Risk of Re-identification from Payment Card Histories in Multiple Domains

Satoshi Ito, Reo Harada, Hiroaki Kikuchi

Meiji University Graduate School

Background

- **The anonymized data has been in enforcement according to the Japanese Act on the Protection of Personal Information (APPI) in 2017.**

- **There is not standard criteria for risk of anonymized data.**

- **We try to reveal the risks of individuals to be identified from given anonymized data.**

Diversity of Anonymized Data

1. **Trajectory Data**

   A. Basu, A. Monreale, R. Trasarti, J. C. Corena, F. Giannotti, D. Pedreschi, S. Kiyomoto, Y. Miyake and T. Yanagihara, "A risk model for privacy in trajectory data", Journal of Trust Management, 2:9, 2015.

2. **Census Data**

   Koot, M. R., Mandjes, M., van't Noordende, G., and de Laat, C., "Efficient probabilistic estimation of quasi-identifier uniqueness", In Proceedings of ICT OPEN 2011, 14-15, pp. 119-126, 2011.

However, what if two distinct dataset are combined?

Our Target Data

The payment card histories data in multiple domains.

This card stores 5 distinct domains records.

(traffic, purchase, deposit, bus charge, and other uses)

User ID	Date	Times	Ent. point	Ali. point	Ent. route	Ali. route	Usage	Location	Fare
1	2016/10/30	2	Ueno	Tokyo	JR-EAST	JR-EAST	Traffic	NA	-194
1	2016/10/30	1	Tokyo	Ueno	JR-EAST	JR-EAST	Traffic	NA	-194
2	2016/10/8	1	NA	NA	NA	NA	Deposit	Ticket vending machine	2000
2	2016/10/1	1	NA	NA	NA	NA	purchase	Vending machine	-120

Objectives

1. **Analysis on the payment card history data that combined multiple domains.**

2. **Evaluation of the risks to be identified in empirical analysis.**

Questions

1. **How many records of histories are necessary to identify individuals uniquely?**

2. **Which risk is high, the traffic or purchase data?**

3. **Is risk increased when multiple domains are combined?**

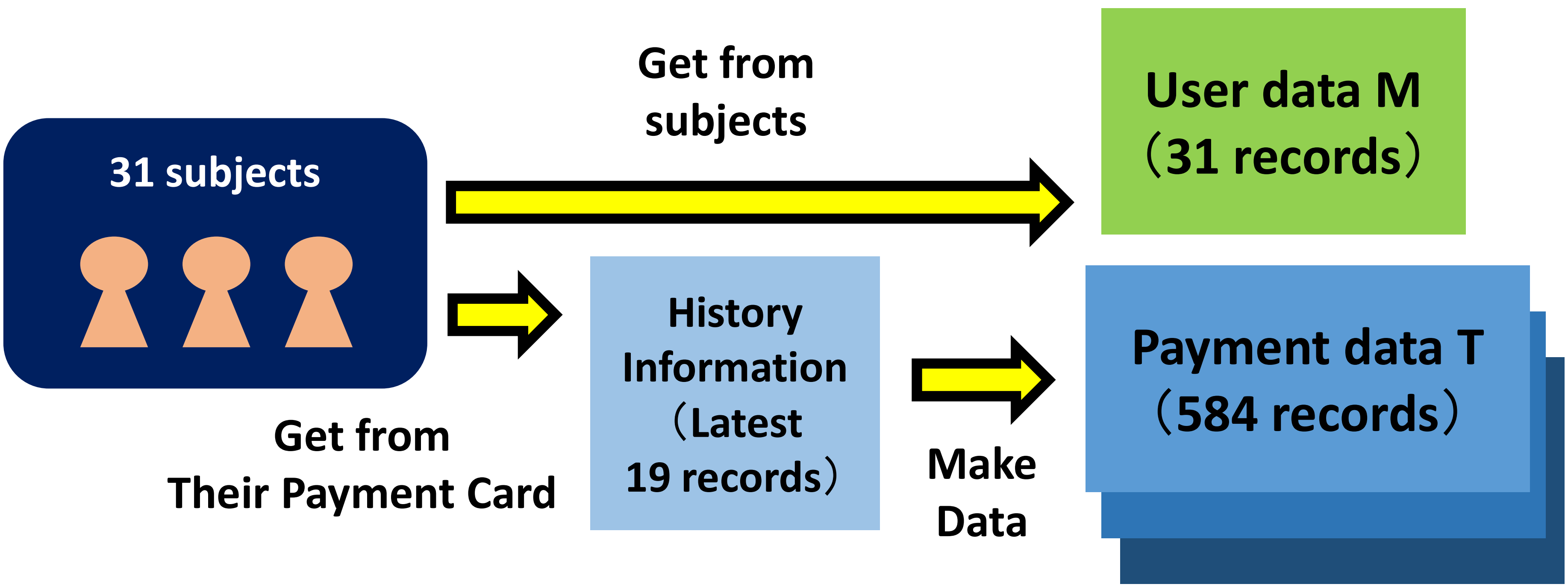
Objectives

1. **Analysis on the payment card history data.**

2. **Evaluation of the risks to be identified.**

The payment card history data

We obtained the payment card history data from payment cards of 31 subjects of our University under each user's consent to our study.



