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

Macarious Abadeer
School of Computer Science
Carleton University
Ottawa, Canada K1S 5B6
macariousabadeer [at] scs.carleton.ca

March 30, 2020

1 Introduction - Why Data Privacy?

- Incidents involving privacy breaches
- Important definitions: Quasi-Identifiers, Sensitive Attributes, Taxonomy Trees

2 k-anonymity theory

- k-anonymous datasets
- Example

3 Top-Down Specialization

- Algorithm overview
- Scoring best anonymization level

4 Preprocessing

- Removing non-QIDs
- Grouping QIDs and SAs together and calculating count
- Build parent-child mapping from taxonomy tree

5 Step 1. Anonymization process

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

6 Step 2. Score calculation

- Parent entropy calculation
- Children entropy calculation

7 Step 3. Determining Top-Scoring Anonymization Level

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

8 Enhancing Performance

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

9 Test Environment setup

- Spark setup on OpenStack cluster
- Cores, memory and disk size used
- Datasets used

10 Test Results

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

11 Comparison with Existing Paper

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