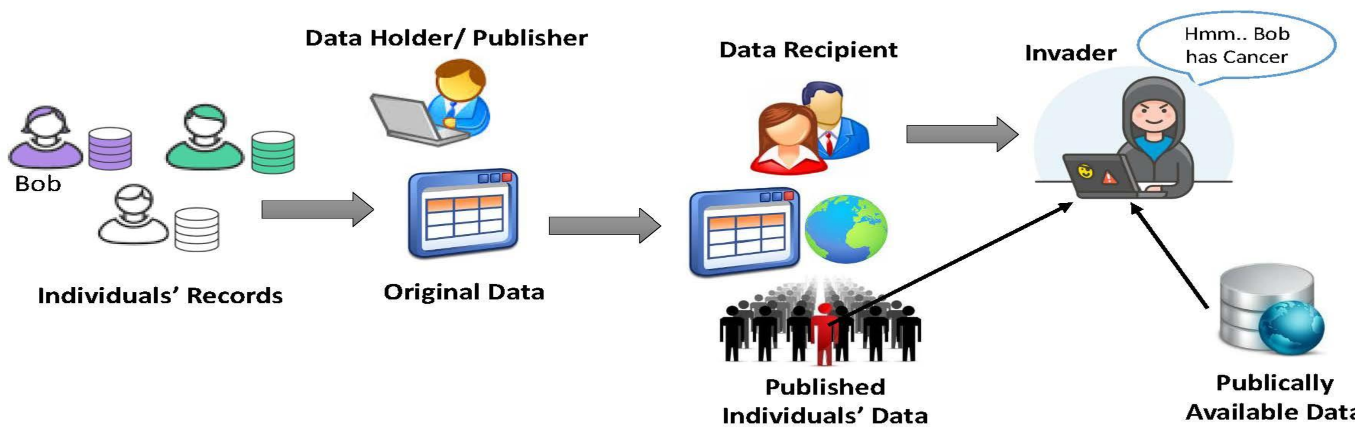
Publishing Differentially Private Datasets via Stable Microaggregation

Masooma Iftikhar, Qing Wang, Yu Lin
{masooma.iftikhar, qing.wang, yu.lin}@anu.edu.au

Australian National University
ANU College of Engineering and Computer Science

Introduction

- Publishing data about individuals poses a privacy threat because data may contain the sensitive information about individuals, e.g., medical history, and publishing them would intrude upon individual privacy.



Challenges and Contributions

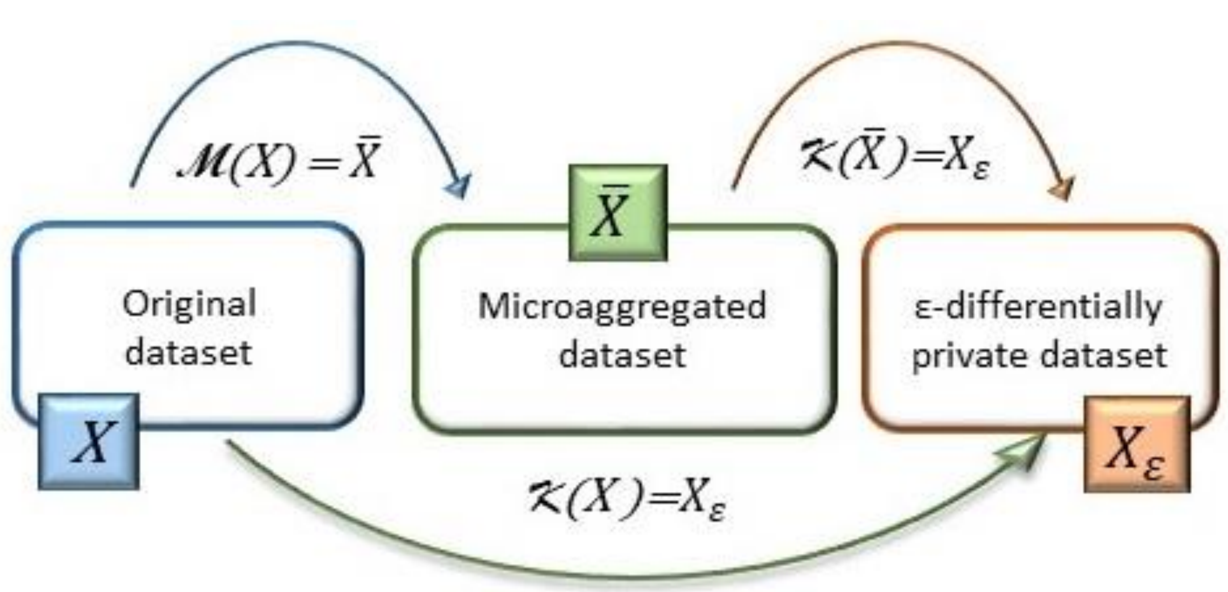
- **Aim**: To generate ε-differentially private datasets by using microaggregation for improving data utility.
- **Key Challenge:** To Enhance utility of published data by providing better within-cluster homogeneity and reducing the amount of noise, in comparison with the state-of-the-art methods.
- **Contributions**:
  a) Developed a microaggregation-based framework for generating ε-differentially private datasets based on a novel notion of *stable microaggregation*;
  b) Design a stable microaggregation algorithm that outperforms the state-of-the-art methods .