The example of data

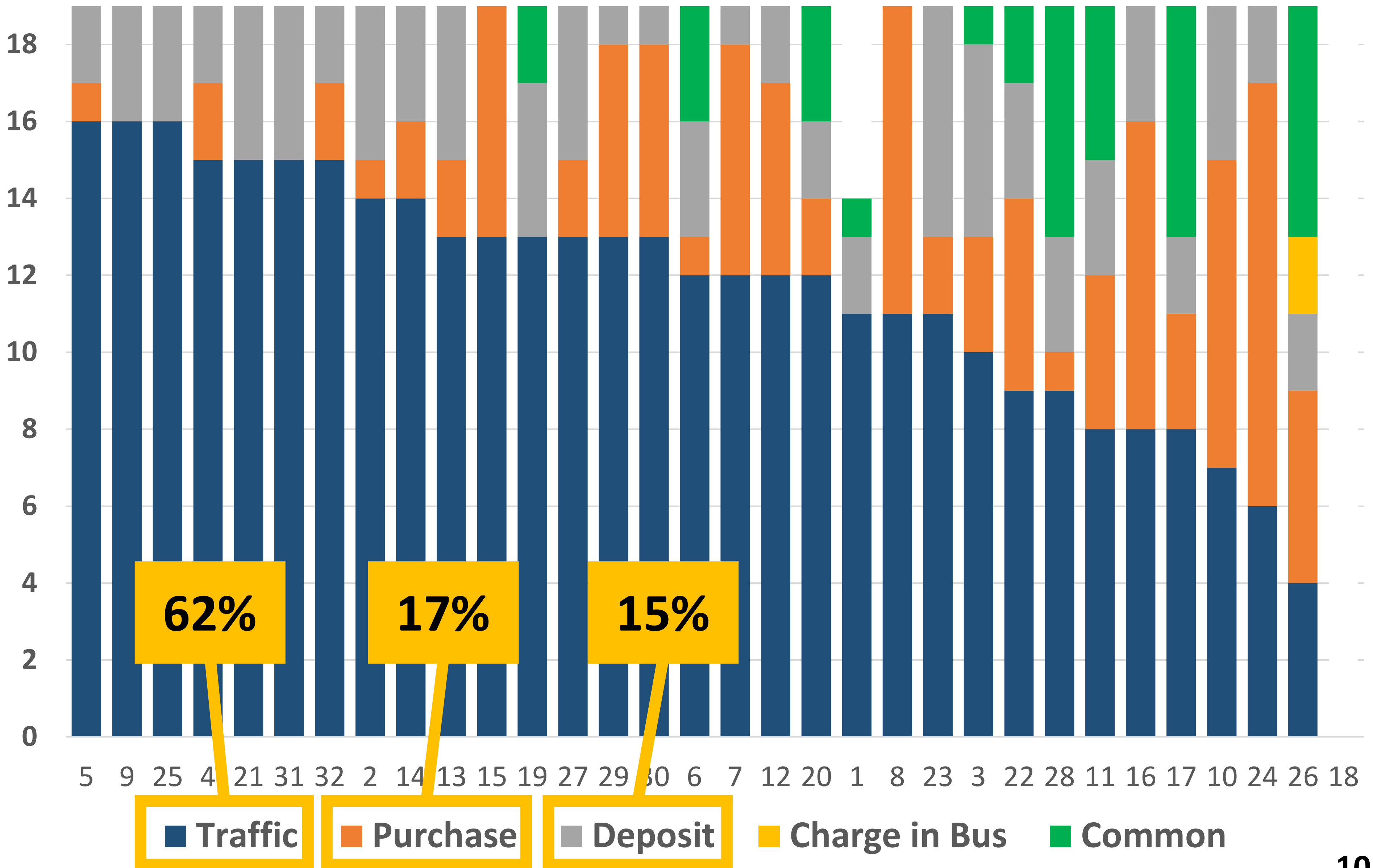
Example of user data M

User ID	Sex	Grade	Address	Range of season ticket 1	Range of season ticket 2
1	M	1	Chiba	NA	NA
2	F	3	Tokyo	Nakano	Shinjuku

