Privacy Preserving Outlier Detection

PI: Jaideep Vaidya
Rutgers University

Key Collaborators: N. Adam, P. Papakonstantinou, B. Shafiq
Ph.D. Students: Hafiz Asif, Tanay Talukdar

This project aims to develop a suite of privacy--preserving tools and techniques that enable outlier detection across different data ownership models, over a variety of multi-modal datasets, while supporting differing tradeoffs of privacy, efficiency, and utility.

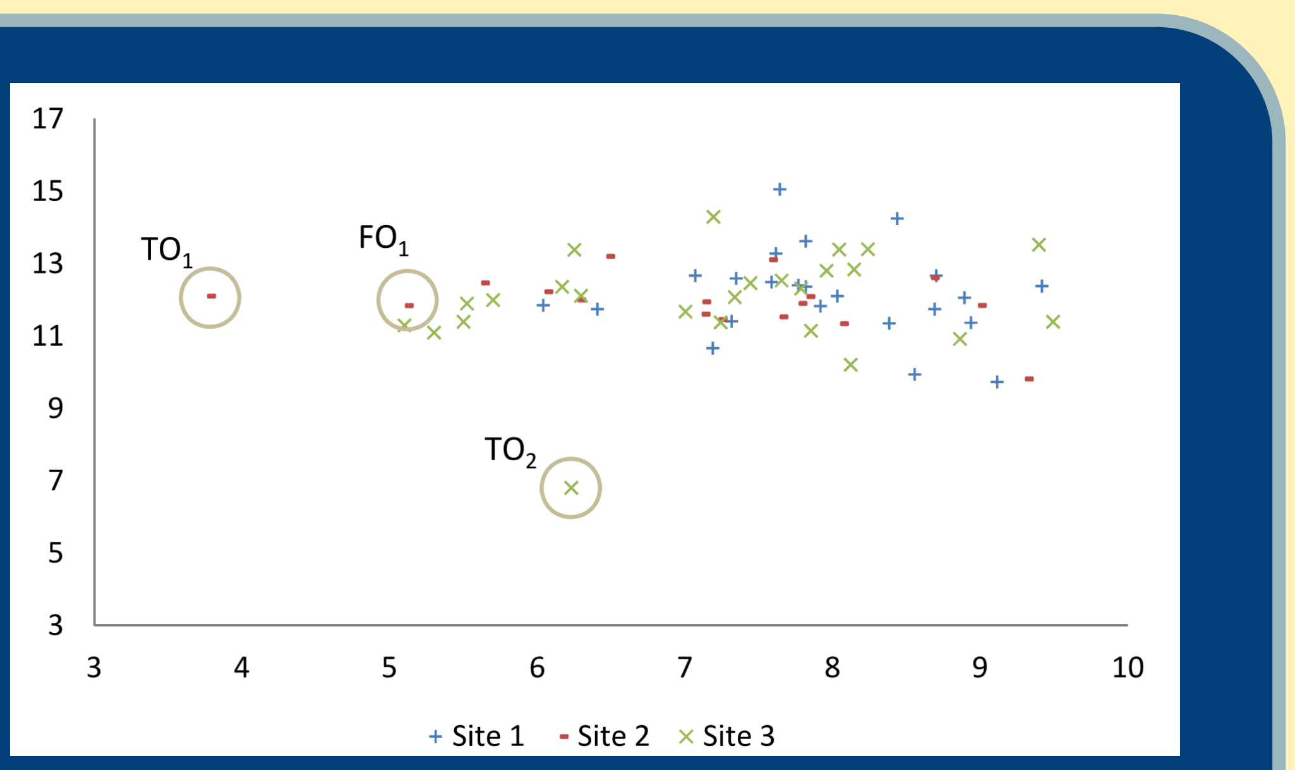
Why is local computation insufficient?

