iPWS Cup 2023 rules

Ver 1.0

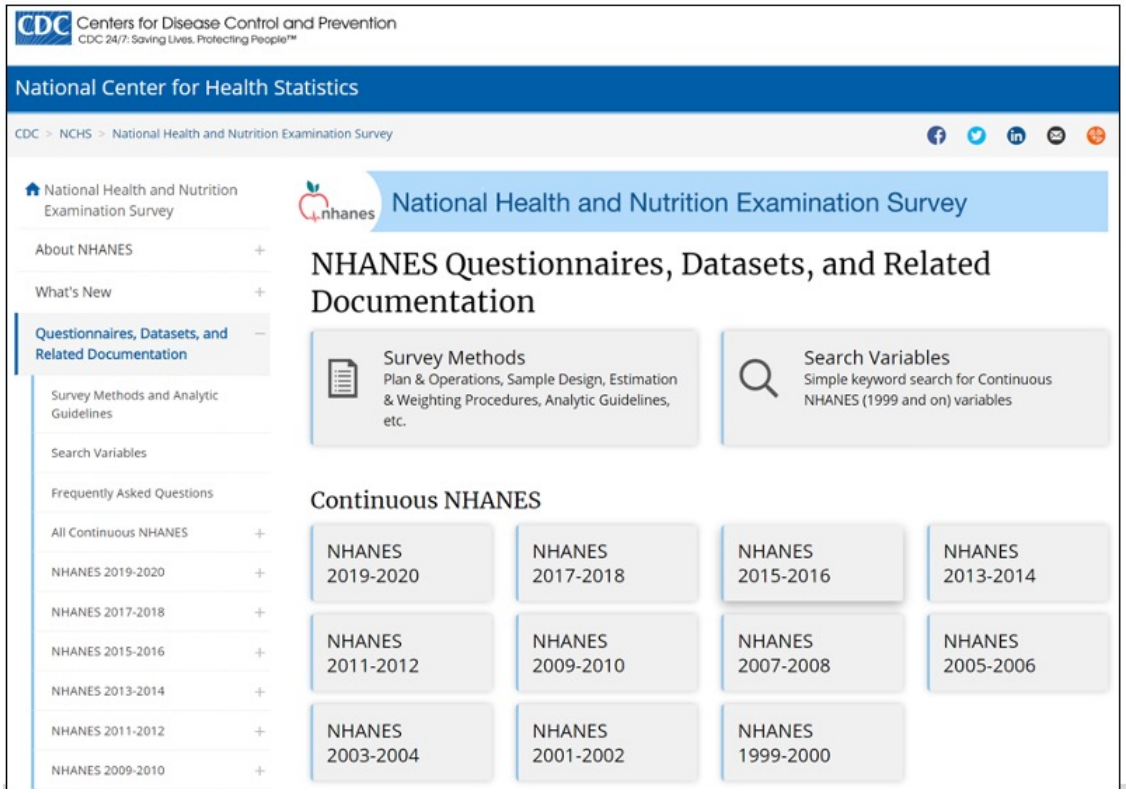
“diabetes challenge”

iPWS Cup committee

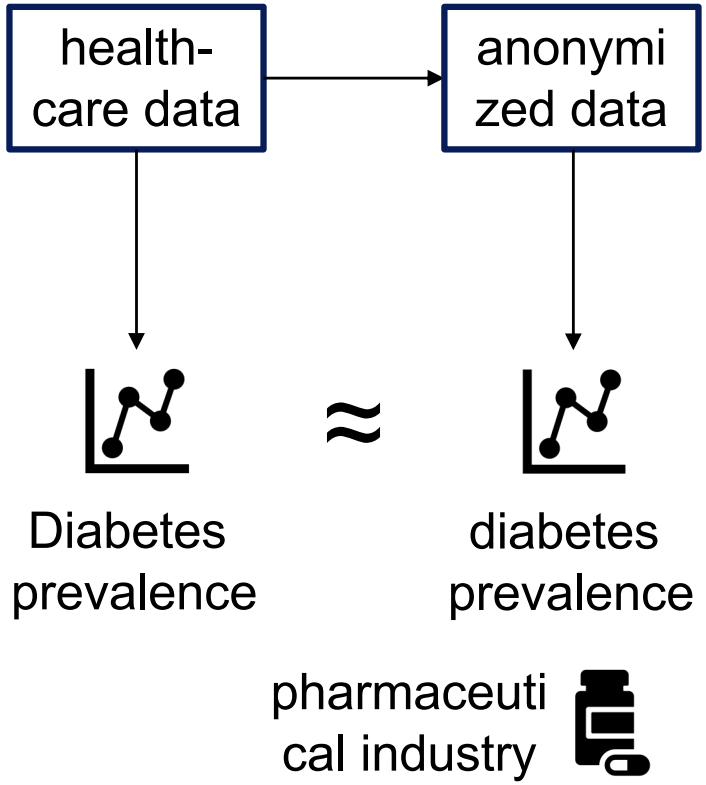
(2023 June 8)

