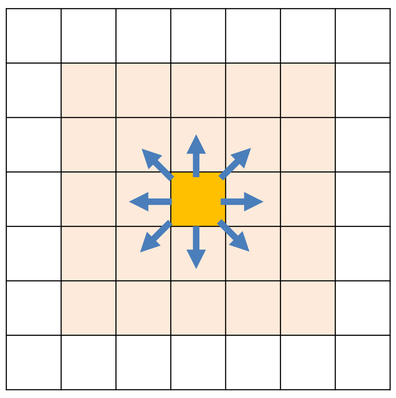
PWS Cup 2019

Team 004 (またぼっち, Botch Again) / Makoto IGUCHI (Kii Corporation)

Data Anonymization

Basic strategy: Shift regions by N

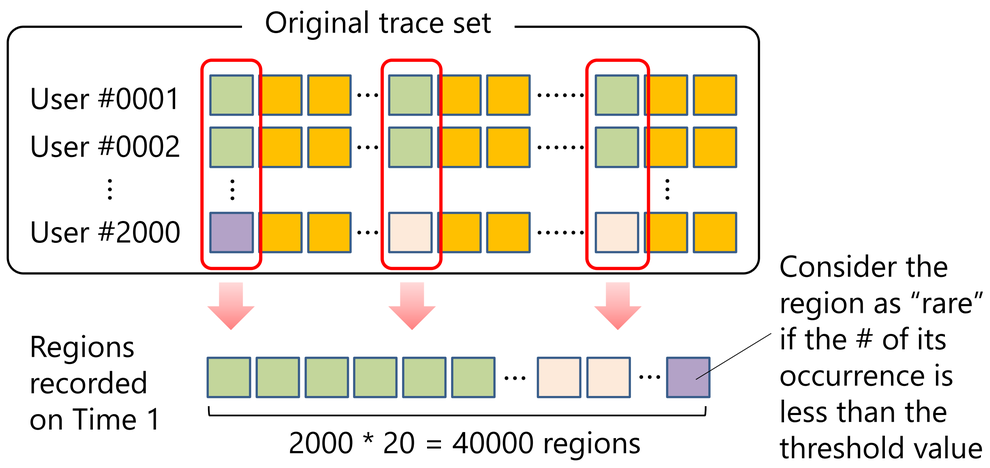
Example: Shift a region by 2



Randomly shift the region (at most 2 blocks) both vertically and horizontally

Additional strategy: Protect **"weak"** regions by shifting

Definition of "weak" regions:
- "Hospital" regions (for TRP)
- "Rare" regions in each time slot (IDP and TRP)



Evaluation/implementation:

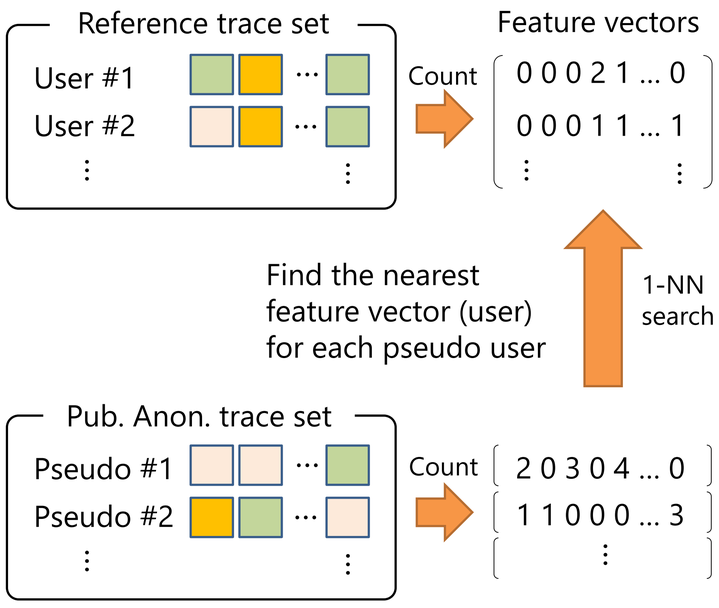
AS-ShiftRareRegions (n, d_1, d_2)
- n : threshold for "rare" region determination
- d_1 : # of maximum shift for hospital regions
- d_2 : # of maximum shift for rare regions

AS (94, 0, 3) for ID disclosure challenge
- Shift as many rare regions as possible by 3
- S_U = 0.70270 (275598 regions untouched)

AS (89, 4, 3) for trace inference challenge
- Shift all hospital regions by 4
- Shift as many rare regions as possible by 3
- S_U = 0.70211 (283,175 regions untouched)

ID Disclosure

Basic strategy: Feature vector comparison



Additional strategy #1:
"TF-IDF" style feature vector generation

When generating the feature vectors, weight "uncommon" regions more than "common" regions

	TF weight	IDF weight
Scheme 1	$f_{r,u}$	$\log U/u_r$
Scheme 2	$\log(1+f_{r,u})$	1
Scheme 3	$f_{r,u}$	$\log U/u_r$
Scheme 4	$\log(1+f_{r,u})$	1

The optimal scheme is to be found through evaluation with sample data.