- Either miss true outliers or find false outliers

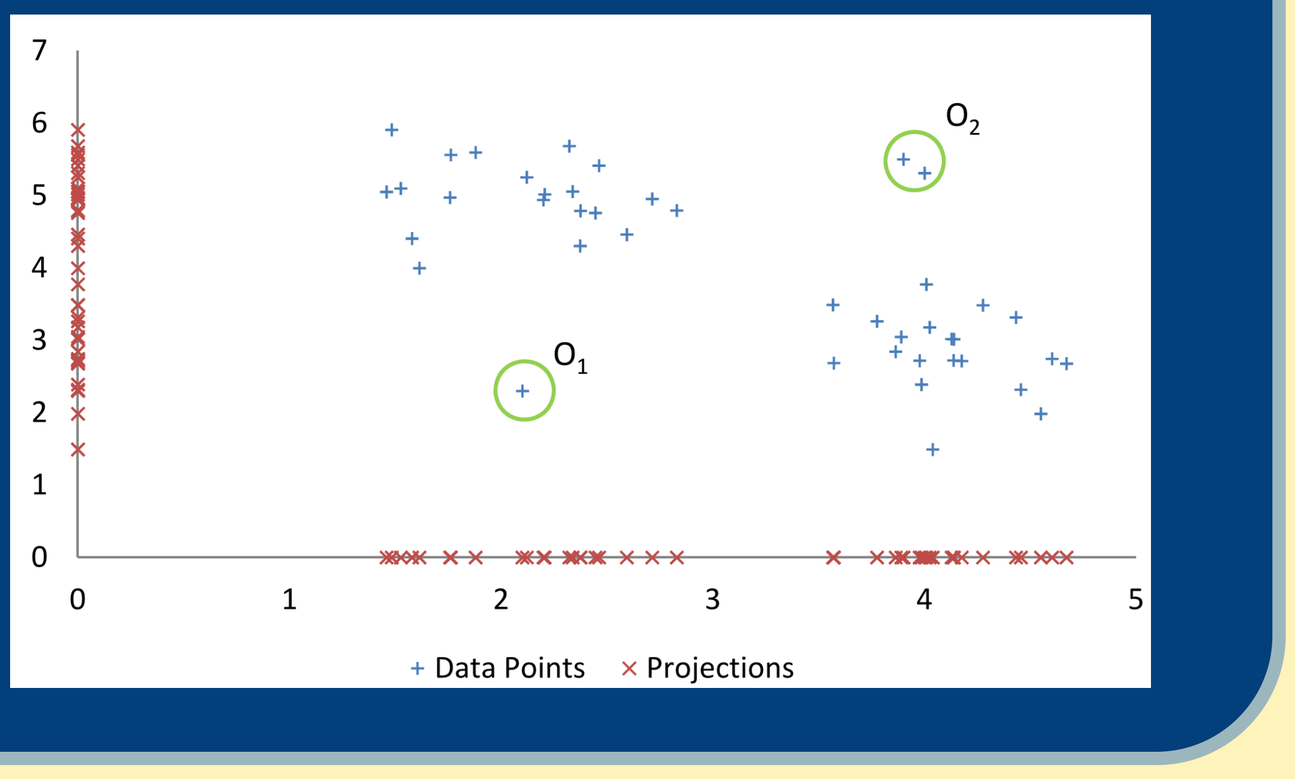
Data Ownership Models

- Horizontally Partitioned Data
- Vertically Partitioned Data
- Centralized Data Warehouse
- Outsourced Database Model

Horizontally Partitioned Data



Vertically Partitioned Data



Approach

Defining Private Outlier Detection

- Privacy of process
  - Secure multiparty computation
- Privacy of results
  - Differential privacy

Defining Private Outlier Detection

- First develop solution for centralized data model
- Develop solutions for distributed model with semi-honest adversaries
- Extend to more powerful adversaries

E.g., Collaborative Differentially Private Outlier detection for Categorical Data