Objectives

- Make healthcare data anonymous so that no re-identification is feasible



- Preserve utility of data good for medical analysis



NHANES

- National Health and Nutrition Examination Survey
  - US CDC study, 5,000 subject per year since 1960

`activ_diabet9_csv.py Csv/B.csv`

12 attr.

gen	age	race	edu	mar	bmi	dep	pir	gh	mets	qm	dia
Male	62	White	Graduate	Married	27.8	0	0	0	0	Q2	1
Male	53	White	HighSchool	Divorced	30.8	0	1	0	0	Q1	0
Male	78	White	HighSchool	Married	28.8	0	0	0	0	Q3	1
Female	56	White	Graduate	Parther	42.4	1	0	0	0	Q3	0
Female	42	Black	College	Divorced	20.3	1	0	0	0	Q4	0
Female	72	Mexican	11th	Separated	28.6	0	0	0	0	Q1	0

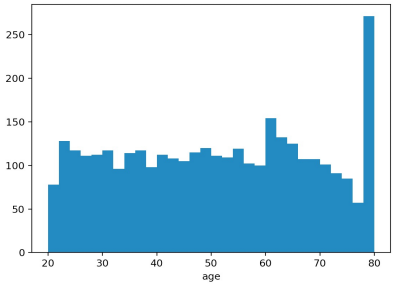
continuous values

dependent variable (diabetes)

n = 4,190 individuals

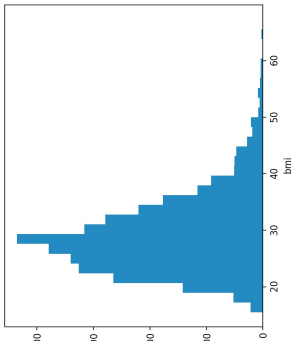
Idiosyncratic data (1)

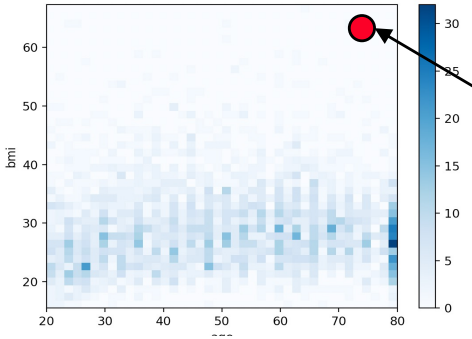
gen	age	race	edu	mar	bmi	dep	pir	gh	mets	qm	dia
Male	62	White	Graduate	Married	27.8	0	0	0	0	Q2	1
Male	53	White	HighSchool	Divorced	30.8	0	1	0	0	Q1	0
Male	78	White	HighSchool	Married	28.8	0	0	0	0	Q3	1
Female	56	White	Graduate	Parther	42.4	1	0	0	0	Q3	0
Female	42	Black	College	Divorced	20.3	1	0	0	0	Q4	0
Female	72	Mexican	11th	Separated	67.3	0	0	0	0	Q1	0



distribution of age
20 < age < 80

distribution of BMI





age = 72,
BMI = 67.3

Idiosyncratic data (1)

