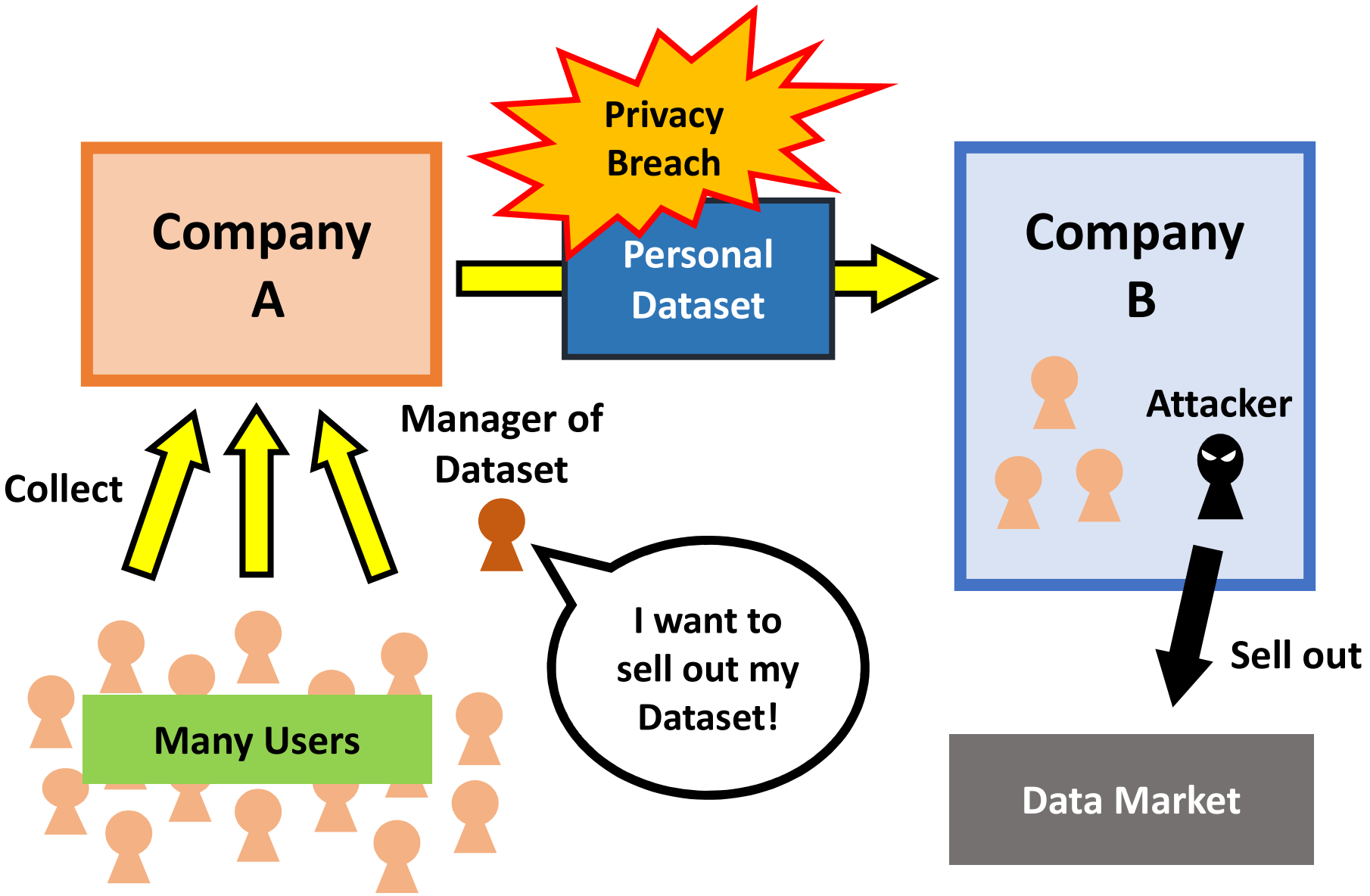
Risk of Re-identification Based on Euclidean distance in Anonymized Data PWSCUP2015

