FAKE INJECTION STRATEGIES FOR PRIVATE PHONETIC MATCHING

A. Karakasidis[1], V. S. Verykios[2] and P. Christen[3]

[1] Department of Computer and Communication Engineering
University of Thessaly
Volos, Greece
akarakasidis@inf.uth.gr

[2] School of Science and Technology
Hellenic Open University
Patras, Greece
verykios@eap.gr

[3] ANU College of Engineering and Computer Science
The Australian National University
Canberra, Australia
peter.christen@anu.edu.au

PRIVACY PRESERVING RECORD LINKAGE (PPRL)

- Approximate matching without common unique identifiers
- Integration without compromising privacy
- Examples:
  - Merging medical data
  - Locating tax evaders

THE PPRL PROBLEM FORMULATION

- *Let* μ *be a privacy metric for PPRL.*
- *A plain text database* D_{pt} *and its ciphered equivalent* D_c.
- μ *represents the ability to infer data from* D_{pt} *using data from* D_c
- Higher values of μ → higher inference ability.

SUFFICIENT PRIVACY GUARANTIES

- *A PPRL method is considered to offer sufficient privacy guaranties, if the value of its privacy metric* μ *does not exceed a predetermined privacy threshold* δ.

PPRL REVISITED

- *Considering data sources A, B, we wish to perform record matching between datasets R_A and R_B in a way that at the end of the process the privacy metric for source A, μ_A will not exceed δ_A.*

- More flexible definition

PRIVACY AND DATA MINING

- Three ways for providing privacy
  - Suppression
  - Perturbation
  - Generalization

SOUNDEX AND PRIVACY

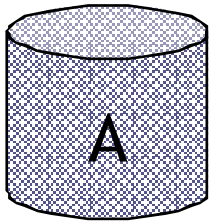
- Inherent Generalization Characteristics
- Retain the first letter of the name and drop all other occurrences of a, e, h, i, o, u, w, y.
- Replace consonants with digits as follows (after the first letter):
  - b, f, p, v => 1
  - c, g, j, k, q, s, x, z => 2
  - d, t => 3
  - l => 4
  - m, n => 5
  - r => 6
- Two adjacent letters with the same number are coded as a single number.
- Continue until you have one letter and three numbers. If you run out of letters, fill in 0s until there are three numbers.

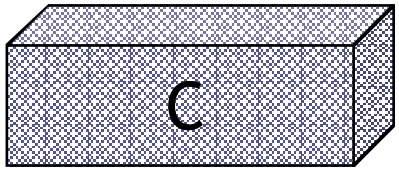
MATCHING PROTOCOL

- Based on Soundex inherent privacy
- Using a trusted third party
- Fake codes to enhance privacy

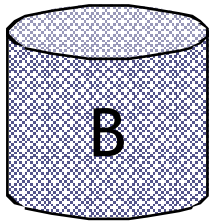
PROTOCOL DESCRIPTION

ID	Sndx	Surname
1	F632	
2	J525	Johnson
3	K364	Miller
4	M460	





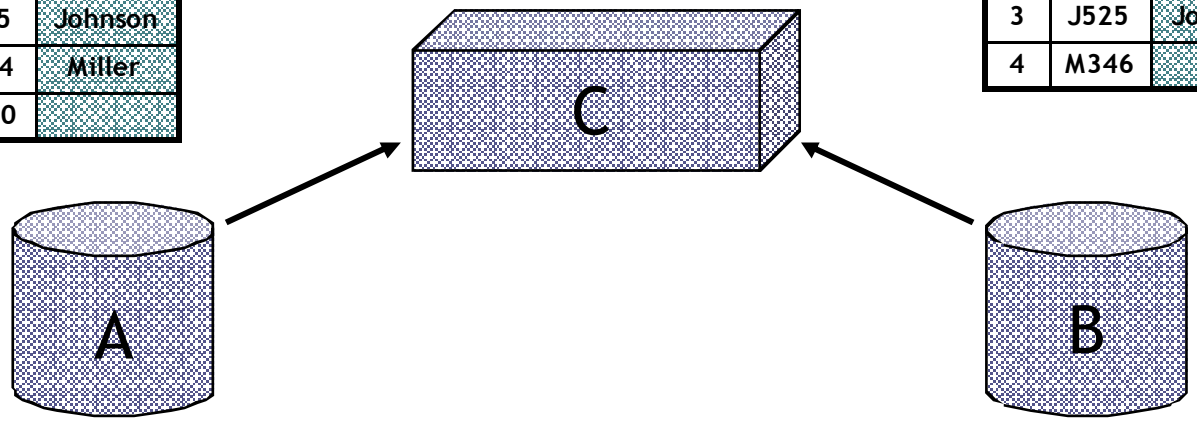
ID	Sndx	Surname
1	A100	
2	F632	Fortson
3	J525	Johnsen
4	M346	