Problem Statement

- Two datasets X, $Y \in \mathcal{D}$ are said to be ***neighboring***, denoted as $X \sim Y$, if $|X| = |Y| = n$, but X and Y differ in one record.
- A randomized mechanism $\mathcal{K}: \mathcal{D} \rightarrow \mathcal{D}$ provides ε-differentially private datasets, if for each pair of neighboring datasets $X \sim Y$, and all possible outputs $\mathcal{D}_\varepsilon \subseteq \text{range}(\mathcal{K})$, it holds

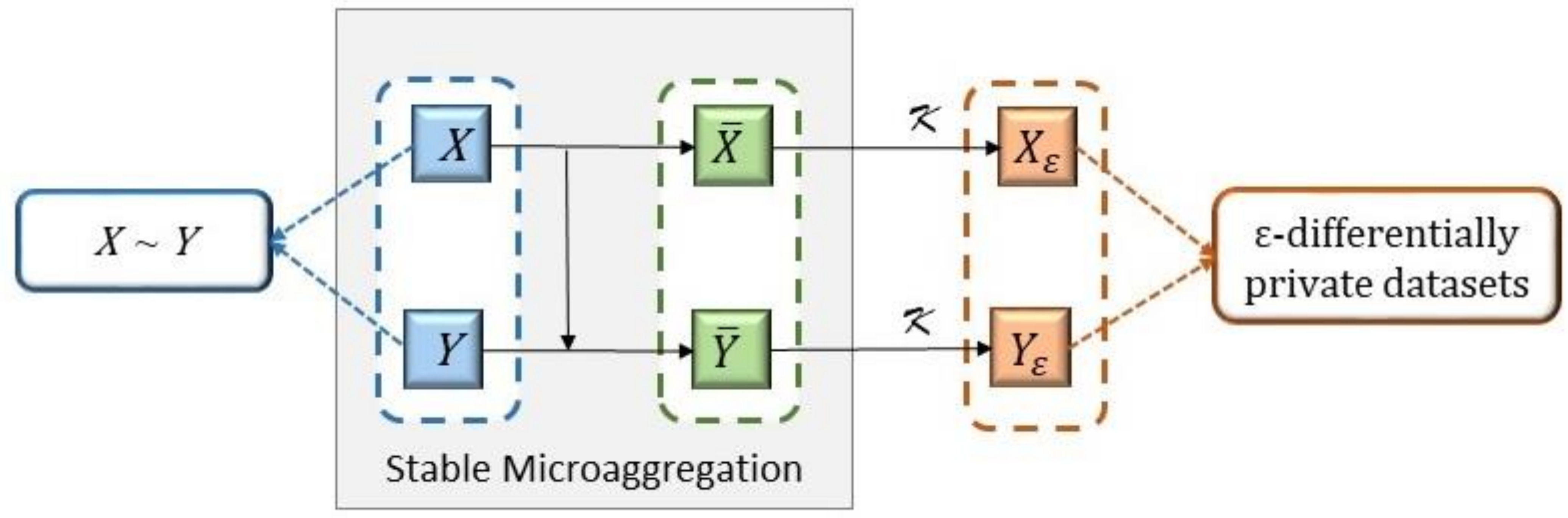
$$\Pr[\mathcal{K}(X) \in \mathcal{D}_\varepsilon] \leq e^{\varepsilon} \times \Pr[\mathcal{K}(Y) \in \mathcal{D}_\varepsilon]$$

- $\varepsilon > 0$ is the differential privacy parameter. Smaller values of ε provide stronger privacy guarantees.