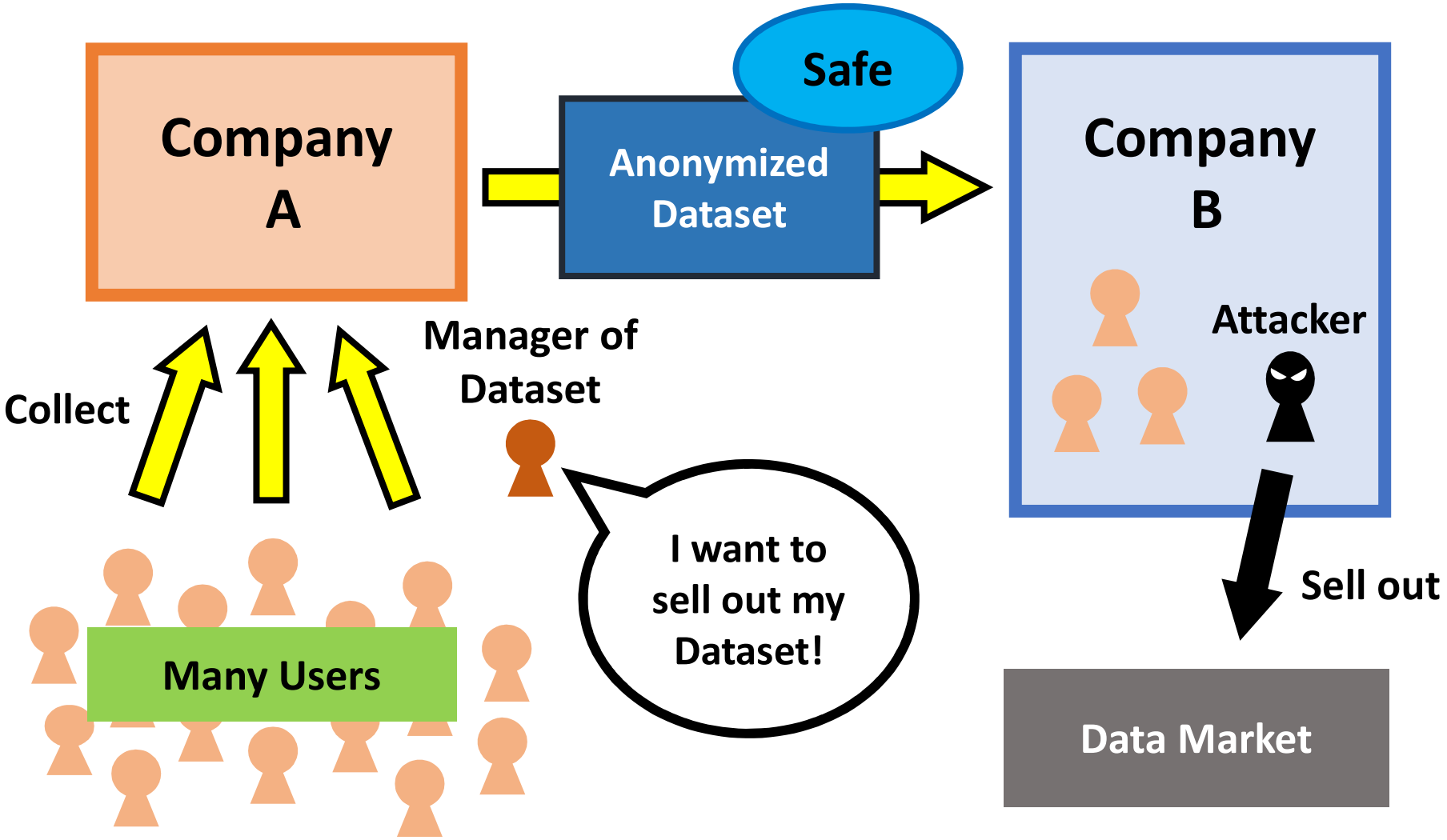
Satoshi Ito, Hiroaki Kikuchi

Meiji University Graduate School

What is Anonymization?



What is Anonymization?



What is Anonymization?

Anonymization: method to modify the personal datasets so that individuals cannot be identified.

