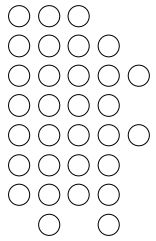
Blocking-Aware Private Record Linkage

Ali Al-Lawati*
Dongwon Lee
Patrick McDaniel

Penn State University, USA

June 17, 2005

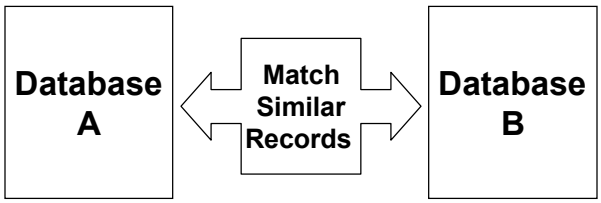


Motivation

- The private record linkage problem ---
  - Alice holds database A
  - Bob holds database B
  - Find common records between A and B, such that
    - A, B remain private
    - A ∩ B is revealed to Alice and Bob
- Applications
  - Patient Information
  - Cooperation between government agencies
  - Sharing of intellectual property
  - Outsourcing

Background: Record Linkage

- Record Linkage
  - Data cleansing
  - Data Integration
  - Duplicate Elimination
- A Common Scenario



3

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. o, 1 (one) vs. l.

4

Record Linkage:Two Steps

- 2. Distance Metrics:
  - Assign a score of similarity: $\mathrm{dist}(\mathrm{r1}, \mathrm{r2})$
  - If score > threshold

    Then r1 and r2 are *matched*

    Ex: Edit distance.

    Ex: TFIDF
- 1. Blocking
  - Only compare tokens with common features

    Ex: Alphabetic Sort

    Ex: Common tokens

5

Token Blocking

Patient Database A

rid	
a1	John Smith, Alzheimer, Chicago
a2	Smith Johns, diabetes, Chicago
a3	John, Alzheimer, Heart, Chicago
a4	Smith, Heart, Chicago

Block(A)

Block id	Database A Blocks
john	a1, a3
smith	a1, a2, a4
alzheimer	a1, a3
chicago	a1, a2, a3, a4
johns	a2
diabetes	a2
heart	a3, a4

6

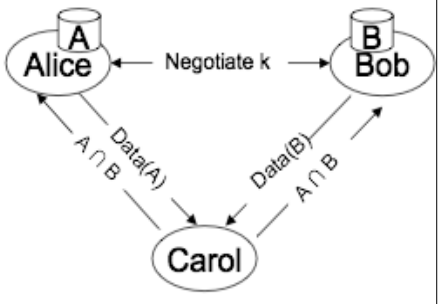
Private Record Linkage

- Added Security Dimension
- Steps:
  - Secure Data Mapping
    Ex: Commutative hashing
  - Security-enhanced distance Metric
    Ex: Secure TFIDF [Ravikumar 04]
- Problem statement: propose adding a new step:
  - Secure Blocking
- More Contributions:
  - Definition of protocol
  - Analysis of privacy
  - Experimental validations

7

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



8

Threat Model

- Loose characterization of sharing:
  - Categories:
    - DB_{size}
    - $Vocab_{size}$
    - Rec_{len}
    - Tok_{frq}
  - Levels of exposure:
    - Yes
    - No
    - **inf**

    [Agrawal et al 03]

9

Threat Model

Patient Database A

Rec_{len}

DB_{size}

rid	
a1	John Smith, Alzheimer, Chicago
a2	Smith Johns, diabetes, Chicago
a3	John, Alzheimer, Heart, Chicago
a4	Smith, Heart, Chicago

Block

$Vocab_{size}$

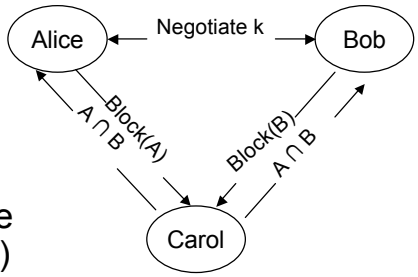
Block id	Database A Blocks
john	a1, a3
smith	a1, a2, a4
alzheimer	a1, a3
chicago	a1, a2, a3, a4
johns	a2
diabetes	a2
heart	a3, a4