PROTOCOL DESCRIPTION

ID	Sndx	Surname
1	F632	
2	J525	Johnson
3	K364	Miller
4	M460	

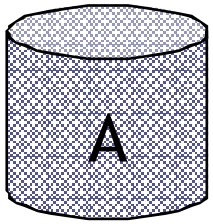
ID	Sndx	Surname
1	A100	
2	F632	Fortson
3	J525	Johnsen
4	M346	

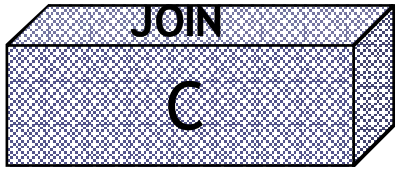


Sources send data to the third party

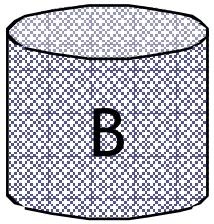
PROTOCOL DESCRIPTION

ID	Sndx	Surname
1	F632	
2	J525	Johnson
3	K364	Miller
4	M460	





ID	Sndx	Surname
1	A100	
2	F632	Fortson
3	J525	Johnsen
4	M346	

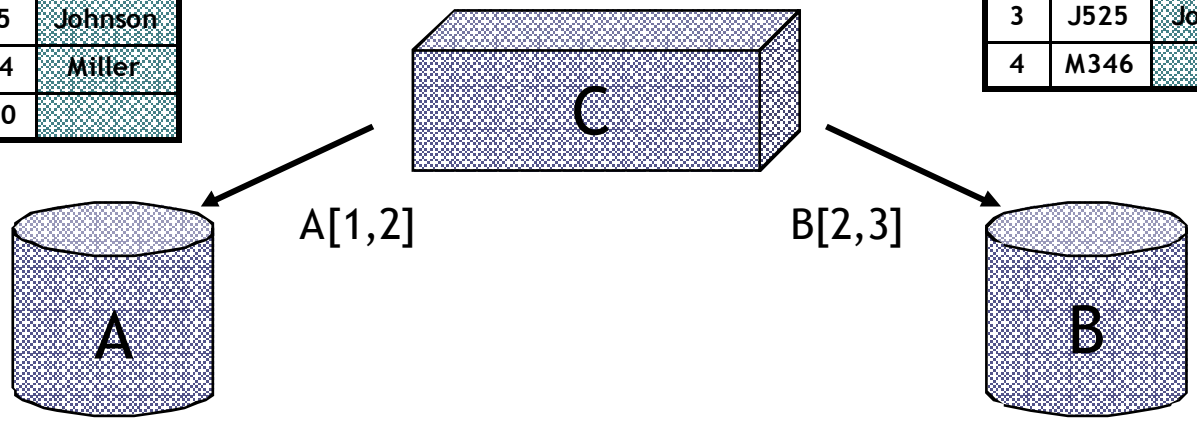


The third party joins the Soundex codes

PROTOCOL DESCRIPTION

ID	Sndx	Surname
1	F632	
2	J525	Johnson
3	K364	Miller
4	M460	

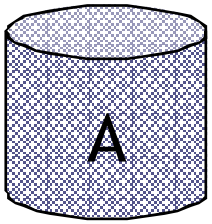
ID	Sndx	Surname
1	A100	
2	F632	Fortson
3	J525	Johnsen
4	M346	