Dataset with personal data

name	age	goods	payment
H. Kikuchi	27	coffee	320
S. Ito	23	tea	280

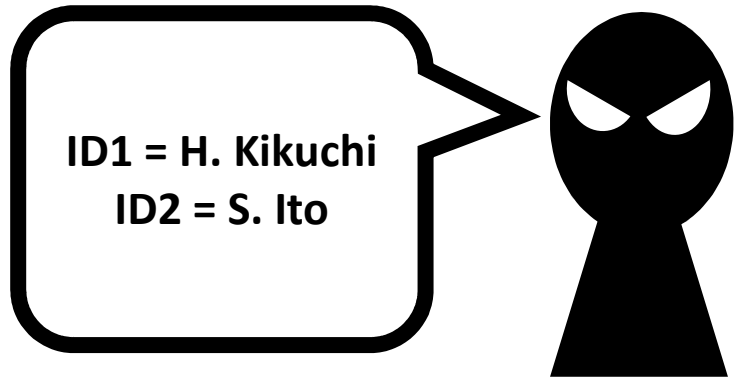
Anonymized dataset

ID	age	goods	payment
1	20s	beverage	300
2	20s	beverage	200

Re-identification: method to identify individuals from the anonymized dataset.

Anonymized dataset

ID	age	goods	payment
1	20s	beverage	300
2	20s	beverage	200



What is Anonymization?

Quasi-Identifier (QI): a discrete attribute that can be used to identify individuals when being combined.

(e.g. age, sex, address)

Sensitive Attribute (SA): a continuous attribute that we should be treated carefully.

(e.g. name of diseases, yearly income, expenses)

Dataset with personal data

name	age	goods	payment
H. Kikuchi	27	coffee	320
S. Ito	23	tea	280

QI **SA**

Anonymized data and PWSCUP

In Japan, the act on the protection of personal information was amended in September 2015.

And the data anonymization competition PWSCUP has been held since 2015.

PWSCUP 2015

PWSCUP 2016

PWSCUP 2017

Problem 1: The existing Re-identification methods

In the PWSCUP 2015, four re-identification methods were used to evaluate the security of anonymized dataset.

Method	Details
identify.rand	Identify user randomly.
identify.sa	Identify user from 1 sensitive attribute (SA) of dataset.
identify.sort	Identify user by sorting sum of SA.
identify.sa21	Identify user from a specific SA.

The qualities of these methods are not good enough to re-identify because these methods use too less attribute of dataset to re-identify.

Problem 2: The de-identified dataset of PWSCUP2015

In the competition, a total of 24 anonymized datasets were submitted from 13 teams.

In our research, we use 12 datasets from 5 teams.

Data	Team	Rank
D_1, D_2	T_A (Meiji University)	
D_3, D_4	T_B	2
D_5, D_6	T_C	
D_7, D_8, D_9	T_D	1
D_{10}, D_{11}, D_{12}	T_E	3

However, since only anonymized data were evaluated without the source code, algorithms used to generate these datasets were unknown.

Our approach

1. **The qualities of the existing methods are not good.**

   →We propose a new Re-identification method based on the Euclidean distance and compare our method with the existing methods for the dataset of PWSCUP2015.

2. **The anonymization methods of the anonymized dataset of PWSCUP2015 are unknown.**

   →We observe the properties of the single method for smaller test data and estimate the algorithm used in the competition based on the known properties.

Our method: identify.euc

identify.euc

Our method identifies individuals by Euclidean distance between values of SA.

Original Data

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	2	300	200
1	1	2	400	500

14.1

322.8

Anonymized Data

QI1	QI2	QI3	SA1	SA2
2	1	1	110	90
2	1	1	220	390
1	1	2	280	210
1	1	2	390	520

Difference between our method and the existing method

The existing method: identify.sa

Original Data

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	110	300
1	1	2	300	200
1	1	2	400	500

✔

50

40

Anonymized Data

QI1	QI2	QI3	SA1	SA2
2	1	1	150	100
2	1	1	160	300
1	1	2	350	200
1	1	2	450	500

Our method: identify.euc

Original Data

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	110	300
1	1	2	300	200
1	1	2	400	500

