Empirical Evaluation of Private Comparison Techniques in Entity Resolution

SBBD2016

THIAGO PEREIRA DA NÓBREGA,
CARLOS EDUARDO SANTOS PIRES
E TIAGO BRASILEIRO ARAÚJO

thiagonobrega@uepb.edu.br
cesp@dsc.ufcg.edu.br
tiagobrasileiro@copin.ufcg.edu.br

04 - 07, October 2016

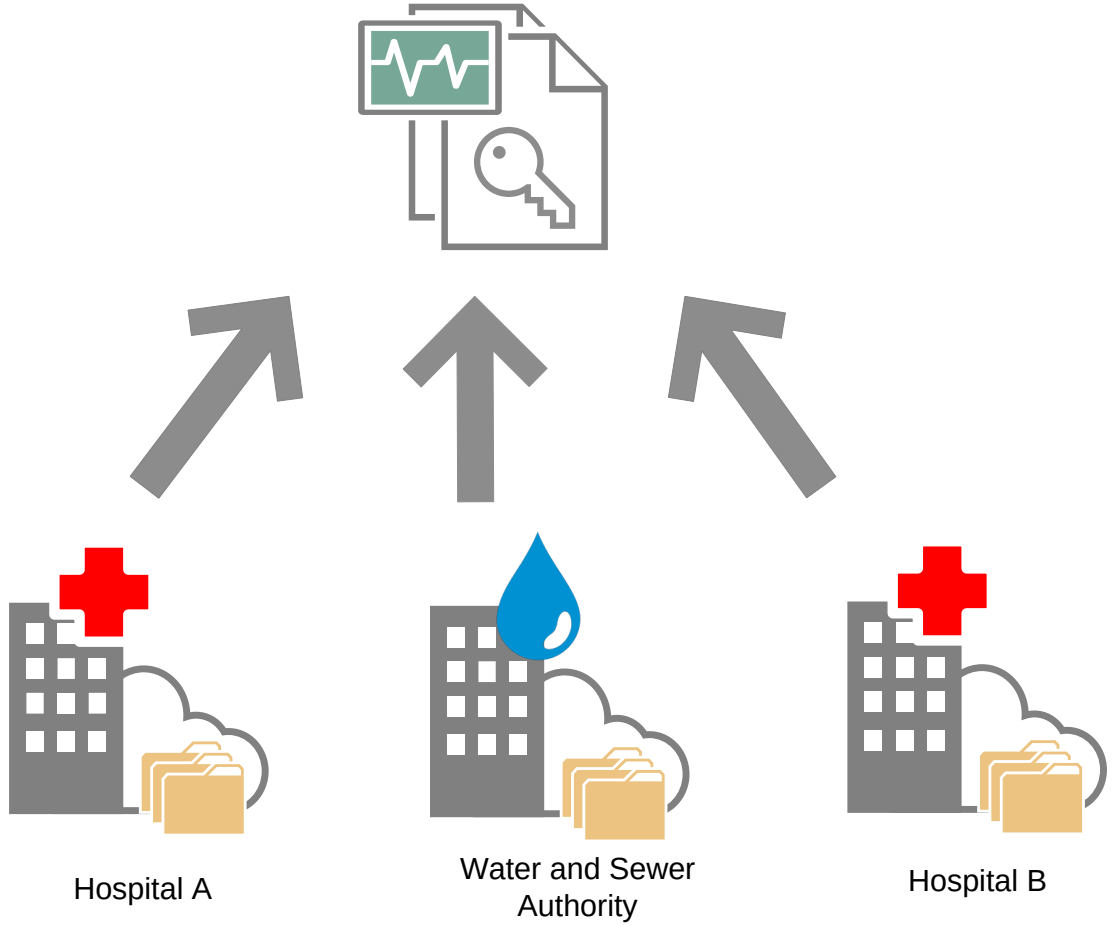
Record Linkage (RL)

HOSPITAL DATABASE			
Name	LastName	City	DOB
Joao	Araujo	Cabedelo	01/01/1994
João	Araújo	Cabedelo	01/01/1944
Joana	Araújo	Campina Grande	21/09/1998
João	Silva	Campina Grande	31/02/1988

IMAGE CENTER DATABASE			
Name	LastName	City	DOB
João	Araújo	Cabedelo	01/01/1944
Joana	Araújo	Campina Grande	21/09/1998
João	Silva	Campina Grande	31/02/1988