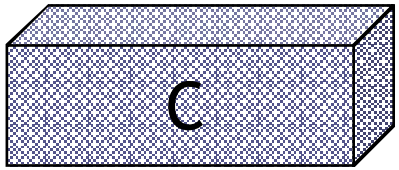


The third party returns the matching identifiers

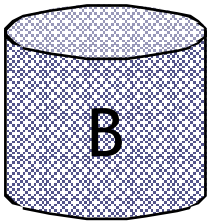
PROTOCOL DESCRIPTION

ID	Sndx	Surname
1	F632	
2	J525	Johnson
3	K364	Miller
4	M460	





ID	Sndx	Surname
1	A100	
2	F632	Fortson
3	J525	Johnsen
4	M346	

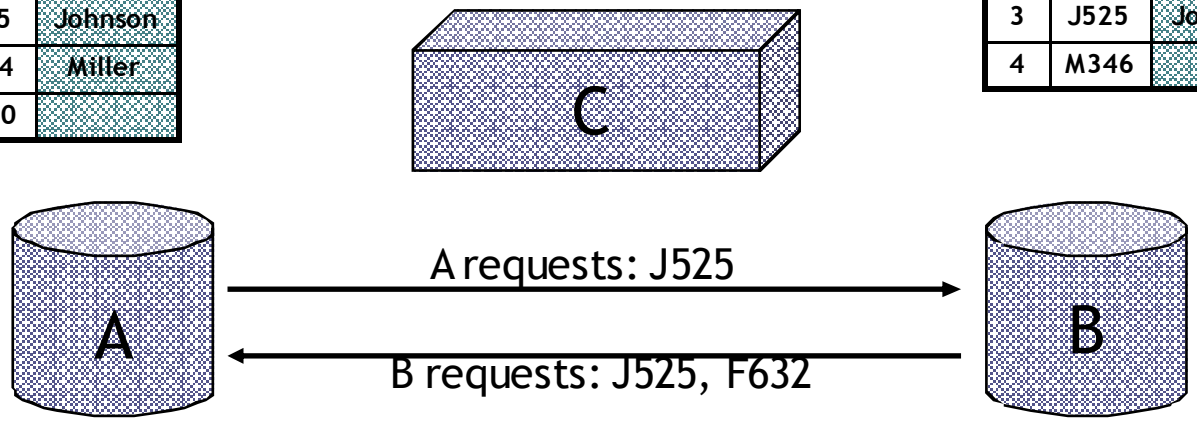


Sources determine identifiers

PROTOCOL DESCRIPTION

ID	Sndx	Surname
1	F632	
2	J525	Johnson
3	K364	Miller
4	M460	

ID	Sndx	Surname
1	A100	
2	F632	Fortson
3	J525	Johnsen
4	M346	

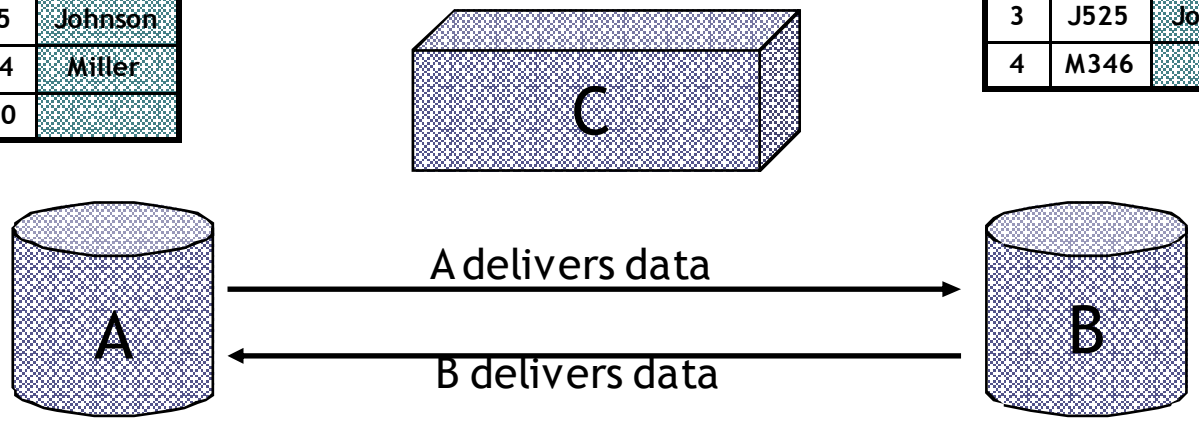


Sources ask directly data from each other

PROTOCOL DESCRIPTION

ID	Sndx	Surname
1	F632	
2	J525	Johnson
3	K364	Miller
4	M460	

ID	Sndx	Surname
1	A100	
2	F632	Fortson
3	J525	Johnsen
4	M346	



Sources deliver data

A QUANTITATIVE MEASURE OF PRIVACY

- Need for a Privacy Metric
- Use of Information Theory
  - Calculation of Entropy
  - Calculation of Information Gain
  - Calculation of Relative Information Gain

ENTROPY

- The amount of information in a message.
- Entropy provides a degree of a set's predictability

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{p(x)}$$

- Low entropy of X means low uncertainty and as a result, high predictability of X's values.