Tok_{frq}

10

Protocol

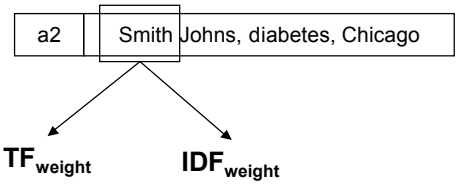
- Participants:
  - Alice holding db A
  - Bob holding db B
  - Third party Carol
- Protocol:
  1. Negotiate k
  2. Alice & Bob pre-generate blocks. (token blocking)
  3. Carol computes private record linkage problem (secure TFIDF)
  4. Carol forwards results to Alice and Bob



11

Secure TFIDF

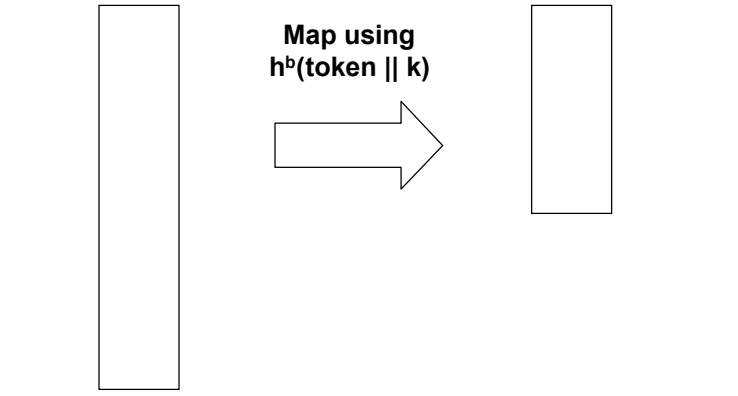
- Based on
  - Token Frequency (TF)
  - Inverse Document Frequency (IDF)
- Per token



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

12

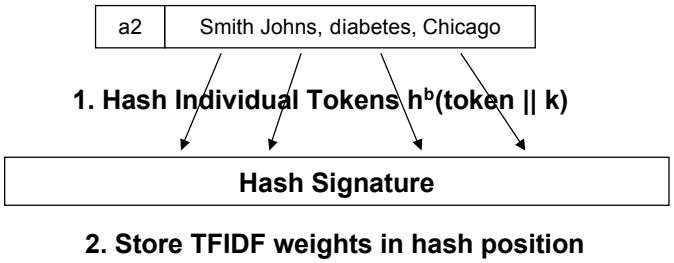
Secure TFIDF



13

Secure TFIDF

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



2. Store TFIDF weights in hash position

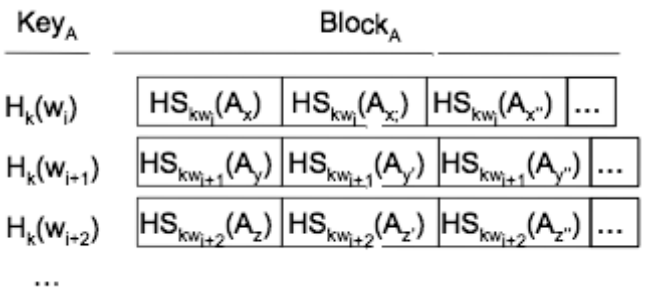
14

4 Blocking Schemes

- Baseline
- Simple
- Record-aware
- Frugal Third Party

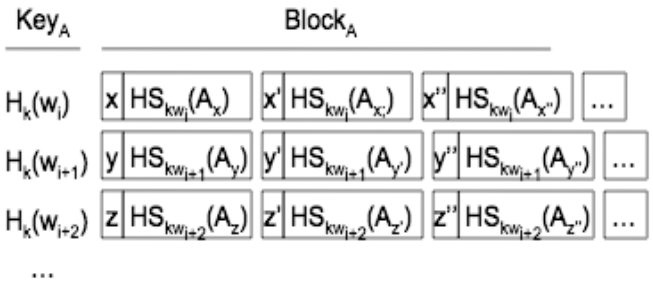
15

Simple Blocking



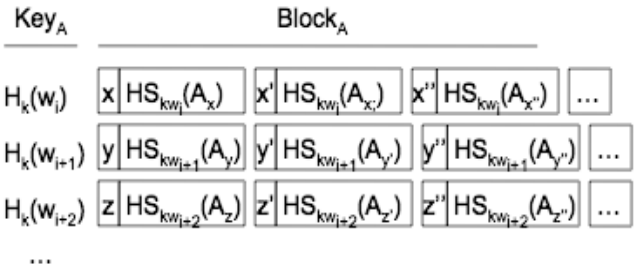
16

Record Aware Blocking



17

Frugal Third Party Blocking



Homomorphic Encryption (Based on Public Key):