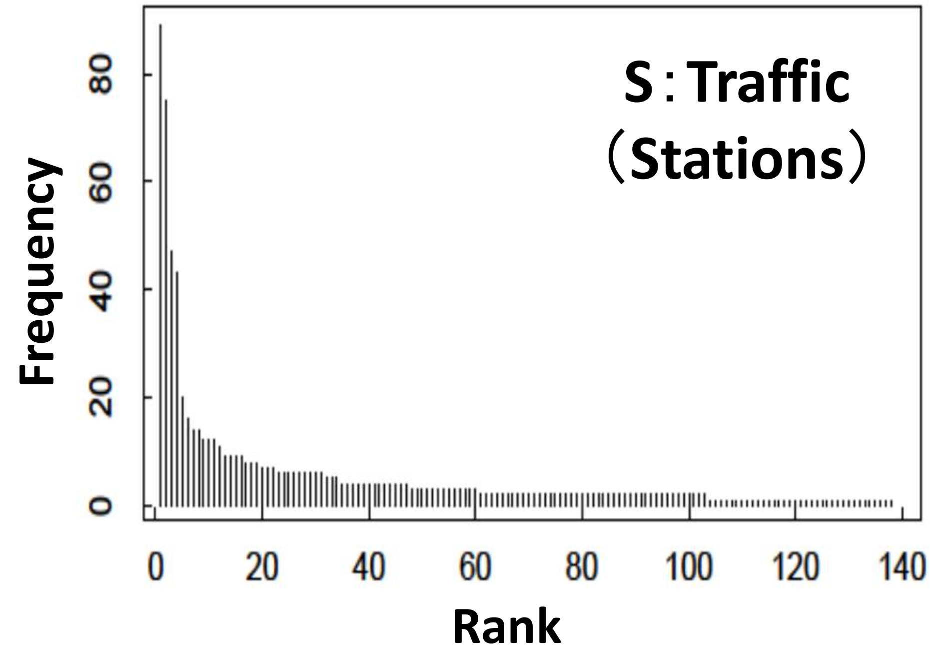
Example of history data T

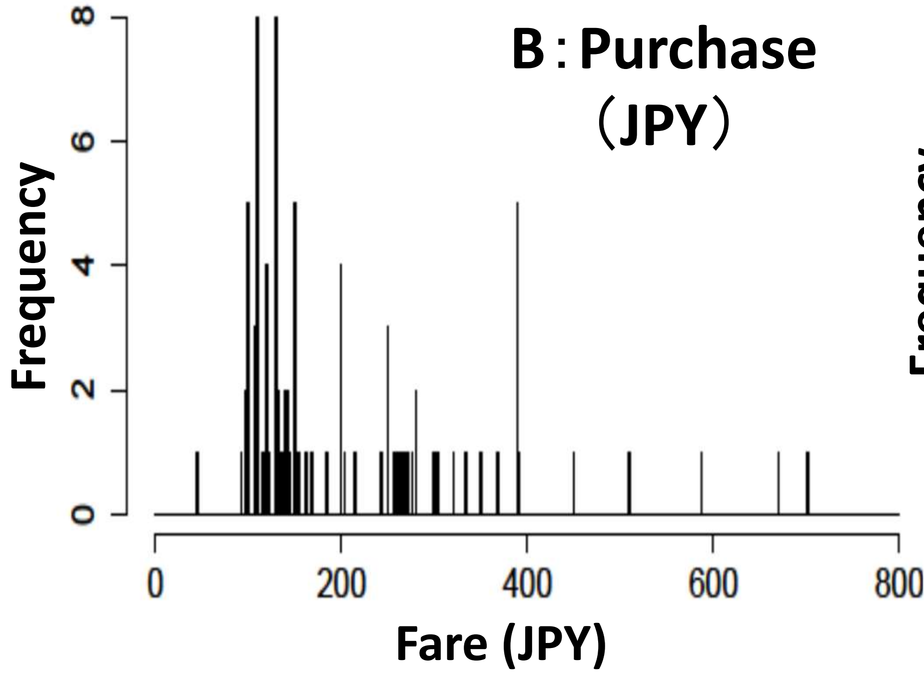
User ID	Date	Times	Ent. point	Ali. point	Ent. route	Ali. route	Usage	Location	Fare
1	2016/ 10/30	2	Ueno	Tokyo	JR-EAST	JR-EAST	Traffic	NA	-194
1	2016/ 10/30	1	Tokyo	Ueno	JR-EAST	JR-EAST	Traffic	NA	-194
2	2016/ 10/8	1	NA	NA	NA	NA	Deposit	Ticket vending machine	2000

