Data Anonymization and Quantifying Risk Competition

Hiroaki Kikuchi

Meiji University, Japan

Issues in Anonymization

- 1. No real dataset
  - Data owner won't publish confidential dataset. Inconsistent Quasi-identifiers
- 2. No standard metrics for quantifying risk
  - Complicated models. Risk depends on many factors, e.g. dataset, technical skill, availability of background data. Utility depends on use case (but which is unknown when collecting data)
- 3. No standard model of adversary
  - "mildly motivated adversary" vs. "highly motivated adversary"

Competition PWSCUP 2015, 2016

- Privacy Workshop
- Organized by IPSJ, CSEC SIG

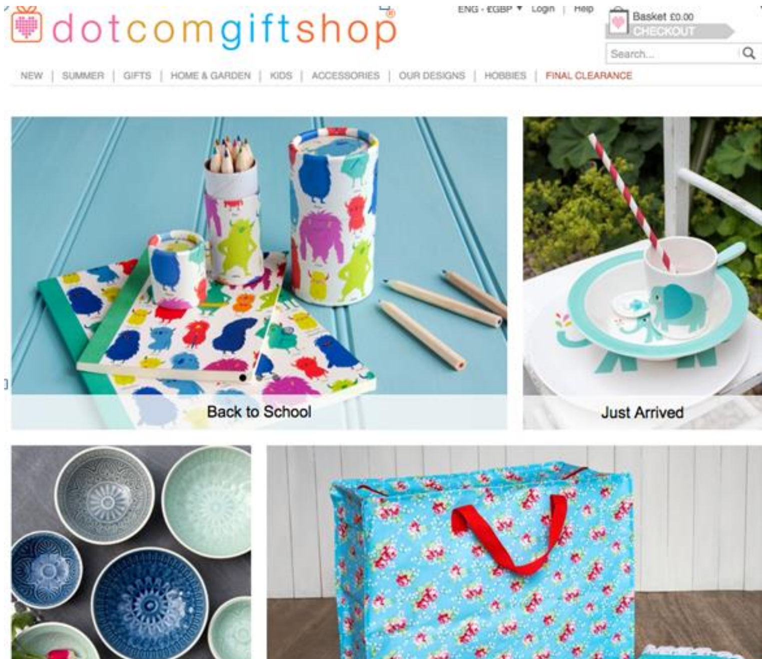
	2015	2016
Venue	Nagasaki (Brick Hall)	Akita (Castel Hotel)
When	Oct. 21, 22	Oct. 11, 12
Partici pants	13 Teams (20 in total)	15 Teams (42 in Total)
Datas et	NSTAC synthesized data	UCI Dataset, Online Retail

Our Approach

- 1. Common Dataset
  - We have used “pseudo microdata” synthesized by governmental agency, NSTAC, in 2015, and UCI Online Retail in 2016.
- 2. Quantifying risk
  - We focus on “records re-identification” risk and defines baseline utility functions and some re-identification algorithms. With arbitrary techniques, the best anonymization dataset is determined.
- 3. Adversary Model
  - We adopt Josef Domingo’s “*maximum-knowledge attacker*” model.

Dataset ‘Online Retail’

- Available from UCI Machine Learning Repository
  https://archive.ics.uci.edu/ml/datasets/Online+Retail
- Real payment transaction of UK Online Shop
  - ❑One year transactions from 2010 Dec.
  - ❑Gift shop
  - ❑540,000 records