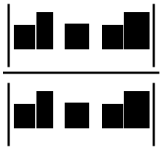
$$E_a (E_b (key)) = E_a (E_b (key))$$

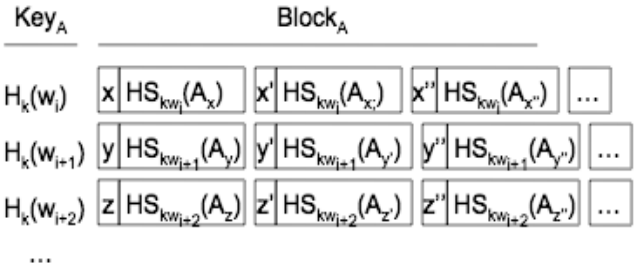
$$Key_A = Key_B$$

18

Level 2 Blocking

PENNSTATE

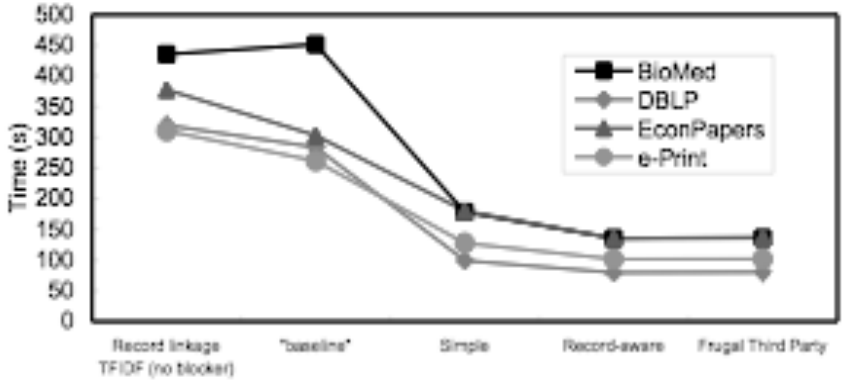
Jaccard($r1$, $r2$) = 



19

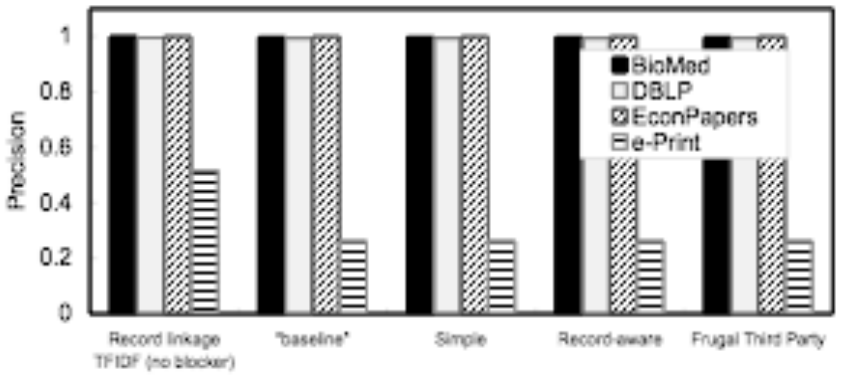
4 datasets - Run times

PENNSTATE



20

4 datasets - Precision



21

Privacy

Scheme	Information Categories			
	DB_{size}	$Vocab_{size}$	Rec_{len}	tok_{freq}
Baseline	Yes	inf	inf	No
Simple	inf	Yes	inf	Yes
Record-aware	Yes	Yes	Yes	Yes
Frugal Third Party	inf	inf	No	inf

22

Dblp datasets

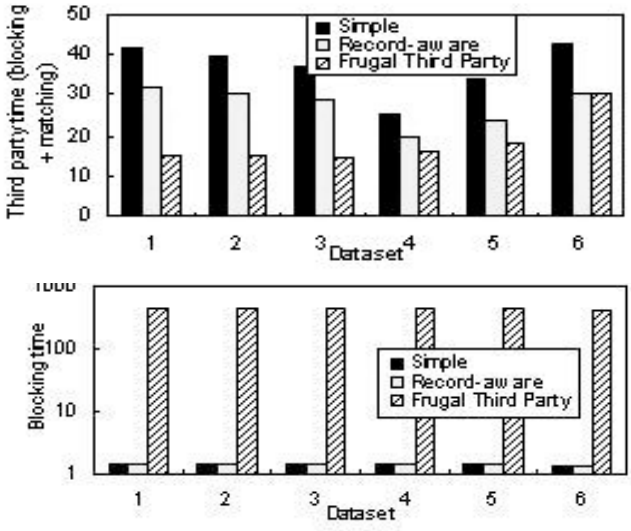
PENNSTATE

Dataset	A size	B size	A ∩ B size
1	2500	2500	0
2	2500	2500	100
3	2500	2500	250
4	2500	2500	500
5	2500	2500	1000
6	2500	2500	2500

23

Comparison of Blocking

PENNSTATE



24

Summary

PENNSTATE

- 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 privatealgorithms.
  - Specify incremental maintenance policy.

