PRESENTATION OUTLINE: Top Down Specialization on Apache Spark$^{\text{TM}}$

Macarious Abadeer
School of Computer Science
Carleton University
Ottawa, Canada K1S 5B6
macariousabadeer@cmail.carleton.ca

November 10, 2019

1 Why Data Privacy?

- Introduction to Sweeney's Paper
- Incidents involving privacy breaches

2 Important Definitions

- Quasi-Identifiers
- Sensitive Attributes
- Taxonomy Trees
- Anonymization Level

3 k-anonymity theory

- Introduction to k-anonymization
- Variations including l-diversity and t-closeness

4 Dataset Example

- k-anonymized dataset

5 Existing solutions

- Bottom-Up Generalization
- Top-Down Specialization
- Combining Top-Down and Bottom-Up
- Differential Privacy

6 Top-Down Specialization

- Algorithm overview
- Information gain
- Privacy loss
- Scoring anonymization levels

7 Preprocessing

- Removing non-QIDs
- Grouping QIDs together and calculating count

8 Parent-Child Taxonomy Mapping

- Algorithm for building parent-child mapping from taxonomy trees

9 Anonymization process

- Generalize all QIDs to root of anonymization levels
- Calculating best score for anonymization levels

10 Score calculation

- Parent entropy calculation
- Children entropy calculation

11 Determining Top-Scoring Anonymization Level

- Building score maps
- Updating parent-child mapping with top scoring anonymization level
- Calculating k

12 Enhancing Performance

- Introduction to Apache Spark
- Spark partitioning
- Using tail recursion

13 Spark Tuning

- Spark configuration

14 Environment setup

- Setting up spark
- Cores, memory and disk size used

15 Test Dataset

- Original dataset
- Enlargement technique
- Sizes tested

16 Tests

- Different values of k
- Number of rows for each test
- Number of nodes
- Number of partitions

17 Results

- Charts by dataset size, values of k, number of nodes, number of partitions

18 Comparison with Existing Paper

- Side-by-side comparison with existing paper's results

19 Personal Observations

- Comments on the algorithm from working in the data privacy industry

20 Future Work

- Areas for further improvement