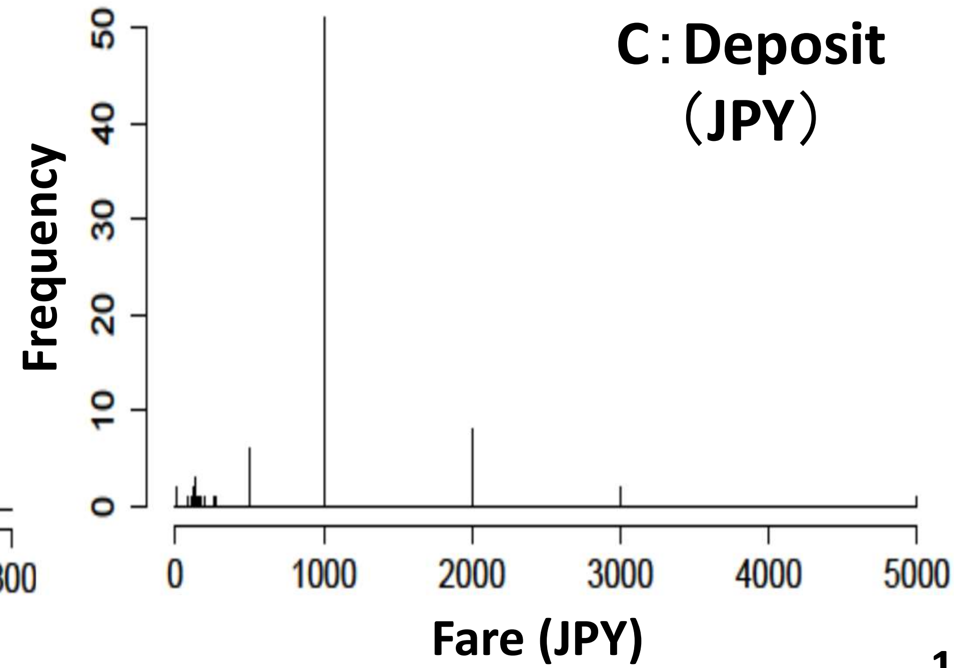
Breakdown of usage of data



Frequency of various values

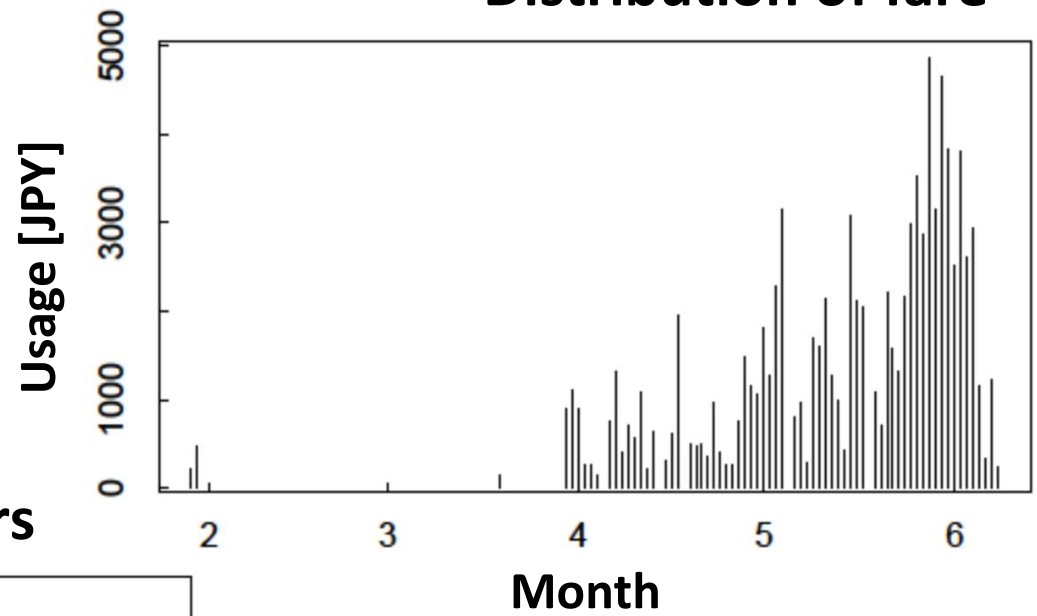




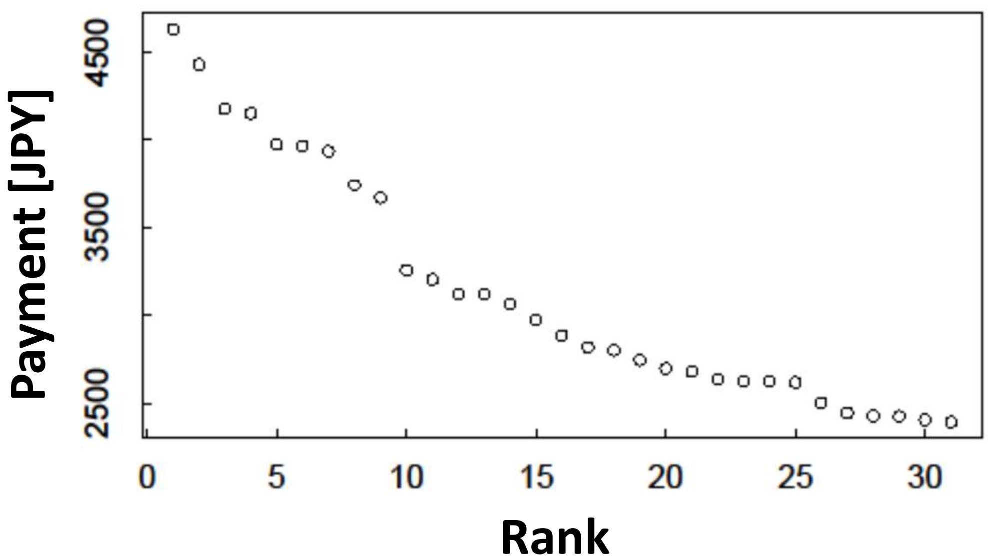


Analysis of Usage

Distribution of fare



Details of fare for all users



Objectives

1. **Analysis on the payment card history data.**

2. **Evaluation of the risks to be identified.**

Risk of identification of Users

Data A

User ID	Station 1	Station 2
1	Tokyo	Nakano
2	Tokyo	Nakano
3	Tokyo	Nakano
4	Tokyo	Nakano
5	Tokyo	Nakano

User's histories are similar.

↓

Users will not
be identified easily.

Data B

User ID	Station 1	Station 2
1	Tokyo	Nakano
2	Shizuoka	Tokyo
3	Gifu	Osaka
4	Shinagawa	Tokyo
5	Osaka	Yokohama

