Quantifying the Risk of Re-identification in Data Anonymization Competition

Takao Murakami (AIST*, Japan)

*AIST: National Institute of Advanced Industrial Science & Technology

Outline

- Data Anonymization Mechanism
  - Plays an important role in balancing users' privacy & data utility.

- PWS (Privacy Workshop) CUP 2016
  - Was held in Japan to understand pros & cons of various mechanisms.

In this talk

We introduce how the privacy level of each mechanism was evaluated.
We introduce some sample re-identification algorithms and their design issue.

Contents

PWS CUP 2016
(Dataset, Anonymization/Re-identification)

Re-identification Sample Algorithms

Conclusion

PWS CUP 2016

- Schedule
  - Preliminary Competition: 2016/08/25 - 201610/03
    - The main purpose of preliminary competition was to see the feasibility of the rule, utility metrics, privacy metrics… before final competition.
  - Final Competition: 2017/10/11
  - Notification of Results: 2017/10/12

Dataset

- Online Retail Data Set (UCI Machine Learning Repository)
  - Publicly available dataset (https://archive.ics.uci.edu/ml/datasets/Online+Retail).
  - Contains transactions between December 2010 and December 2011 for a UK-based and registered non-store online retail.
  - We performed **data cleansing**.
    - E.g. deleted cancel receipts, deleted records who had missing values.
  - We performed **data sampling** (due to the limited computational resource).
    - 4333 customer IDs → 400 customer IDs.

Description	Value
#Records	38,087
#Customer IDs	400
#Receipts	1,763
#Items	2,781
#Countries	30

Dataset

- Master Data & Transaction Data
  - We divided the data set into **master data** & **transaction data**.
  - We artificially generated gender & birthday.

Master M

Customer ID	Gender	Birthday	Country
12346	**f**	**1960/12/25**	**UK**
12347	f	1957/5/15	Iceland
12348	m	1947/2/19	Finland

Transaction T

Customer ID	Receipt	Date	Time	Item ID	Unit Price	Quantity
12347	544203	2011/2/17	10:30	21913	3.75	4
12347	544203	2011/2/17	10:30	22431	1.95	6
12346	**545017**	**2011/2/25**	**13:51**	**22630**	**1.95**	**12**
12346	**545017**	**2011/2/25**	**13:51**	**22555**	**1.65**	**12**
12346	**551346**	**2011/4/28**	**9:12**	**21866**	**1.25**	**8**
12348	554132	2011/5/23	9:43	21094	0.85	12

Anonymization/Re-identification

Attacker estimates, for each line in M', the corresponding line no. in M.

Master M

Customer ID	Gender	Birthday	Country
12346	f	1960/12/25	UK
12347	f	1957/5/15	Iceland
12348	m	1947/2/19	Finland

Transaction T

Customer ID	Date	Item ID
12347	2010/12/7	85116
12347	2010/12/7	22375
12346	2011/1/18	23166

Anonymization (pseudonymize, perturb, shuffle, delete record, dummy transaction record)

Q	P
3	3
2	1
2	2

Anonymized Master M'

Nym	Gender	Birthday	Country
10	m	**1947/1/1**	Finland
20	f	**1960/1/1**	UK
30	f	**1960/1/1**	**UK**

Anonymized Transaction T'

Nym	Date	Item ID
10	**2010/12/1**	**85123A**
30	**2010/12/1**	**85123A**
30	2010/12/7	**20000**
20	2011/1/18	**20000**

Line no. in M

Line no. estimated by attacker (re-identification result)

Re-identification rate: Re-ID(P,Q) = (#correct lines) / |P| = 2/3

Data Anonymization/Re-identification Phase

- Data Anonymization Phase:
  - Each team submits anonymized data M' & T' (and line P)
  - Utility (resp. privacy) are evaluated using 4 (resp. 13) algorithms.
    - U_i $(0 \le U_i \le 1)$: utility score based on the i-th algorithm $(1 \le i \le 4)$.
    - E_i $(0 \le E_i \le 1)$: re-identification rate based on the i-th algorithm $(1 \le i \le 13)$.
  - Total score S (the smaller is the better) is calculated as follows:

$$S = \max_{1 \le i \le 4} U_i + \max_{1 \le i \le 13} E_i$$

Worst utility score

Worst privacy score
(max of re-identification rate)

Utility evaluation algorithms (4 algorithms in total)

Cross table (gender x country)-based algorithm,
RFM (Recency Frequency Monetary)-based algorithm, etc.

Re-identification algorithms (13 algorithms in total)

transaction number-based algorithm,
total price-based algorithm, etc.