CONDITIONAL ENTROPY

- Quantification of the amount of uncertainty in predicting the value of the discrete random variable Y given X.

$$H(Y|X) = -\sum_{x \in \mathcal{X}} p(x) H(Y|X = x).$$

INFORMATION GAIN

$$IG(Y|X) = H(Y) - H(Y|X).$$

- The difficulty of inferring the original text (Y), knowing its enciphered version (X)
- How the knowledge of X's value can reduce the uncertainty of inferring Y.
- Lower Information Gain means that it is difficult to infer the original text from the cipher.

RELATIVE INFORMATION GAIN

$$RIG(Y|X) = \frac{IG(Y|X)}{H(Y)}$$

- Information Gain depends on the size of the measured dataset.
- Relative Information Gain on the other hand, provides a normalized scale.

FAKE RECORD GENERATION METHODOLOGIES

- Uniform Ciphertext / Uniform Plaintext
- Uniform Ciphertexts by Swapping Plaintexts
- k-anonymous Ciphertexts

UNIFORM CIPHERTEXT/UNIFORM PLAINTEXT (UCUP)

- Intuitive approach
- To reduce RIG, plaintexts and ciphertexts appear equal number of times
- Inject fake records so that all ciphers map to an equal number of surnames

UNIFORM CIPHERTEXTS BY SWAPPING PLAINTEXTS (UCSP)

- Calculate the average number of plaintext occurrences $\lceil K \rceil$ for each Soundex code
- For Soundex codes with more than $\lceil K \rceil$ *occurrences, remove the* plaintexts redundant occurrences
- Add an equal number of fake occurrences for Soundex codes with less than $\lceil K \rceil$ *appearances,*
- *E*ach Soundex code appears exactly $\lceil K \rceil$ *times.*

UNIFORM CIPHERTEXTS BY SWAPPING PLAINTEXTS (UCSP)

- For:
  - Avoid oversized datasets
- Against:
  - Removed plaintexts will have to be separately matched.

K-ANONYMOUS CIPHERTEXTS

- Same intuition with Sweeney's k-anonymity
- Create datasets so that each Soundex code reflects to at least k Surnames.
- Parametric approach with k as its tuning parameter.
- For each Soundex code with less than k Surnames we inject fake surnames.
- Tunable by means of the k parameter.