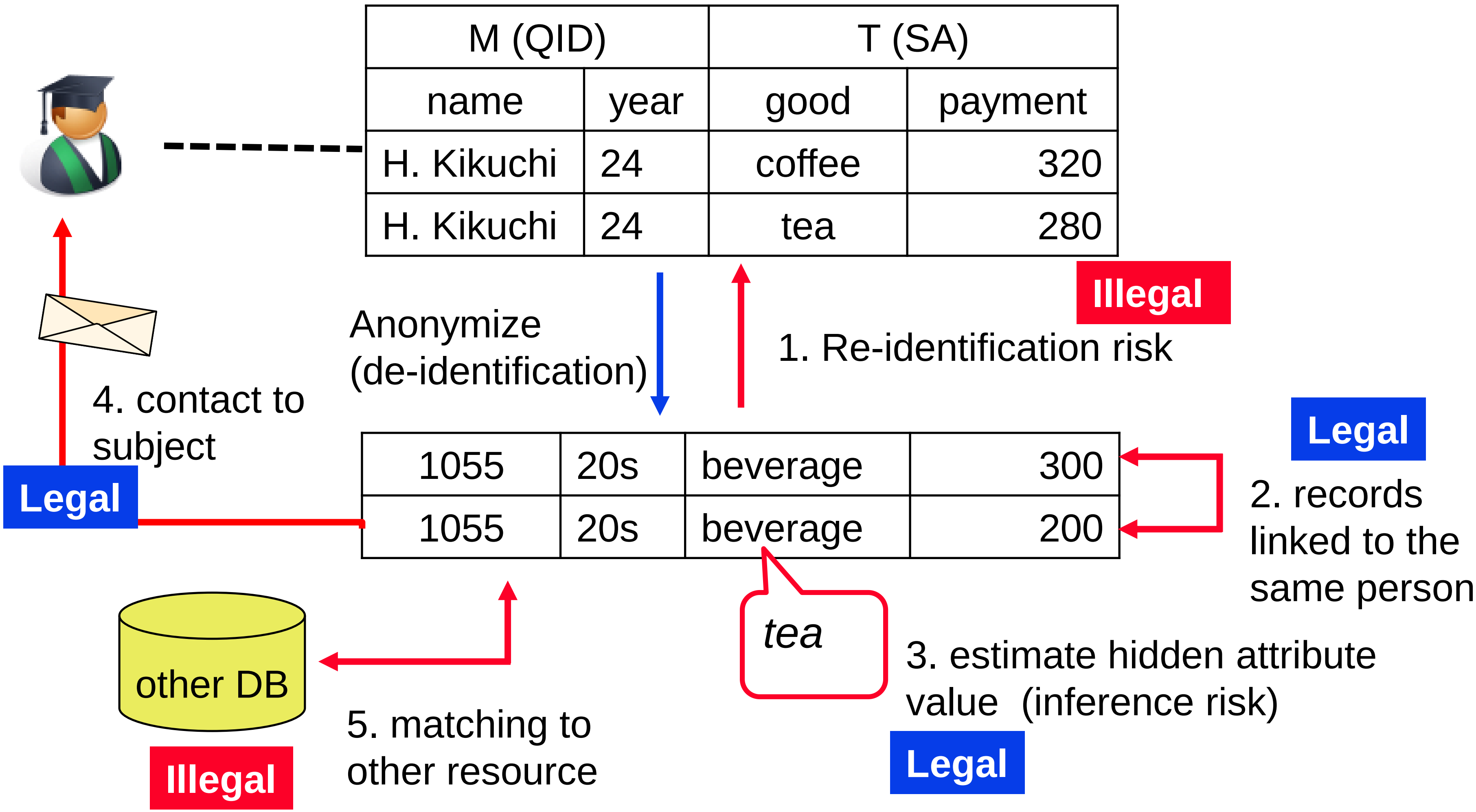
Dataset 'Online Retail'

- Master *M*
  - n = 400 customers
  - From 36 countries
- Transaction *T*
  - m = 38,087 records
  - 2,781 goods (stock code)

Customer ID	Sex	Birthday	Nationality
Online retail	synthesized		Online retail
12360	M	1876/2/24	Australia
12361	F	1954/2/14	Belgium
12362	F	1963/12/2	Belgium
12364	F	1960/9/16	Belgium

Customer ID	Invoice ID	Data	Time	Stock Code	Unit Price	Quantity
12362	544203	2011/2/17	10:30	21913	3.75	4
12362	544203	2011/2/17	10:30	22431	1.95	6
12361	545017	2011/2/25	13:51	22630	1.95	12
12361	545017	2011/2/25	13:51	22326	2.95	6

Privacy Risks (in Japan)



M (QID)		T (SA)	
name	year	good	payment
H. Kikuchi	24	coffee	320
H. Kikuchi	24	tea	280