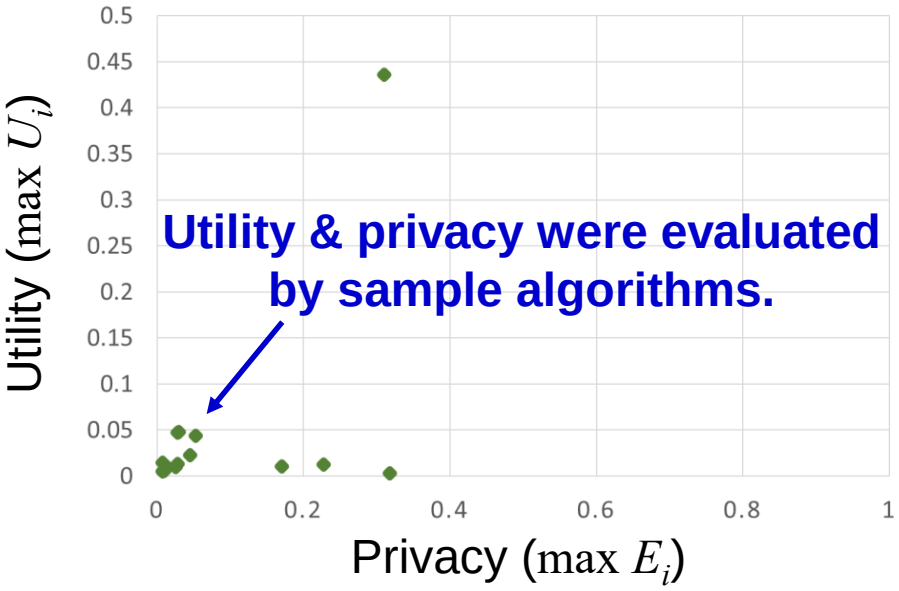
Data Anonymization/Re-identification Phase

- Re-identification Phase:
  - Each team tries to re-identify other teams' data.
  - Privacy was evaluated again based on max of re-identification rate.

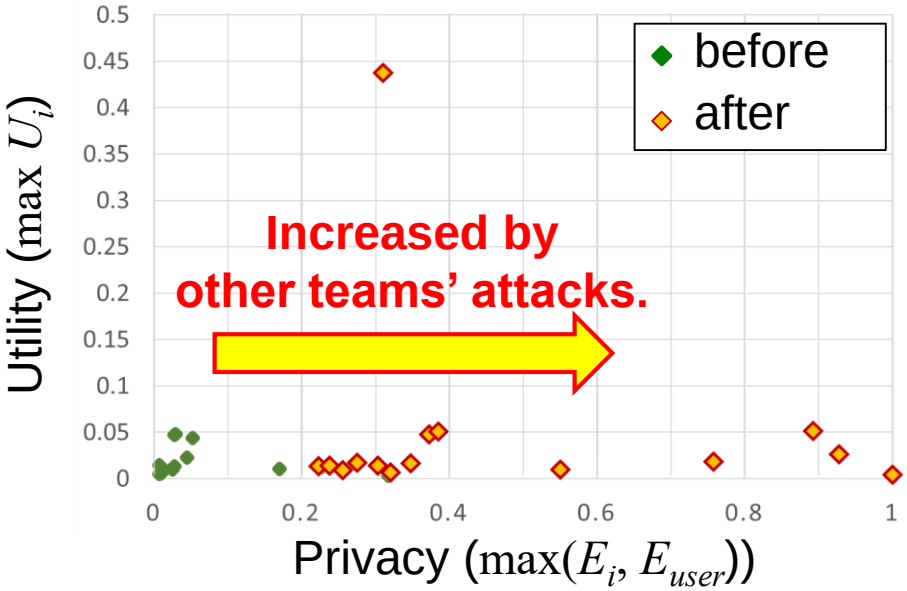
Re-identification rate by other teams

$$S = \max_{1 \le i \le 4} U_i + \max_{1 \le i \le 13} (E_i, E_{user})$$

Anonymization Phase
(PWS CUP 2016 Final)



Re-identification Phase
(PWS CUP 2016 Final)



Contents

PWS CUP 2016

(Dataset, Anonymization/Re-identification, Interface)

