Microdata masking as permutation

Krish Muralidhar
Price College of Business
University of Oklahoma

Josep Domingo-Ferrer
UNESCO Chair in Data Privacy
Dept. of Computer Engineering and Mathematics
Universitat Rovira i Virgili

Diversity of microdata masking mechanisms

- A wide variety of microdata masking mechanisms are available
  - Rounding
  - Microaggregation
  - Noise infusion
    - Additive
    - Multiplicative
  - Model based
  - Data swapping
  - Data shuffling
  - And many others

Diversity is good, but …

- Diversity in mechanisms also means that comparing across mechanisms can be very difficult

Traditional approaches for comparison

- Based on parameters
- Based on performance

Comparison based on parameters

- Syntactic approach
  - This approach has been criticized since it does not reflect the true security offered by the mechanism
- Difficulty of comparing across different mechanisms
  - How do you compare microaggregation with aggregation parameter = 5, noise addition with 10% of variance, multiplicative perturbation with parameter = 10%?
- Difficulty of comparing across data sets even for the same mechanism
  - Two data sets with different characteristics may yield completely different levels of protection for the same parameter selection

Impact of data characteristics

- Two data sets masked using multiplicative perturbation drawn from UNIFORM(0.9, 1.1) range.
- $Y = X \times e$
  - e is the masking value
- Value in data set 2 = value in data set 1 + 1000
- Same procedure but completely different results for two closely related data sets

ID	Data set 1	Data set 2	Masking value	Masked data set 1	Masked data set 2	Rank of masked data set 1	Rank of masked data set 2
1	1	1001	1.091	1.091	1092.361	1	7
2	11	1011	0.907	9.981	917.356	2	1
3	21	1021	1.004	21.077	1024.742	3	4
4	31	1031	1.003	31.088	1033.931	4	5
5	41	1041	1.090	44.676	1134.341	5	10
6	51	1051	1.051	53.601	1104.594	6	8
7	61	1061	0.929	56.677	985.814	7	2
8	71	1071	0.985	69.964	1055.371	8	6
9	81	1081	0.912	73.901	986.258	9	3
10	91	1091	1.015	92.334	1106.988	10	9

Comparison based on performance

- Analyze the masked data for disclosure risk
  - Identity disclosure
  - Value disclosure
- Comparison based on results
- Many different approaches for assessing identity and value disclosure
  - One alternative is to aggregate the different measures
    - Raises more questions about how to aggregate

Example of comparison based on performance

- An empirical evaluation
  - Score = 0.5(IL)+ 0.125(DLD)+0.125(PLD)+ 0.25 (ID)
- While this is a reasonable approach, it can be argued that the weights should be different
- If we remove (or modify the weight of) a criterion, the results may be different
  - If we use only disclosure risk measures, the results would be different
- What about alternative measures of information loss and disclosure risk?
- These are typical problems with any empirically based evaluation

"A Quantitative Comparison of Disclosure Control Methods for Microdata" in *Confidentiality, Disclosure, and Data Access*