1055	20s	beverage	300
1055	20s	beverage	200

The Game

Master *M*

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

Transaction *T*

C. ID	Date	Stock
12347	2010/12/7	85116
12347	2010/12/7	22375
12346	2011/1/18	23166

Anonymization (pseudonym, perturbation, suppression)

Record index

Q	P
3	3
2	1
2	2

Estimated index Q

Anonymized M'

C. ID	Sex	Birthday	Country
10	m	1947/01/01	UK
20	f	1960/01/01	UK
30	f	1960/01/01	UK

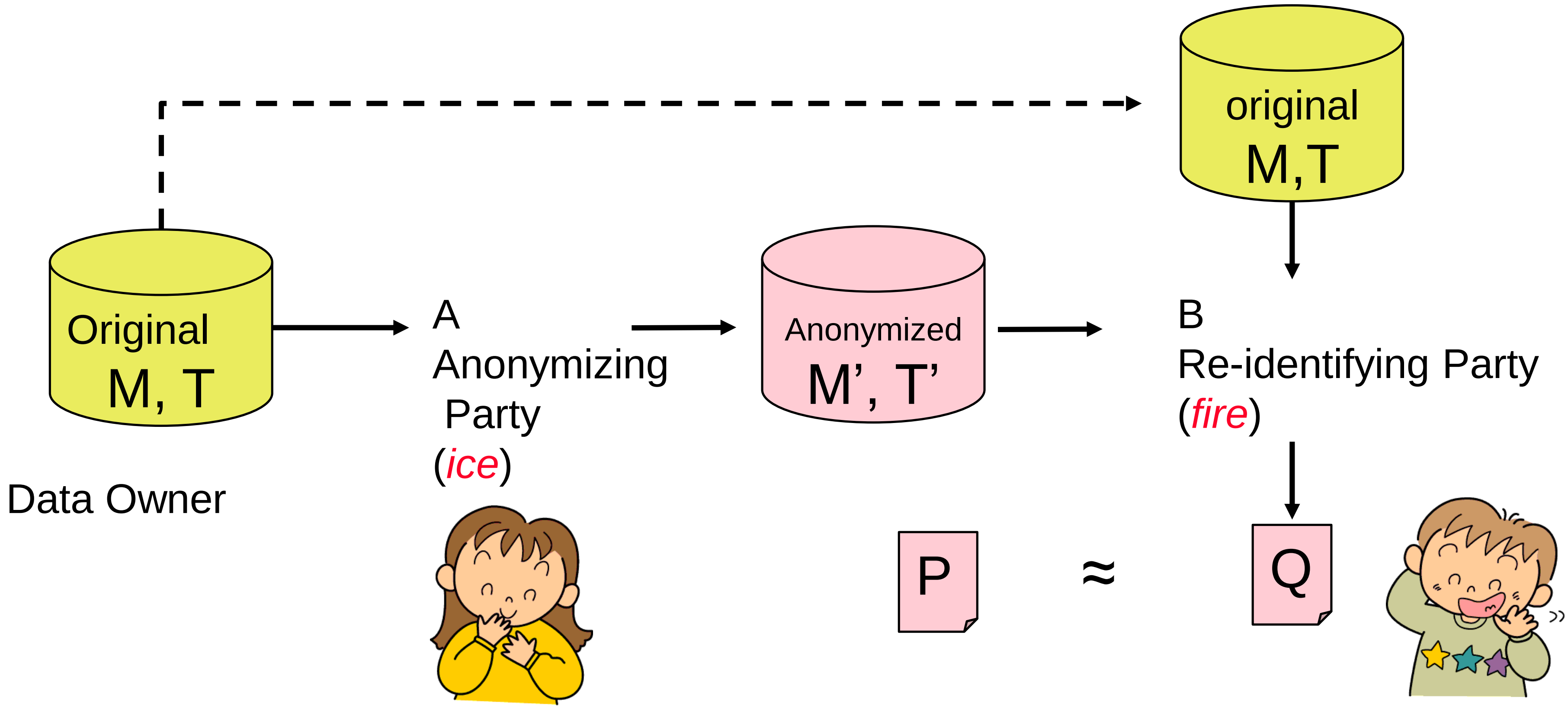
Anonymized T'

C. ID	Date	Stock
10	2010/12/1	85123A
30	2010/12/1	85123A
30	2010/12/7	20000
20	2011/1/18	20000

$$\text{Re-identification rate Re-ID(P,Q)} = \frac{\text{\# Correct records}}{n'} = 2/3$$

Adversary Model

- Maximum Knowledge Adversary Model



Use cases and Utility

- 1. RFM Analysis
  - Classification of customers based on **R**ecency (last purchase), **F**requency (of puchase), **M**onetary（Amount of payment）

U3: ut-rfm

- 2. Association Rule mining
  - Association rule of stock code

U4: ut-top_item

- 3. Cross tabulation
  - Accumulation of payment for several categories, sex, age, countries.