Re-identification Sample Algorithms

Conclusion

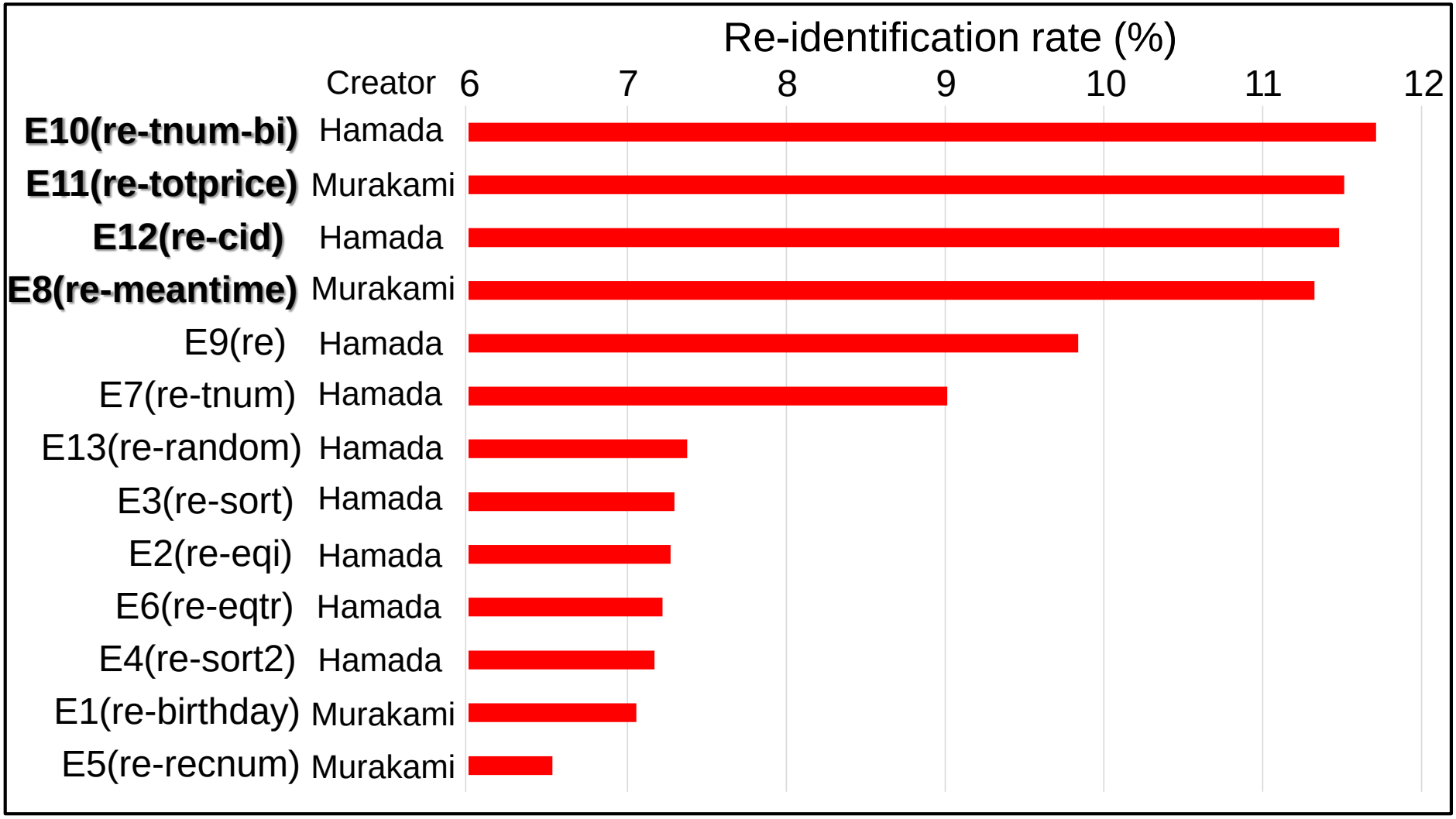
Basic Design Strategy

- We designed the following sample algorithms:
  - (1) **Simple** (so that everyone can easily understand them).
  - (2) **Modestly accurate** (but there is a lot of room for improvement).
    - In the identification phase, each team develops more sophisticated algorithms.
  - (3) **Fast** ($O(m^2)$ (m: #customers) may be slow → O(mlogm) is better).

ID	Name	Master Data				Transaction Data						
		ID	Gender	Birthday	Country	ID	Receipt	Date	Time	Item	Unit Price	Quantity
E1	re-birthday			✓ ←		“E1:re-birthday” used the birthday attribute.						
E2	re-eqi	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
E3	re-sort		✓	✓	✓							
E4	re-sort2			✓								
E5	re-recnum					✓						
E6	re-eqtr	✓				✓	✓	✓	✓	✓	✓	✓
E7	re-tnum	✓				✓						
E8	re-meantime							✓	✓			
E9	re											
E10	re-tnum-bi	✓		✓		✓						
E11	re-totprice										✓	✓
E12	re-cid	✓										
E13	re-random											

Re-identification Rate at Preliminary Competition

I calculated the average re-identification rate over all anonymized data.

