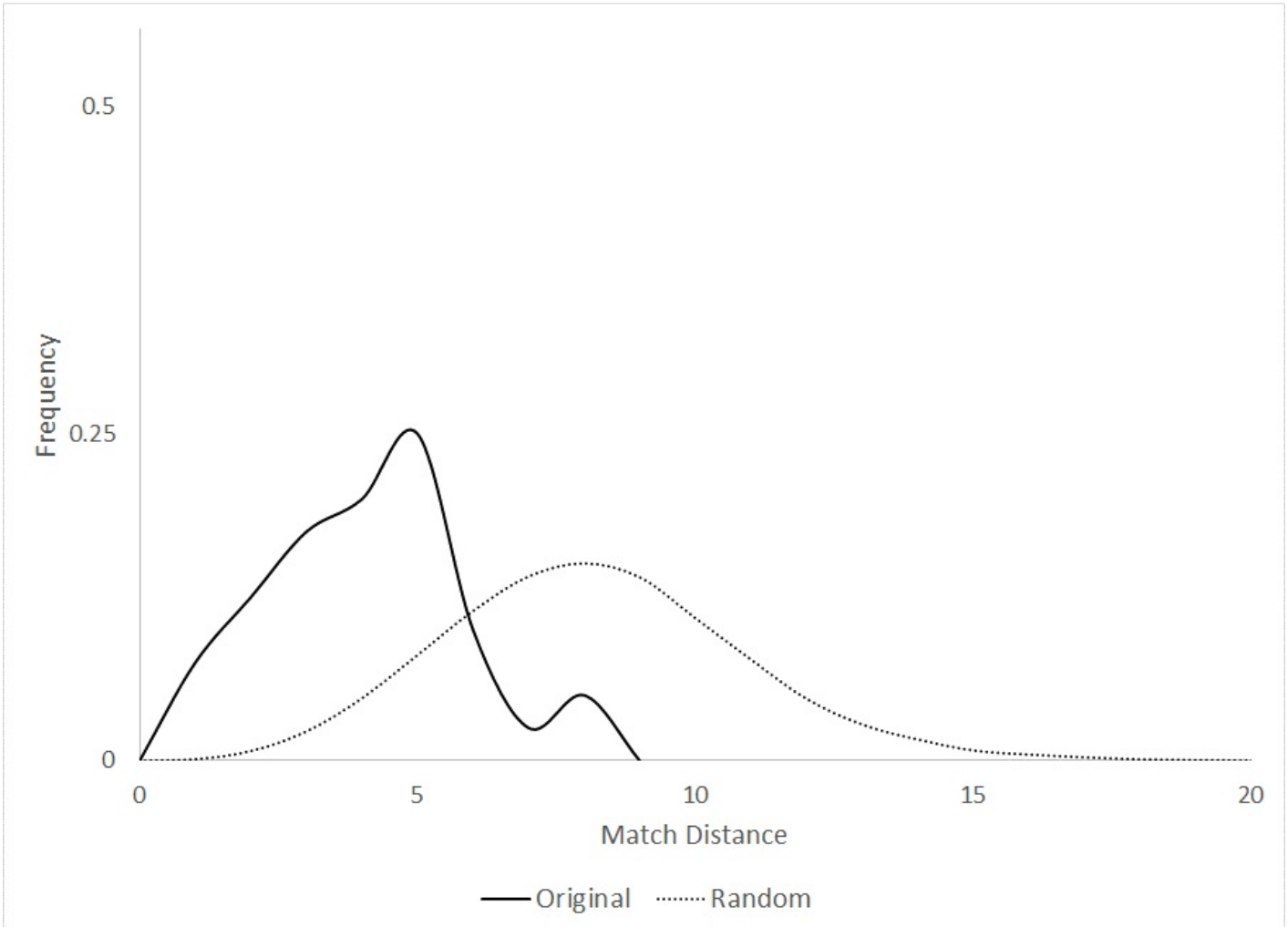
Evaluation of methods: additive noise (II)



Evaluation of methods: multiplicative noise

- We use the same $\mathbf{X}$ 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 $\mathbf{X}$ (original) and $\mathbf{T}$ (random) turn out to be different (but less different than for additive noise) $\Longrightarrow$ linkages by the adversary are still not plausibly deniable by the administrator.
- The administrator needs to increase the noise until both distributions are more similar/overlap more.

