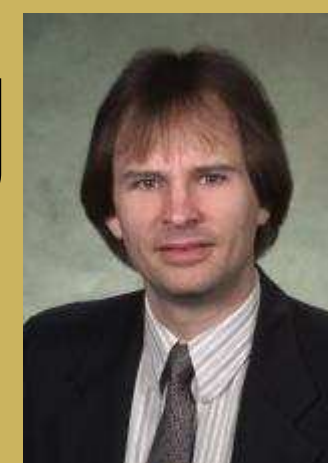


Privacy-Preserving Data Integration and Sharing



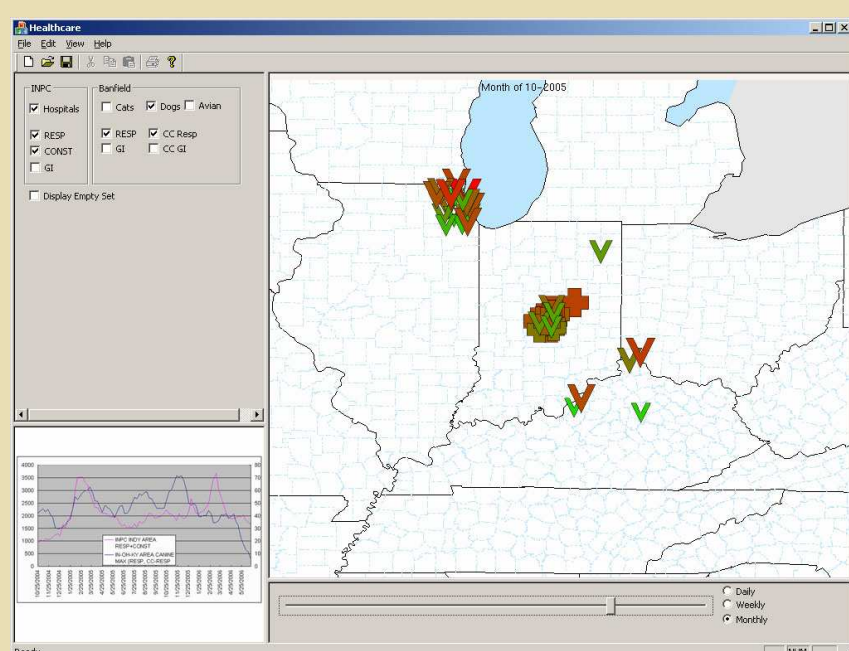
Chris Clifton, Ahmed Elmagarmid, Dan Suciu, AnHai Doan, and Gunther Schadow

Data Sharing is Good but so is Privacy

Goal Support use and analysis of integrated data without risks inherent in an omniscient data store