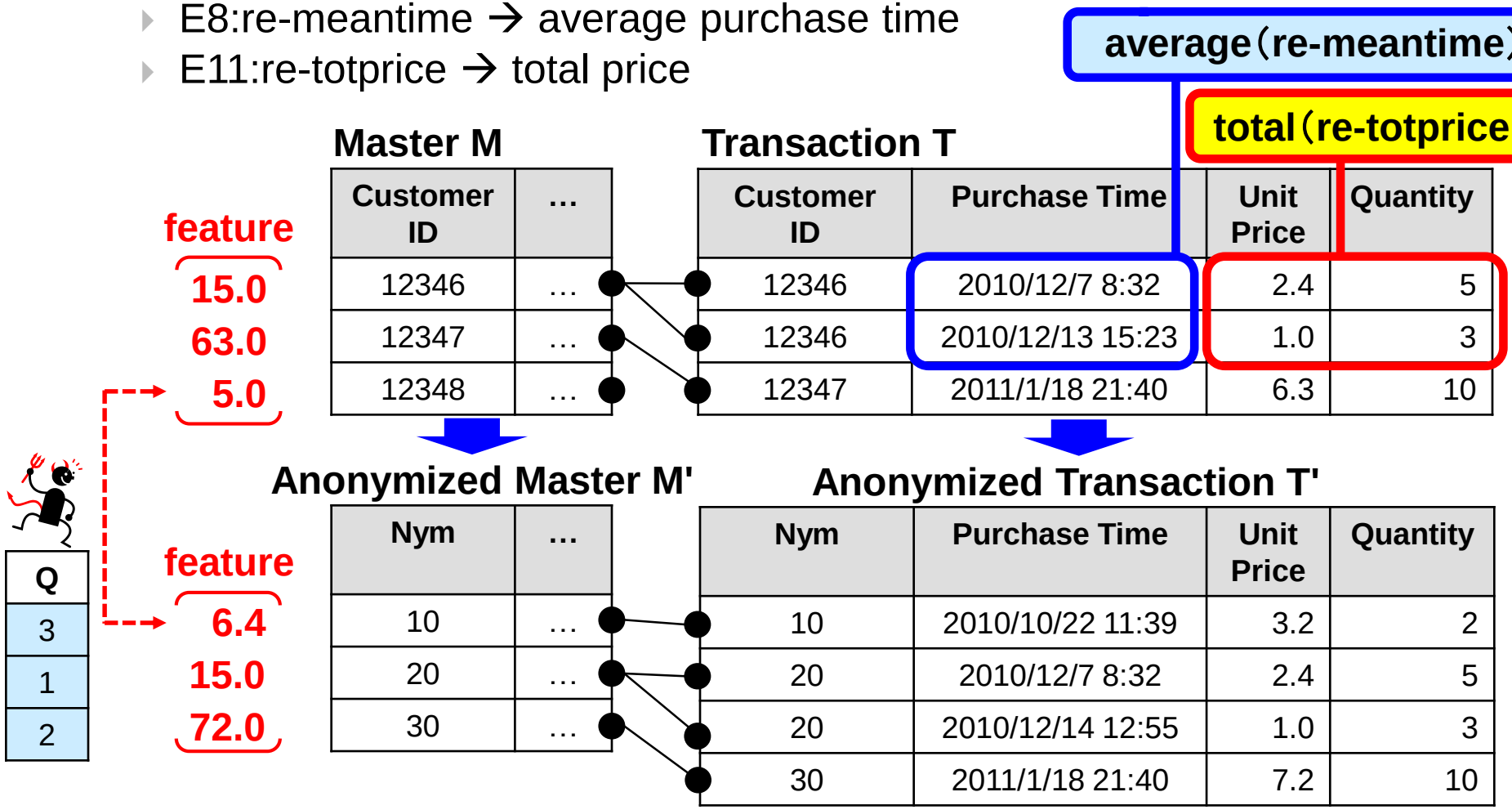


I will introduce E10,11,12, and 8, which achieved the 1st to 4th places.

E8:re-meantime (4th) & E11:re-totprice (2nd)

E8:re-meantime (4th) & E11:re-totprice (2nd)

- Scalar Feature
  - These algorithms extract a **scalar feature** for each customer ID/pseudonym.
    - E8:re-meantime → average purchase time
    - E11:re-totprice → total price

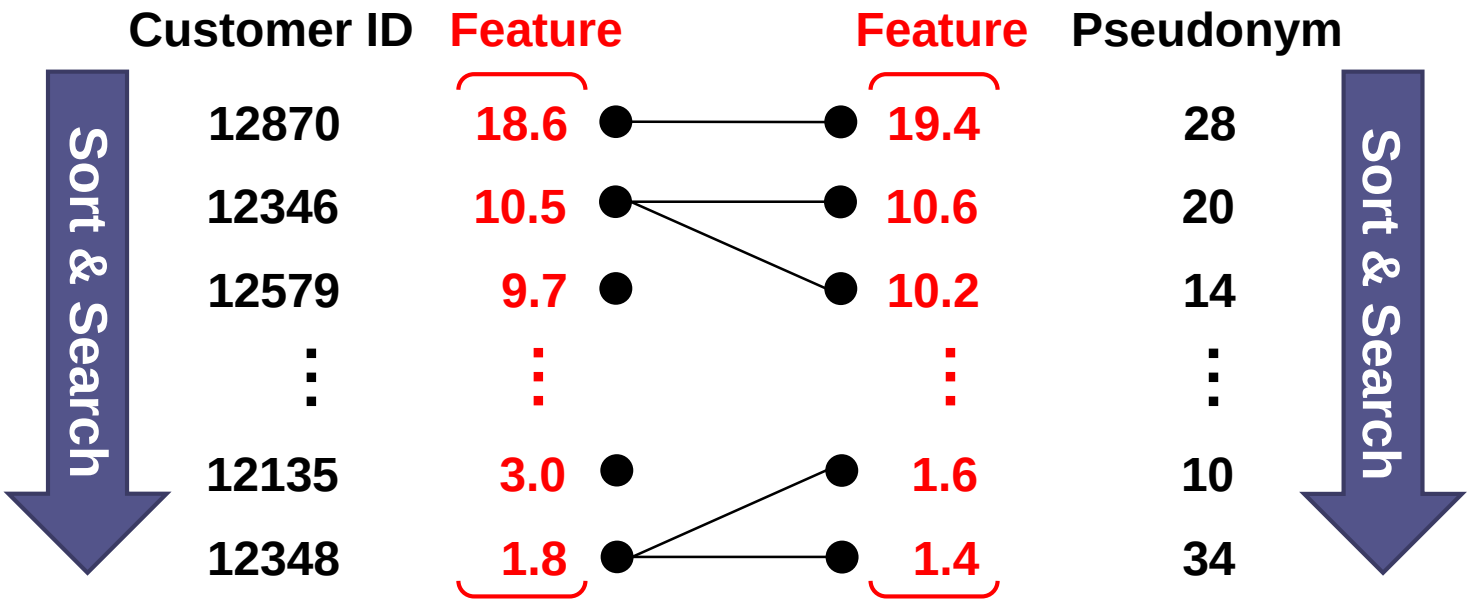


Attacker searches, for each feature in M', the closest feature in M.

