大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Human and Social Data Sharing
for Research and Education

Utility-enhanced k-Anonymity Algorithm for Non-Uniform Datasets

ISSI 2012, Tokyo

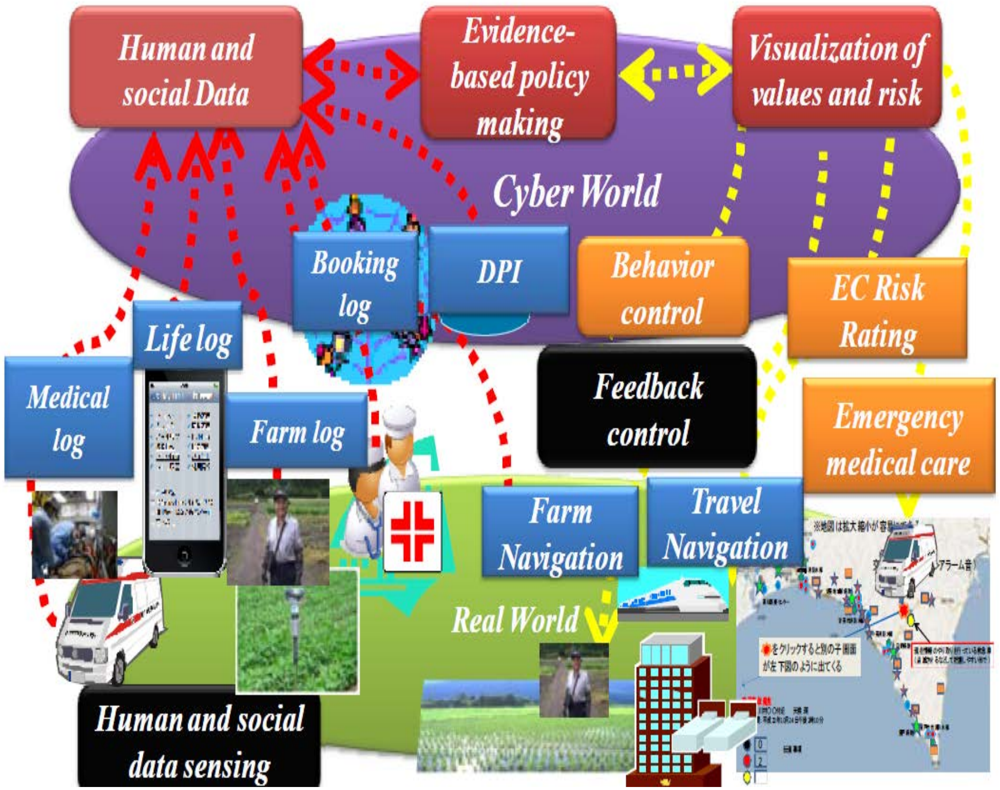
Lei Zhong, Ph. D.
Transdisciplinary Research Integration Center (TRIC),
National Institute of Informatics (NII), Japan
Email: lei@nii.ac.jp

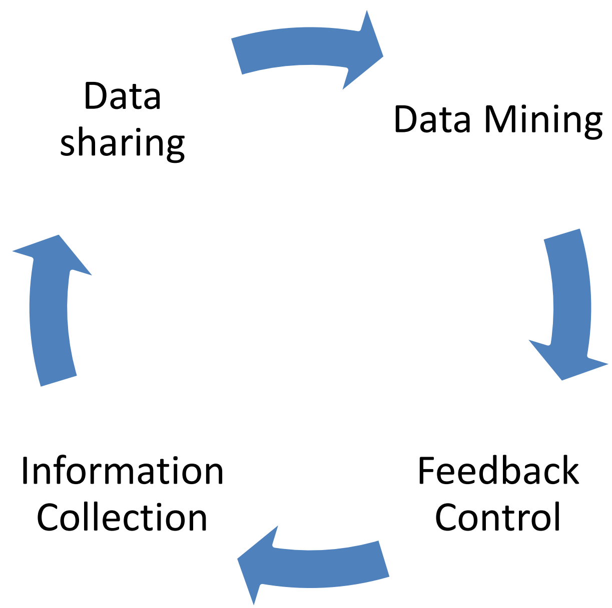
Outline

- Introduction
- Preliminaries
- Spatial Representation Model
- M&S Algorithm and Some Results
- Conclusions

Introduction

Information circulation in Cyber-Physical Systems





Data Sharing Example

- Suppose a hospital has some person-specific patient data which it will publish such that:
  - Information remains practically truthful and useful
  - Identity of an individual record cannot be determined

Identifier	Quasi-identifier (QI)		Sensitive Attribute (SA)
Name	Country	Gender	Disease
Allen	U.K.	M	prostate cancer
Bob	Spain	M	diabetes
Calvin	Hungary	M	heart disease
David	Poland	M	diabetes
Eve	U.S.	F	HIV
Grace	Canada	F	HIV