✔

50

203.96

Anonymized Data

QI1	QI2	QI3	SA1	SA2
2	1	1	150	100
2	1	1	160	300
1	1	2	350	200
1	1	2	450	500

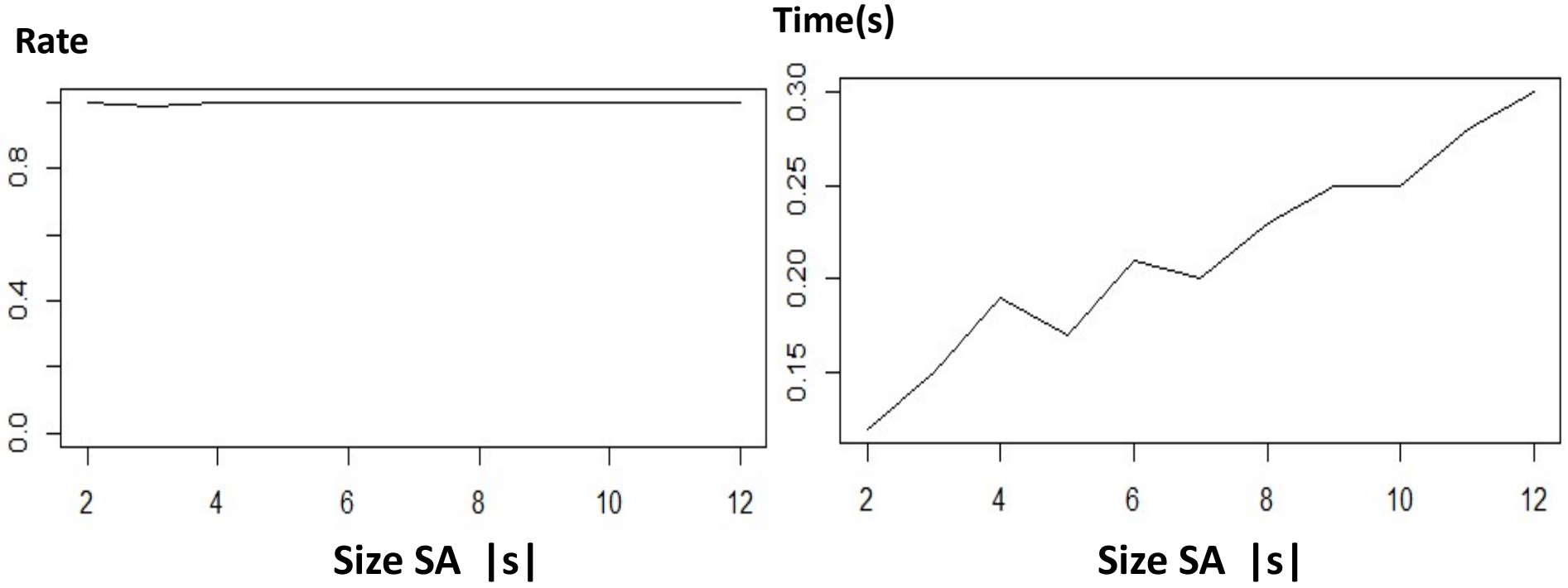
Result of re-identification rate

	Existing				Our
Data	id-rand	id-sa	id-sort	id-sa21	EUC1
D_1	0.033	0.824	1.000	0.186	0.301
D_2	0.649	0.651	0.001	0.002	0.478
D_3	0.199	0.241	0.248	0.051	0.207
D_4	0.189	0.240	0.253	0.045	0.211
D_5	0.000	0.022	0.000	0.000	0.074
D_6	0.000	0.022	0.000	0.000	0.074
D_7	0.002	0.022	0.009	0.001	0.876
D_8	0.000	0.022	0.000	0.000	0.001
D_9	0.000	0.000	0.000	0.000	0.002
D_{10}	0.006	0.007	0.000	0.000	0.004
D_{11}	0.018	0.016	0.000	0.000	0.008
D_{12}	0.021	0.021	0.000	0.000	0.008
Average	0.093	0.172	0.126	0.024	0.187
Standard Deviation	0.174	0.258	0.268	0.050	0.243
Best Score	2	3	3	0	5

Our proposed method is the best average rate for these methods and re-identify well for most of data.

Performance of our proposed method

We show the performance of our proposed method with small data.



Analysis about anonymized data

We guess what anonymization methods were used in D_1,…,D_{12} based on the result of known datasets dataD_A,…,D_H.