E8:re-meantime (4th) & E11:re-totprice (2nd)

Scalar Feature → Simple, Modestly Accurate, and Fast (O(mlogm)).

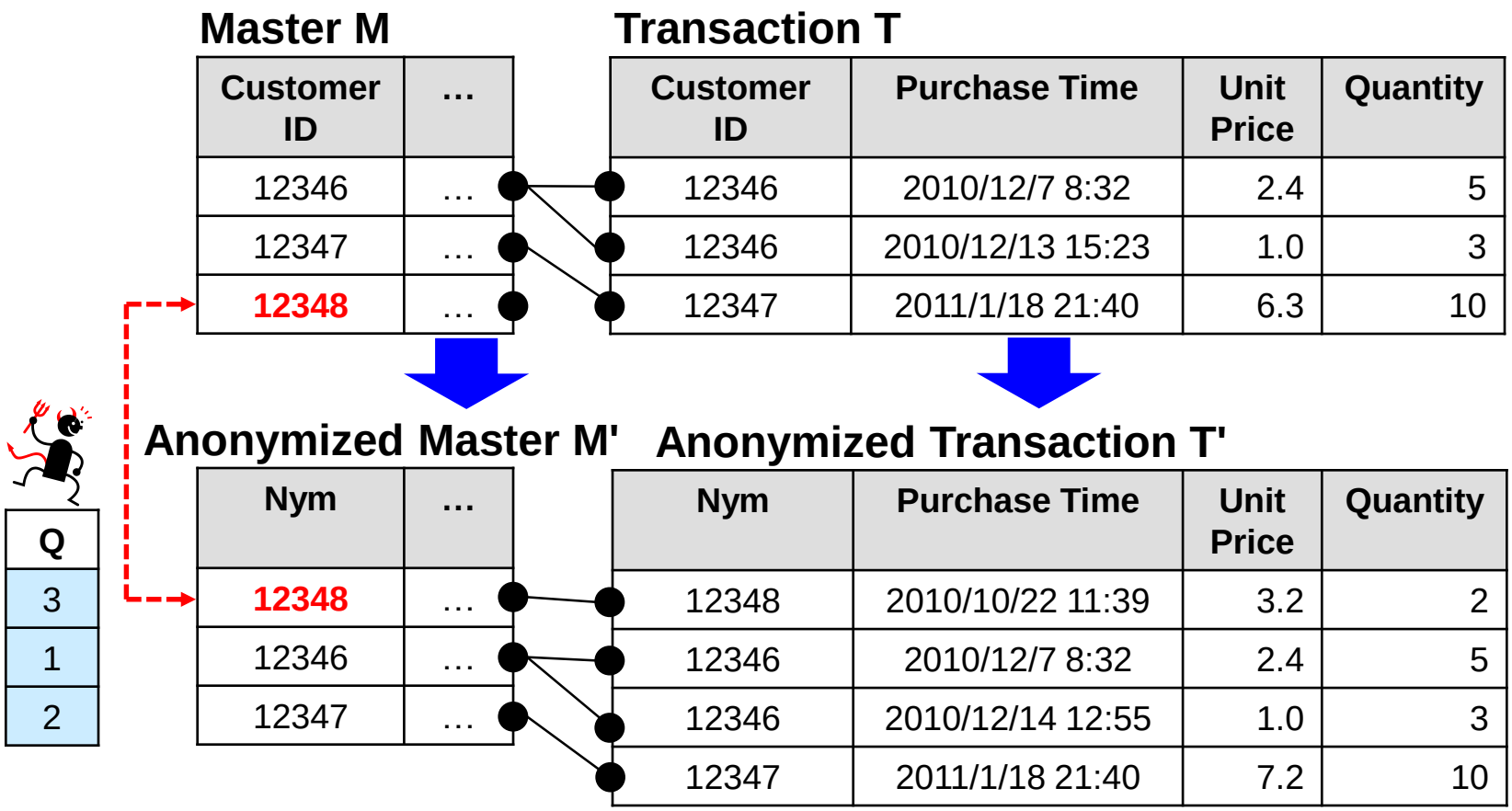
- Re-identification Algorithm
  - Step 1: Sort customer IDs/pseudonyms in descending order of features.
  - Step 2: For each pseudonym, find a customer ID whose distance is the smallest (we can find all pairs by sequential search).
  - Step 3: Re-identify each pseudonym as the corresponding customer ID.
  - **→ Average time complexity is O(mlogm) (m: #customers).**



E12:re-cid (3rd)

E12:re-cid (3rd)

- Re-identification Algorithm
  - Step 1. For each pseudonym, find **the completely same customer ID**.
  - Step 2. Output the corresponding line no.
    (If there is no such customer IDs, output random value from 1 to M.)



This is just an algorithm to eliminate data not even pseudonymized.