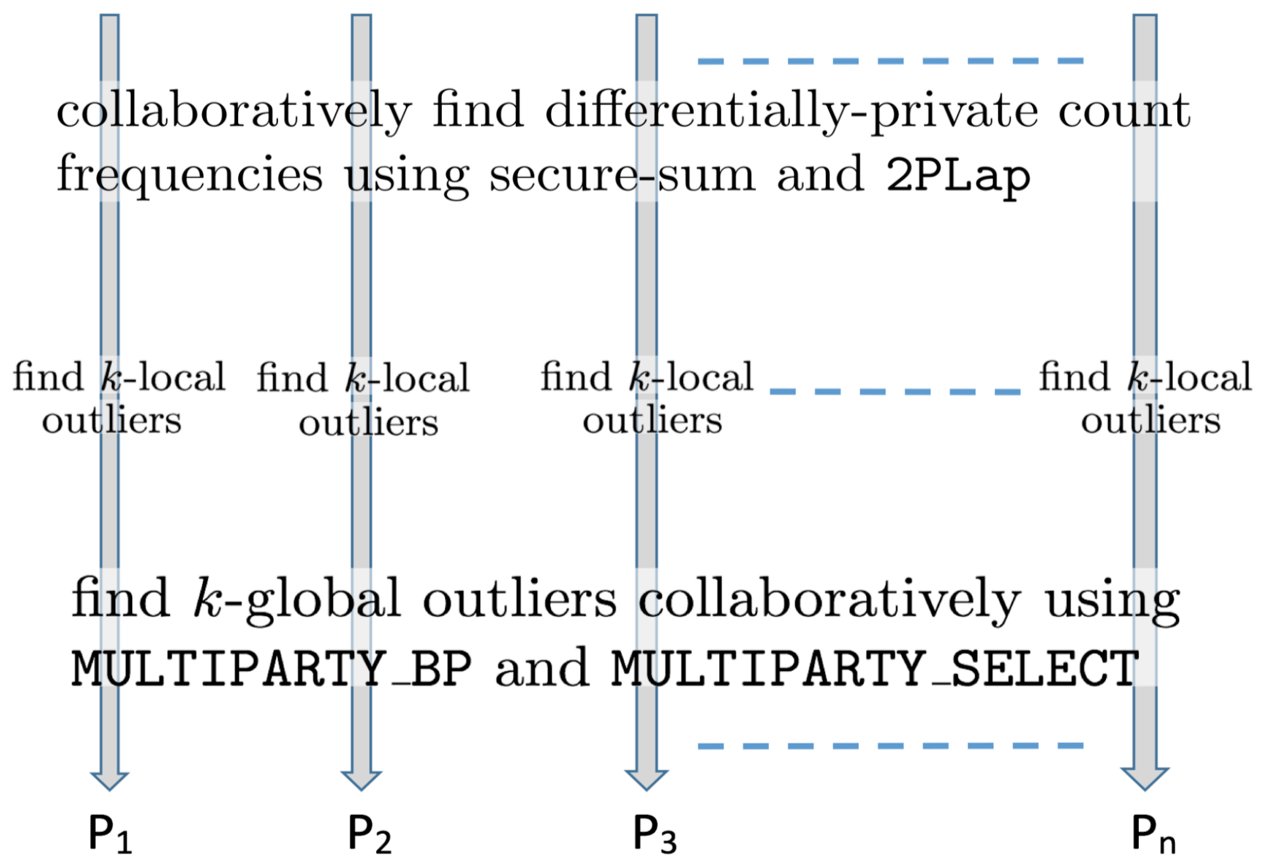
Based on the notion of Attribute Value Frequency Score

Attribute Count Frequency (ACF): Number of times an attribute value appears in the data

Attribute Value Frequency Score: Sum of ACF of all attribute values in a record

Outliers: Records with the smallest AVF score

Protocol Overview



collaboratively find differentially-private count frequencies using secure-sum and 2PLap

find k-local outliers
find k-local outliers
find k-local outliers
find k-local outliers