User's histories are distinct.

↓

Users will
be identified easily.

Risk of identification from Stations

Example of totalization table

User/Station	Tokyo	Osaka	Kyoto
u_1	2	1	0
u_2	4	0	4
u_3	4	4	0

All users have been to Tokyo station. Therefore, the risk of identification from this station is low.

Only u_2 went Kyoto station. Therefore, the risk of identification from this station is high.

Risk by Station: Tokyo＜Osaka＜Kyoto

Conditional Entropy

User/Station	Tokyo	Osaka	Kyoto	Sum	$P(U = u_i)$
u_1	2	1	0	3	3/19
u_2	4	0	4	8	8/19
u_3	4	4	0	8	8/19
$H(U\|S = s_i)$	1.52	0.72	0		
$P(S = s_i)$	10/19	5/19	4/19		

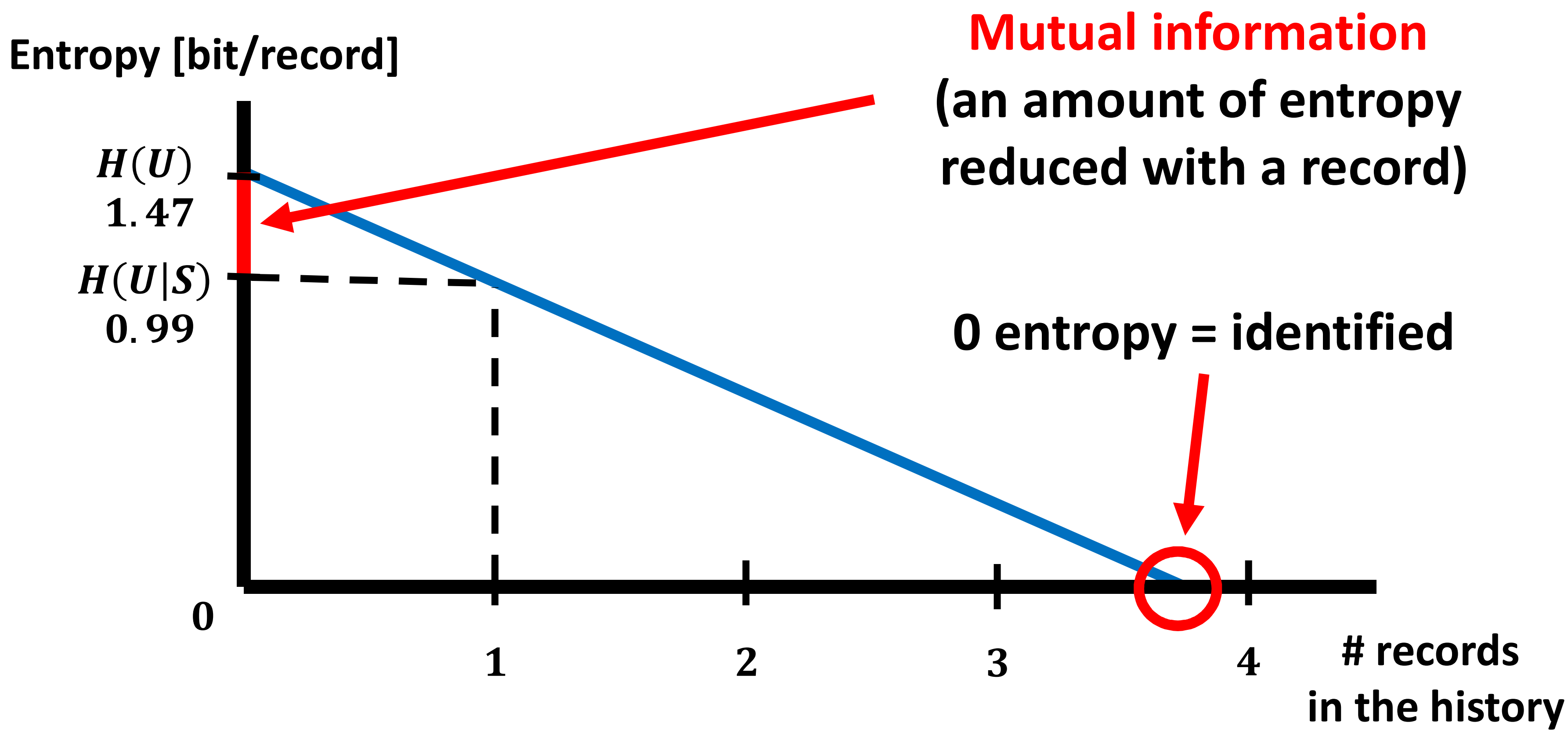
The entropy of users, given the history of use of stations S $[bit/record]$