$f_{r,u}$: raw count of Region r in User u's traces
U: total # of users (2000)
u_r: # of users whose traces contain Region r

Additional strategy #2:
"Fuzzy" feature vector generation

When generating the feature vectors, count the target region and its surrounding regions "fuzzily."

Example:

0	0	0	0	0
0	0.099	0.121	0.099	0
0	0.121	0.20	0.121	0
0	0.099	1.121	0.099	0
0	0	0	0	0

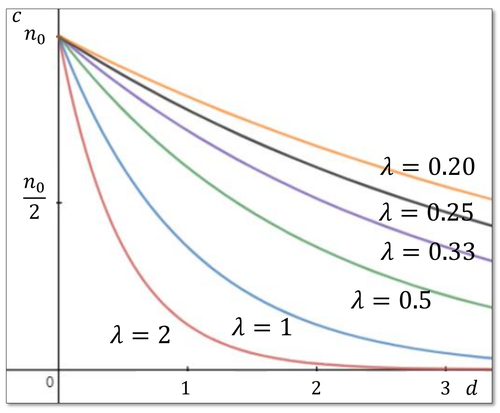
Fuzzy counting ($n_0 = 0.2, \lambda = 0.5$)

0	0	0	0	0
0	0	0	0	0
0	0	1	0	0
0	0	0	0	0
0	0	0	0	0

Raw counting (equiv. to $n_0 = 1, \lambda = \infty$)

"Fuzzy" counting is realized by an exponential decay function: $c = n_0 e^{-\lambda d}$

(n_0: initial quantity, λ: exponential decay constant, d: distance from the target region)



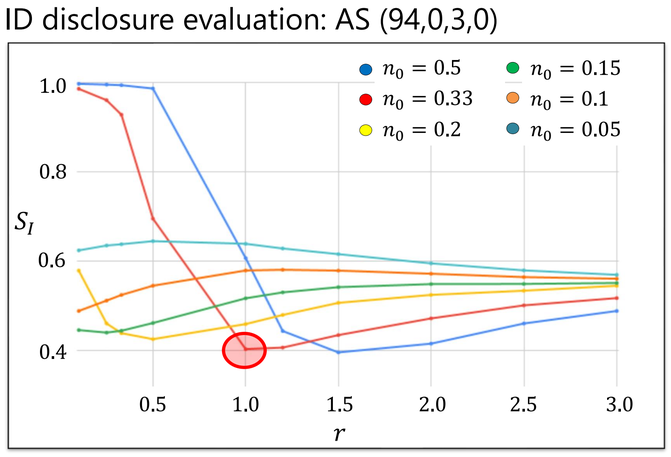
The optimal curve (i.e., n_0 and λ) is to be found through evaluations with sample data

Evaluation/implementation:

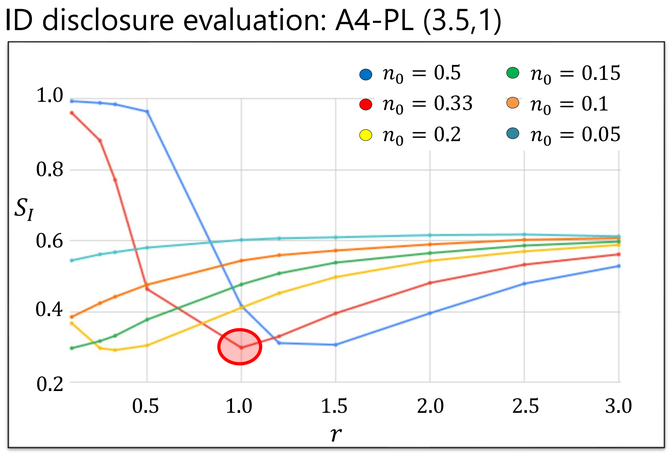
IF-FuzzyVisitVector (n_0, λ)
- n_0 : initial quantity
- λ : exponential decay constant

Experiments with sample data sets revealed that **IF (0.33,1)** with **Scheme 2** (TF weight: $log(1+f_{r,u})$, IDF weight: 1) yields to the optimal ID disclosure result S_I.

