

Progress on the PORTIA Project in Privacy-Preserving Data Mining

Rebecca Wright

*Computer Science Department
Stevens Institute of Technology
www.cs.stevens.edu/~rwright*

2nd Japan/US workshop on CIIP

26 June, 2005

Overview

- Introduction to PORTIA
- Privacy-preserving data mining
- Beyond privacy-preserving data mining
- Implementation and experimentation
- Lessons learned, conclusions

The Data Revolution

- The current data revolution is fueled by advances in technology, as well as the perceived, actual, and potential usefulness of the data.
- Most electronic and physical activities leave some kind of data trail. These trails can provide useful information to various parties.
- However, there are also concerns about appropriate handling and use of sensitive information.
- Privacy-preserving methods of data handling seek to provide sufficient privacy as well as sufficient utility.

The PORTIA Project

Privacy, Obligations, and Rights in Technologies of Information Assessment

A five-year multidisciplinary project focusing on the technical challenges of handling sensitive data and the policy and legal issues facing data subjects, data owners, and data users.

Funded for US\$12.5 million by the National Science Foundation as a Large ITR (Information Technology Research) grant, Oct 2003 - Sept 2008.

PORTIA Personnel

- Academic investigators:
 - **Dan Boneh**, Hector Garcia-Molina, John Mitchell, Rajeev Motwani, *Stanford*
 - **Joan Feigenbaum**, Ravi Kannan, Avi Silberschatz, *Yale*
 - **Stephanie Forrest**, *University of New Mexico*
 - **Helen Nissenbaum**, *NYU*
 - **Rebecca Wright**, *Stevens Institute of Technology*

PORTIA Personnel

- Research partners
 - Jack Balkin, *Yale Law School*
 - Greg Crabb, *Secret Service*
 - Cynthia Dwork, Brian LaMacchia, *Microsoft*
 - Sam Hawala, *US Census Bureau*
 - Kevin McCurley, *IBM Research*
 - Perry Miller, *Yale Center for Medical Informatics*
 - John Morris, *Center for Democracy and Technology*
 - Benny Pinkas, *HP Labs*
 - Marc Rotenberg, *Electronic Privacy Information Center*
 - Alejandro Schaffer, *DHHS/National Institutes of Health*
 - Dan Schutzer, *Citigroup*

PORTIA Team at Stevens

- Prof. Rebecca Wright
- Postdocs: Sheng Zhong
- PhD Students:
 - Geetha Jagannathan
 - Zhiqiang Yang
 - Michael de Mare
 - Onur Kardes
 - Mike Engling
- US/Japan collaboration with Eiji Okamoto and his group, University of Tsukuba

Additional support is provided by the WiNSeC Center and the Technogenesis program at Stevens.

PORTIA Goals

- Produce a next generation of technology for handling sensitive information that is qualitatively better than the current generation's.
- Enable end-to-end handling of sensitive information over the course of its lifetime.
- Formulate an effective conceptual framework for policy making and philosophical inquiry into the rights and responsibilities of data subjects, data owners, and data users.

Major Technical Themes

- privacy-preserving data mining
- identity theft and identity privacy
- database policy enforcement tools
- managing sensitive information in P2P systems
- using trusted platforms to provide trusted privacy-preserving services
- contextual integrity

Privacy-Preserving Data Mining

Allow multiple data holders to collaborate to compute important information while protecting the privacy of other information.

- Security-related information
- Public health information
- Marketing information
- etc.