EMERGENCY ROOM (E.R) DATABASE			
FirstName	Surname	County	Birth
Joao	Araujo	Cabedelo	01/01/1994
Joanna	Araujo	Campina Grande	21/09/1998
Joao	Araujo	Campina Grande	31/01/1988

Private Preserving Record Linkage (PPRL)



Private Comparison Techniques

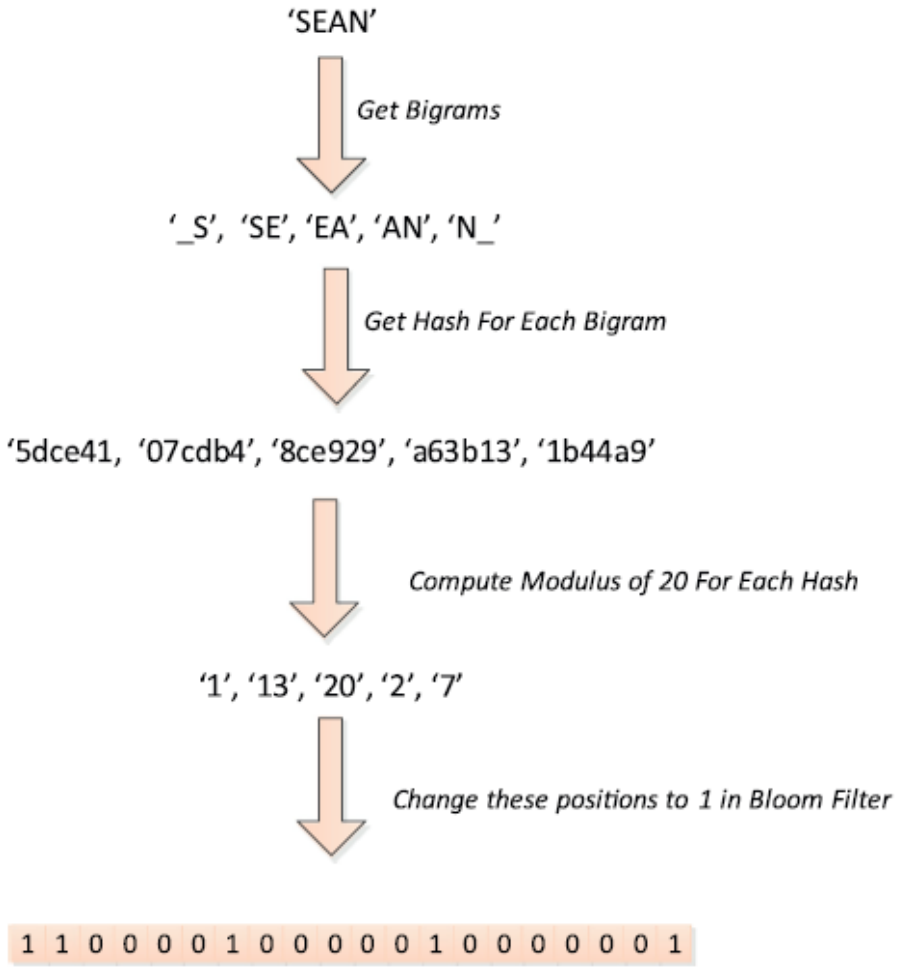
Exact comparison

- One-Way-Hash cryptography (e.g. md5, sha1,etc)

Approximate comparison

- Bloom Filters (Textual)
- Homomorphic cryptography (Data and Numeric)
  - PHE
  - FHE

Bloom Filter



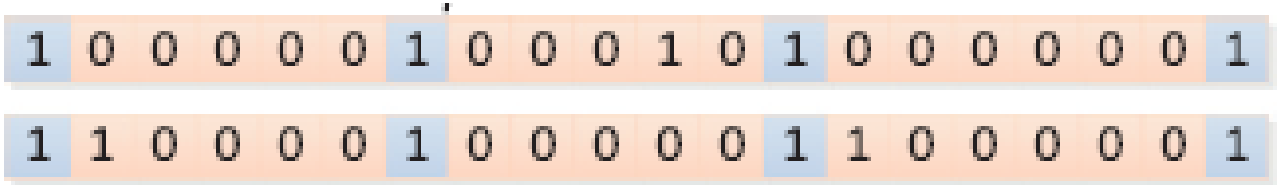
Randall, et al(2014)