- Single-out data
  - unique combination of values

gen	age	race	edu	mar	N
Male	50	White	Graduate	Widowed	2
Male	50	White	Graduate	Widowed	
Male	20	White	Graduate	Married	2
Male	20	White	Graduate	Married	
Male	**20**	White	Graduate	**Widowed**	1
Female	60	White	Graduate	Widowed	2
Female	60	White	Graduate	Widowed	
Female	50	White	Graduate	Divorced	3
Female	50	White	Graduate	Divorced	
Female	50	White	Graduate	Divorced	
Female	20	White	Graduate	Married	2
Female	20	White	Graduate	Married	

Story

- Players

- Data processor (anonymizer)
  - » owns demographic data and medical examination data

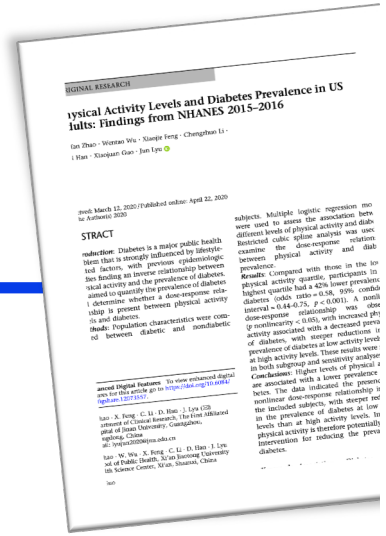
- Data subject (attacker)
  - » curious to know if my data was disclosed

- Data consumer (utility)
  - » wish to use the anonymized data to estimate diabetes prevalence risk given medical examination results.

- Judge (organizer)
  - » determines which anonymization is most robust against re-identification

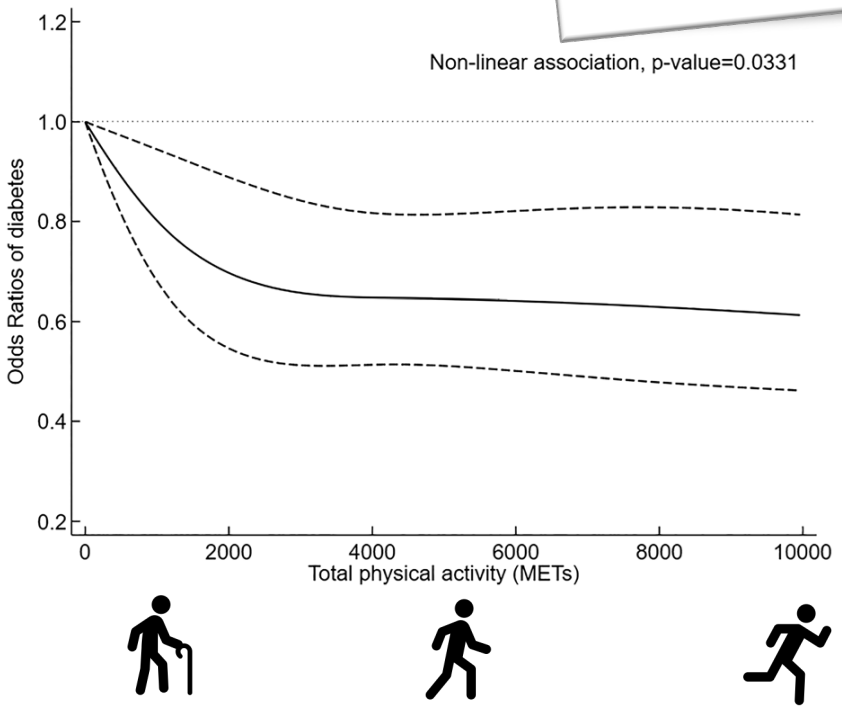
Diabetes prevalence

- ORs (Model III, adjusted)


ysical Activity Levels and Diabetes Prevalence in US
lults: Findings from NHANES 2015–2016