U1: ut-cmae
U2: ut-cmae2

Sample Re-identification

No	Algorithm	Description	M	T
E1	Re-birthday.py	Find the shortest birthday	✓	
E2	Re-eqi.rb	Find exact match	✓	✓
E3	Re-sort.rb	Sort and match	✓	
E4	Re-sort.rb	Sort by M and match	✓	
E5	Re-recnum.py	Find the shortest # recipients		✓
E6	Re-eqtr.rb	Find the same T		✓
E7	Re-tnum.rb	Sort by # records		✓
E8	Re-voting.py	Voting by birth, mean time, payment		✓
E9	Re-meantime.py	Find the shortest mean time		✓
E10	Re-ret.jar	Find similar set of goods		✓
E11	Re-sort2.tb	Sort by time and match		✓
E12	Re-search.rb	Find the shortest total payments		✓
E13	Re-totprice.py	Find the nearest set of goods		✓

E7 re-tnum-bi (best re-id score)

- Step 1: count # records in T for each customer
- Step 2: sort C-ID and P-ID by # records and birthday
- Step 3: match two sorted sequence and output Q

M

# records	C-ID	birthday	･･･
2	12346	1960/12/25	･･･
1	12347	1957/5/15	･･･
0	12348	1947/2/19	･･･

T

C-ID	･･･
12346	･･･
12346	･･･
12347	･･･

M’

# records	P-ID	Birthday	･･･
1	10	1947/01/01	･･･
2	20	1960/01/01	･･･
0	30	1960/01/01	･･･

T’

Pseudo	･･･
10	･･･
20	･･･
20	･･･

Competition rule

- Rule Ver. 1.3
  (1) Each team submits one anonymized data.
  (2) Reject cheating anonymization
  (3) Each team is allowed to re-identify the anonymized data submitted by others in hour.
  (4) Winner is determined by grade defined by U + E, the sum of minimum utilities and the minimum security (max re-identification rate).
  (5) Best Re-identification is award to team who succeeds to re-idetentify the winner’s data.

The "Cheating"

- Cheating anonimization

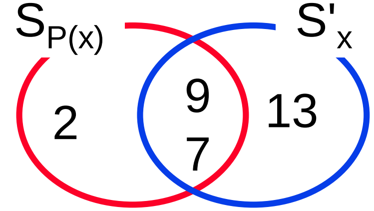
	M		
X= 1	12346	f	GE
	12347	f	UK
P(X)= 3	12348	m	UK