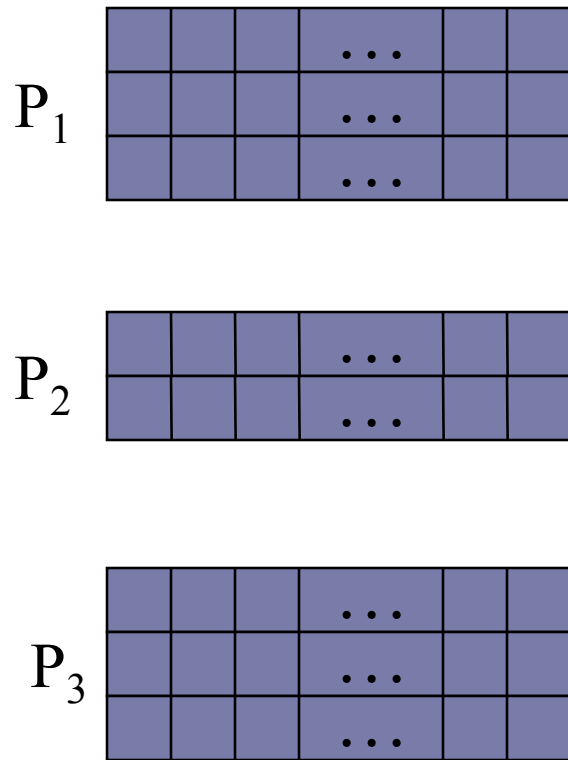
Technological tools include cryptography, data perturbation and sanitization, access control, inference control, trusted platforms.

Advantages of privacy protection

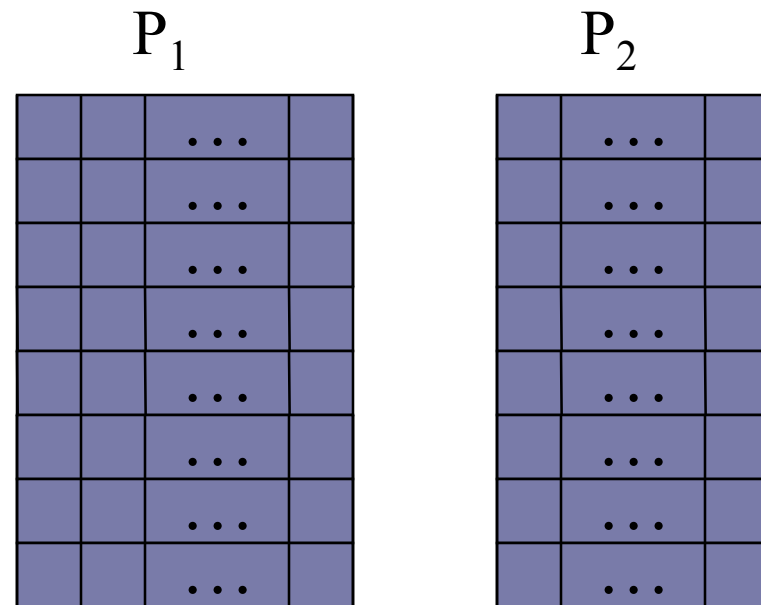
- protection of personal information
- protection of proprietary or sensitive information
- enables collaboration between different data owners (because they may be more willing or able to collaborate if they need not reveal their information)
- compliance with legislative policies