$$H(U|S = Tokyo) > H(U|S = Osaka) > H(U|S = Kyoto)$$

1.52 **0.72** **0**

low risk **high risk**

Mutual information

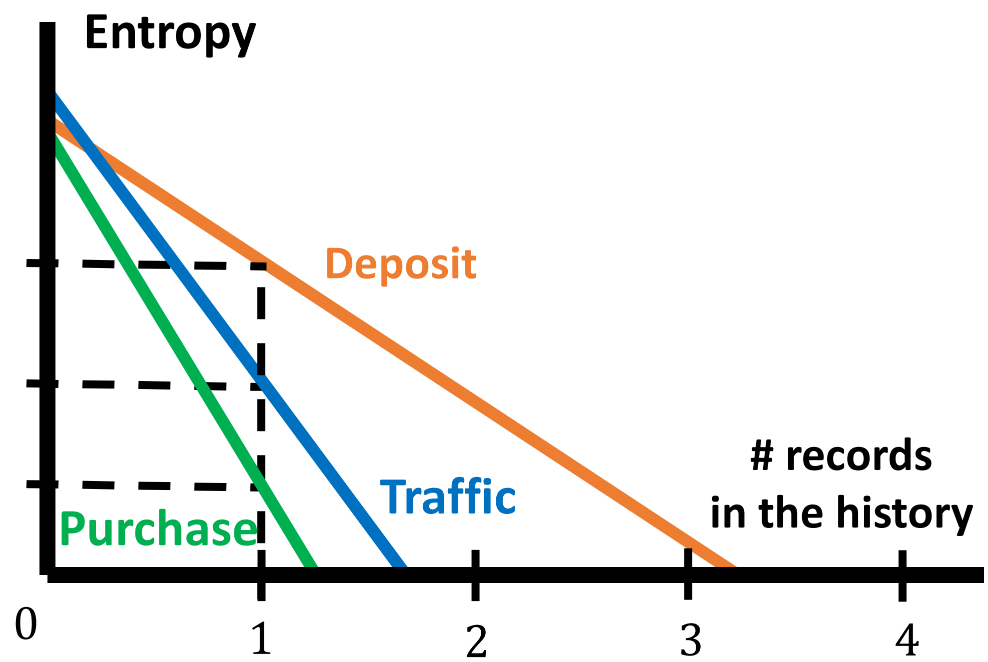


The average probability of identifying is $\frac{1}{2}^{H(U)}$

The actual values of risk

Risk of identification depends on domains.

	Traffic (S)	Purchase (B)	Deposit (C)
$H(U)$	4.900	4.338	4.736
$H(U\|x)$	1.814	0.948	3.256
$I(U;x)$	3.085	3.389	1.479
$P(U\|x)$	0.284	0.518	0.105



- **The histories of purchase are the highest risk factor.**
- **Individuals can be identified from 2 records of traffic or purchase.**

Obtaining history of two domains

The example of cross-tabulation table of traffic history

User/Station	s_1	s_2	s_3
u_1	2	1	0
u_2	4	0	4
u_3	4	4	0

The example of cross-tabulation table of purchase history

User/Fare	b_1	b_2
u_1	2	0
u_2	1	3
u_3	0	1

Cross-tabulation table when combination of traffic and purchase

	s_1, b_1	s_1, b_2	s_2, b_1	s_2, b_2	s_3, b_1	s_3, b_2
u_1	4	0	2	0	0	0
u_2	4	12	0	0	4	12
u_3	0	4	0	4	0	0