ID disclosure evaluation: AS (94,0,3,0)

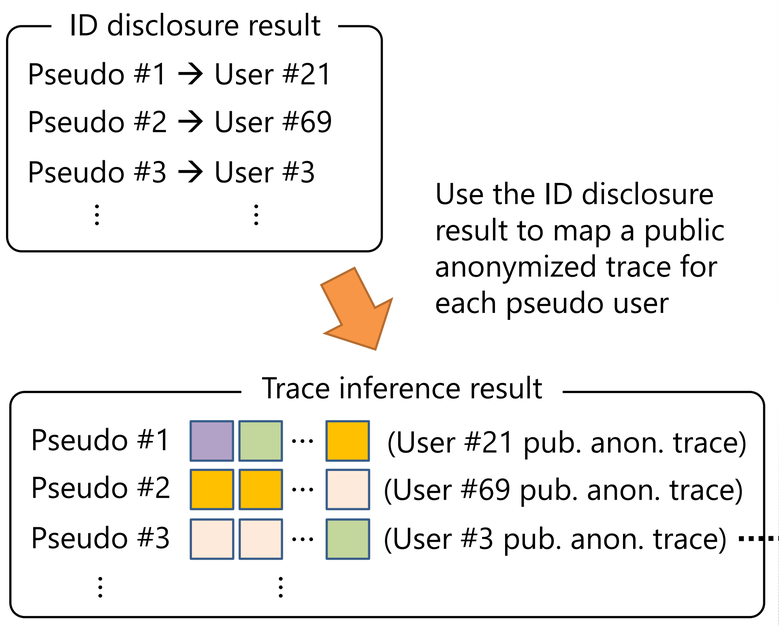


ID disclosure evaluation: A4-PL (3.5,1)

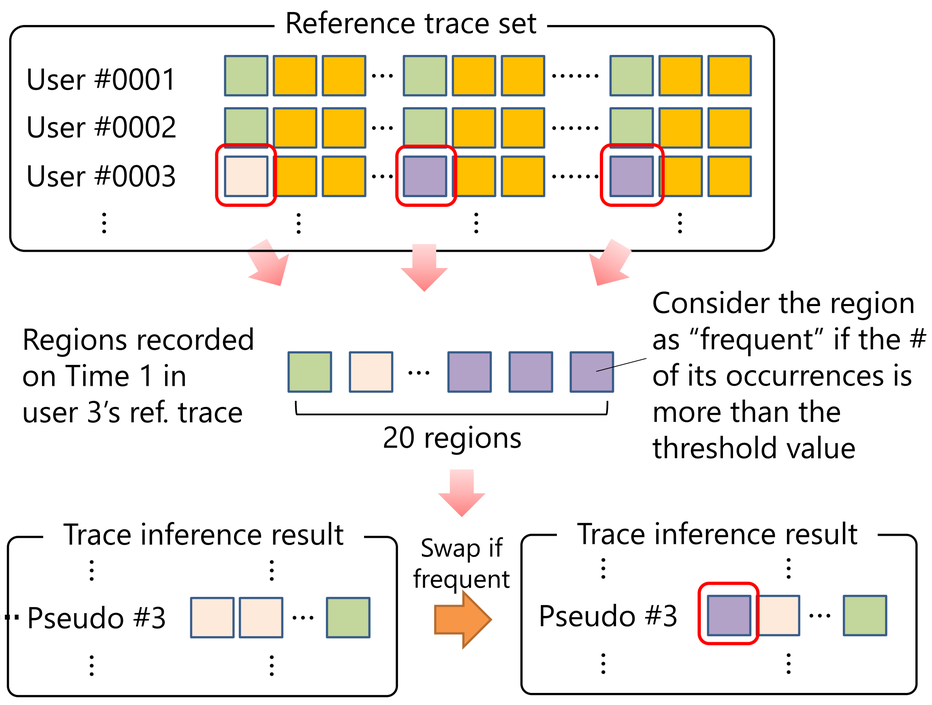


Trace Inference

Basic strategy:
List-up public anonymized traces on the basis of the ID disclosure result



Additional strategy: Use the reference trace sets to replace **"frequent"** regions



Evaluation/implementation:

TS-SwapFrequentRegions (n)
- n : threshold for "frequent" region determination

Experiments with sample data sets indicated that **TS(5)** yields to the optimal trace inference score S_T.

Trace inf. evaluation: AB(89,4,3,0)+IF(0.33,1)