Records Linkage Risk

Hospital Patient Data

DOB	Gender	Zip code	Disease
1/21/76	Male	53715	Heart Disease
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Bronchitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Flu
2/28/76	Female	53706	Hang Nail

Vote Registration Data

Name	DOB	Gender	Zip code
Beth	1/10/81	Female	55410
Carol	10/1/44	Female	90210
Dan	2/21/84	Male	02174
Andre	1/21/76	Male	53715
Ellen	4/19/72	Female	02237

Andre has heart disease

Most promising solution: k-anonymity (Sweeney, 2002)

L. Sweeney. K-anonymity: A model for protecting privacy. International Journal on Uncertainty Fuzziness and Knowledge based Systems, 2002

k-anonymity

- *Definition*: A data set is called k-anonymity, if and only if each record on its QI appears at least k times.
- *Methods*: Generalization, Suppression

DOB	Gender	ZIP	Disease
76	Male	537**	Heart Disease
76-86	Female	537**	Hepatitis
76	Male	537**	Brochitis
76	Male	537**	Broken Arm
76-86	Female	537**	Flu
76-86	Female	537**	Hang Nail

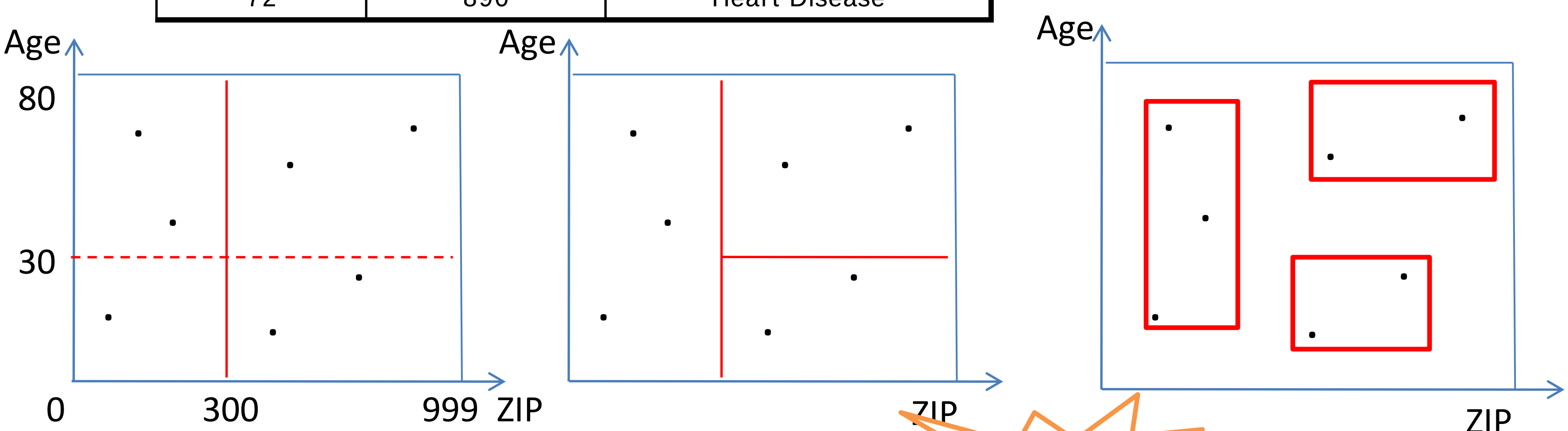
Example: 3-anonymity

Classification of Generalization Method

- A group of records on a QI attribute mapping to the same domain

Age	ZIP	Disease
71	103	Heart Disease
41	210	Hepatitis
19	090	Bronchitis
18	400	Broken Arm
67	405	Flu
28	550	Hang Nail
72	890	Heart Disease

