CERIAS

k-Anonymity Privacy Protection: Questions without Answers

M. Ercan Nergiz & Chris Clifton

k-Anonymity

- **What:** A table is *k*-anonymous if every tuple over the identifying attributes occurs at least *k* times
- **Why:** Limits the ability to link released data to other external information
- **Proposed algorithms:** Single Dimensional or Cell based
- **How:** Use of suppressions and generalizations "generally" over a user defined DGH (Domain Generalization Hierarchies)
- **Cost Metrics:** Measure of the quality or information value of the *k*-anonymous dataset
  - LM: Get penalized for data cells that stand for many values
    e.g. Indiana has higher cost than Lafayette
  - AM, CM, DM…

Experiments

Alg Type	Generalization Type	Optimized Against
Clustering Genetic	DGH NDGH	LM AM CM DM

Datasets	Applications
Adult Mushroom	Classification Asc. Rule Mining

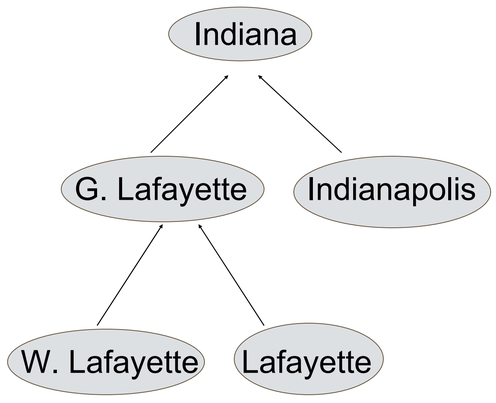
Original Dataset

Age	Sex	Address	Disease
19	M	W. Lafayette	Flu
18	M	Lafayette	Tetanus
23	F	Lafayette	Flu
25	M	Indianapolis	Cancer

2-Anonymized Dataset

Age	Sex	Address	Disease
10-20	M	G. Lafayette	Flu
10-20	M	G. Lafayette	Tetanus
20-30	*	Indiana	Flu
20-30	*	Indiana	Cancer

Domain Generalization Hierarchy (DGH) for Address Attribute



In experiments, we discovered that the existing cost metrics failed to explain the usefulness of the output dataset in terms of some applications' accuracy. The findings bring in the following k-anonymity related questions

What is a good anonymization?

- A good anonymization may not come with good precision
  - CDGH had the best classification accuracy with the highest LM, AM, CM, DM costs
- Reason
  - User input
  - Variety in attribute importance

What is a good metric?

- Optimizing against some metrics is better than optimizing against others
  - CM works better than LM in classification.
  - CM is the worst in explaining rule mining results
  - AM is better in classification for small k than LM
  - LM is better in rule mining than AM
  - No perfect metric

What is a good algorithm?

- Some algorithm types perform better in some applications
  - DGH algorithms are better in classification
  - NDGH algorithms are better in rule mining
  - SDA algorithms can't mine low level rules
- What is a good input?
  - Intuitive DGH inputs help in classification