Bloom Filter

$$DiceCoefficien_{A,B} = \frac{2h}{a+b}$$

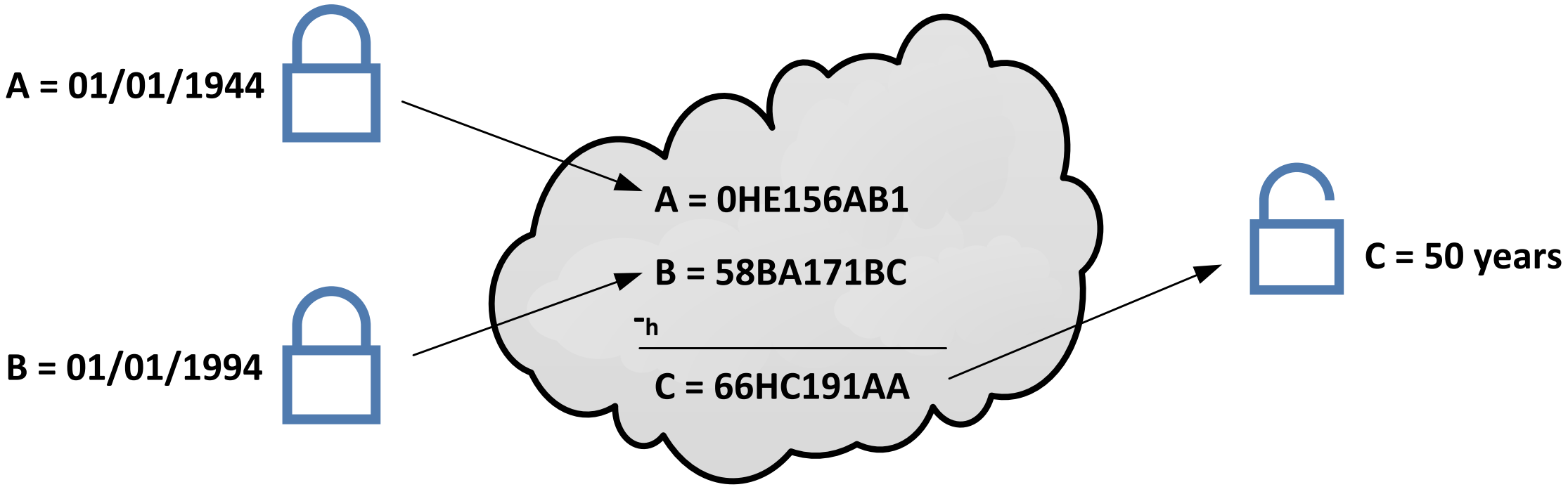
A = 01/01/1944

B = 01/01/1994



$$\Rightarrow \frac{2 \times 4}{5+6} = \mathbf{0.72}$$

Homomorphic cryptography



Research Question

The utilization of homomorphic encryption could improve the quality of private preserving entities resolution (PPRL)?

Related Work

- M. G. Kaosar, R. Paulet, and X. Yi, “Fully homomorphic encryption based two-party association rule mining,” Data Knowl. Eng., vol. 76–78, pp. 1–15, Jun. 2012.

- D. Vatsalan and P. Christen, “Privacy-preserving matching of similar patients,” J. Biomed. Inform., vol. 59, no. December, pp. 285–298, 2016.

Experiment Setup

Comparison Strategies :

Strategy 1 : Bloom Filter in all attributes

Strategy 2 : Bloom Filter in textual attributes and Homomorphic Encryption in dates and numeric attributes