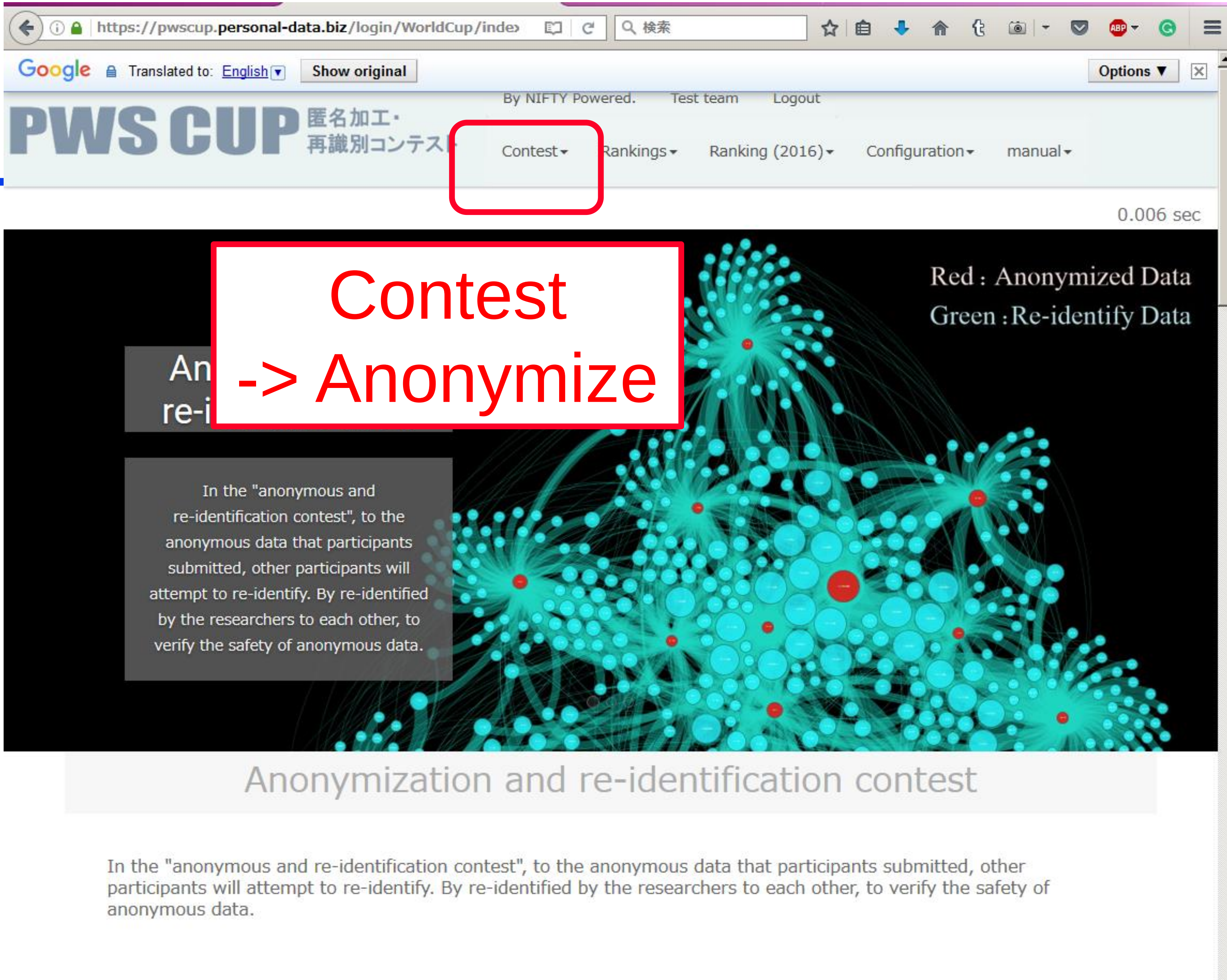
M' (=M)			P		Estimate Q
12346	f	GE	3	≠	1
12347	f	UK	1	≠	2
12348	m	UK	2	≠	3

- Cheating detection
  - ❑ Y1 (subset) > 50,000
  - ❑ Y2 (Jaccard) > 0.7

Y1:
$\mu_{P(x)}$= Total monthly payment of P(X) = 305
$\mu_{P(x)}$= Total monthly payment of X = 405

Y2:
S'_x = set of goods paid by X
$S_{P(X)}$ = set of goods paid by P(X)

