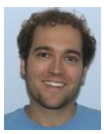
PURDUE UNIVERSITY

Brandeis University

Parallel Composition Revisited

Chris Clifton

23 October 2017



This is joint work with Keith Merrill and Shawn Merrill

This work supported by the U.S. Census Bureau under Cooperative Agreement CB16ADR0160002

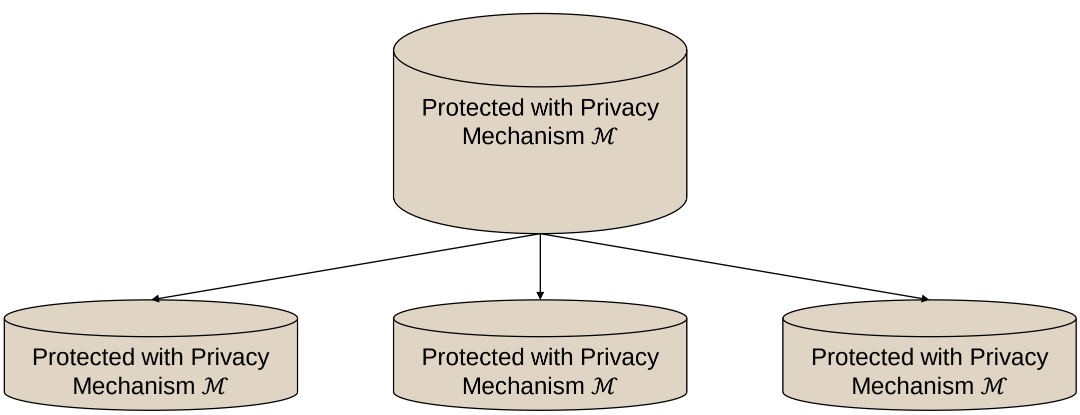
CERIAS

Center for Education and Research in Information Assurance and Security

---

PURDUE UNIVERSITY

Partitioning and Privacy



- When can we treat the databases independently?

2

Definition: Parallel Composition

We say that a sanitization scheme *A* satisfies **parallel composition** if, given disjoint datasets $D_1, \ldots, D_n$, with corresponding outputs $A(D) = A(D_1), \ldots, A(D_n)$ satisfies the privacy guarantee of the original scheme.

- Satisfied by:
  - Differential Privacy *(McSherry SIGMOD'09)*
    - Privacy budget treated independently for each dataset
  - Generalization-based *k*-anonymity, *l*-diversity with local recording
- Not satisfied by
  - Generalization-based anonymization with global recording
  - Differential Privacy *(Dwork, McSherry, Nissim, Smith TCC'06)*: 2ϵ

5

Parallel Composition: Differential Privacy

Dwork, McSherry, Nissim, Smith TCC'06

Let D be partitioned into d disjoint regions, let $f : D^n \to \mathbb{R}^d$ be a function whose output coordinates $f(x)_i$ depend only on those elements in the ith region. We can bound $S(f) \leq 2 \max_i S(f_i)$.

McSherry SIGMOD'09

Let M_i each provide ϵ-differential privacy. Let D_i be arbitrary disjoint subsets of the input domain D. The sequence of $M_i(X \cap D_i)$ provides ϵ-differential privacy.

6

Why the discrepancy?

- Definition of “differ on a single entry”
  - Deletion (TAMC’08) - easy to show ϵ
  - Substitution (TCC’06) - easy to show 2ϵ
  - Modifying values – Is this ϵ or 2ϵ?
- Disjoint datasets (’09 ϵ) vs. Partitioned dataset (’06 2ϵ)
  - We narrow this gap

8

Definition: Partitioned Preprocessing

Choose a random partition $\{d_i\}$ of $|D|$ into positive integers, then partition D into pieces D_i of size d_i uniformly at random. We call $\bigcup_{i=1}^{n} A(D_i)$ a **partitioned preprocessing** dataset.