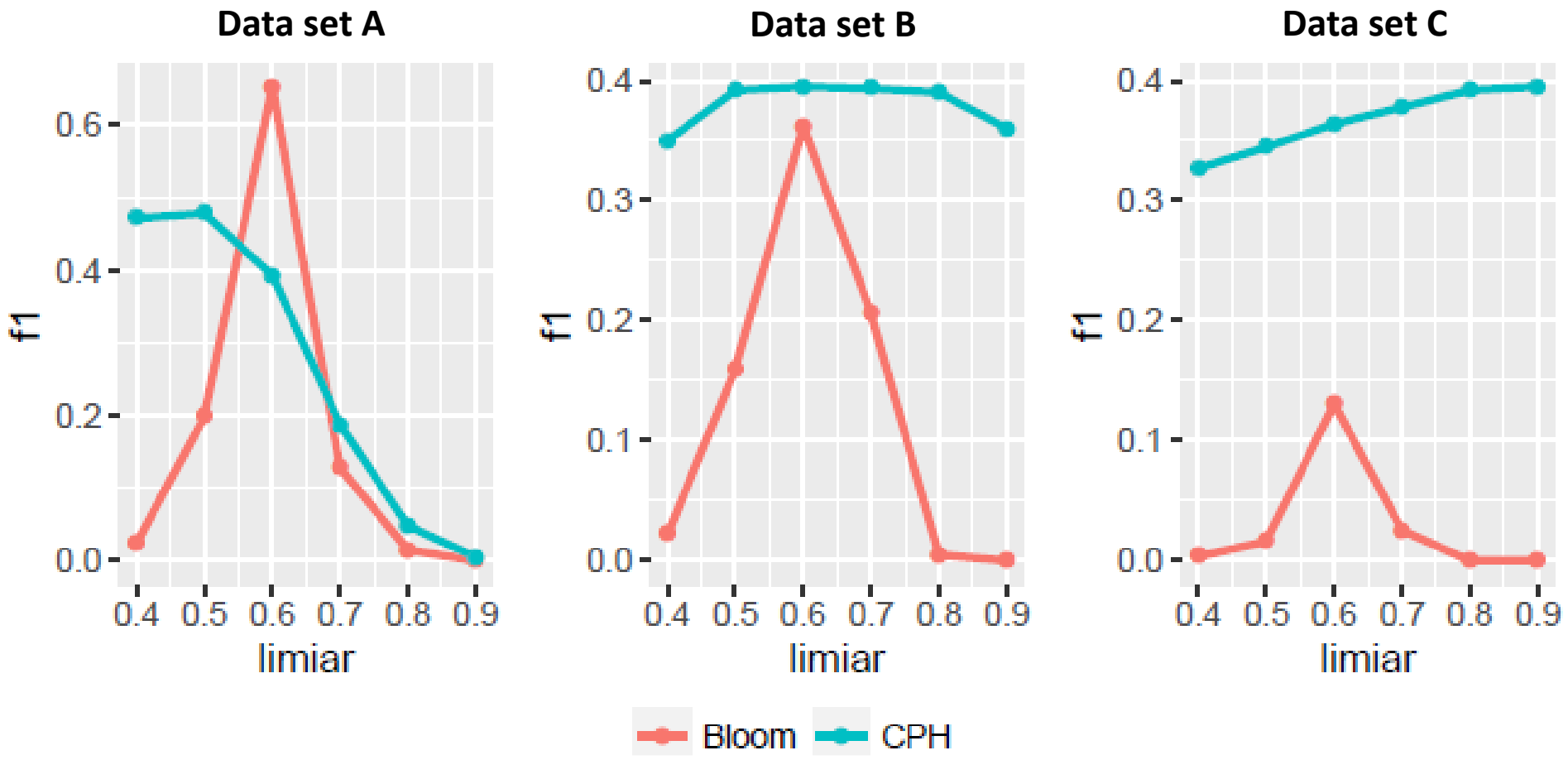
Synthetic data Generator and Corruptor (Tran et al. 2013)