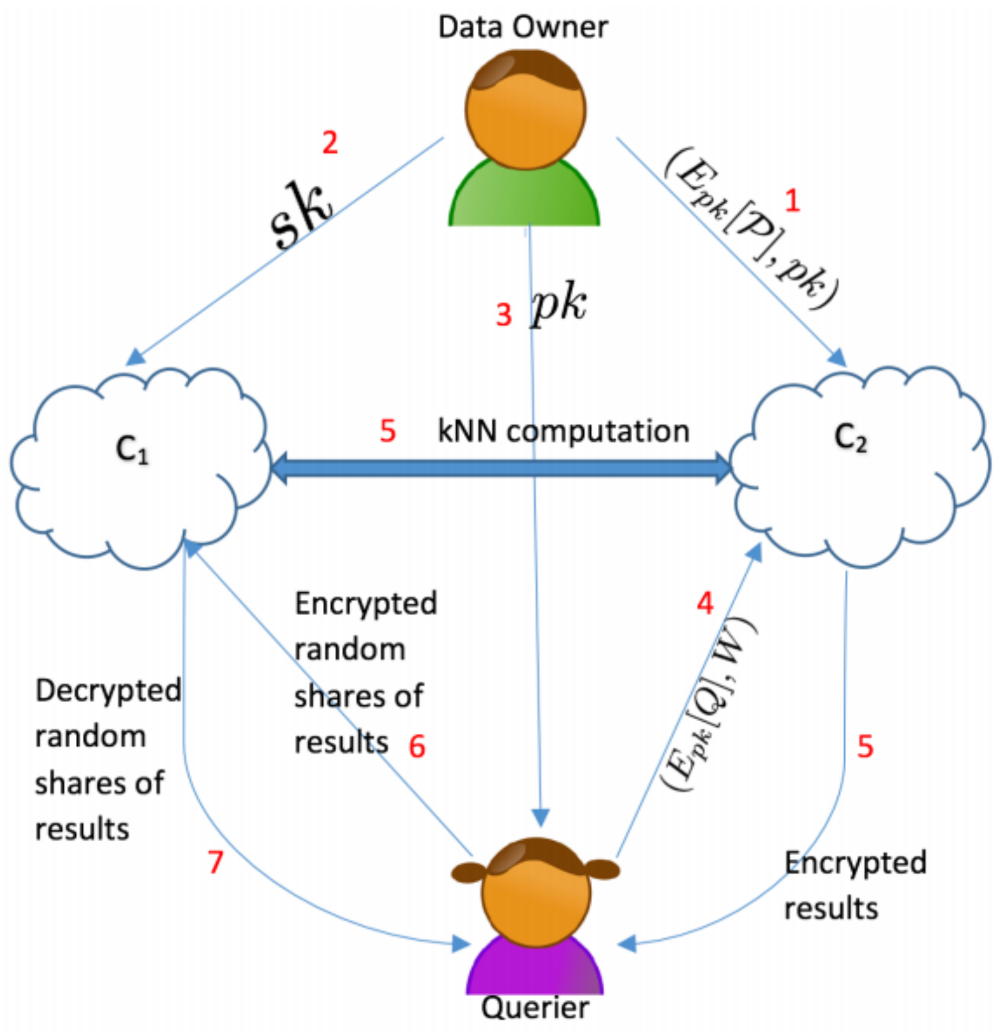
find k-global outliers collaboratively using MULTIPARTY_BP and MULTIPARTY_SELECT

E.g., Secure and Efficient k-NN queries

Propose a notion of semantic awareness for distance metrics, allowing hierarchical distance computation

Develop a novel two-party k-NN computation protocol that is based on record splitting

Protocol can be extended to multiple parties and to outsourcing environment


Data Owner
2
sk
1
$(E_{pk}[\mathcal{P}], pk)$
3
pk
C_1
5
kNN computation
C_2
Encrypted random shares of results
6
4
$(E_{pk}[Q], W)$
Decrypted random shares of results
7
5
Encrypted results
Querier

Interested in meeting the PIs? Attach post-it note below!

The 3rd NSF Secure and Trustworthy Cyberspace Principal Investigator Meeting
January 9-11, 2017
Arlington, Virginia

National Science Foundation
WHERE DISCOVERIES BEGIN

RUTGERS
UNIVERSITY