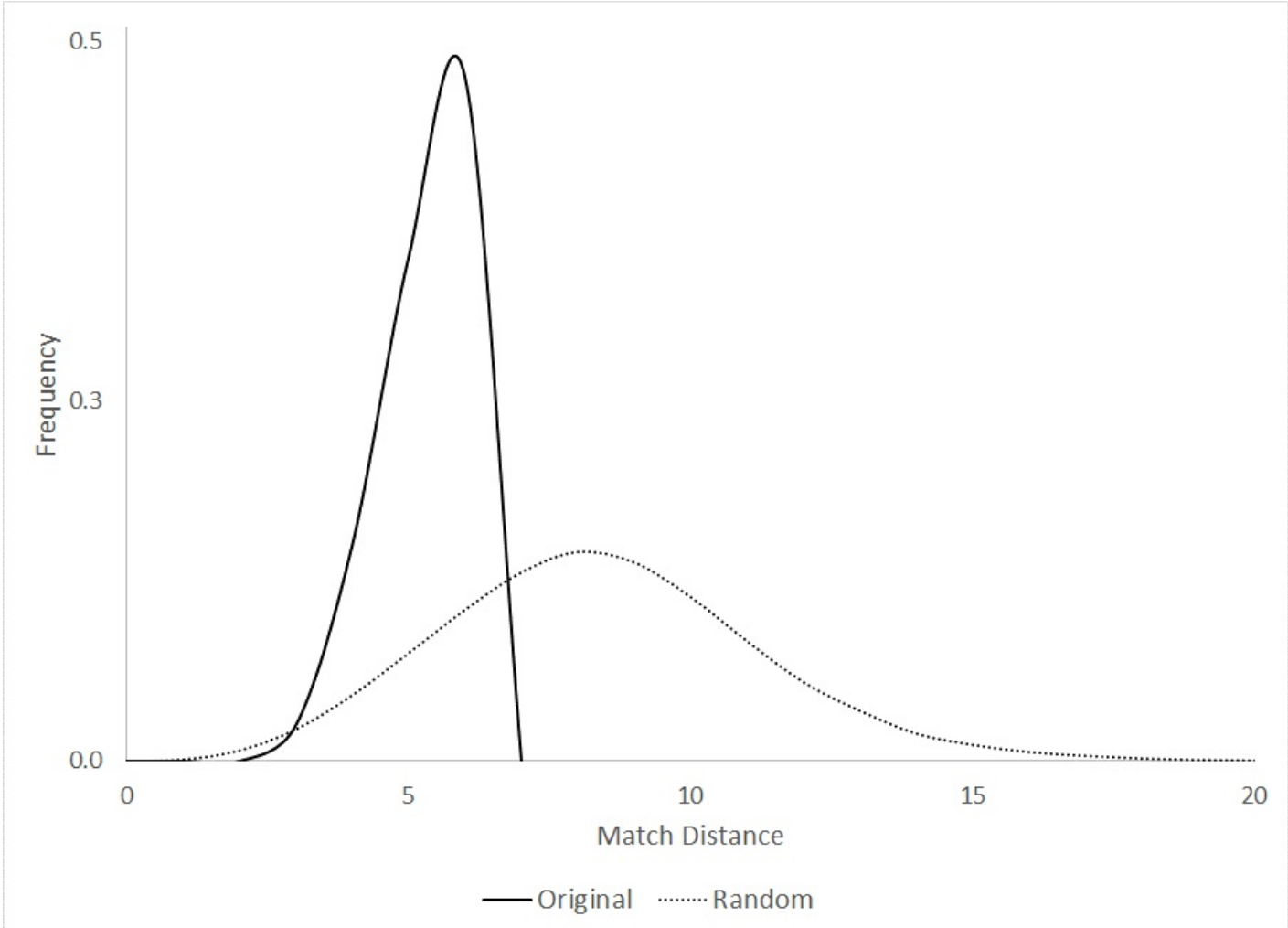
Evaluation of methods: multiplicative noise (II)



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 adversary are still not plausibly deniable by the administrator.
- The administrator possibly needs to increase the swapping parameter until both distributions are more similar.