25

Summary

PENNSTATE

- 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 private algorithms.
  - Specify incremental maintenance policy.

26

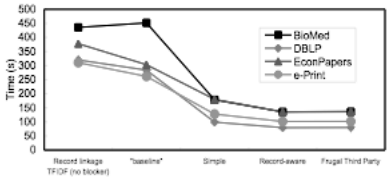
Outline

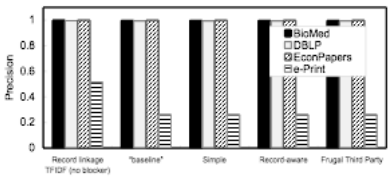
- What is Web Services?
- Motivation
- Main Idea: MISQ
- Illustration
- Conclusion

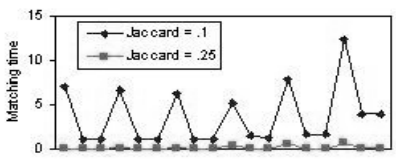
27

Blocking Summary

Scheme	Information Categories			
	DB_{size}	$Vocab_{size}$	Rec_{len}	tok_{freq}
Baseline	Yes	inf	inf	No
Simple	inf	Yes	inf	Yes
Record-aware	Yes	Yes	Yes	Yes
Frugal Third Party	inf	inf	No	inf

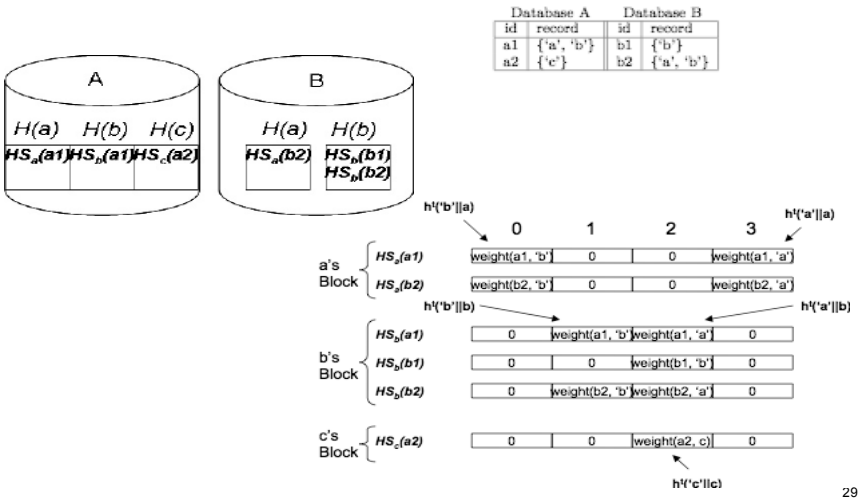






28

Example



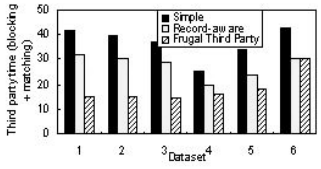
Example (cont.)

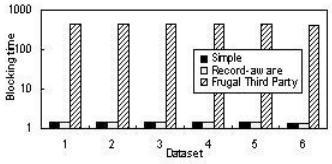
	0	1	2	3
$HS_a(a1)$	1	0	0	1
$HS_a(b2)$	1	0	0	1
$HS_b(a1)$	0	1	1	0
$HS_b(b1)$	0	0	1	0
$HS_b(b2)$	0	1	1	0
$HS_c(a2)$	0	0	1	0

Comparison of Blocking & Matching Time

PENNSTATE

Dataset	A size	B size	A ∩ B size
1	2500	2500	0
2	2500	2500	100
3	2500	2500	250
4	2500	2500	500
5	2500	2500	1000
6	2500	2500	2500





31

Related Work

PENNSTATE

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

32

Ohio State
Penn State
Pennsylvania State
Penn state university
Michigan State
Illinois
Purdue university
Wisconsin