Table 1: Comparison Results for Continuous Microdata

Method	IL	DLD	PLD	ID	Score	IL-Rank	DLD-Rank	PLD-Rank	ID-Rank
Rank15	19.01	1.19	0.15	35.05	18.44	53	6	7	21
Rank19	22.95	0.93	0.08	28.04	18.61	59	2	2	2
Rank16	20.91	1.39	0.11	32.18	18.69	56	8	5	16
Rank13	16.77	2.17	0.12	40.35	18.76	48	12	6	28
Rank14	19.72	1.92	0.07	37.00	19.36	55	10	1	25
Rank11	14.32	2.43	0.25	47.81	19.45	44	13	14	39
Rank12	16.37	2.50	0.25	43.73	19.46	47	14	11	35
Rank20	25.81	0.69	0.09	26.83	19.71	64	1	3	1
Rank18	25.74	0.95	0.09	29.25	20.31	63	4	4	6
Rank10	13.37	3.90	0.38	53.17	20.51	41	24	17	45
Rank17	25.12	1.52	0.20	30.95	20.51	61	9	9	10
Rank09	11.66	5.01	0.52	57.58	20.91	38	37	29	49
Rank08	11.60	6.07	0.85	63.37	22.51	37	39	39	56
Rank07	9.25	7.51	1.08	68.71	22.87	30	41	43	63
Rank06	7.87	9.02	2.79	73.80	23.86	26	43	56	71
Mic3mul07	11.06	19.34	4.70	72.34	26.62	36	68	65	69
Rank05	6.78	16.80	13.60	78.89	26.91	22	58	70	77
Mic3mul09	13.46	19.22	3.44	69.91	27.04	42	67	60	65
Mic3mul10	14.84	17.99	3.44	68.61	27.25	46	64	59	62
Mic4mul04	12.14	19.76	6.67	71.85	27.33	39	69	68	68
Mic4mul05	14.50	17.43	5.45	69.09	27.39	45	61	66	64
Mic3mul08	13.51	20.81	4.15	70.68	27.54	43	71	63	66
Mic4mul08	18.89	17.78	3.35	62.84	27.80	52	62	58	55
Mic3mul06	10.24	20.41	13.90	74.00	27.91	33	70	71	72
Mic4mul07	19.36	17.10	2.08	64.41	28.18	54	60	53	58
Mic4mul06	17.91	17.82	3.98	66.41	28.28	50	63	62	60
Mic4mul09	21.35	15.93	2.00	61.66	28.33	58	57	52	54
Mic4mul10	22.98	16.85	2.37	60.56	29.03	60	59	55	51
Mic3mul05	9.73	23.78	18.29	76.59	29.27	31	76	73	74
Mic3mul04	7.45	23.49	22.75	79.14	29.29	24	75	75	79

Desiridata

- A common basis of comparison for microdata masking mechanisms that is
  - Applicable to all mechanisms,
  - Meaningful,
  - Independent of the
    - parameters of the mechanism
    - risk assessment measure
    - characteristics of the data

Our proposal: The permutation model

- All microdata masking mechanisms can be viewed as permutations of the original data
- The permutation model is
  - Meaningful
  - Independent of the
    - parameters of the masking mechanism
    - risk assessment measures
    - characteristics of the data

Traditional view of microdata masking

ID	X		Masked
1	44		24.76
2	14		21.51
3	42		53.97
4	24	Masking	25.93
5	93	mechanism	94.36
6	41	→	36.66
7	94		84.38
8	54		58.22
9	16		34.35
10	26		22.80

Reverse mapping

- (Reverse) Map the masked data back to the original data
  - Compute rank of masked value
  - Replace the masked value with the value of the original data with the same rank
    - Rank of the first masked observation is 3
    - Replace this value with the value of X with rank of 3
    - Repeat for all masked records
- The reverse mapped values represent the permuted version of the original data

ID	X	Rank of X	Masked	Rank of Masked		Permuted
1	44	7	24.76	3		24
2	14	1	21.51	1		14
3	42	6	53.97	7		44
4	24	3	25.93	4	Reverse mapping →	26
5	93	9	94.36	10		94
6	41	5	36.66	6		42
7	94	10	84.38	9		93
8	54	8	58.22	8		54
9	16	2	34.35	5		41
10	26	4	22.80	2		16

Permuted + Residual Noise = Masked

ID	X	Rank of X	Masked	Rank of Masked	Reverse mapping →	Permuted	Noise	Masked
1	44	7	24.76	3		24	0.76	24.76
2	14	1	21.51	1		14	7.51	21.51
3	42	6	53.97	7		44	9.97	53.97
4	24	3	25.93	4		26	-0.07	25.93
5	93	9	94.36	10		94	0.36	94.36
6	41	5	36.66	6		42	-5.34	36.66
7	94	10	84.38	9		93	-8.62	84.38
8	54	8	58.22	8		54	4.22	58.22
9	16	2	34.35	5		41	-6.65	34.35
10	26	4	22.80	2		16	6.80	22.80

The permutation model

- Any masking mechanism can be represented by the permutation model
- The masked output from any microdata masking mechanism is conceptually viewed as (or functionally equivalent to) permutation plus residual noise.
  - We are not suggesting a new masking mechanism.