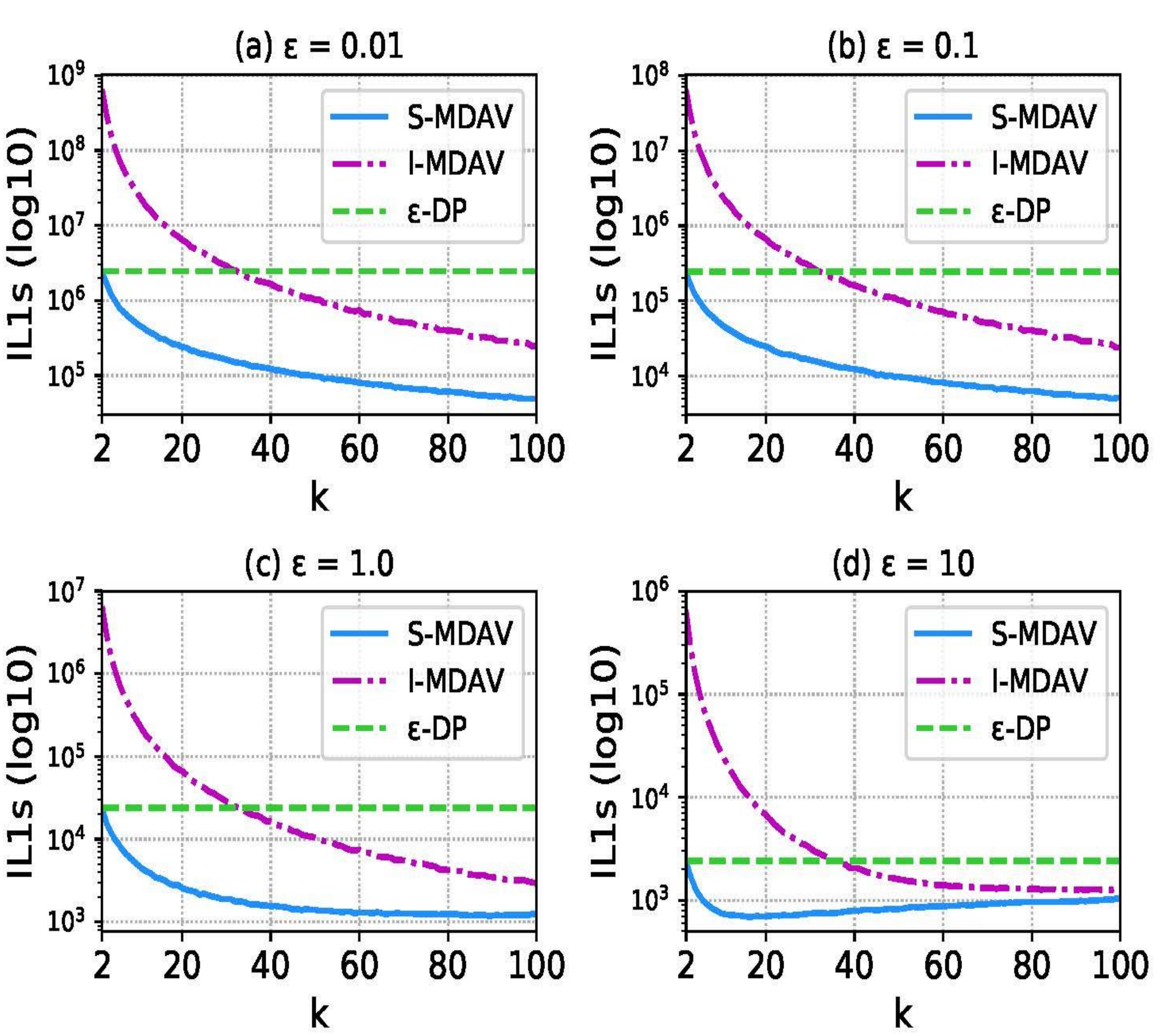
Proposed Framework

- ***Stable microaggregation:*** Let $\mathcal{M}$ be a microaggregation algorithm, $C_x = \{c_1, \ldots, c_n\}$ be the set of clusters that results from running $\mathcal{M}$ on X , and $C_Y = \{c'_1, \ldots, c'_n\}$ be the set of clusters that results from running $\mathcal{M}$ on Y, such that each cluster in C_x and C_Y has at least k records. $\mathcal{M}$ is ***stable*** if, for any $X \sim Y$, there is a bijection between C_x and C_Y such that at most two pairs of corresponding clusters in C_x and C_Y differ in a single record.
- By approximating a query f to $f \circ \mathcal{M}$ via stable microaggregation, the utility of ε-differentially private datasets is enhanced due to significant reduction in sensitivity as compared to the state-of-the-art methods. The addition of noise can always be reduced in comparison with directly applying $\mathcal{K}$ over X , regardless of the size of a dataset, when $k \geq 2$.