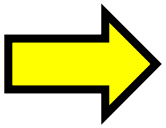
Data	Method	Target
D_A	K-anonymization	QI
D_B	Adding noise to SA	SA
D_C	Cheating attack	ID
D_D	Unification QI 1	QI
D_E	Unification QI 2	QI
D_F	Averaging SA	SA
D_G	Swapping QI	SA
D_H	Deleting records	Record

Data	Method
D_1	?
D_2	?
D_3	?
D_4	?
D_5	?
D_6	?
D_7	?
D_8	?
D_9	?
D_{10}	?
D_{11}	?
D_{12}	?

Examples of de-identification methods

- **K-anonymization**

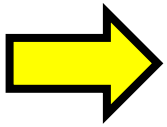
QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	2	200	400
1	1	1	300	200
1	1	2	400	500



QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	1	300	200
1	1	1	400	500

- **Averaging SA**

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	1	300	200
1	1	1	400	500



QI1	QI2	QI3	SA1	SA2
2	1	1	150	250
2	1	1	150	250
1	1	1	350	350
1	1	1	350	350

Effect of combination methods

	known		unknown
	D_A	D_F	D_{10}
Method	K-anony	Averaging	K-ano + Ave
U1	-	-	-
U2	negative	-	negative
U3	negative	-	negative
U4	-	negative	negative
U5	-	negative	slightly
U6	-	-	-
S1	positive	negative	positive
S2	positive	negative	positive
E1	slightly	negative	slightly
E2	slightly	negative	slightly
E3	negative	positive	positive
E4	negative	positive	positive
EUC1	slightly	negative	positive

Result 2: Evaluation and Prediction of data of PWSCUP2015

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}
$U1$	-	-	-	-	-	-	-	-	-	-	-	-
$U2$	negative	-	negative	negative	-	-	negative	-	-	negative	negative	negative
$U3$	negative	-	slightly	slightly	-	-	slightly	-	-	negative	negative	slightly
$U4$	-	slightly	-	slightly	slightly	slightly	slightly	slightly	slightly	negative	negative	negative
$U5$	-	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly
$U6$	-	-	-	-	-	-	-	-	-	-	-	-
$S1$	-	-	slightly	slightly	-	-	slightly	-	-	positive	positive	slightly
$S2$	-	-	slightly	slightly	slightly	positive	positive	positive	positive	positive	positive	positive
$E1$	slightly	negative	negative	negative	positive	positive	positive	positive	positive	slightly	slightly	slightly
$E2$	negative	negative	negative	negative	slightly	slightly	slightly	positive	positive	slightly	slightly	slightly
$E3$	negative	positive	negative	negative	positive	positive	positive	positive	positive	positive	positive	positive
$E4$	negative	positive	slightly	slightly	positive	positive	positive	positive	positive	positive	positive	positive
$EUC1$	negative	negative	negative	negative	slightly	slightly	negative	positive	positive	positive	positive	positive
D_a	-	-	×	×	-	-	×	-	-	×	×	×
D_b	-	-	-	-	-	-	-	-	-	-	-	-
D_c	-	-	-	-	×	×	-	×	×	-	-	-
D_d	-	-	-	-	×	×	×	-	-	-	-	-
D_e	×	-	-	-	-	-	-	×	×	-	-	-
D_f	-	×	-	-	-	-	-	-	-	×	×	×
D_g	-	-	×	×	-	-	×	×	×	-	-	-
D_h	-	-	-	-	-	-	-	-	-	-	-	-