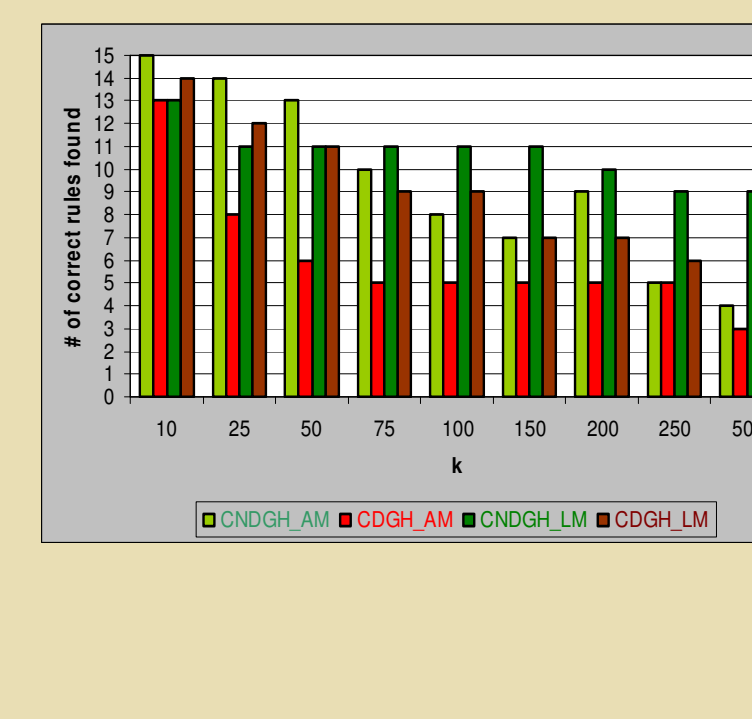
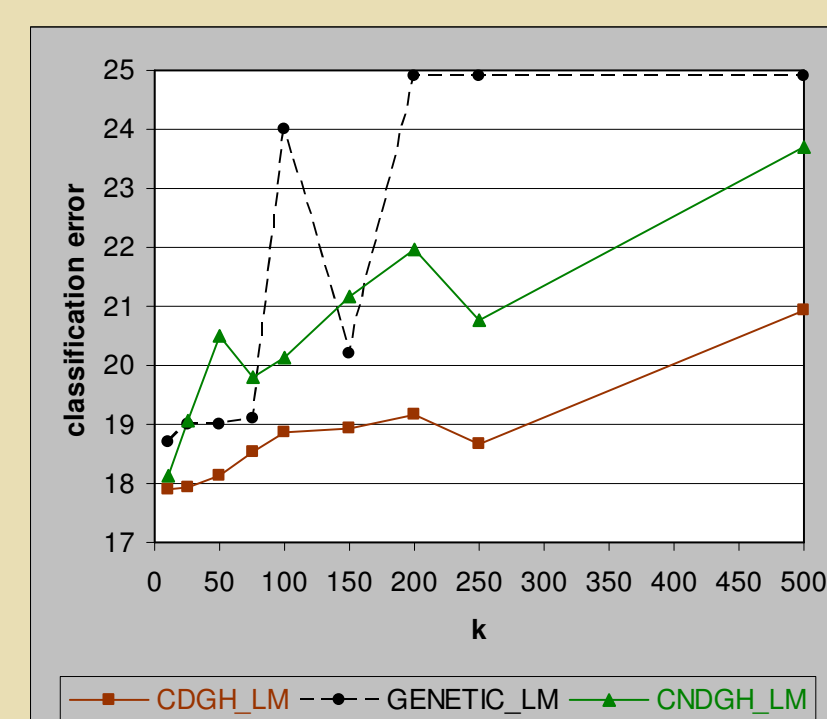
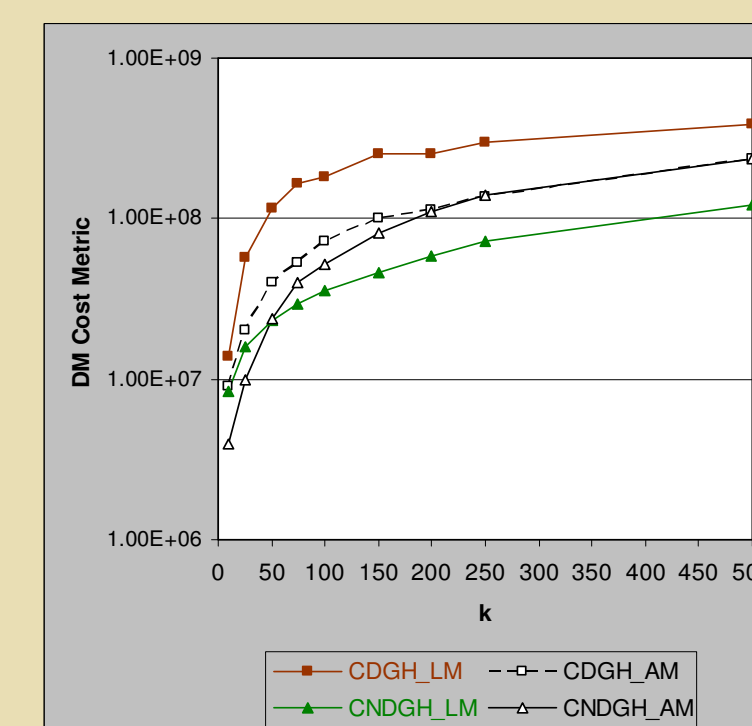