- Dirty attributes (Table 1)
- 3 Data sets
- 2.400 Entities

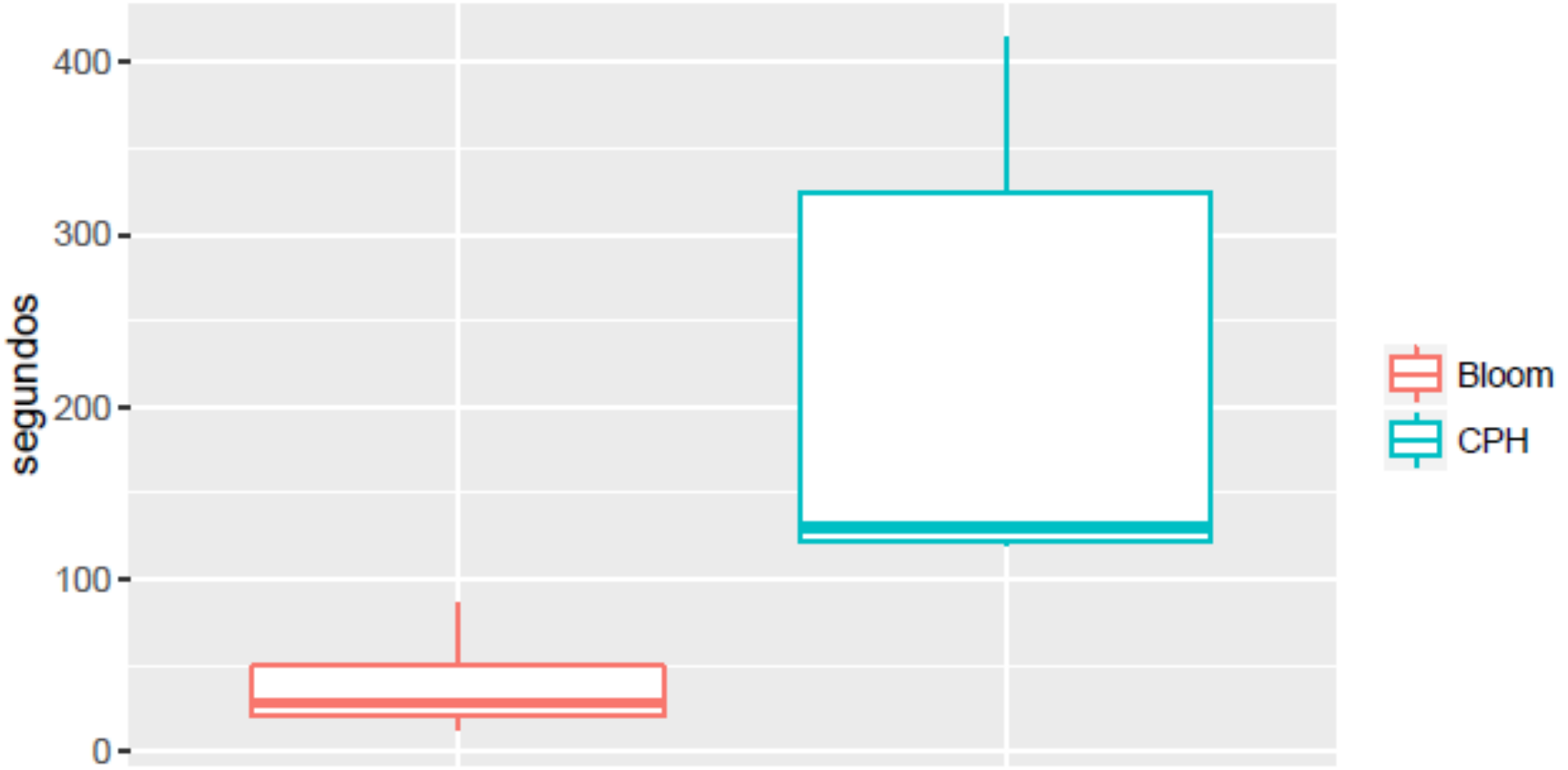
	lastname	name	D.O.B	D.O.D	salary
Data set A	20%	15%	15%	15%	35%
Data set B	x	10%	20%	30%	40%
Data set C	x	x	30%	25%	45%

Table 1: Error probability per attribute

Results


Data set A
Data set B
Data set C
f1
limiar
0.0
0.1
0.2
0.3
0.4
0.6
0.4 0.5 0.6 0.7 0.8 0.9
Bloom
CPH

Results



Conclusion and Future Work

Use of Homomorphic Encryptions

- Better quality in non-textual data resolution
- Higher computational cost

Future work

- Evaluate linkage quality in real data
- Evaluate the usage of homomorphic encryption in spatial data

References

S. M. Randall, A. M. Ferrante, J. H. Boyd, J. K. Bauer, and J. B. Semmens, “Privacy-preserving record linkage on large real world datasets,” *J. Biomed. Inform.*, vol. 50, pp. 205–212, 2014.

M. G. Kaosar, R. Paulet, and X. Yi, “Fully homomorphic encryption based two-party association rule mining,” Data Knowl. Eng., vol. 76–78, pp. 1–15, Jun. 2012.

D. Vatsalan and P. Christen, “Privacy-preserving matching of similar patients,” J. Biomed. Inform., vol. 59, no. December, pp. 285–298, 2016.

P. Christen, Data Matching. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

Tran, K.-n.,Vatsalan, D., and Christen, P. (2013). GeCo. Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM ’13, pages 2473–2476.

Empirical Evaluation of Private Comparison Techniques in Entity Resolution

SBBD2016

Thank you