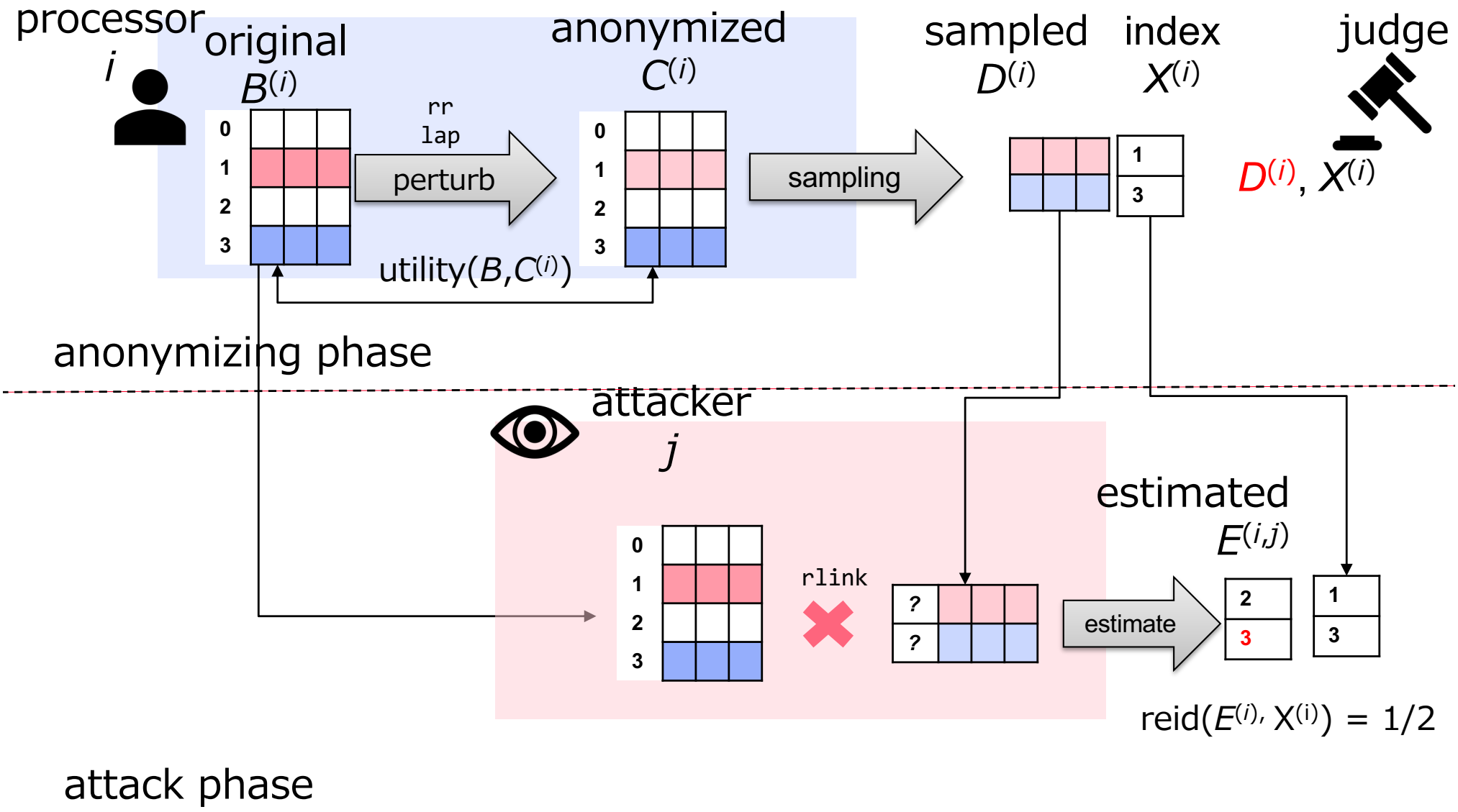
STRACT

METs	OR	95% Conf. int.	P value
Q1	1		
Q2	0.71	0.56-0.89	0.003
Q3	0.66	0.52-0.84	0.001
Q4	0.58	0.44-0.75	< 0.001



Fanfan Zhao, et. al (The First Affiliated Hospital of Jinan, Xi'an Jiaoton Univ.), "Physical Activity Levels and Diabetes Prevalence in US Adults: Findings from NHANES 2015–2016", Diabets Ther 2020.