- Works for parallel composition techniques
  - Including ϵ-DP under substitution
- Potentially stronger against some types of attacks on generalization
  - Minimality
  - deFinetti
- Attack resistance arguments hold for non-parallel decomposable techniques
  - E.g., global recoding (and potential utility benefits)

10

Theorem: Parallel Composition on Random Partitions

- Let D be a dataset, $|D| = n$. Choose a decomposition $\mathbf{n}$ of n and a permutation π on n elements uniformly at random, and partition the dataset D into n pieces $\{D_{\pi,i}\}_{1\leq i\leq j}$. Let $\mathcal{A}_1, \ldots, \mathcal{A}_j$ be differentially private mechanisms with privacy budgets $\epsilon_1, \ldots, \epsilon_j$.
  The mechanism $\mathcal{A} = \left(\mathcal{A}_1\left(D_{\pi,1}\right), \ldots, \mathcal{A}_j\left(D_{\pi,j}\right)\right)$ satisfies ϵ-differential privacy, where $\epsilon = \max_{1\leq i\leq j} \epsilon_i$.

12

Proof Idea

- Partitions determined in advance, independent of data
  - Substituting a tuple affects only one partition
- For partitions without the changed tuple, $D_{\pi,k}$=$D'_{\pi,k}$, so $P\left(\mathcal{A}_k(D_{\pi,k}) \in T_k\right) = P\left(\mathcal{A}_k(D'_{\pi,k}) \in T_k\right)$
- The changed partition j has a difference bounded by ϵ_j
  - This bounds the total difference between $\mathcal{A}(D)$ and $\mathcal{A}(D')$

14

More Differences between Deletion and Substitution

- What is the sensitivity of
  - |*D*| ?
    - Deletion: 1
    - Substitution: 0
  - Average
- Amplification *(Li, Qardaji, Su '12)*
  - Defined under deletion

Is there a difference in the privacy semantics?

15

Partitioned Preprocessing: Potential Utility Benefit

Age	Gender	Zip	Cancer	Age	Gender	Zip	Cancer
40-50	Male	92***	Yes	40-60	Male	925**	No
40-50	Male	92***	No	40-60	Male	925**	No
40-50	Male	92***	No	40-60	Male	925**	Yes
40-50	Male	92***	Yes	40-60	Male	925**	No

- Some benefits of local recoding
  - "Outliers" only force over-generalization in a single partition
- Each partition satisfies global recoding
  - Difficulty identifying which partition an item belongs to provides defense against attacks

21

Partitioned Preprocessing: Example

Semantic Attacks: Determine likely distribution of sensitive values in an equivalence class

- Individual may belong to many equivalence classes
  - Attack gives information on one equivalence class
- Attack increases $Pr(x.S = S_i)$ by only a (weighted) proportion of the increase in probability for that class

k=20	Underlying Partitions	Visible Partitions	Distribution of Partitions	% of Population
Average 25,000 size	20	6 + Suppressed Class	6, 5, 6, 1, 1, 1	.244, .30, .295, .062, .048, .024 Suppress: .016

24

Partitioned Preprocessing: Example

- Original Record:

ZIP	YOB	GEN	VISIT	HOSPITAL	COMP	CAT	Possible Matches
43125	1967	F	2005-08-31	Riverside Methodist	Mosquito Bite	Other	7,916

- Anonymized Versions:

ZIP	YOB	Visit Date	Hospital	Matches
43000 - 43240	1940 - 1979	2004-01-01 - 2005-12-31	Riverside Methodist Hospital	2520
43068 - 43156	1940 - 1979	2004-01-01 - 2005-12-31	Medium & Large Hospitals	3497
43068 - 43156	1900 - 1992	2004-01-01 - 2005-12-31	Riverside Methodist Hospital	1068
43119 - 43156	1940 - 1979	2004-01-01 - 2008-02-31	Large Hospitals	421
43119 - 43156	1900 - 1992	2005-07-01 - 2005-12-31	Medium & Large Hospitals	169
43068 - 43156	1900 - 1992	2004-01-01 - 2005-12-31	Large Hospitals	241

25

PURDUE
UNIVERSITY

Still working…

- Implications of partitioned preprocessing on differential privacy
  - Near-optimal use of privacy budget
    - Use noise from random partitioning to satisfy differential privacy
  - Potential operational value?
  - Amplification of privacy budget through sampling
- Thank You
  - Chris Clifton, clifton@cs.purdue.edu
  - Keith Merrill, merrill2@brandeis.edu
  - Shawn Merrill, smerrill@cs.purdue.edu