The risk of combining of two domains

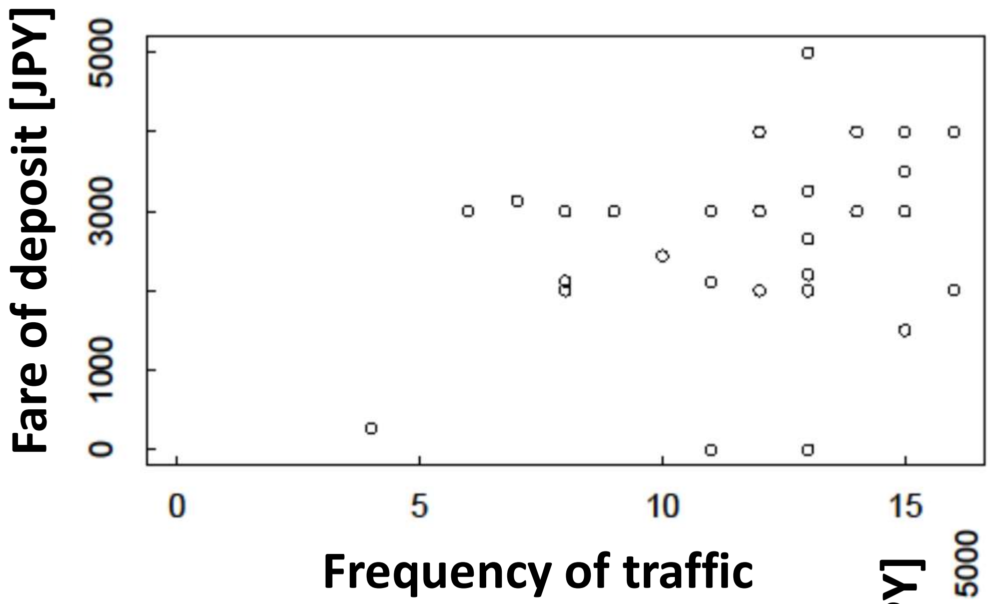
	traffic･purchase(S,B)
$H(U)$	4.412
$H(U\|x)$	0.182
$I(U;x)$	4.230
$P(U\|x)$	0.881

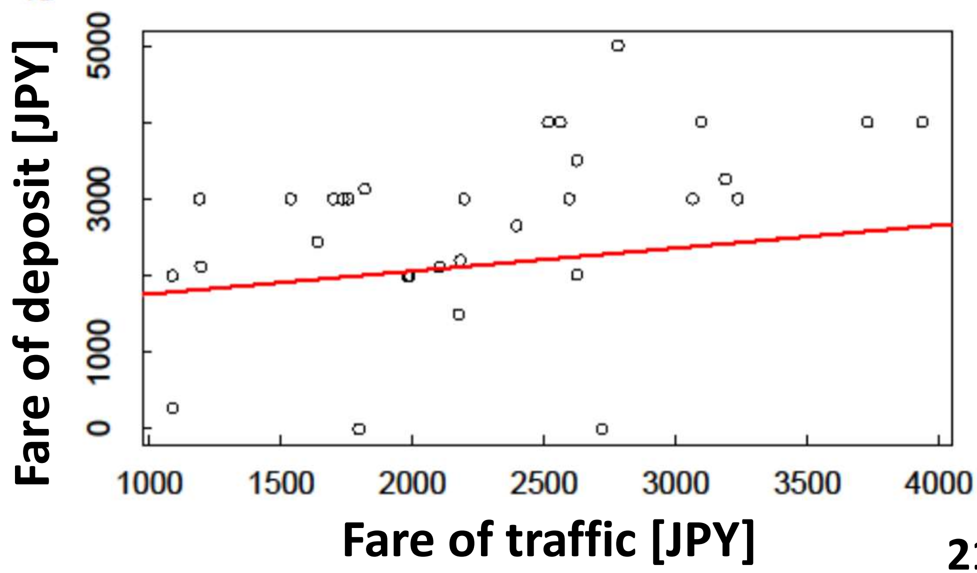
The risk to identify individual rises to 88.1% when two records are given, one from traffic and one from purchase.

The histories of traffic are not independent from histories of purchase because

$$I(U;S,B) = 4.230 < 6.474 = I(U;S) + I(U;B)$$

Correlations between histories of traffic and deposit





Questions

1. **How many records of histories are necessary to identify individuals uniquely?**

   →2 records of traffic and purchase. 4 record of deposit.

2. **Which risk is high, the traffic or purchase data?**

   →The purchase is.

3. **Is risk increased when multiple domains are combined?**

   →Yes.

Conclusion

1. **We reported the statistics of the payment cards obtained from 31 students.**
   As the result, the payment card data contain 5 domains.
   （traffic：62%, purchase：17%, deposit：15%）

2. **2 records of traffic history or purchase to be identified individual and the mutual information of histories of purchase is largest.**
   The risk to identify individual rise to 88.1% when one history of traffic and one history of purchase are given.