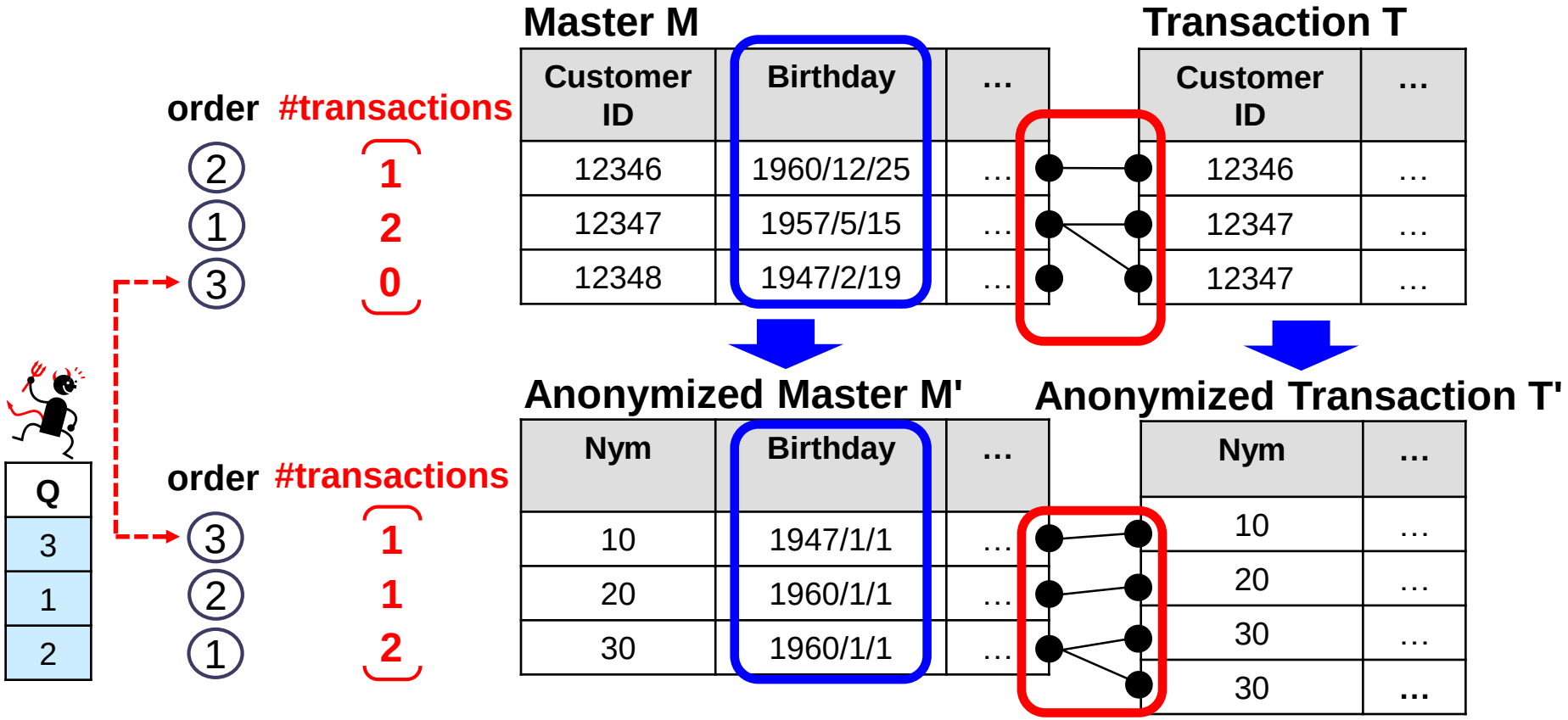
re-cid(3rd)

- Why did this algorithm achieve the 3rd place?
  - **Many teams did not even pseudonymize their own data at the preliminary competition.**
  - I was shocked to see that this algorithm took the 3rd place. (many of my algorithms were worse than this…)
  - → At the final competition, I asked everyone to pseudonymize the data.

E10:re-tnum-bi (1st)

re-tnum-bi(1st)

- Re-identification Algorithm
  - Step 1: Compute **#transactions** for each customer ID/pseudonym.
  - Step 2: Sort customer IDs & pseudonyms by **(#transactions, birthday)**.
  - Step 3: Make a pair of customer ID & pseudonym in the sorted order.



Re-identification rate can be increased by using multiple features.

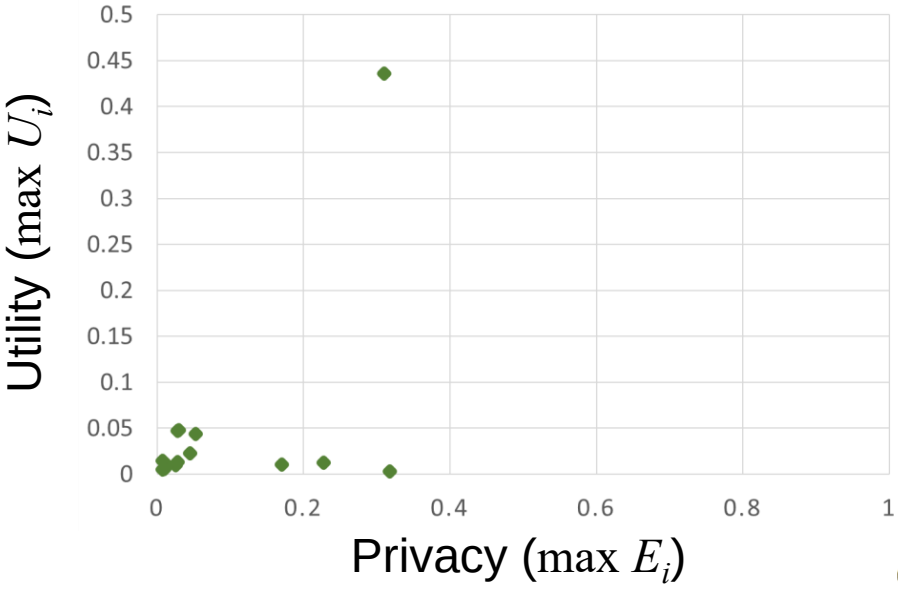
Design Strategy for Re-identification Phase

