CERIAS

Murat Kantarcioglu
Chris Clifton

Privacy-preserving Mining of Horizontally Partitioned Data

Why do we need this?

- Data often partitioned among parties that are restricted from sharing data
  - Medical records kept by different hospitals
- Useful Data Mining can be done on the joined data
  - Learning a classifier to predict the type of the disease
- Can we do the useful data mining without joining data and giving "provable" security assurances?

YES!!!

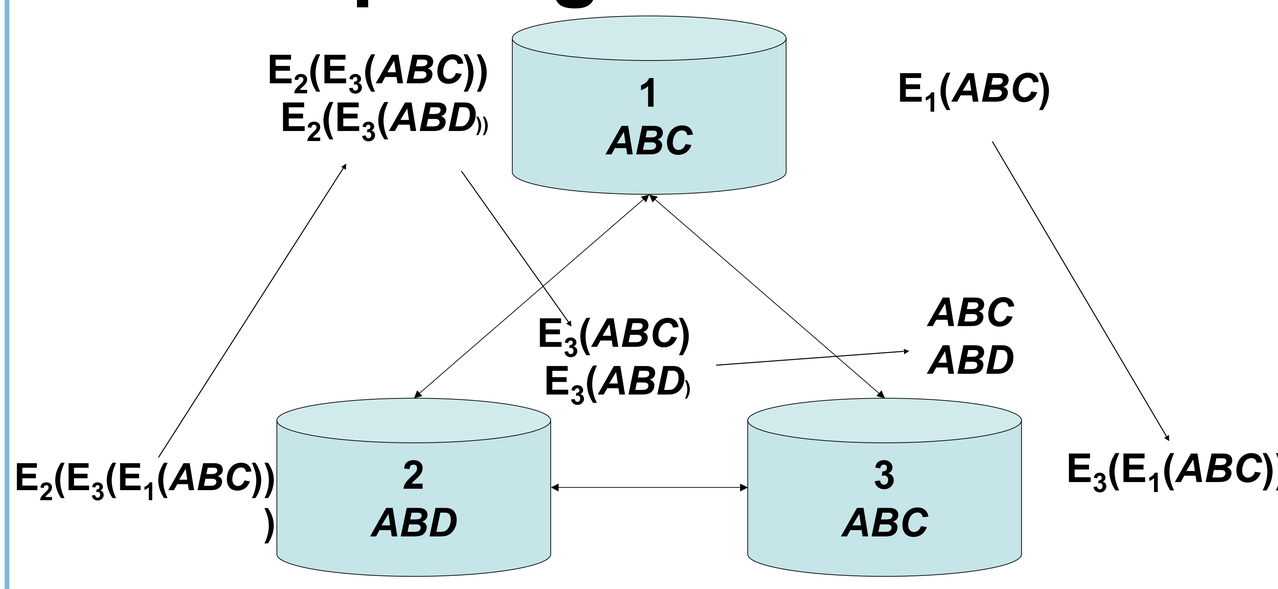
Example: Privacy-preserving Distributed Association Rule Mining

- Association rules a common data mining task
  - Find *A*, *B*, *C* such that $AB \Rightarrow C$ holds frequently
  - Fast algorithms for centralized and distributed computation
    - Basic idea: For $AB \Rightarrow C$ to be frequent, *AB*, *AC*, and *BC* must all be frequent
    - Requires sharing data
- Distributed Association Rule Mining: Easy without sharing the individual data *[Cheung+'96]*
  - *(Exchanging support counts is enough)*
- What if we do not want to reveal which rule is supported at which site, the support count of each rule, or database sizes?
  - Hospitals want to participate in a medical study
    - But rules only occurring at one hospital may be a result of bad practices
  - *Is the potential public relations / liability cost worth it?*

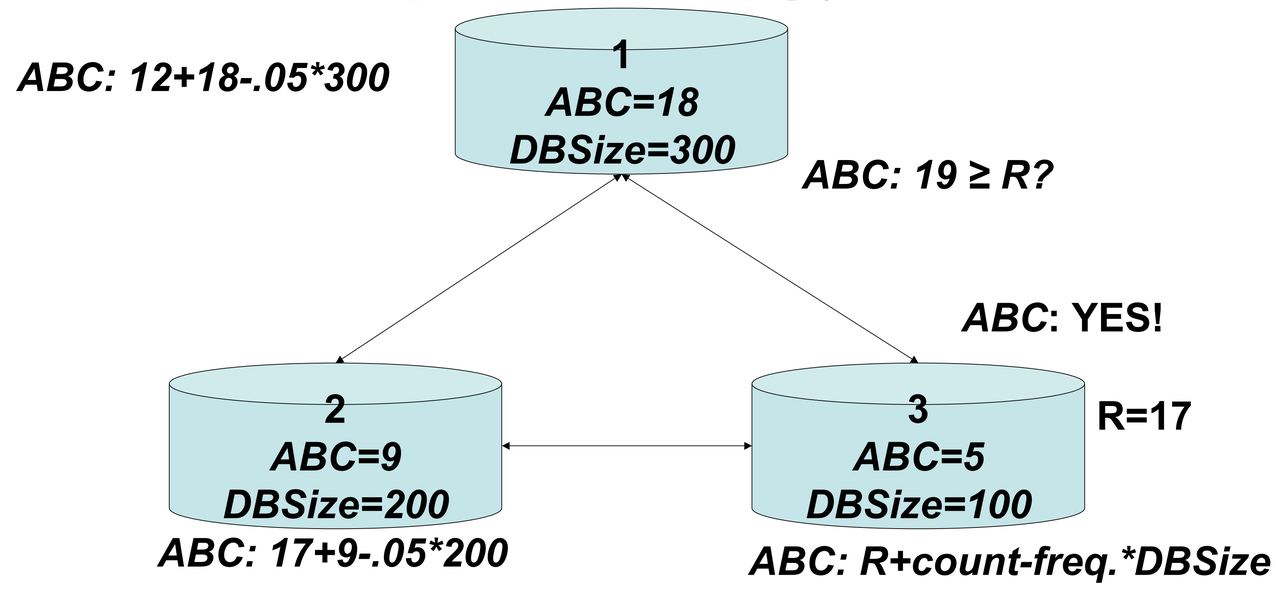
Overview of the Method

- Find the union of the locally large candidate itemsets securely
- After the local pruning, privately compute globally supported large itemsets
- Securely check the confidence of the potential rules

Computing Candidate Sets



Computing Frequent: Is *ABC* ≥ 5%?



Other Examples:

- Privacy-Preserving Distributed EM Clustering: How to find cluster centers of distributed data while revealing as little as possible *(Lin & Clifton 2003)*
- Privacy-Preserving Distributed Top-*K* Queries: Given the query instance, answer top-*k* queries on the union of the distributed data without revealing distances and site information *(Kantarcioglu & Clifton 2003)*

Effect of Data Mining Results:

- Assuring Privacy when big Brother is Watching: How to use the models privately?(Kantarcioglu & Clifton,2003)
- When Do Data Mining Results Violate Privacy?: What is the privacy loss when data mining results are revealed? (Kantarcioglu, Jin, Clifton 2004)

PURDUE UNIVERSITY

e-Enterprise Center