Models for Distributed Data Mining, I

- Horizontally Partitioned

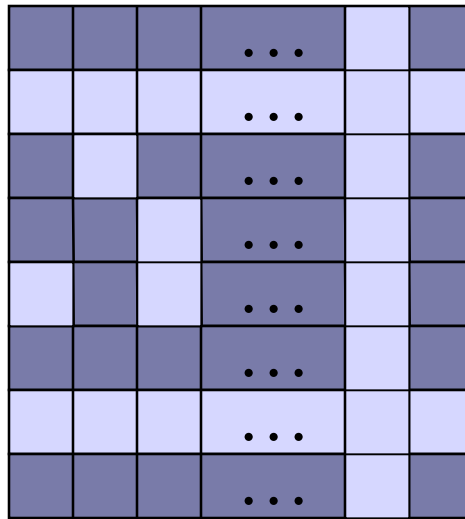


- Vertically Partitioned



Models for Distributed Data Mining, II

- Arbitrarily Partitioned



P_1 ■ P_2 □

Models for Distributed Data Mining, III

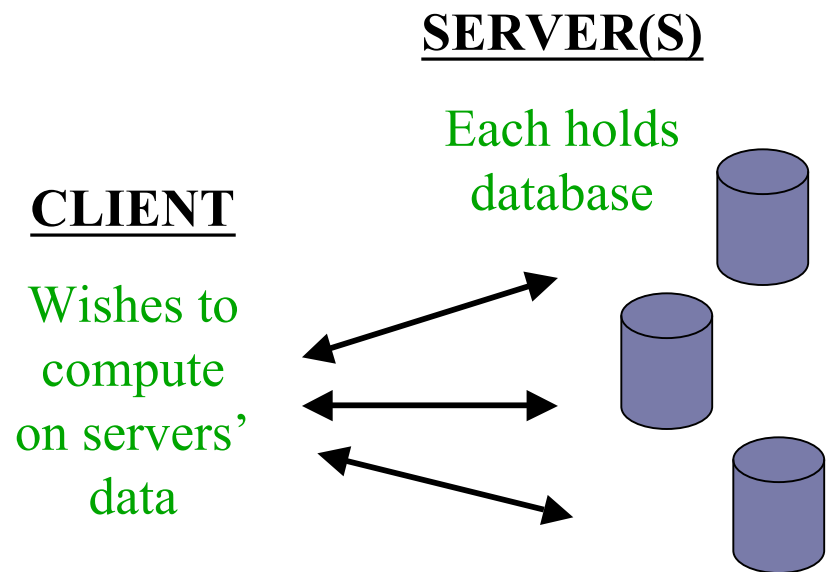
- Fully Distributed



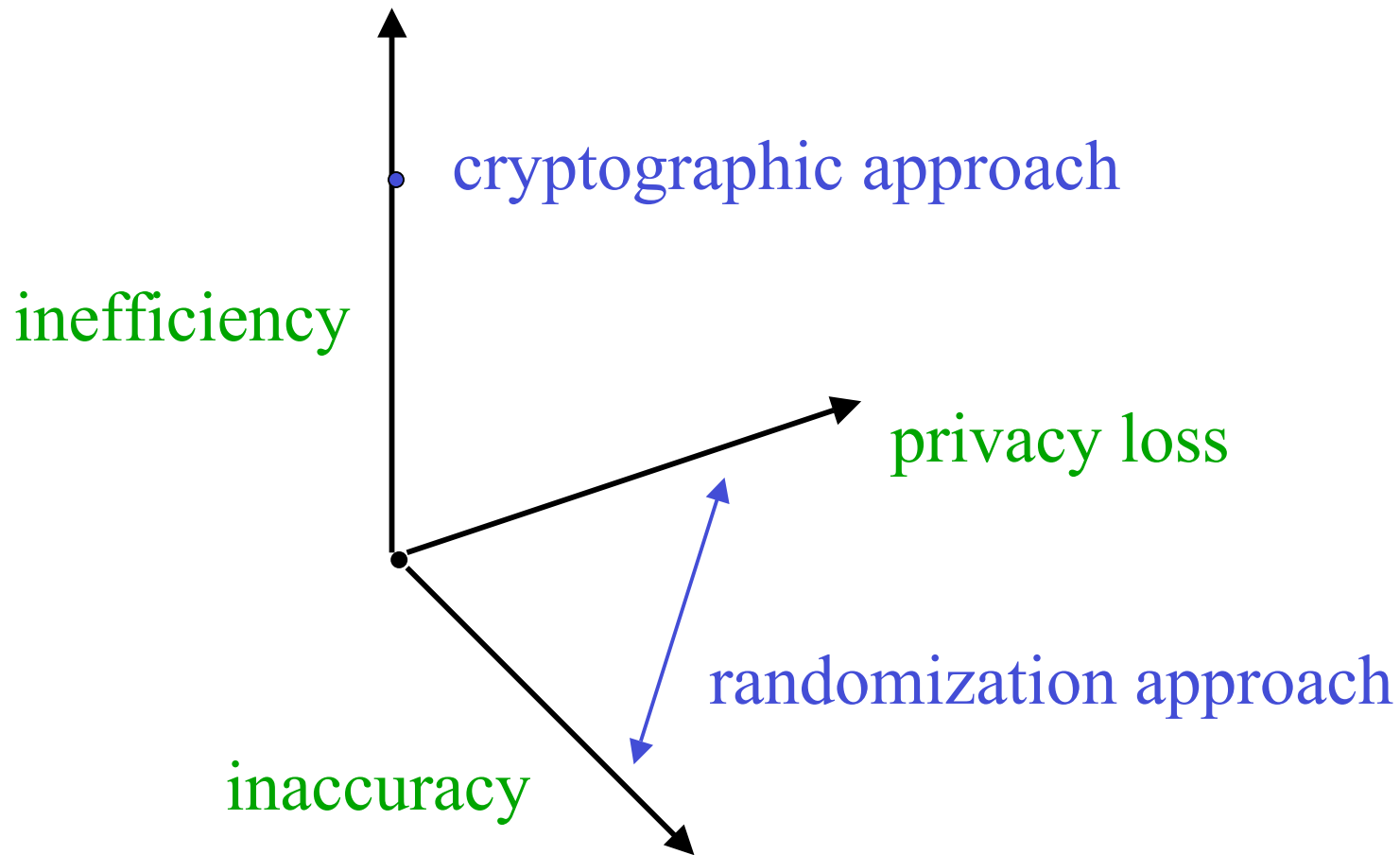
⋮



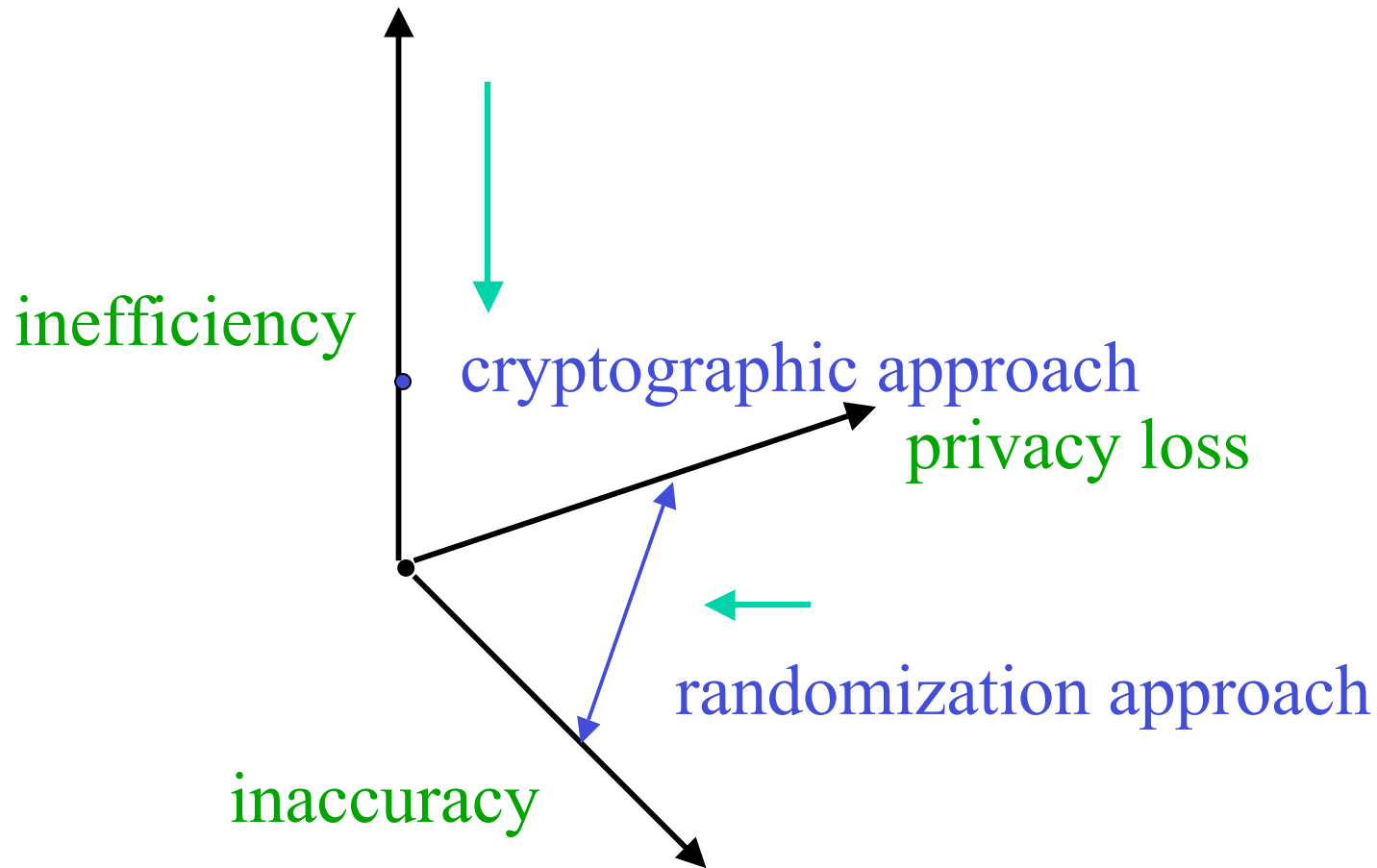
- Client/Server(s)