- Anonymization Phase ← sample algorithms were used.
- Re-identification Phase ← Each team re-identifies other teams' data.

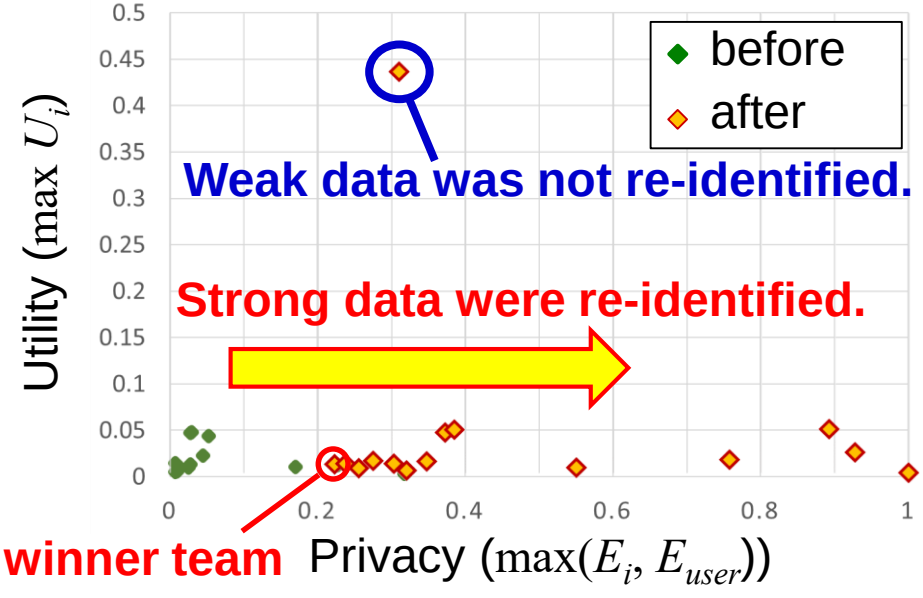
We gave a "Re-identification Award" to a team who achieved the highest re-identification rate for the "winner team".

To make everyone re-identify the strongest data.

Anonymization Phase
(PWS CUP 2016 Final)



Re-identification Phase
(PWS CUP 2016 Final)



It also made the final competition interesting.

Contents

PWS CUP 2016
(Dataset, Anonymization/Re-identification)

Re-identification Sample Algorithms

Conclusion

Conclusion

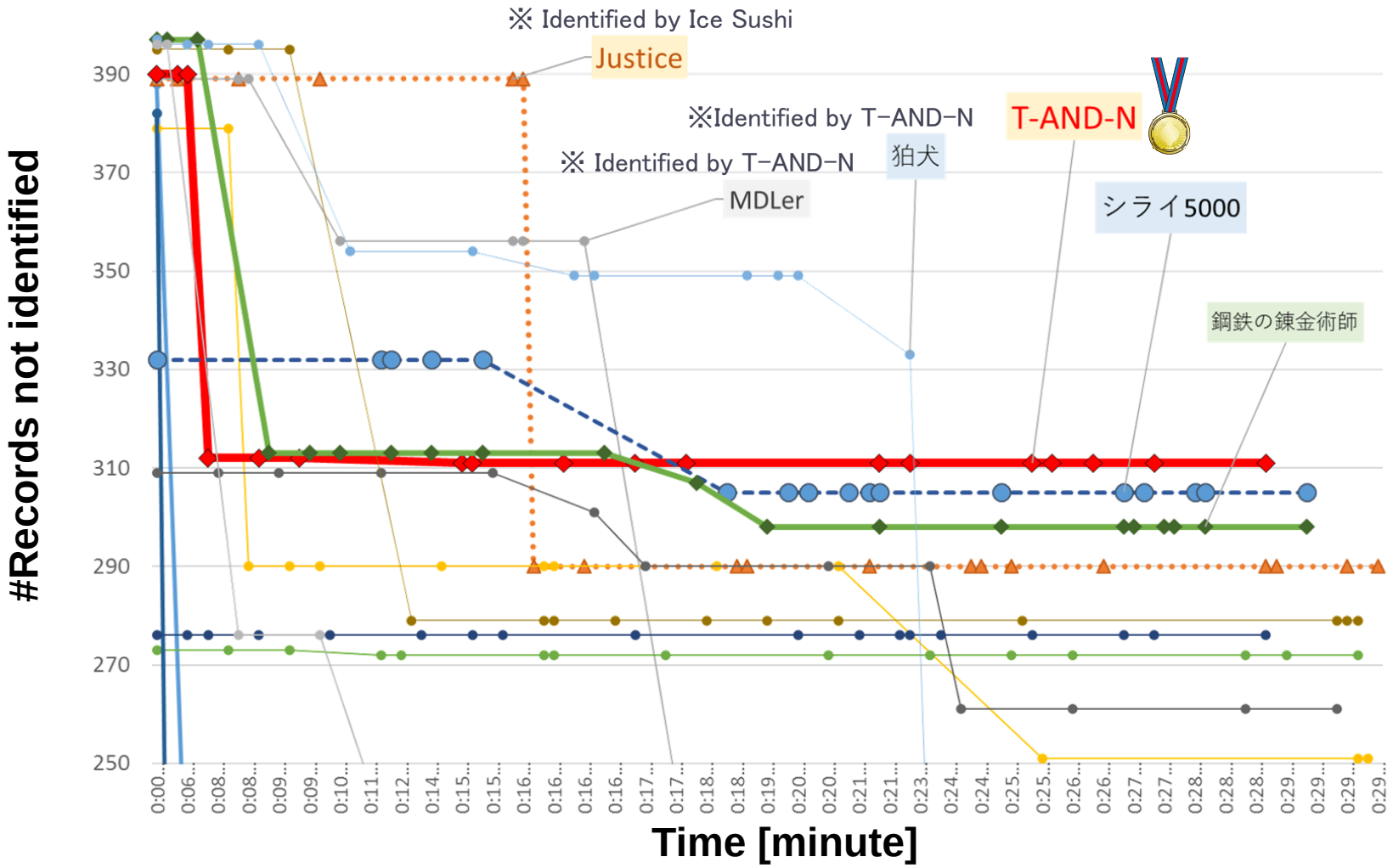
- Design Strategy for Anonymization Phase
  - We designed the following sample algorithms:
    (1) **Simple**.
    (2) **Modestly accurate** (but there is a lot of room for improvement).
    (3) **Fast** ($O(m^2)$ (m: #customers) may be slow → O(mlogm) is better).