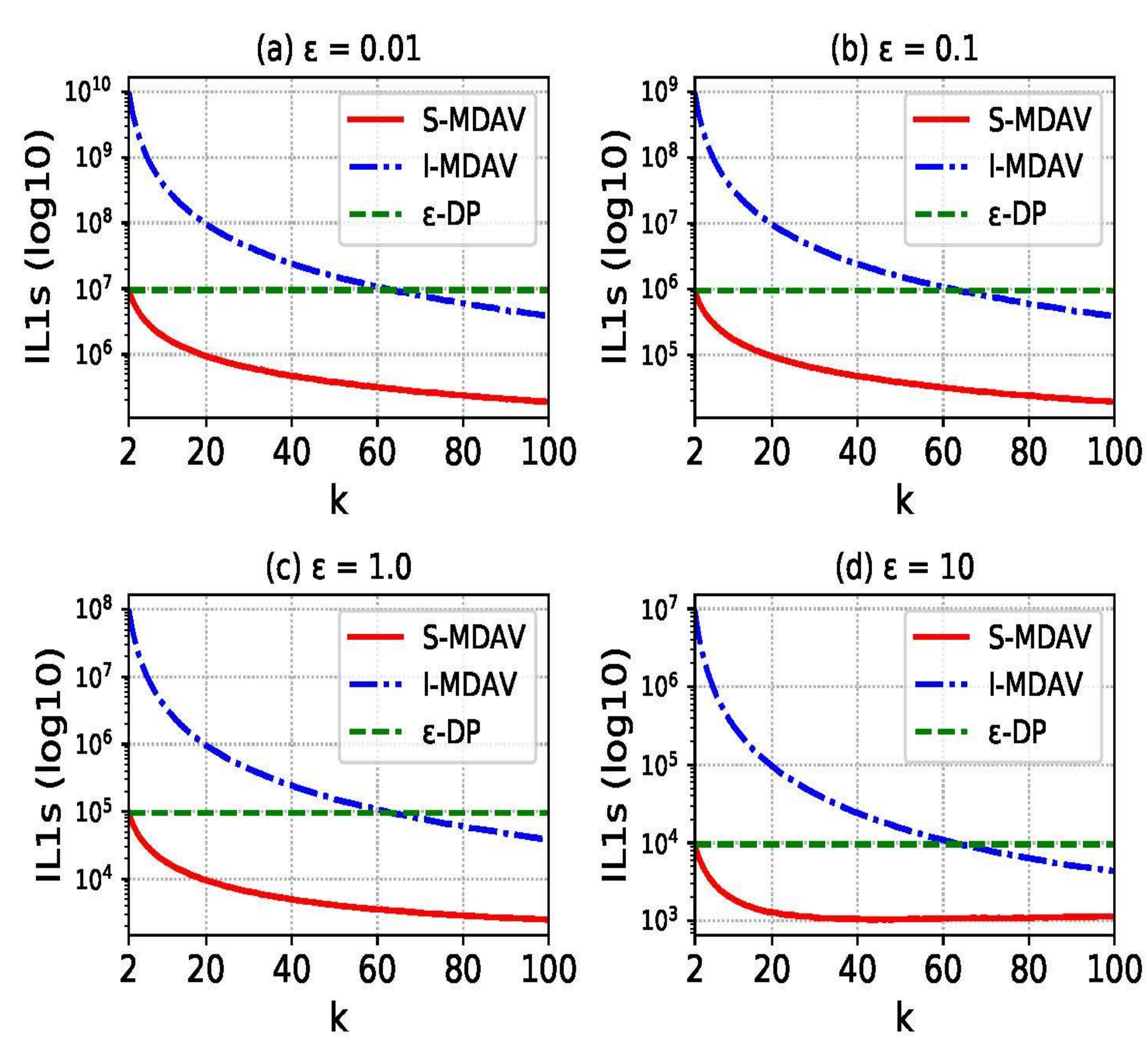


Experimental Evaluation

- Our proposed stable microaggregation algorithm *S-MDAV* outperforms the state-of-the-art methods: *I-MDAV* and *ε-DP,* for different values of k and ε in two real word datasets: *CENSUS* and *EIA*.



CENSUS



EIA

Conclusion

- The proposed framework outperforms the state-of-the-art methods by providing better within-cluster homogeneity and also reducing noise added into differentially private datasets significantly, regardless of the size of a dataset.

Research School of Computer Science, College of Engineering and Computer Science,
The Australian National University, Canberra ACT 0200, Australia