Cryptography vs. Randomization

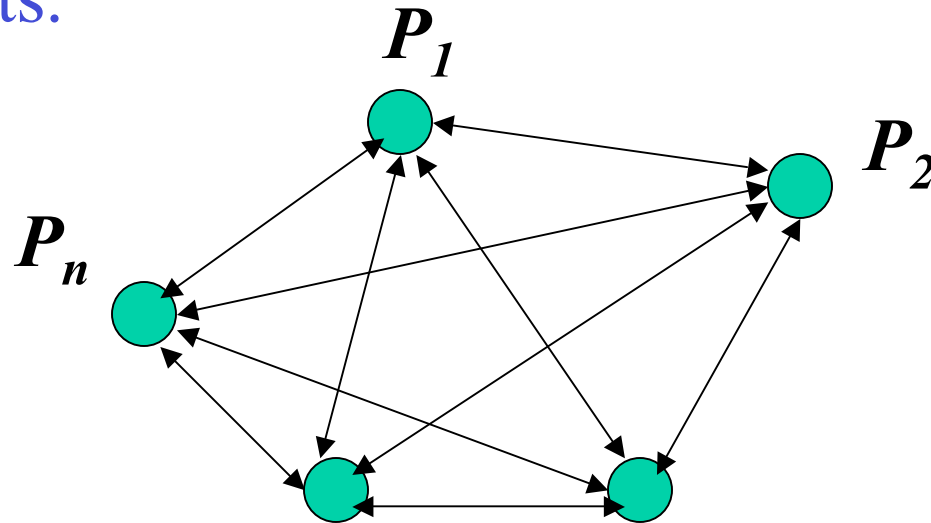


Cryptography vs. Randomization



Secure Multiparty Computation

- Allows n players to privately compute a function f of their inputs.