Result 2: Evaluation and Prediction of data of PWSCUP2015

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}
$U1$	-	-	-	-	-	-	-	-	-	-	-	-
$U2$	negative	-	negative	negative	-	-	negative	-	-	negative	negative	negative
$U3$	negative	-	slightly	slightly	-	-	slightly	-	-	negative	negative	slightly
$U4$	-	slightly	-	slightly	slightly	slightly	slightly	slightly	slightly	negative	negative	negative
$U5$	-	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly
$U6$	-	-	-	-	-	-	-	-	-	-	-	-
$S1$	-	-	slightly	slightly	-	-	slightly	-	-	positive	positive	slightly
$S2$	-	-	slightly	slightly	slightly	positive	positive	positive	positive	positive	positive	positive
$E1$	slightly	negative	negative	negative	positive	positive	positive	positive	positive	slightly	slightly	slightly
$E2$	negative	negative	negative	negative	slightly	slightly	slightly	positive	positive	slightly	slightly	slightly
$E3$	negative	positive	negative	negative	positive	positive	positive	positive	positive	positive	positive	positive
$E4$	negative	positive	slightly	slightly	positive	positive	positive	positive	positive	positive	positive	positive
$EUC1$	negative	negative	negative	negative	slightly	slightly	negative	positive	positive	positive	positive	positive
D_a	-	-	×	×	-	-	×	-	-	×	×	×
D_b	-	-	-	-	-	-	-	-	-	-	-	-
D_c	-	-	-	-	×	×	×	×	×	-	-	-
D_d	-	-	-	-	×	×	×	-	-	-	-	-
D_e	×	-	-	-	-	-	-	-	×	-	-	-
D_f	-	×	-	-	-	-	-	-	-	×	×	×
D_g	-	-	×	×	-	-	×	×	×	-	-	-
D_h	-	-	-	-	-	-	-	-	-	-	-	-

Group 1
EUC1 is effective

Group 2
Cheating + Other

Group 3
K-anonymization
+ Averaging SA

Result 2: Evaluation and Prediction of data of PWSCUP2015

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}
$U1$	-	-	-	-	-	-	-	-	-	-	-	-
$U2$	negative	-	negative	negative	-	-	negative	-	-	negative	negative	negative
$U3$	negative	-	slightly	slightly	-	-	slightly	-	-	negative	negative	slightly
$U4$	-	slightly	-	slightly	slightly	slightly	slightly	slightly	slightly	negative	negative	negative
$U5$	-	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly
$U6$	-	-	-	-	-	-	-	-	-	-	-	-
$S1$	-	-	slightly	slightly	-	-	slightly	-	-	positive	positive	slightly
$S2$	-	-	slightly	slightly	slightly	positive	positive	positive	positive	positive	positive	positive
$E1$	slightly	negative	negative	negative	positive	positive	positive	positive	positive	slightly	slightly	slightly
$E2$	negative	negative	slightly	slightly	slightly	slightly	slightly	positive	positive	slightly	slightly	slightly
$E3$	negative	negative	slightly	slightly	positive	positive	positive	positive	positive	positive	positive	positive
$E4$	negative	positive	slightly	slightly	positive	positive	positive	positive	positive	positive	positive	positive
$EUC1$	negative	negative	negative	negative	slightly	slightly	negative	positive	positive	positive	positive	positive
D_a	-	-	×	×	-	-	×	-	-	×	×	×
D_b	-	-	-	-	-	-	-	-	-	-	-	-
D_c	-	-	-	-	×	×	-	×	×	-	-	-
D_d	-	-	-	-	×	×	×	-	-	-	-	-
D_e	×	-	-	-	-	-	-	×	×	-	-	-
D_f	-	×	-	-	-	-	-	-	-	×	×	×
D_g	-	-	×	×	-	-	×	×	×	-	-	-
D_h	-	-	-	-	-	-	-	-	-	-	-	-

D_8 won the best anonymized data in the PWSCUP2015.

All data in the group 2 are ranked higher in PWSCUP 2015.

Conclusion

- **We have proposed a new Re-identification method based on Euclidean distance. Our method works best in 5 out of 12 anonymized data of PWSCUP2015 and better than any the existing methods in re-identification rate.**

- **We guess unknown algorithms used to process 12 data of PWSCUP2015. Our analysis reveals that the Cheating anonymization with other methods performs better.**

Cheating attack

Cheating attack:

De-identification method exchange ID of data.

Original Data

ID	QI1	QI2	QI3	SA1	SA2
1	2	1	1	100	100
2	2	1	1	200	400
3	1	1	2	300	200
4	1	1	2	400	500

Anonymized data

ID	QI1	QI2	QI3	SA1	SA2
2	2	1	1	100	100
3	2	1	1	200	400
4	1	1	2	300	200
1	1	1	2	400	500