Overview


processor
i
original
B(i)
rr
lap
perturb
anonymized
C(i)
utility(B,C(i))
sampling
sampled
D(i)
index
X(i)
judge
D(i), X(i)
anonymizing phase
attacker
j
rlink
estimated
E(i,j)
estimate
reid(E(i), X(i)) = 1/2
attack phase

Utility metrics

rate

diabetes prevalence (33x2)

Table 1 [Zhao 2020]

	cnt		rate	
diabetes	0	1	0	1
Female	1710	407	0.408	0.097
Male	1607	466	0.384	0.111
(19, 44]	1574	110	0.376	0.026
(44, 64]	1033	379	0.247	0.090
(64, 80]	710	384	0.169	0.092
Black	647	208	0.154	0.050
Hispanic	443	127	0.106	0.030
Mexican	518	200	0.124	0.048
Other	537	112	0.128	0.027
White	1172	226	0.280	0.054
11th	373	120	0.089	0.029
9th	386	168	0.092	0.040
College	975	239	0.233	0.057
Graduate	851	156	0.203	0.037
HighSchool	732	190	0.175	0.045
Divorced	342	103	0.082	0.025
Married	1669	502	0.398	0.120
Never	652	87	0.156	0.021
Parther	331	57	0.079	0.014
Separated	101	32	0.024	0.008
Widowed	222	92	0.053	0.022
(15.0, 18.5]	64	4	0.015	0.001
(18.5, 25.0]	1027	114	0.245	0.027
(25.0, 30.0]	1149	241	0.274	0.058
(30.0, 70.0]	1075	514	0.257	0.123
dep_0	2666	647	0.636	0.154
dep_1	651	226	0.155	0.054
pir_0	2656	650	0.634	0.155

cor

covariance matrix (30x30)

	0_Male	0_Female	1	2_White	2_Black	2
0_Male	0.00	0.00	0.00	0.00	0.00	
0_Female	-1.00	0.00	0.00	0.00	0.00	
1	0.01	-0.01	0.00	0.00	0.00	
2_White	0.02	-0.02	0.12	0.00	0.00	
2_Black	-0.02	0.02	-0.05	-0.37	0.00	
2_Mexican	-0.02	0.02	-0.03	-0.34	-0.23	
2_Other	0.04	-0.04	-0.11	-0.30	-0.20	
2_Hispanic	-0.03	0.03	0.02	-0.29	-0.19	
3_Graduate	0.01	-0.01	-0.06	0.09	-0.07	
3_HighSchool	0.03	-0.03	-0.01	0.02	0.07	
3_College	-0.06	0.06	-0.05	0.11	0.07	
3_11th	0.04	-0.04	0.01	-0.10	0.02	
3_9th	-0.01	0.01	0.16	-0.21	-0.10	
3_nan	-0.02	0.02	0.02	-0.01	-0.01	
4_Married	0.10	-0.10	0.14	0.04	-0.15	
4_Divorced	-0.06	0.06	0.14	0.03	0.04	
4_Parther	0.02	-0.02	-0.21	-0.03	0.00	
4_Separated	-0.03	0.03	0.02	-0.06	0.04	
4_Never	0.00	0.00	-0.37	-0.08	0.16	
4_Widowed	-0.14	0.14	0.34	0.07	-0.01	
5	-0.08	0.08	0.06	-0.04	0.09	
6	-0.08	0.08	-0.03	0.03	-0.04	
7	-0.03	0.03	0.00	-0.21	0.04	
10_Q2	-0.06	0.06	0.09	0.00	0.01	
10_Q1	-0.11	0.11	0.21	-0.08	-0.03	
10_Q3	0.02	-0.02	-0.07	0.07	-0.04	

OR

odds retio (21x2)

Table 2 [Zhao 2020

	Coef	OR	pvalue
Intercept	-7.319	0.001	0.0
gen[T.Male]	0.380	1.463	0.0
race[T.Hispanic]	-0.302	0.740	0.0
race[T.Mexican]	0.084	1.088	0.5
race[T.Other]	0.020	1.020	0.9
race[T.White]	-0.844	0.430	0.0
edu[T.9th]	-0.151	0.860	0.4
edu[T.College]	-0.073	0.930	0.6
edu[T.Graduate]	-0.150	0.861	0.3
edu[T.HighSchool]	-0.192	0.825	0.2
mar[T.Married]	0.278	1.320	0.0
mar[T.Never]	0.326	1.386	0.1
mar[T.Parther]	0.316	1.372	0.1
mar[T.Separated]	0.081	1.084	0.7
mar[T.Widowed]	-0.134	0.875	0.5
qm[T.Q2]	-0.228	0.796	0.0
qm[T.Q3]	-0.328	0.720	0.0
qm[T.Q4]	-0.401	0.670	0.0
age	0.057	1.058	0.0
bmi	0.097	1.102	0.0
dep	0.440	1.552	0.0
pir	0.147	1.158	0.1