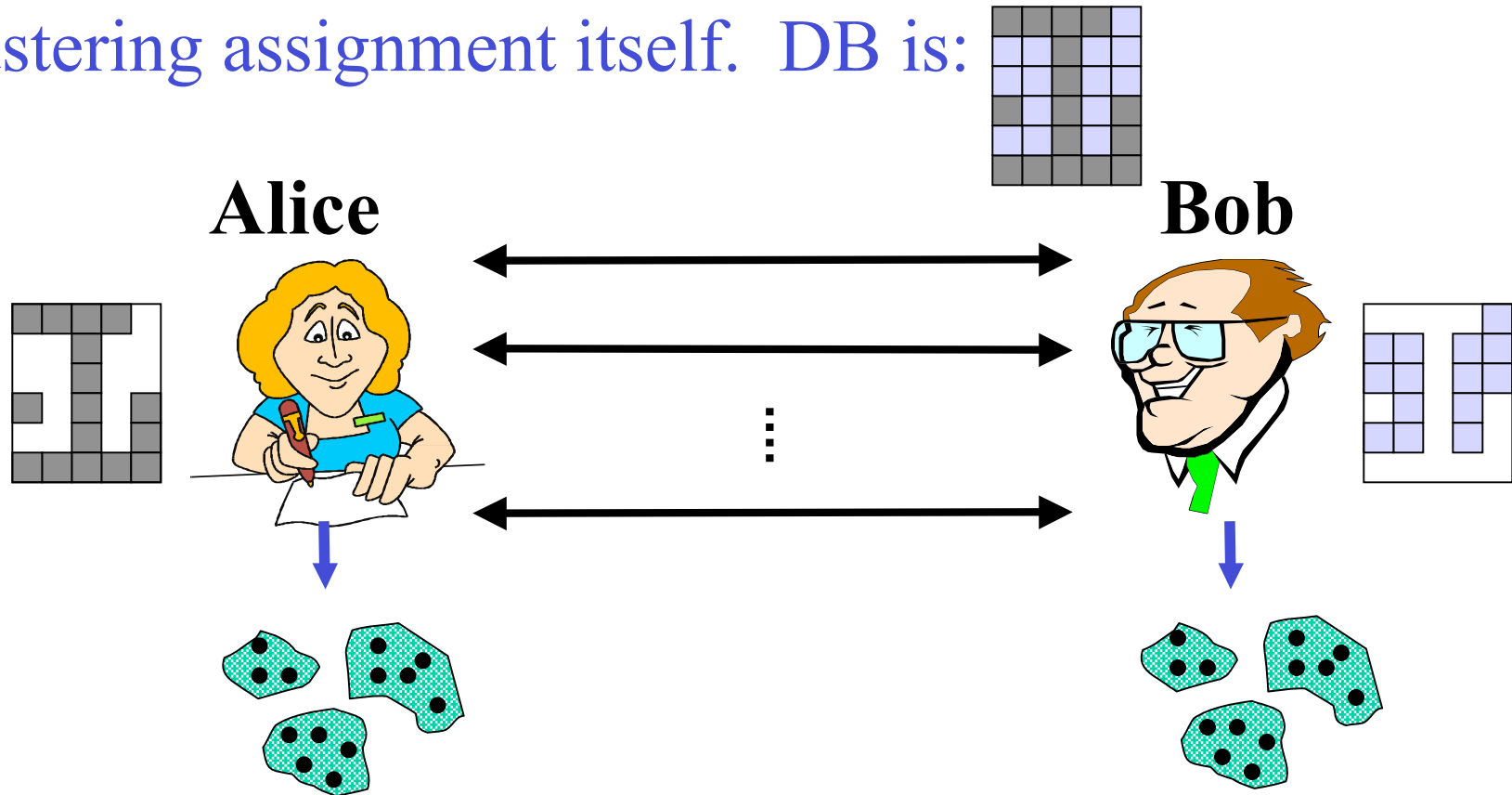
- Overhead is polynomial in size of inputs and complexity of f [Yao86, GMW87, BGW88, CCD88, ...]
- In theory, can solve any private distributed data mining problem. In practice, not efficient for large data.

PORTIA PPDM work

- [WY04,YW05]: privacy-preserving construction of Bayesian networks from vertically partitioned data.
- [YZW05]: frequency mining and classification in fully distributed model (naïve Bayes classification, decision trees, and association rule mining).
- [JW05]: privacy-preserving k -means clustering for arbitrarily partitioned data.
- [AST05]: privacy-preserving computation of multidimensional aggregates on vertically or horizontally partitioned data using randomization.