EMPIRICAL EVALUATION

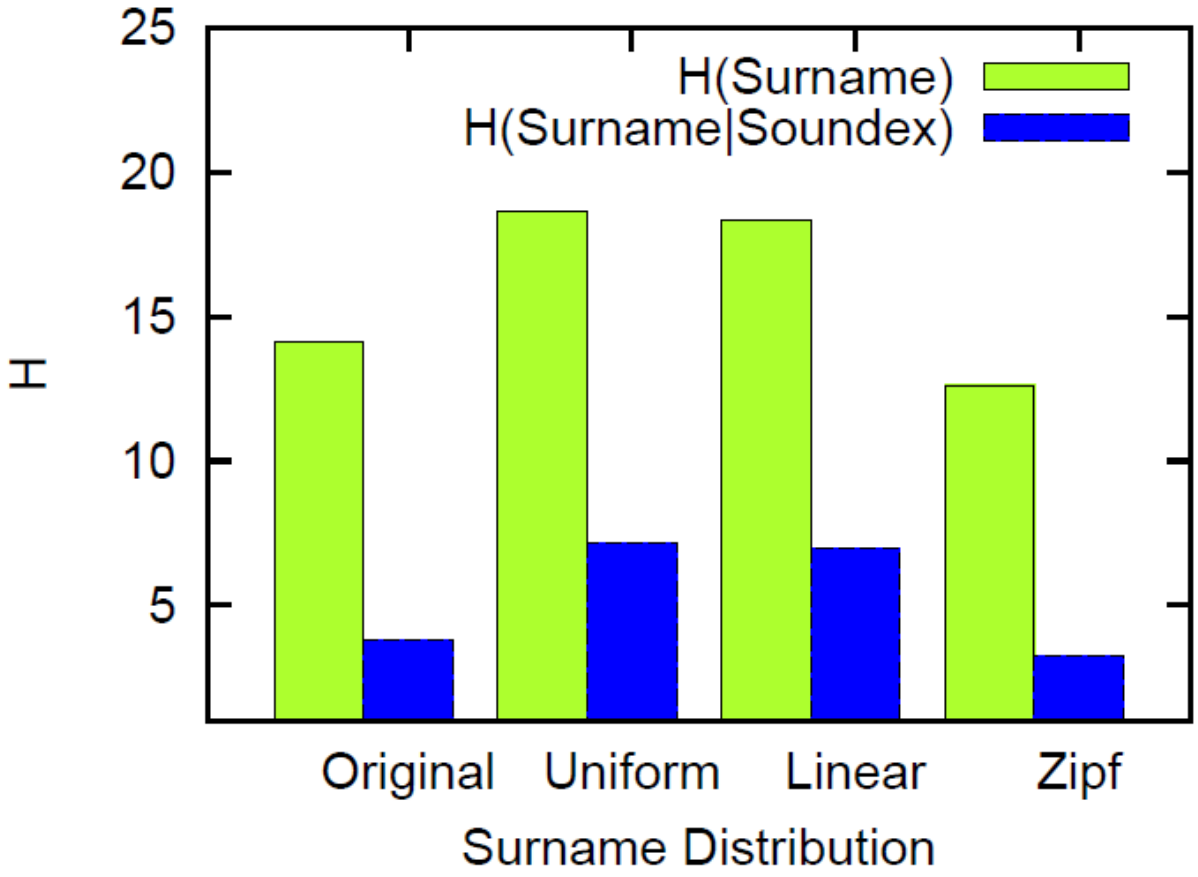
- Four datasets with different distributions
- Real world and synthetic data
- Study on a single (Surname) field

Dataset	Distribution	Number of records	K
O	Original	6917514	75087
L	Linear	81867776403	364928606
U	Uniform	404642	1462
Z	Zipf	5443039	410007

SOUNDEX INHERENT INFORMATION GAIN

- Assess the amount of information hidden by Soundex
- Calculate
  - Entropy H(Surname) and
  - Conditional Entropy H(Surname|Soundex)