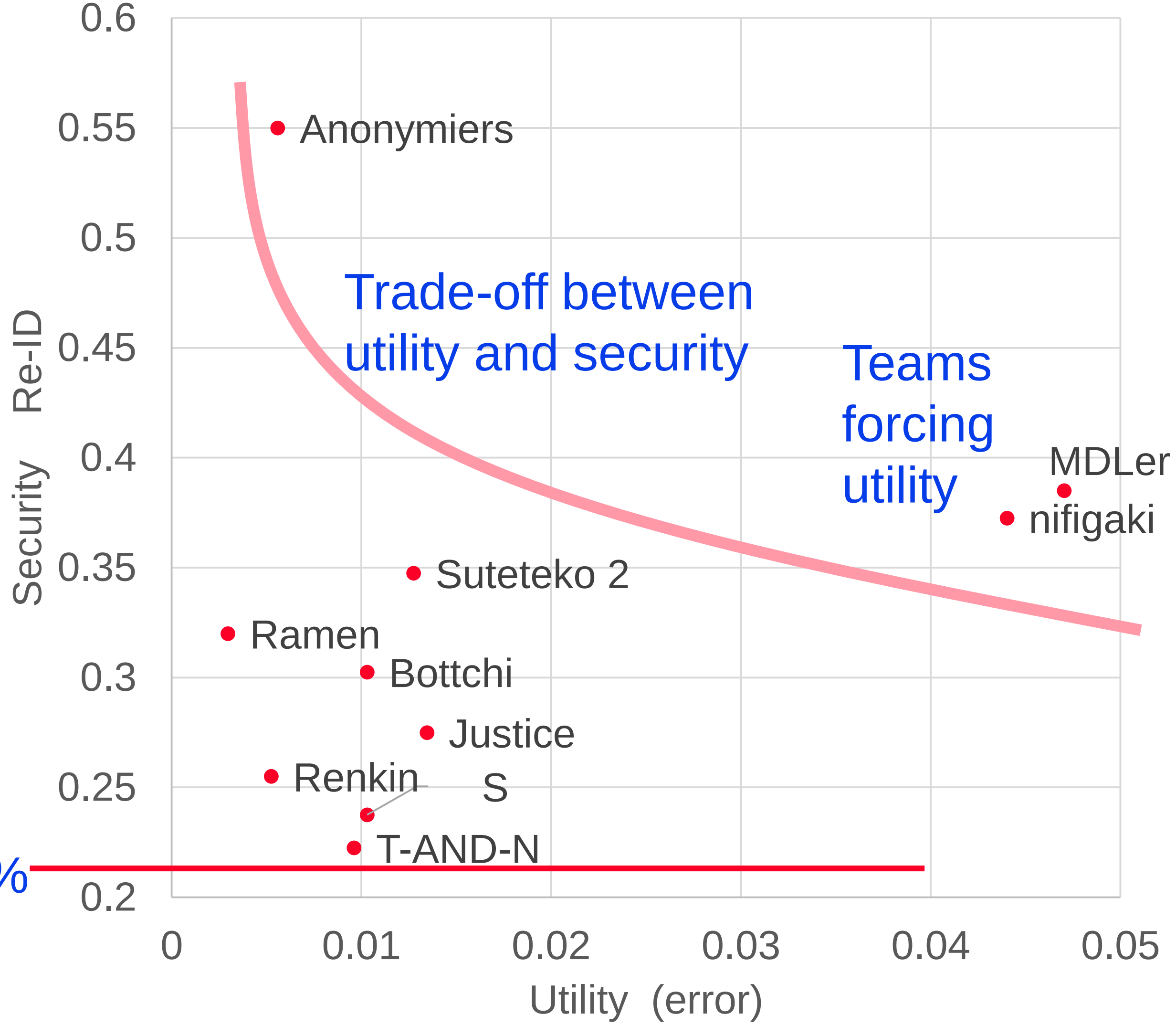



https://pwscup.personal-data.biz/login/WorldCup/inde
Google Translated to: English Show original Options
By NIFTY Powered. Test team Logout
PWS CUP 匿名加工・再識別コンテスト
Contest Rankings Ranking (2016) Configuration manual
0.006 sec
Contest
-> Anonymize
Red : Anonymized Data
Green :Re-identify Data
In the "anonymous and re-identification contest", to the anonymous data that participants submitted, other participants will attempt to re-identify. By re-identified by the researchers to each other, to verify the safety of anonymous data.
Anonymization and re-identification contest
In the "anonymous and re-identification contest", to the anonymous data that participants submitted, other participants will attempt to re-identify. By re-identified by the researchers to each other, to verify the safety of anonymous data.