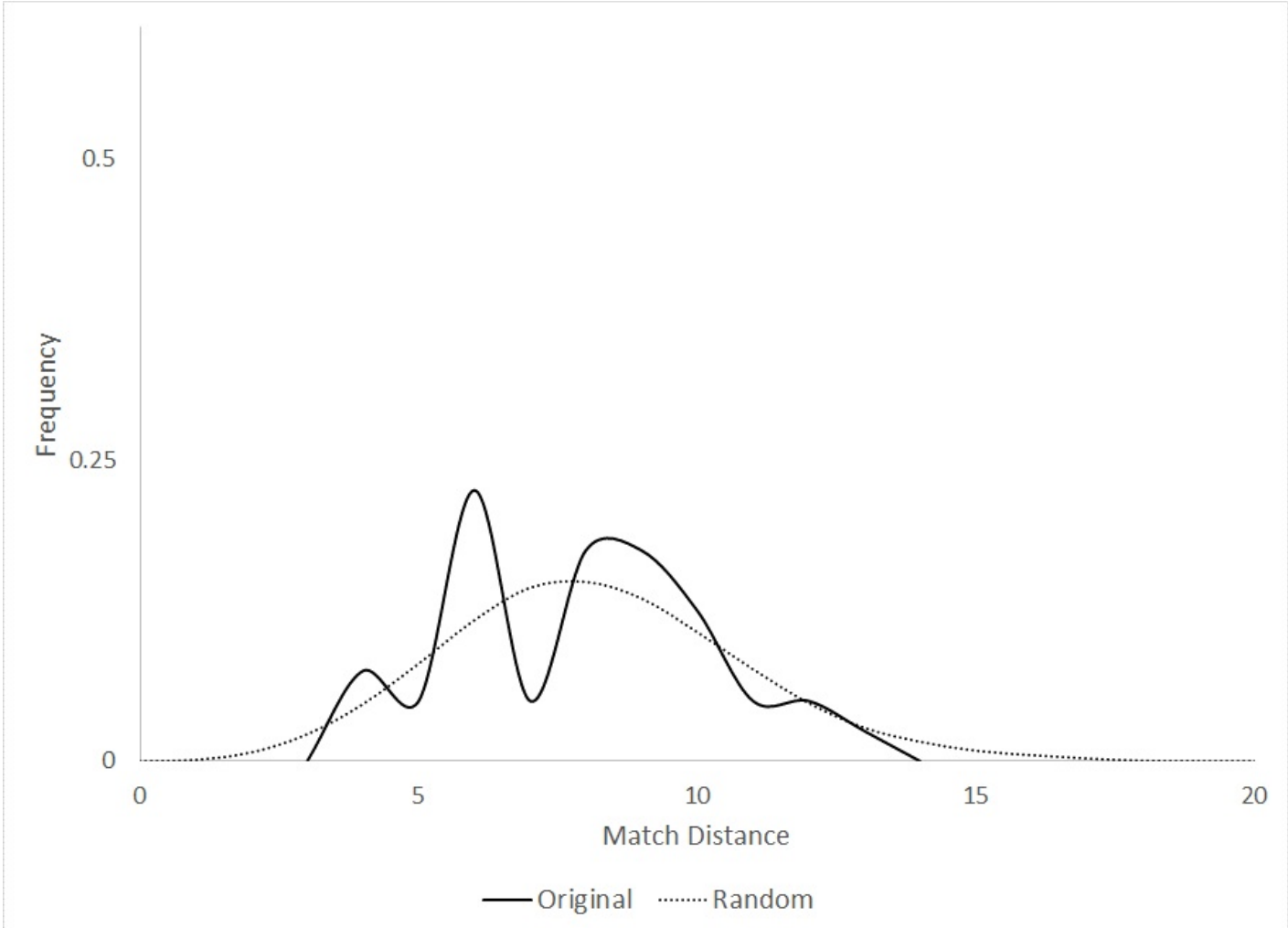
Evaluation of methods: rank swapping (II)



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 adversary can be plausibly denied by the administrator.

Evaluation of methods: synthetic data (II)



The data administrator: data set-level privacy measure

- Measure the similarity between the match distance distributions from **X** and **T** using some metric, *e.g.*, the Hellinger distance $H(P, Q)$.
- If the distribution from **X** is $P = (p_0, p_1, \cdots, p_n)$ and the distribution from **T** is $Q = (q_0, q_1, \cdots, q_n)$

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=0}^{n} (\sqrt{p_i} - \sqrt{q_i})^2}$$

- $H(P, Q) \in [0, 1]$ with 0 meaning $P = Q$ and 1 meaning no overlap at all between P and Q.

The data administrator: record-level privacy measure

Identity disclosure For each record $\mathbf{x}_i \in \mathbf{X}$ we measure the permutation distance for that record as

$$d_i = \max_{j=1,\cdots,n} |Rank(x_{ij} - y_{ij})|$$

Attribute disclosure Let k represent the rank of x_{ij} with respect to attribute X_j. The attribute disclosure of x_{ij} regarding X_j is measure as

$$Var\{y_{[k+l],j} | -d_i \leq l \leq d_i\}$$

where $y_{[k+l],j}$ is the value of Y_j with rank $k + l$.

Conclusions

- We make a worst-case assumption on our adversary, in order to avoid assumptions on background knowledge.
- Anonymization basically amounts to permutation.
- The permutation model allows verifying anonymization based on the permutation/match distance of linkages.
- Verification can be done by the adversary and the administrator (who can use it to tune the anonymization parameters).
- Synthetic data can achieve a good protection while offering pre-selected utility guarantees.

Further details

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`