

Design for a data Anonymization Competition 2018

Hiroaki Kikuchi (Meiji Univ.)

PETS 2017, Minneapolis, US

Criticize to past PWSCUP

1. Hidden algorithm

- Players submit the anonymized data without showing source or algorithm. Not able to analyze the process for details.
- 2. Max-knowledge assumption is too strong.
 It is far from reality.
- 3. Record-linkage challenge is problematic.
 Instead, why don't us to attribute estimation?
- 4. Synchronized fashion of games
 Arbitrarily attack and defense is more exciting, like the CTF style.

Open-Source style

IDash Privacy and Security WS

IDASH PRIVACY & SECURITY WORKSHOP 2016



Important time points

02/12/2016 Competition registration open

03/15/2016 (delayed to 3/18/2016) Competition start (data release to registered teams)

03/31/2016 Early competition registration

05/31/2016 Workshop registration deadline (onsite attendance)

09/07/2016 Solutions submission registration due (No more registration allowed after 09/07/2016). Here

1. Pros and Cons for Open-Source style

Pros

□ Allows deep analysis

- Can be re-used for anonymizing other dataset.
- Fair and reliable.
 Allows to trace the steps one by one.
- "cheating" can be denied.
- No need highperformance

Cons

- Revealing method is prohibited by Japanese low
- Most companies does not allow to submit their source since it has IP.
- Not processed in a single source. Often used internal library.

Our Suggestion to 1.

- We should have a closed-source (PWSCUP) style so that industry teams can participate.
- Alternatively, we may have an additional open-source style completion as well as the closed-style.

2. Why we assume the Maxknowledge adversary

Reasons

It is simple. If some algorithm was better than others in the Max-knowledge adversary, it could be safe against a moderate adversary.

- Many requests to join both anonymizing and re-identifying. (including committee members)
- It is hard to provide exactly equal knowledge to all parties. The risk may quite depend on the (partial) knowledge.

3. Why we did not study attribute estimation in the past PWSCUP



Our new competition Update PWSCUP 2017

PWS CUP 2017

- Oct. 23-25
- Yamagata Int. Hotel
- Call (July 24-Au
 - (July 24-Aug. 21)
- Privacy Workshop
 2017 (IPSJ, Sig.
 CSEC)



2017 Outline

Anonymize: submit T'1, T'2, T'3, ... Identification: given T'1, T'2, guess IDs

Μ					M	T ₁				T_2			
	ID	Se	ex	С		ID	Date		good		ID	date	good
	12346	; 1	F	UK		12347	2010/1/7		85		12347	2010/3/7	85
	12347	, I	F	UK	•	12347	2010/2/7		22		12347	2010/4/7	22
	12348	n	n	DE	•	12346	2011/1/18		66		12346	2011/3/3	30
							Anon	ymizati	0	n			
								7	T' ₁				T'
	ID	ID				Pse	Date		good		Pse	Date	good
1	2346	20	1]		30	2010/1/7		85		60	2010/3/7	85
1	2347	30	1	 ←	-	30	2010/2/7		22		60	2010/4/7	22
1	2346	40	1]		20	2011/1/18		66		40	2011/3/3	30
1	2347 50												
Re-id = 75					12347	2010/1/7		85	Partial knowledge of Ts				

2011/1/18

66

12346

1-year History divided



Changes in 2017

- 1. anonymization of long history
 - Allows multiple pseudonyms per one person so that re-identification becomes harder
 - □The more pseudonym, the more secure. But, it accordingly loses the utility.
- 2. weaken the adversary's knowledge
 Given (some) partial transaction records, try to estimate model and guess the assignment

Some plans for Competition

Proposal of completion 2018

- Plan A. NSTAC synthesized data
- Plan B. Online Retail
- Plan C. Online Retail with pseudonyms
- Plan D. Open Algorithms completion
- Plan E. Trajectory Data

Plan A "Pseudo Micro Data"

- NSTAC (National Statistics Center)
 - Real statistics about income and expenditure for Japanese household in 2004.

Dataset	# of records	QI	SA	
	n	m	(exp)	(inc)
Full	32,027	14	149	34
Simple	8,333	14	11	N/A

新たの	→Home →English	文字サイズ ^大 中 小 →サイトマップ →お問い合わせ	 ○ www を検索 ● 約 	Google 充計センターサイト						
†センターについて	業務案内	調達情報	情報公開	採用						
me > 業務案内 > <u>公前統計基盤サービス</u> > <u>公的統計のミクロデータ利用</u> > 擬似ミクロデータの利用										
業務案内統計作成実務	> 業務案内	9								
	-									
○統計利用者同分与ビス及び公的基盤サービス	擬似ミクロデータの利用									
・公的統計のミクロデータ利用	↓ 統計センターでは、公的統計のミクロデータの利用を図るため、ミクロデータを用いた実影									
… 匿名データの利用	ブログラムテストヤ	っ大学・高等学校等の授業で利	利用できるような擬似ミクロ	データの試行提供						
"オーダーメード集計の利用	を17っています。 セットであることを	なめ、擬似ミンロテータは、朱言 理解してご利用ください。	iT表がら1FMLしたミンロナー	「ジボケン(リノ族)以日う						
利用実績	また、擬似ミクロデータをご利用の際は、アンケートにご協力いただきますようよろしくお願 す。 報									
… イベント情報										
4.オンサイト利用	(参考)競队ミンロ: 製表技術参考)	テータの11F成力法 資料「平成24年7月「教育用擬	似ミクロデータの 開発とその	の利用 ~平成16						
^{LI} LISのデータベース利用	費実態調査を例と	<u>・して~」」</u> をご覧ください。								
* 擬似ミクロデータの利用										
- 利用例リンク集										
^に 利用に関するFAQ	■利用可能な擬	以ミクロデータ								
… 学術研究機関等との連携	■利用要件									
… 他府省における提供状況										

http://www.nstac.go.jp/services/giji-microdata.html#P

Pseudo Micro Data (Tbl. VII)

No	Attribute	# of value	Average	Example	Туре
1	Туре	1	1	1 (empied)	QID
2	# of people	1	4	4	QID
3	# of employed	1	1.504	1	QID
4	Accom. Type	5	1	1 (wooden)	QID
5	Bldg. type	7	1	1 (detached)	QID
6	Owner	8	1	1 (owned)	QID
7	Sex	1	1	1 (male)	QID
8	Age	11	5	1 (1-18 Y/O)	QID
					QID
14	Weight	8333	15.741	13.2	SA
15	Total Expenditures	8333	324,525	155,006	SA
16	Foods	8333	74,639	25,227	SA
17	Accom.	8333	14,686	2000	SA
14	Lightning	8333	19,733	18,333	SA
					SA
25	Others	8333	62,227	20,455	SA

Record Re-identification



Re-identification Ratio: Re-id ${}^{IE}(I^{Y}, I^{E}) = |\{j \text{ in } \{1, ..., n'\} | i_{j}^{Y} = i_{j}^{E}\}/n'$

Plan B: Online Retail

Dataset

□UCI Machine Learning, "Online Retail"

Task

□ Identify secret permutation P(M) from anonymized data M' and T'

Limitation

□Assign one pseudonym to one customer

Plan C: Online Retail with Many Pseudonyms

- Dataset
 - □UCI Machine Learning, "Online Retail"
- Task

Identify owners of records from anonymized history T' using partial knowledge

Limitation

□Assign one pseudonym to one customer

Plan D: Open-source style competition

Data:

Plan E: Trajectory Data Competition