Main goals of privacy-preserving data mining

Privacy of individual user or user group—privacy of database records

- Statistical perturbation techniques.
- No solid theoretical foundations, inapplicable for operational data.
- Unexpected surprises, filtering attacks (ICDM 2003).
- PSM games might be a partial solution.

Secure merge of different data sources—secure multi party computation

- Possible in polynomial time, but still infeasible in practice.
- Solid theoretical foundations, but many insecure protocols.
- Missing or too inefficient basic primitives: matrix operations, set operations, max and argmax.
- Need for oblivious data structures: trees, DAG-s, stacks etc.

Main goals of privacy-preserving data mining

Privacy-preserving queries—*private information retrieval*

- Strong cryptographic foundations (1997)
- High computational complexity. Lack of search structures.
- Predicates and summaries *vs.* information retrieval.
- Tradeoff between online and off-line efficiency.

Outsourcing storage or computational power

- Searching in encrypted data—several new articles.
- Outsourcing computations seems tricky. Several protocols for matrix operations are proposed. Some of those are weak.
- Oblivious data mining and data anonymization.

Cryptographic tools

Homomorphic encryption $E(a)E(b) = E(a+b)$.

- Cornerstone of efficient blinded computations.
- Usually introduces big message bloat (up to 1024).
- Oblivious polynomial evaluation.

Oblivious transfer and garbaged circuit evaluation

- If everything else fails...
- Quite costly—one OT per input bit.
- Currently the only known way to evaluate $\max$ and $\leq$.

Private equality testing

- Only in-extendable efficient protocols are known.

Error correction codes

- Might be useful in approximate private queries.

Simple and hard tasks

Simple tasks

- Counting, adding and low degree polynomial evaluation.
- Message bloat and actual workload might be a problem.
- Examples: pattern counting, averages, (general)linear regression.

Relatively hard tasks

- Protocols with relatively large communication but simple operations.
- Examples: Frequent patter mining, k-means clustering etc.

Hard tasks

- Minimum or maximum finding. Statistical model driven inference.
- Examples: k-nearest neighborhood algorithm, approximate queries.