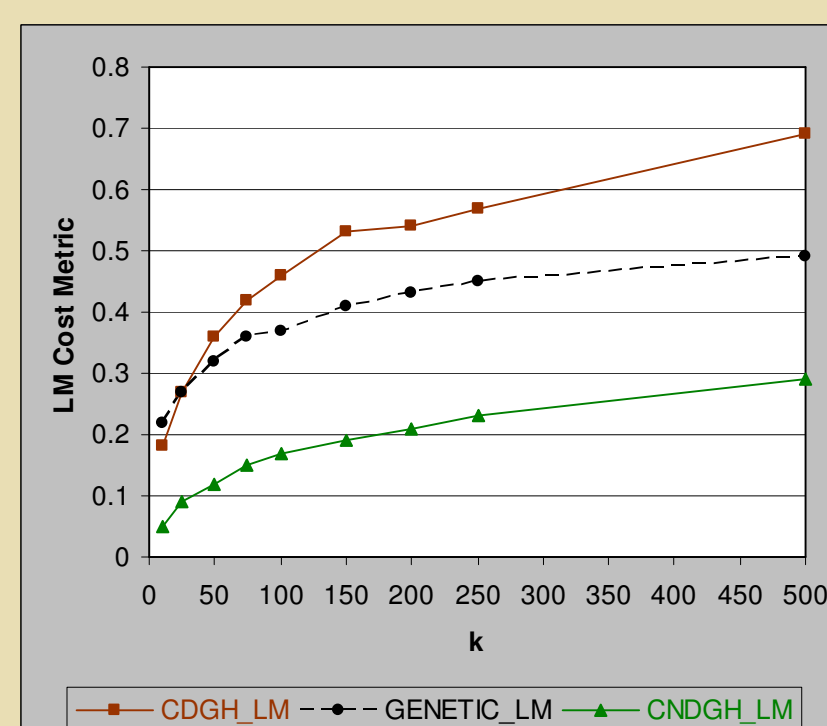
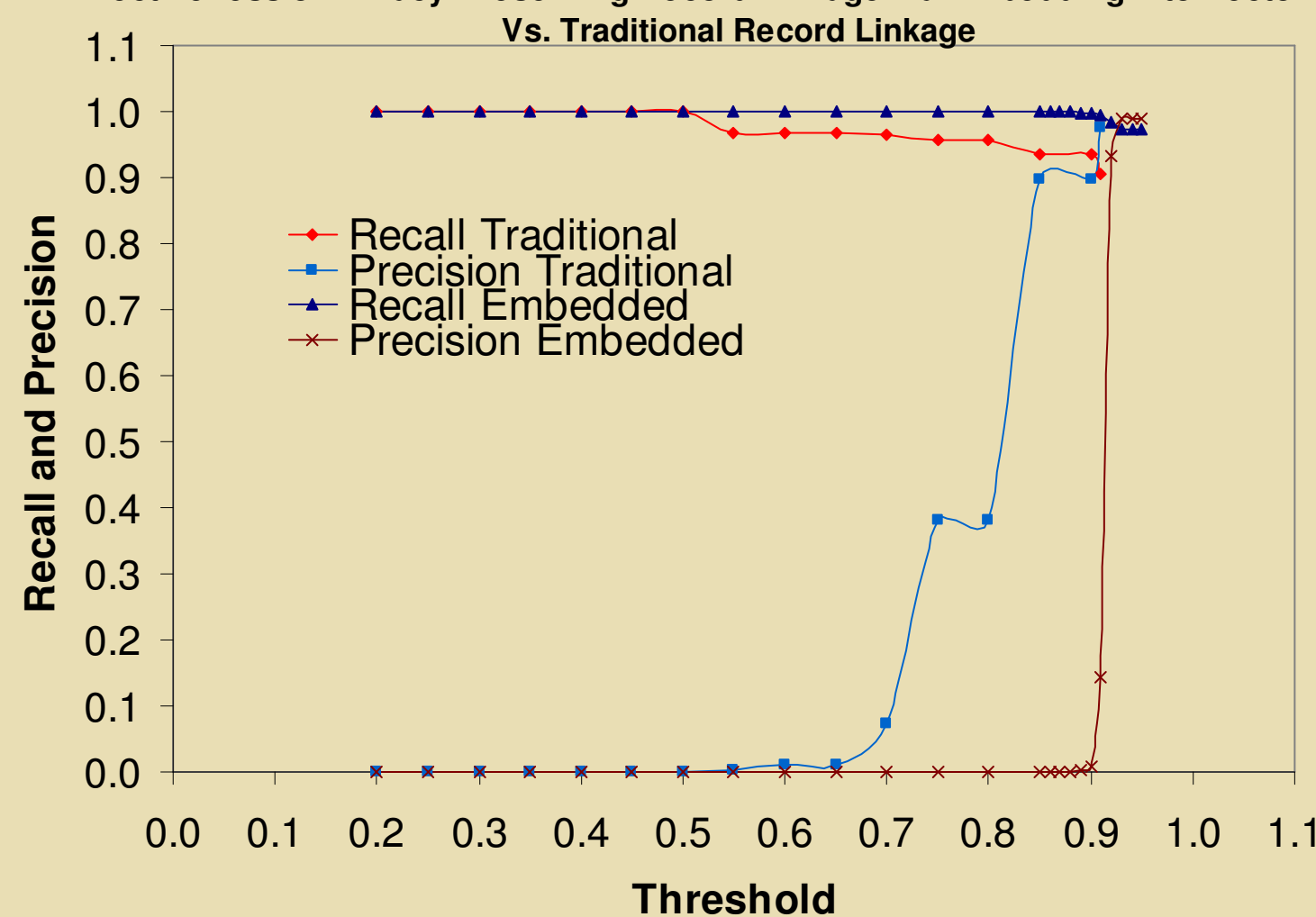
Benefit Societal and scientific advances through increased knowledge from data