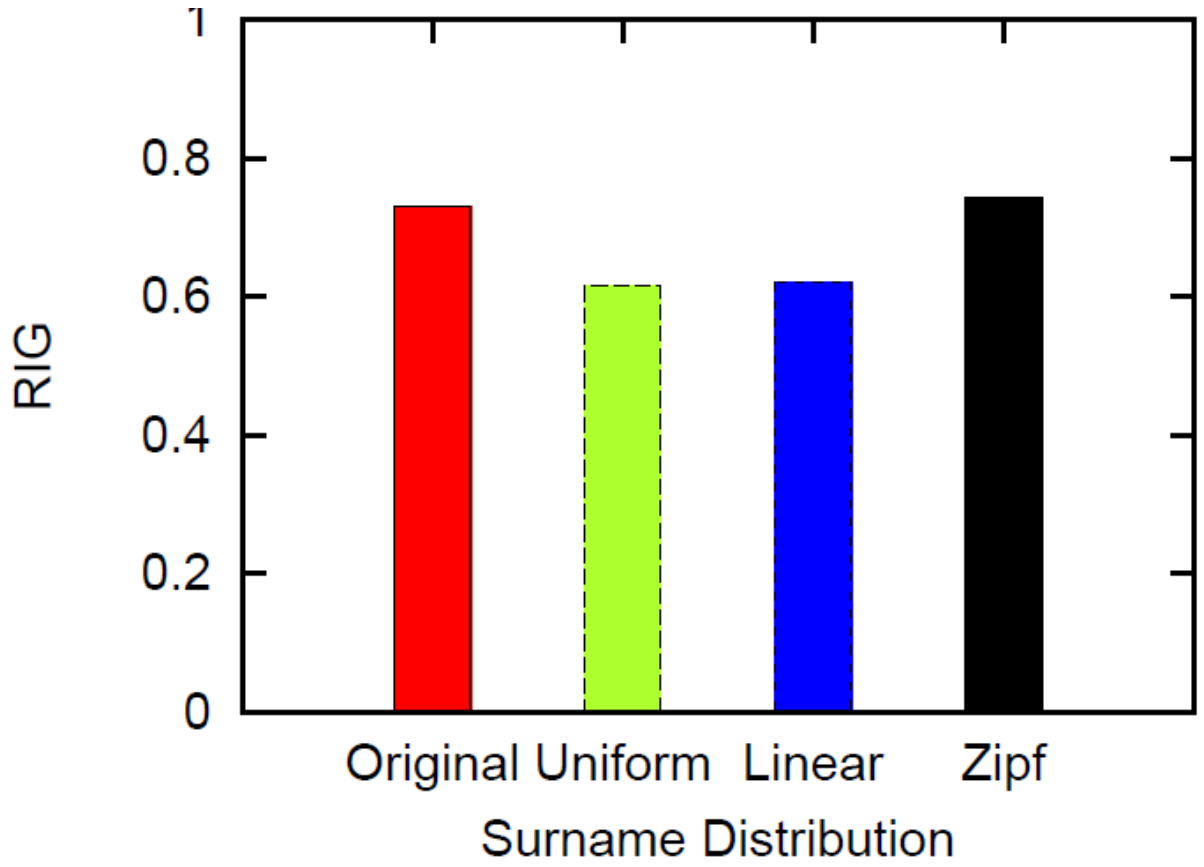
ENTROPIES FOR SURNAME DISTRIBUTIONS



SOUNDEX INHERENT INFORMATION GAIN

- Drop in RIG represents how much privacy we gain.
- Quantitatively measure the inherent reduction in RIG that the Soundex algorithm provides

RIG FOR EACH SURNAME DISTRIBUTION



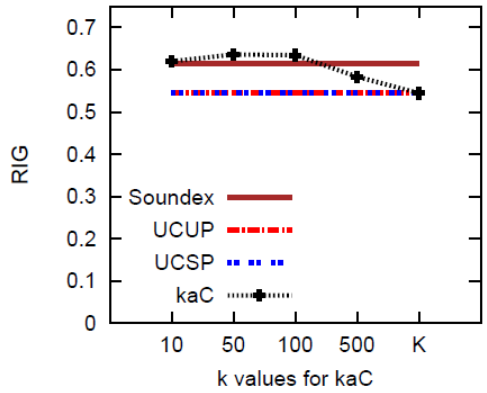
HIGHER PRIVACY BY FAKE INJECTION

- Use fake records in order to further reduce RIG
- Results for
  - UCUP
  - UCSP
  - kaC

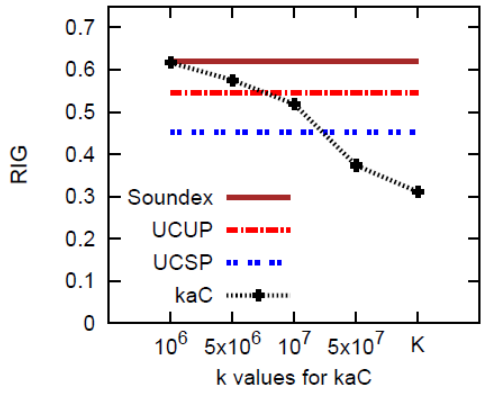
RIG DROP ASSESSMENT

- Determine privacy gain by each fake injection strategy
- Measure results for all four distributions