- Design Strategy for Re-identification Phase
  - We gave a "Re-identification Award" to a team who achieved the highest re-identification rate for the "winner team"
  - **→ everyone tried to re-identify the strongest data.**

Thank you for listening.

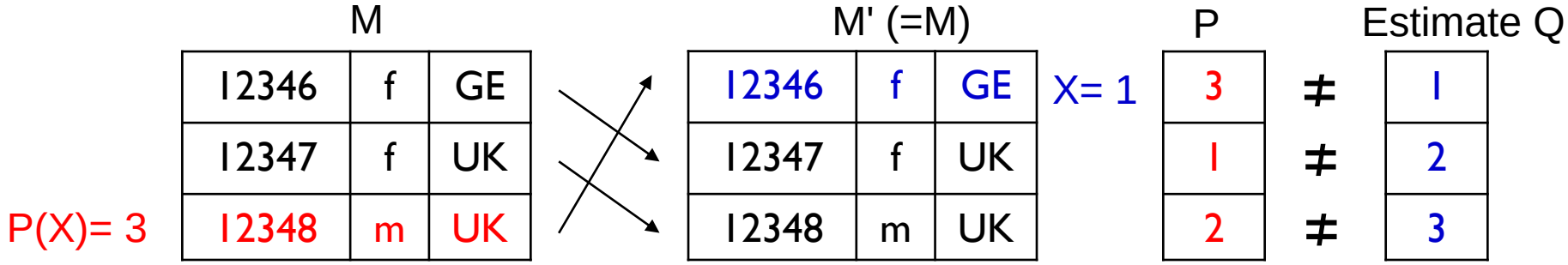
Appendix: Re-identification Rate v.s. Time

From 10 minutes to 16 minutes, team “Justice” kept the 1st place.
However, “Justice” was re-identified and the 1st team was changed as follows:
“Justice” → “MDLer” → “狛犬(Komainu)” → “T-AND-N”. “T-AND-N” won the cup.



Appendix: The “Cheating”

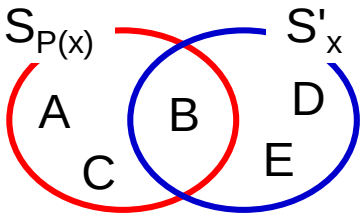
- Cheating Anonymization
  - Each record is anonymized **too much**.



- Cheating Detection Based on Jaccard Distance
  - We regarded anonymized data as cheating data if Jaccard distance is larger than 0.7 on average.

S'_x = set of goods paid by X

$S_{P(X)}$ = set of goods paid by P(X)



Jaccard Distance = 1 – |{B}| / |{A,B,C,D,E}| = 0.8 > 0.7

Appendix: Re-identification Based on Jaccard Distance

- Re-identification Based on Jaccard Distance
  - For each record in M', search a record whose Jaccard distance is the smallest.
  - Is very strong against the anonymized data whose Jaccard distance is smaller than 0.7 on average.

Master M

Customer ID	Set of Items
12346	A, B, C, D, E
12347	B, C, E, F
12348	A, B, D, E, G

Anonymized Master M'

Nym	Set of Items
1001	A, B, D, E, G
1002	A, B, C, D, E
1003	B, C, E, F

Q
3
1
2