ID	X	Permuted X	Noise	Masked
1	44	24	0.76	24.76
2	14	14	7.51	21.51
3	42	44	9.97	53.97
4	24	26	-0.07	25.93
5	93	94	0.36	94.36
6	41	42	-5.34	36.66
7	94	93	-8.62	84.38
8	54	54	4.22	58.22
9	16	41	-6.65	34.35
10	26	16	6.80	22.80

Magnitude of the residual noise

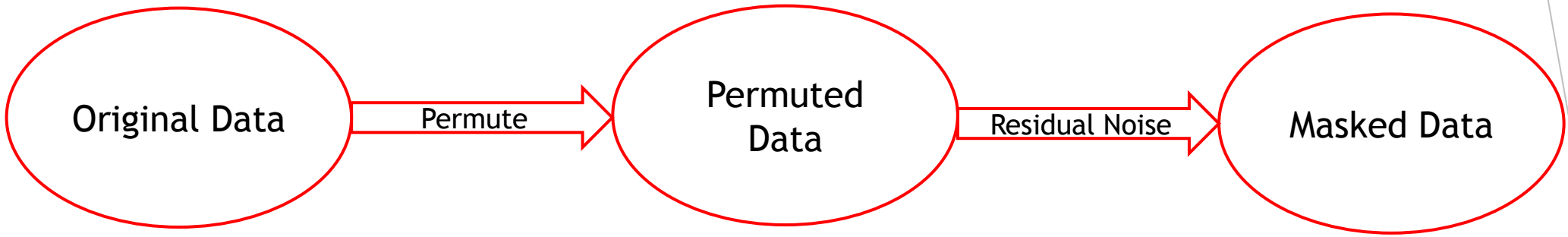
- The magnitude of the residual noise will be small
- Why?
  - The residual noise cannot change the permutation. Hence, the permutation automatically imposes a bound on the residual noise
    - Record ID 6: Permuted value = 42
    - Bounds for masked value: [41, 44]
    - Residual noise bound: [-1, 3]
- Magnitude of the residual noise inversely related to n

ID	X	Permuted X	Noise	Masked
1	44	24	0.76	24.76
2	14	14	7.51	21.51
3	42	44	9.97	53.97
4	24	26	-0.07	25.93
5	93	94	0.36	94.36
6	41	42	-5.34	36.66
7	94	93	-8.62	84.38
8	54	54	4.22	58.22
9	16	41	-6.65	34.35
10	26	16	6.80	22.80

For large data sets …

- Disclosure prevention is achieved primarily through permutation
- The residual noise provides additional (but small level of) masking to prevent the original values from being released
  - With procedures such as swapping and shuffling, there is no residual noise since the original values are released unmodified

The permutation model



Protection level

- Meaningful interpretation of protection
  - No permutation = No protection
  - Randomly sorted data = Maximum protection
  - Simple, meaningful explanation of the protection level
- Actual: Level of permutation resulting from the masking mechanism

The adversary model

- The permutation model also leads to a natural maximum knowledge adversary
- We assume that the adversary has the ability to perform reverse mapping on the masked data
- Reverse mapping can be performed if the adversary has access to the entire original data set
- This assumption is the same as that used in record linkage – the adversary has access to both the original and masked data set (but not the individual record linkages)
  - Consistent with Kerckhoff's principle that the adversary knows everything but the "key"