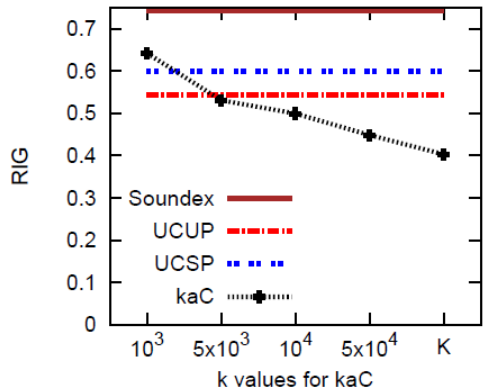
RELATIVE INFORMATION GAIN



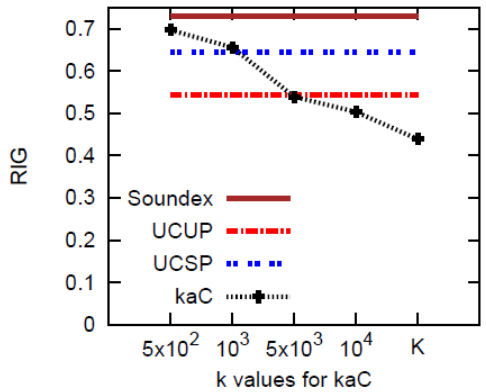
(a) Uniform Distribution



(b) Linear Distribution



(c) Zipf Distribution

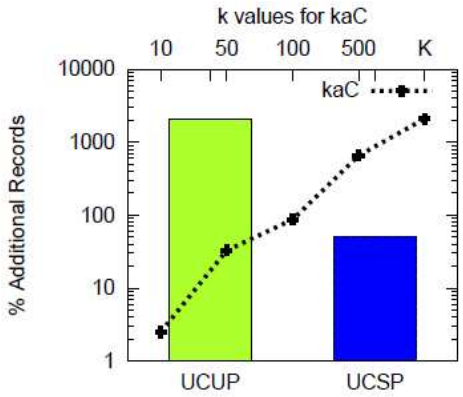


(d) Original Distribution

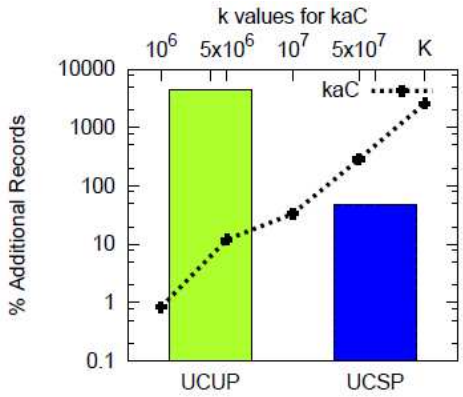
DATA QUALITY ASSESSMENT

- Determine impact on data quality
- Estimate the number of additional records required by each strategy
- Gather results for all four distributions

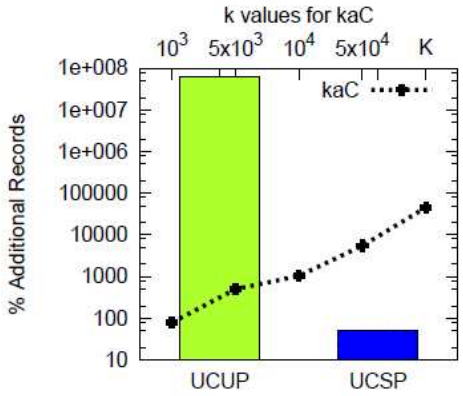
FAKE RECORDS OVERHEAD



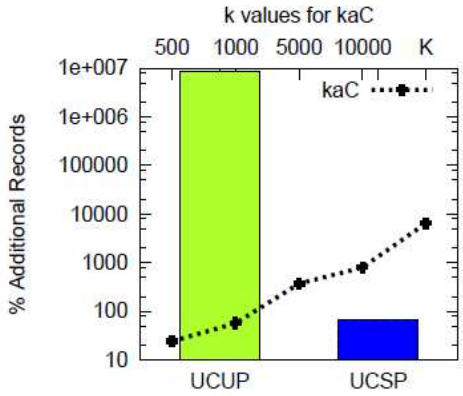
(a) Uniform Distribution



(b) Linear Distribution



(c) Zipf Distribution



(d) Original Distribution

RELATED WORK

- Private record matching using differential privacy, Inan et al (2010)
- Privacy-preserving record linkage using Bloom filters, Schnell et al (2009)
- Privacy preserving schema and data matching, Scannapieco et al (2007)

CONCLUSIONS AND FUTURE WORK

- Privacy without complicated encryption schemes
- Use more fields
- Probabilistic alternative of Soundex
- Experiment with more phonetic algorithms
- And many more…

ACKNOWLEDGMENTS

- This research is partially supported by the FP7 ICT/FET Project MODAP (Mobility, Data Mining, and Privacy) funded by the European
- Visit us here: www.modap.org.

THANK YOU

For your attention