Q & A

	traffic(S)	purchase(B)	deposit(C)
$H(U)$	4.900	4.338	4.736
$H(U\|x)$	1.814	0.948	3.256
$I(U;x)$	3.085	3.389	1.479
$P(U\|x)$	0.284	0.518	0.105

	traffic· purchase(S,B)	traffic· deposit(S,C)	purchase· deposit(B,C)
$H(U)$	4.412	4.677	4.149
$H(U\|x)$	0.182	1.065	0.529
$I(U;x)$	4.230	3.612	3.620
$P(U\|x)$	0.881	0.478	0.692

	traffic(S)	purchase(B)	deposit(C)
$H(U)$	4.900	4.338	4.736
$H(U\|x)$	1.814	0.948	3.256
$I(U;x)$	3.085	3.389	1.479
$P(U\|x)$	0.284	0.518	0.105
n_x	31	25	29
m_x	138	58	17

	traffic・purchase(S,B)	traffic・deposit(S,C)	purchase・deposit(B,C)
$H(U)$	4.412	4.677	4.149
$H(U\|x)$	0.182	1.065	0.529
$I(U;x)$	4.230	3.612	3.620
$P(U\|x)$	0.881	0.478	0.692
n_x	31	31	31
m_x	8004	2346	986

Cheating anonymization

Cheating anonymization:

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

Objectives

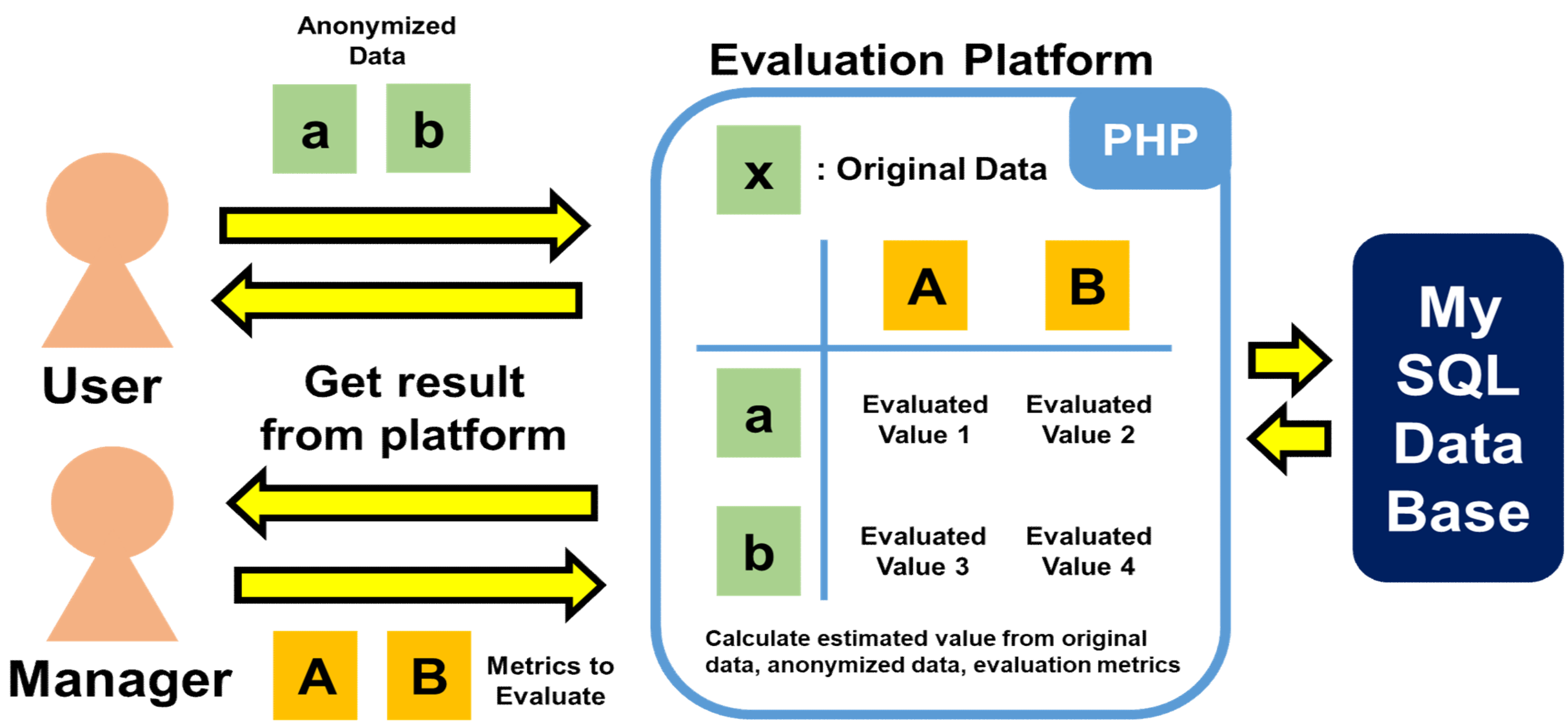
1. Analysis on the payment card history data.

2. Evaluation of the risks to be identified.

3. **Study on anonymization method of this data.**

Evaluation experiment

We developed a web-based platform on Linux to evaluate anonymized data automatically.



Experimental results

We made 47 anonymized data of payment card data in many methods.