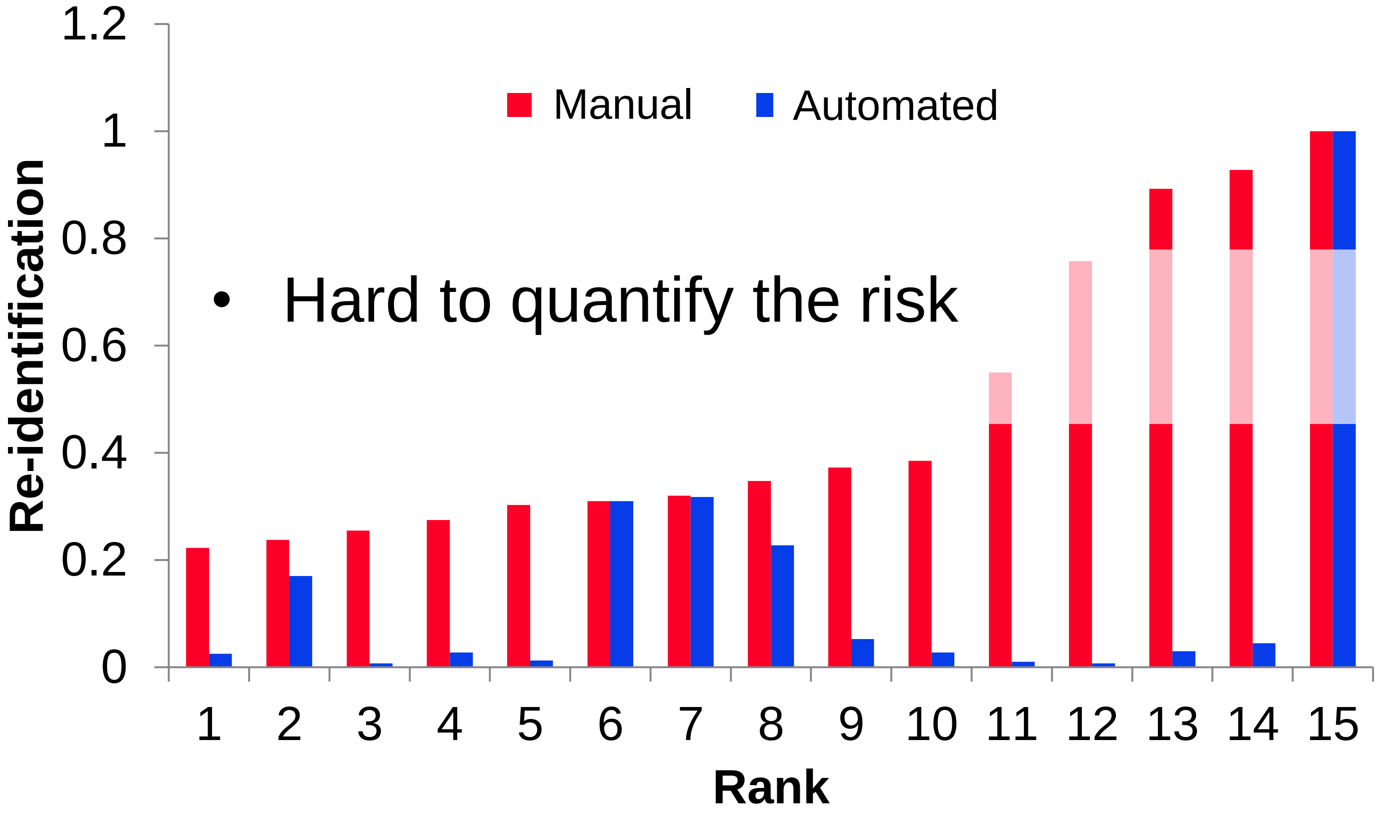
Competition Result (Top 10 teams)

No	team
1	T-AND-N
2	Shirai 5000
3	Renkin
4	Justice
5	Bottchi
6	Ramen
7	Suteteko 2
8	nifigaki
9	MDLer
10	Anonymers

Min re-id = 22.25%



Automated and Manual re-id.



- Hard to quantify the risk

Conclusions

- Data anonymization competition 2016 with real online retail data was done successfully.
- Average re-identification is 188 (47%) out of 400 customers. The best (minimum) re-identification ratio is 22%.
- Mean Automated re-identification was 18%, manual re-identification was 47%.
  - Kikuchi, et.al, “A Study from the Data Anonymization Competition Pwscup 2015”, DPM 2016, LNCS 9963.
  - Kikuchi, et. Al, “Ice and Fire: Quantifying the Risk of Re-identification and Utility in Data Anonymization”, IEEE AINA 2016.