CERIAS

the center for education and research in information assurance and security

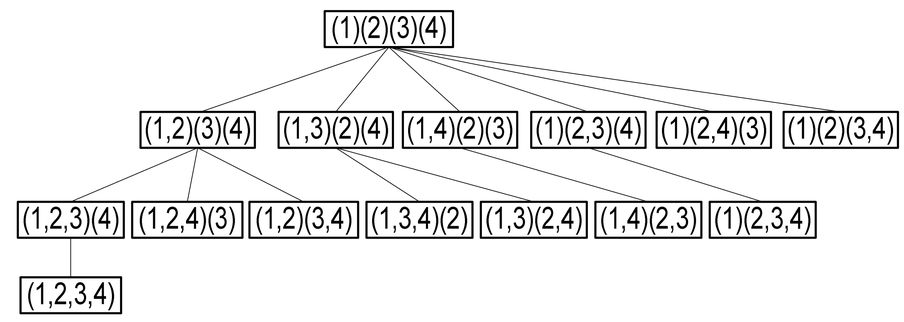
A Taxonomy of Generalization Schemes for Data Anonymization

Tiancheng Li, Ninghui Li

K-Anonymity

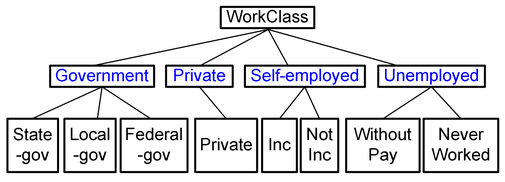
- **Scenario**
  - A data holder wants to publish data to share with researchers
  - Individuals' privacy may be at risk
- **Objective**
  - Maximize data utility while limiting privacy risk
- **Approaches**
  - Data swapping, adding noise, etc
  - *k*-Anonymization
    - Remove identifiers, generalize quasi-identifying attributes
    - Each record is identical to at least *k-1* other records

Enumeration Algorithm



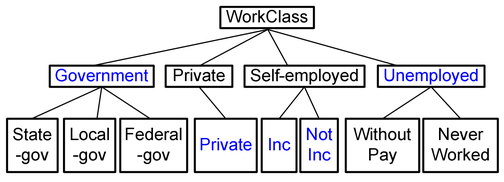
A bottom-up algorithm that enumerates all ways to partition an unordered set exactly once

Generalization Schemes



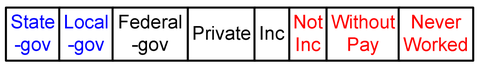
All values are generalized to the same level of the hierarchy

BHS: Basic Hierarchical Scheme



A generalization is a cut across the hierarchy

GHS: Group Hierarchical Scheme

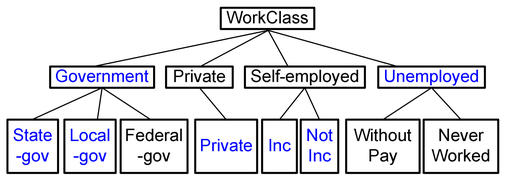


A generalization is an order-preserving partition of the attribute domain

OPS: Ordered Partitioning Scheme

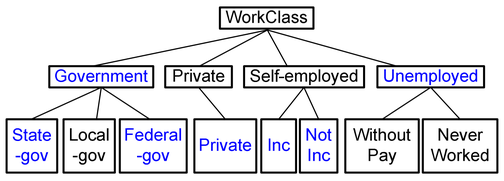
A generalization is a partition of the attribute domain without a total order or a hierarchy

SPS: Set Partitioning Scheme



A generalization is an order-preserving partition with the guidance of a hierarchy

GOPS: Guided Ordered Partitioning Scheme

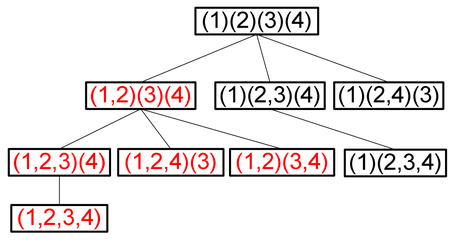


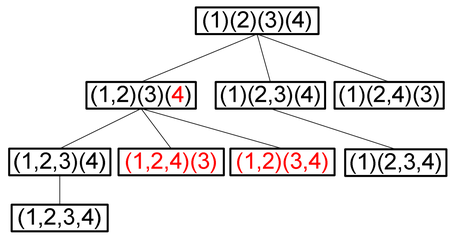
A generalization is a partition with the guidance of a hierarchy

GSPS: Guided Set Partitioning Scheme

Cost-based Pruning

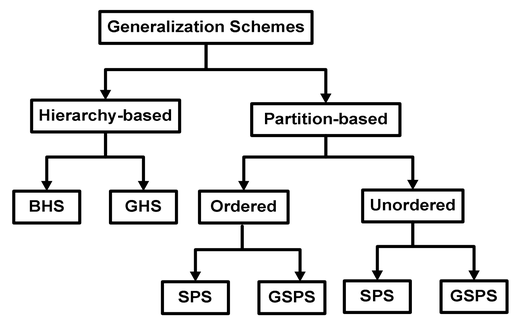
- **Cost Metrics**
  - Discernibility Metric
  - Hierarchical Discernibility *Metric*
- **Cost-based Pruning**
  - Node Pruning
  - Applicator Pruning



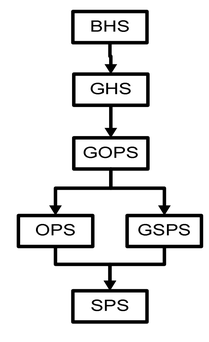


- **Optimizations**
  - Seeding
  - Modified BFS Search Strategy
  - Node Rearrangement
  - Applicator Rearrangement

A Taxonomy



A Taxonomy of Generalization Schemes



Relationship

Evaluation

- **Bottom-up V.S. Top-Down**
  - k Values: Bottom-up approaches work better for small k values
  - Dimensionality: Bottom-up approaches work better when the data contains a small number of attributes.
- **Efficiency V.S. Utility**
  - A more sophisticated generalization scheme produces better "optimal" anonymizations at the cost of performance degradation.

PURDUE UNIVERSITY

CERIAS

Discovery Park e-Enterprise Center