Privacy-Preserving Clustering [JW05]

Goal: Cooperatively learn k -means clustering on database arbitrarily partitioned between Alice and Bob, ideally without either party learning anything except the clustering assignment itself. DB is:



k -means Clustering [Llo82]

Input: Database D , integer k .

Output: Assignment of database objects to k clusters.

- Randomly select k objects from D as initial cluster centers.
 - Iteratively try to improve clusters:
 - For each object d_i , determine the closest cluster center and assign d_i to that cluster.
 - Recompute the new cluster centers.
- until the change is sufficiently small.

Privacy-Preserving Clustering

Input: Database D , integer k .

Output: Assignment of database objects to k clusters.

- Randomly select k objects from D as initial cluster centers. Alice and Bob share these centers.
 - Iteratively try to improve clusters:
 - For each object d_i , determine the closest cluster center and assign d_i to that cluster.
 - Recompute shares of the new cluster centers.
- until the change is sufficiently small.

Computing Closest Cluster

- For an object d , compute distance to each shared cluster center:
 - Alice owns some attributes and Bob owns some attributes.
 - Distance can be written as a quadratic function of these attributes and Alice and Bob's shares of the cluster center.
 - Can be computed as shares using local computation and secure scalar products.
- Use Yao's secure 2-party computation on the k shared distances to determine which is minimum.

Overall Performance

| | |
|--|--|
| <i>k</i> number of clusters <i>c</i> bits for encrypted attribute | <i>m</i> number of attributes <i>s</i> number of iterations |
|--|--|

- **Computation:** $O(kmns)$ encryptions and multiplications for each party.
- **Communication:** $O(ckmns)$ bits
- In vertically partitioned case, similar to two-party restriction of [VC03].

Beyond Privacy-Preserving Data Mining

Enforce policies about what kind of queries or computations on data are allowed.

- [JW#]: Extends private inference control of [WS04] to work with more complex query functions. Client learns query result if and only if inference rule is met (and learns nothing else).
- [KMN05]: Simulatable auditing to ensure that query denials do not leak information.
- [ABG+04]: P4P: Paranoid Platform for Privacy Preferences. Mechanism for ensuring released data is usable only for allowed tasks.

Implementation and Experimentation

- secure scalar product protocol [SWY04]
- MySQL private information retrieval (PIR) [BBFS#]
- Fairplay: a system implementing Yao's two party secure function evaluation [MNPS04]
- Bayesian network implementation [KRFW#]
- secure computation of surveys using Fairplay and use for Taulbee survey [FPRS04]

Survey Software [FPRS04]

- User-friendly, open-source, free implementation using Fairplay [MNPS04], suitable for use with CRA's Taulbee salary survey.
Not adopted.
- CRA's reasons:
 - Need for data cleaning, multiyear comparisons, unanticipated use
 - “Perhaps most member departments will trust us.”
- Provost Offices' reasons:
 - No legal basis for using this privacy-preserving protocol on data that we otherwise don't disclose
 - Correctness and security claims are hard and expensive to assess, despite open-source implementation.
 - All-or-none adoption by Ivy+ peer group. Can't make decision unilaterally.

Future Directions

- More experiments, especially with real data and real user communities.
- Preprocessing of data for PPDM.
- Privacy-preserving data solutions that use both randomization and cryptography in order to gain some of the advantages of both.
- Policies for privacy-preserving data mining: languages, reconciliation, and enforcement.
- Incentive-compatible privacy-preserving data mining.

Conclusions

- Increasing use of computers and networks has led to a proliferation of sensitive data.
- Without proper precautions, this data could be misused.
- Many technologies exist for supporting proper data handling, but much work remains, and some barriers must be overcome in order for them to be deployed.
- Cryptography is a useful component, but not the whole solution.
- Technology, policy, and education must work together.

THANK YOU!