Cryptographic equivalent

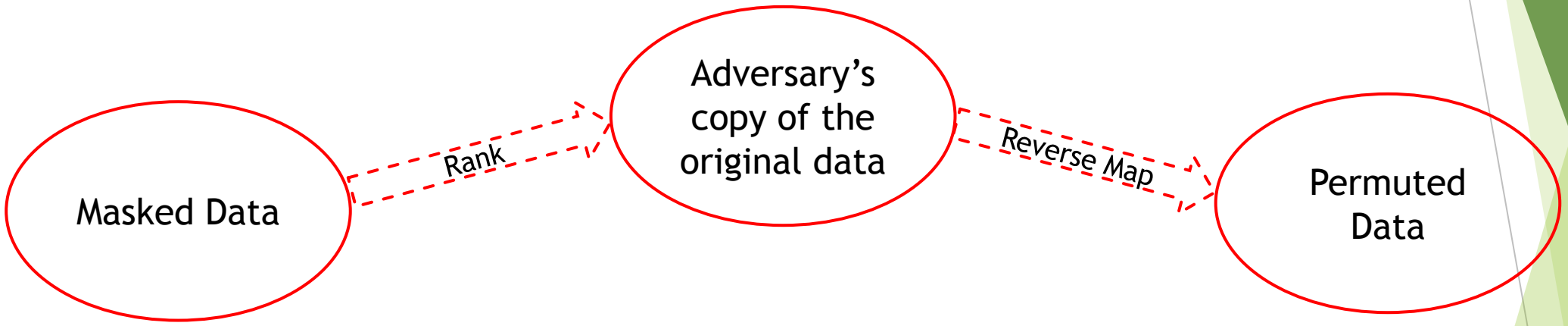
- Ciphertext-only
  - Adversary has access only to ciphertext (i.e. masked records).
- Known-plaintext
  - Adversary has access to pairs plaintext/ciphertext (i.e. pairs original and masked records)
  - In a non-interactive setting (microdata release), known-plaintext is the strongest possible attack
  - Our adversary model
- Chosen-plaintext
  - Adversary can choose a plaintext (original records) and get the corresponding ciphertext (masked records)
- Chosen-ciphertext
  - Adversary can choose a ciphertext (masked records) and get the corresponding plaintext (original records)

Adversary with malicious intent

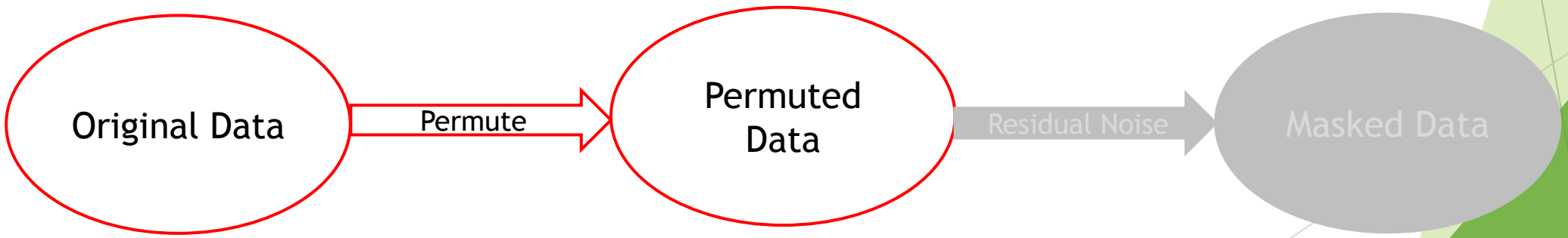
- One of the difficulties with microdata release was the inability to distinguish between the user and adversary
- A practical way of thinking of this adversary is that the intent of this adversary is purely malicious
- Since the adversary has access to the entire original data set, they cannot learn anything new from the data set
- Our adversary model differentiates the malicious adversary (who does not learn anything from the released data) from the user (who learns something from the data)

Adversary model

- The adversary is able to eliminate residual noise through reverse mapping



- The only protection against this adversary is permutation



Adversary objective

- The objective of the adversary is to break the key (recreate the linkage between the original and permuted data)
- The adversary wishes to show provable linkage
- Provable linkage eliminates plausible deniability

Auxiliary information

- One of the advantages of our adversary model is that it eliminates the need to consider auxiliary information
- Our adversary has maximum knowledge (has access to the entire original data)
- No auxiliary information (other than the random number seed) will help the adversary improve the linkage

Important clarification

- We are suggesting the adversary model for comparison benchmarks
- For risk assessment; NOT necessarily risk mitigation

On-going work

- Formalizing a measure of the permutation level
- Formalizing a measure of disclosure
- Multivariate scenario

Conclusion

- The permutation model offers a new approach for evaluating the efficacy and effectiveness of masking mechanisms
- It allows the data administrator to compare different masking mechanisms using the same benchmark
- More work remains

Questions, comments, or suggestions?

Thank you