Healthcare – Early response to pandemic without disclosing individual health records



Homeland Security – Enable “connecting the dots” without enabling intrusion into our private lives

Effectiveness of Privacy-Preserving Record Linkage via Embedding Into Vector Space Vs. Traditional Record Linkage



Approach and Impact

New approach

- Information Disclosure as function of cardinality/fan-out
- Distributed k -anonymity
- Secure bi-gram score, jaro score, and hamming distance

Research Impact

- Better privacy metrics
- New privacy proof method
- Tools for record linkage

Privacy Definitions

- Formal methods to express information disclosure as a function of cardinality statistics and fan-out statistics in relational databases
- New measure for anonymity based on privacy risk of detecting an individual in a database
- New adversary model for secure multiparty computation

Anonymization

- Distributed k -anonymity method that guarantees privacy without using traditional Secure Multiparty Computation definitions
- Empirically demonstrated poor privacy/utility tradeoff of existing k -anonymity metrics and algorithms

Record Linkage

- Techniques for secure calculation of bi-gram score, jaro score, and hamming distance score.
- Embedding of records into vector space that preserves privacy and provides accurate record linkage

Schema Matching

- Techniques to automatically tune schema matching software and to maintain mappings over time

Query evaluation

- Efficient algorithm for retrieving the top k answers in probabilistic databases that is provably optimal (as a function of k) has good performance in practice