Motivation

- The private record linkage problem ---
  - Alice holds database A
  - Bob holds database B
  - Find common records of A and B, such that A and B remain private.
  - Common records only revealed to Alice and Bob

Applications

- Patient Information
- Cooperation between government agencies
- Sharing of intellectual property
- Outsourcing

Background: Record Linkage

- Heterogeneous data
  - Multiple Object Naming / Representations
    Ex: The Pennsylvania State University vs. Penn State
    Ex: John A. Smith vs. Smith, J. A.
  - Spelling Mistakes
    Ex: were hear vs. where here
  - Object Character Recognition (OCR)
    Ex: 0 (zero) vs. 0, 1 (one) vs. 1.

Record Linkage: Two Steps

1. Blocking
   - Only compare tokens with common features
     Ex: Alphabetic Sort
     Ex: Common tokens

2. Distance Metrics:
   - Assign a score of similarity: \( \text{dist}(r_1, r_2) \)
   - If score > threshold
     Then \( r_1 \) and \( r_2 \) are matched
     Ex: Edit distance.
     Ex: TFIDF

Token Blocking

<table>
<thead>
<tr>
<th>Block id</th>
<th>Patient Database A</th>
<th>Database A Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>John Smith, Alzheimer, Chicago</td>
<td>a1, a2</td>
</tr>
<tr>
<td>a2</td>
<td>Smith John, diabetes, Chicago</td>
<td>a1, a3</td>
</tr>
<tr>
<td>a3</td>
<td>John, Alzheimer, Heart, Chicago</td>
<td>a2</td>
</tr>
<tr>
<td>a4</td>
<td>Smith, Heart, Chicago</td>
<td>a3, a4</td>
</tr>
</tbody>
</table>
Private Record Linkage

- Added Security Dimension
- Steps:
  - Secure Data Mapping
    Ex: Commutative hashing
  - Security-enhanced distance Metric
    Ex: Secure TFIDF
- Propose adding a new step:
  - Secure Blocking
- More Contributions:
  - Definition of protocol
  - Analysis of privacy
  - Experimental validations

Related Work

- Medical Field
- Minimal Information Sharing (Agrawal et al 2003)
- Rivakumar et al 2004
  - Secure distance metrics
  - Secure intersection algorithm

Communication / Threat Model

- Three parties
  - 2 collaborating parties
  - A third party
- All parties semi-trusted
- Follow protocol precisely
  - Provide accurate data
  - Do not collude with other parties
- However, try to find as much other information
  - Dictionary attacks
  - Statistical analysis

Threat Model

- Loose characterization of sharing:
  - Categories:
    - DBsize
    - Vocabsize
    - Reclen
    - Tokfrq
  - Levels of exposure:
    - Yes
    - No
    - inf

Threat Model

![Diagram](image)

Protocol

- Participant:
  - Alice holding db A
  - Bob holding db B
  - Third party Carol
- Protocol:
  1. Negotiate k
  2. Alice & Bob pre-generate blocks. (token blocking)
  3. Carol computes private record linkage problem (secure TFIDF)
  4. Carol forwards results to Alice and Bob

![Diagram](image)
Secure TFIDF

- Based on
  - Token Frequency (TF)
  - Inverse Document Frequency (IDF)

- Per token

\[
TFIDF_{weight} = \log(TF_{weight}) + \log(1/IDF_{weight})
\]

Secure TFIDF

- Hash Signature
  - Normalize size of weight vectors
  - Compact representation

\[
\text{Hash Signature} = 2^t \quad \text{float array}
\]

4 Blocking Schemes

- Phase 1:
  - Baseline
  - Simple
  - Record-aware
  - Frugal Third Party

- Phase 2:
  - Jaccard

Two Phase Blocking

- Phase 1:
  - Baseline
  - Simple
  - Record-aware
  - Frugal Third Party

- Block

Two Phase Blocking

- Phase 1:
  - Baseline
  - Simple
  - Record-aware
  - Frugal Third Party

- Block
### Two Phase Blocking

- **Phase 2 Blocking**

\[
\text{Jaccard}(r1, r2) = \frac{|r1 \cap r2|}{|r1 \cup r2|}
\]

### 4 datasets - Run times

![Run times graph](image)

### 4 datasets - Precision

![Precision graph](image)

### Comparison of Blocking (1/2)

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

### Comparison of Blocking (2/2)

![Blocking comparison graph](image)

### Summary

- Private Record Linkage Protocol that supports blocking.
- Secure & efficient representation of TFIDF weight vectors using hash signatures.
- Two phase blocking, characterization of information leakage, and three blocking schemes.

- Future Work:
  - Apply concepts to related algorithms.
  - Specify incremental maintenance policy.
Summary

- Private Record Linkage Protocol that supports blocking.
- Secure & efficient representation of TFIDF weight vectors using hash signatures.
- Two phase blocking, characterization of information leakage, and three blocking schemes.

- Future Work:
  - apply concepts to related algorithms.
  - Specify incremental maintenance policy.

Outline

- What is Web Services?
- Motivation
- Main Idea: MISQ
- Illustration
- Conclusion

Blocking Summary

Example

Example (cont.)

Comparison of Blocking & Matching Time
<table>
<thead>
<tr>
<th>University</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ohio State</td>
</tr>
<tr>
<td>Penn State</td>
</tr>
<tr>
<td>Pennsylvania State</td>
</tr>
<tr>
<td>West Virginia State</td>
</tr>
<tr>
<td>Michigan State</td>
</tr>
<tr>
<td>Nebraska</td>
</tr>
<tr>
<td>Purdue University</td>
</tr>
<tr>
<td>Wisconsin</td>
</tr>
</tbody>
</table>