Utility metrics

- Utility loss score

$$U(B,C) = \left(\prod_{u \in \{rate,cor,or,age,bmi,cat\}} \Delta(u(B), u(C)) \right)^{1/6}$$

  - where distances are defined

$$\Delta(x,y) = \max |u(x) - u(y)|$$

$$\Delta(x,y) = \max \frac{\min(|u(x) - u(y)|, 20)}{20}$$

$$\Delta(x,y) = \max \frac{|\{u(x) \neq u(y)\}|}{8}$$

for utility metrics (rate, cor, or)
for numerical values (age, bmi)
for categorical values (gender, marital, …)

Risk metrics

- Re-identification ratio
  - a fraction of correctly identified record of i-th anonymized data estimated by j-th attacker, defined by

$$R(E^{(i,j)}, X^{(i)}) = \frac{|\{k \in \{1, ..., N\} | e_k = x_k\}|}{N}$$

  - where E and X are N-dimensional vectors for record indexes

$$E^{(i,j)} = (e_1, \ldots, e_N), X^{(i)} = (x_1, \ldots, x_N)$$

Winner

judge

- Utility loss score
  - U = U(B,C)
- Risk score
  - Re-identification rate R = # correct record indexes/ n (# records)
  - Maximum re-id for i-th anonymized data against j-th attacker $R^{(i)} = \max_{j \neq i} R^{(i,j)}$
  - F1 score of U and R (F1 = 2/(1/U + 1/R))
- Attack score
  - Attacker j-th has attacking score as the mean risk score for m-1 teams (except self j-th data) T as $A^{(j)} = \frac{1}{m-1} \sum R^{(i,j)}$
- Overall score
  - weighted sum of preliminary and final phases with 1:9.

Example

$$A^{(j)} = \frac{1}{4} \sum_{i \in T_3} R^{(i,j)}$$

Proc. / estimators	01	02	03	04	05	A	rank
01		0.520	0.005	0.016	0.010	0.022	3
02	0.068		0.003	0.003	0.002	0.016	4
03	**0.093**	**0.970**		0.004	0.049	0.013	5
04	0.010	0.007	**0.011**		0.002	0.027	2
05	0.005	0.005	0.005	0.005		0.043	1
R	0.093	0.970	0.011	0.067	0.165	0.024	
R(Max)	0.093	0.970	0.011	0.067	0.165	1.000	
rank	3	5	1	2	4		

$R^{(i)} = \max_{j \neq i} R^{(i,j)}$

Processor: anonymizes data to minimize Risk

Reference

- Zhao, Fanfan, Wu, Wentao, Feng, Xiaojie, Li, Chengzhuo, Han, Didi, Guo, Xiaojuan, et al. Physical Activity Levels and Diabetes Prevalence in US Adults: Findings from NHANES 2015-2016. Adis Journals. Figure. DOI 10.6084/m9.figshare.12073557.v1 (2020):（NHANES）
- Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Hu- man Services, Centers for Disease Control and Prevention.
- Takao Murakami, Hiromi Arai, Koki Hamada, Takuma Hatano, Makoto Iguchi, Hiroaki Kikuchi, Atsushi Kuromasa, Hiroshi Nakagawa, Yuichi Nakamura, Kenshiro Nishiyama, Ryo Nojima Hidenobu Oguri, Chiemi Watanabe, Akira Yamada, Takayasu Yamaguchi, and Yuji Yamaoka, "Designing a Location Trace Anonymization Contest ", Proceedings on Privacy Enhancing Technologies (PoPETs), vol. 2023, no. 1, pp.225-243, 2023.
DOI: https://doi.org/10.56553/popets-2023-0014