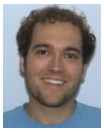
PURDUE UNIVERSITY

Brandeis University

Practical Issues in Anonymity


Chris Clifton, Shawn Merrill (Purdue University)

Keith Merrill (Brandeis University)



CERIAS Center for Education and Research in Information Assurance and Security

Indiana Center for Database Systems™

PURDUE UNIVERSITY

Problems with Anonymization

RE-IDENTIFICATION! Many anonymization schemes keeps being "broken", eventually people find sufficient data to link/re-identify (e.g. *k*-anonymity)

But, there is still a use case: Private use under a data use agreement

- Want to provide protection against accidental (or low resource) re-identification
- Contractual data use agreement to "pull back" data if linking datasets found

PURDUE UNIVERSITY

Problems with Anonymization

Even if we aren't concerned about re-identification

- Anonymization algorithm impacts practical utility more than value of "utility metric" *(Nergiz & Clifton 2007)*
- Choice of (user-defined) Generalization Hierarchy has even greater impact on utility
- Difficulties with global generalization scaling on large datasets
  - Efficiency
  - Utility
  - Outliers

Need to sanitize the data in a way that preserves its use for the recipient.

PURDUE UNIVERSITY

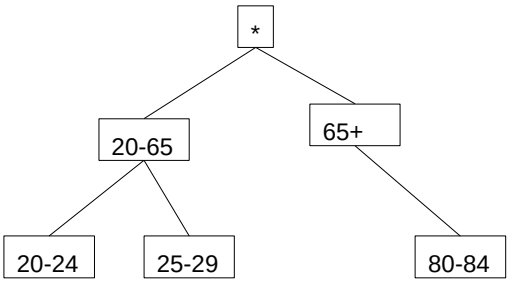
Issue: Poor Utility from Bad Generalization Hierarchies

- Example: Issue with poor generalization hierarchy
  - Million-record anonymization of health data
  - Initial hierarchy (straightforward splits): minimum group size of 48, even with k=2
  - Improved hierarchy (data-depended) showed significantly better granularity
    - Differences each level of k [2, 4, 6, 8, 10, 20]
- Similar issues arise with differential privacy
  - Higher *relative* noise for small groups, even for histogram

Hierarchy Example: How Anonymization Can Go Awry

A college town will have a different age distribution than a retirement community.

- Given this Hierarchy:



- The presence of few (<*k*) 80-84 year olds forces everything to be generalized to “working age” and “retirement age”

User-Defined Hierarchy: Issues

- Relies on a curator’s knowledge of the data
- Too data-driven causes significant information leaks
  - Similar problems to local recoding, clustered anonymization
- Context-insensitivity can lead to issues like semantic similarity among attributes
  - deFinetti Attack
- Can vary greatly based on the attribute [age vs. zip code vs. car type] and specifics [Lafayette, IN vs. Lafayette, LA]

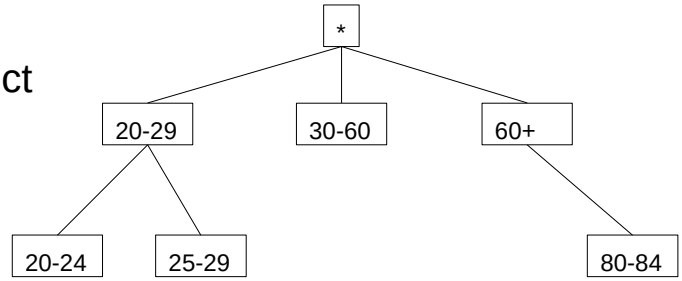
PURDUE UNIVERSITY

Hierarchy Example

A better generalization hierarchy:

- Must be made without direct use of the data
  - Minimality attacks
- Can be done with relative frequency of the values in the population

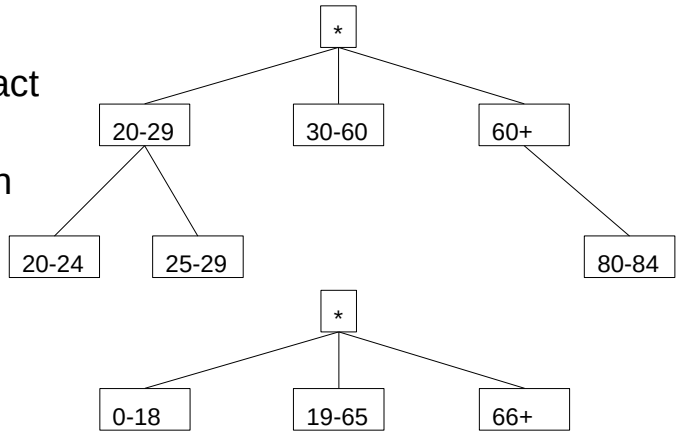
A case for differential privacy?



PURDUE UNIVERSITY

Further Issues

- Release of data at multiple levels
  - Potential interactions impact privacy
  - Inter-level consistency can improve results
- Lattice rather than hierarchy



PURDUE UNIVERSITY

Issue: Scaling

- Challenge: Difficulty in scaling generalization-based anonymization to million record dataset
  - Many techniques fail
  - Few that succeed result in significant record suppression
- Idea: Independently anonymize partitions
  - Potential for different generalizations for different partitions
    - *Will this reduce suppression?*
  - Agnostic to algorithm, privacy definition

13

PURDUE UNIVERSITY

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 $A(D_i)$, $\bigcup_{i=1}^{n} A(D_i)$ 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
  - *t*-closeness

14

PURDUE UNIVERSITY

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 $\cup_{i=1}^{n} A(D_i)$ a **partitioned preprocessing** dataset.

- Works for parallel composition techniques
- 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)

15

PURDUE UNIVERSITY

Partitioned Preprocessing: Potential Utility Benefit

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

Age	Gender	Zip	Cancer
40-60	Male	925**	No
40-60	Male	925**	No
40-60	Male	925**	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

19

PURDUE UNIVERSITY

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

PURDUE UNIVERSITY

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

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
- Implications of hierarchies on a differentially private census
  - Appropriate hierarchies, top-coding
  - Any “non-histogram” analyses?

26