2-anonymity



Single-dimensional Global Generalization

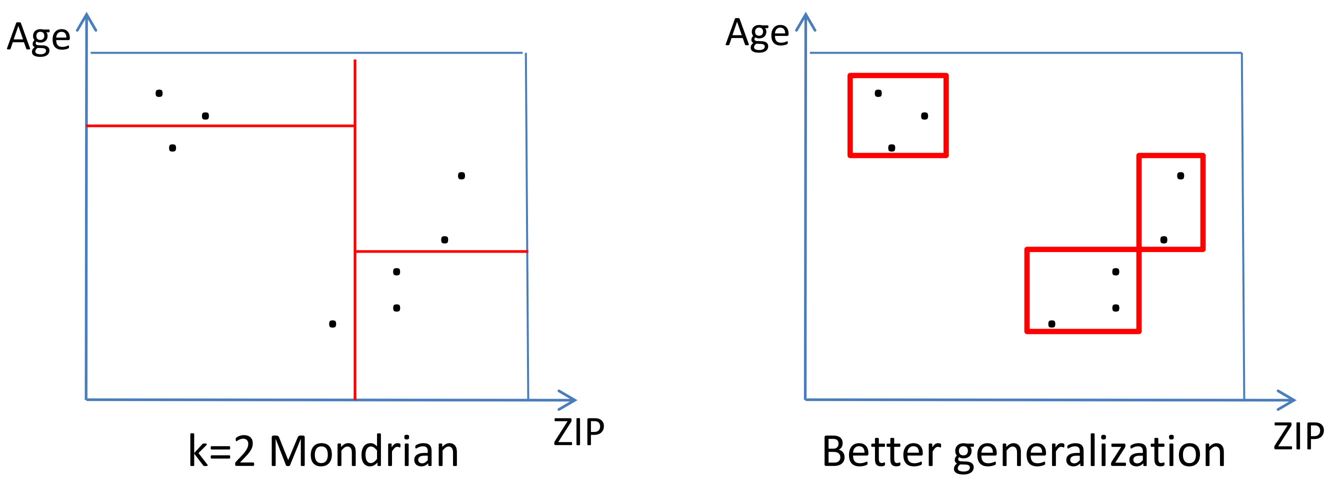
Single-dimensional Local Generalization

NP-hard

Multi-dimensional Local Generalization

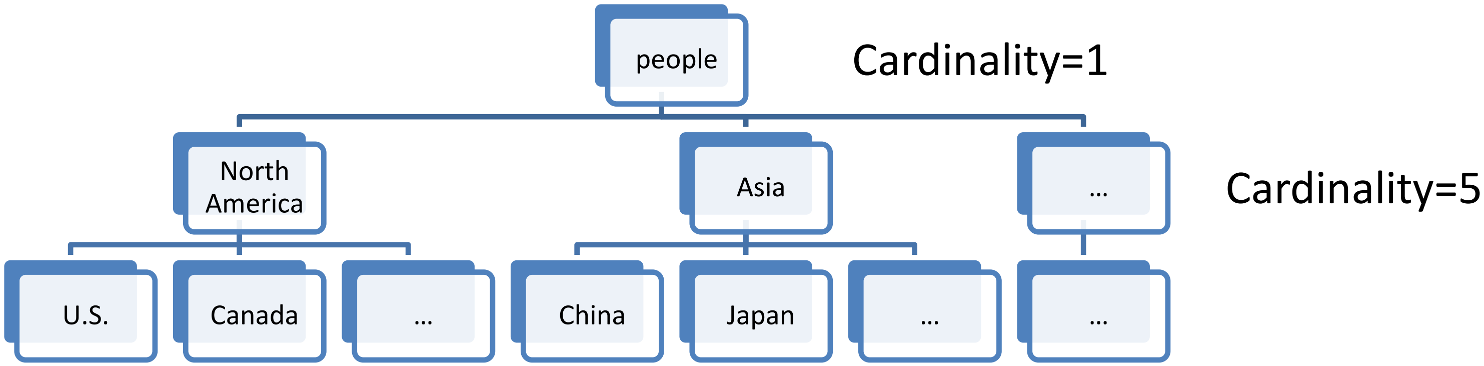
Issues for Existing Algorithms

- Mondrian—A greedy algorithm based on Multi-dimensional local Generalization Model
- More efficient performance and higher-quality results than Optimal Single-dimensional Global Generalization
- Drawbacks:
  - large group size (upper bound is 2*k*-1)
  - Equally partition cause utility loss for non-uniform dataset



Spatial Representation Model

- A dataset *T* with *n* records and *m* attributes can cast as *n* dots distributed in a *m*-dimensional Euclidean space
- Unified measurement (spatial distance) for both numerical and categorical attributes (dimensions)
  - Numerical attributes such as Age, Salary can be normalized with [0,1]
  - Categorical attributes such as Nationality, Profession can also be numerated and normalized



Normalized distance between two Nationality is defined as one of cardinality their same parent node

- Higher distance means higher generalization cost, i.e. information lost

Merge and Split Algorithm

Input: original Table *T*, hierarchies on categorical attributes

Output: a *k*-anonymous table *T'*

Initialization: every record in *T* forms a single-record E_i , a merge set $M = [E_i$,
a final equivalent class set $E = ;$

WHILE $M \neq ;$ {

FOR each E in M {

Scan all neighbor equivalent class to find a $E^{\prime}$ such that $IPG(E [E^{\prime})$ is the largest,

Merge $E^{\prime}$ into E , and delete $E^{\prime}$ from M

IF $IPG(newE) \neq IPG(oldE)$

Split_Flag=1

WHILE Split_Flag=1{

Scan E to find a record r such that $IPG(E \circ f r g)$ is the largest,

IF $IPG(E \circ f r g) > IPG(E)$

Split r form E , and move r to M

ELSE

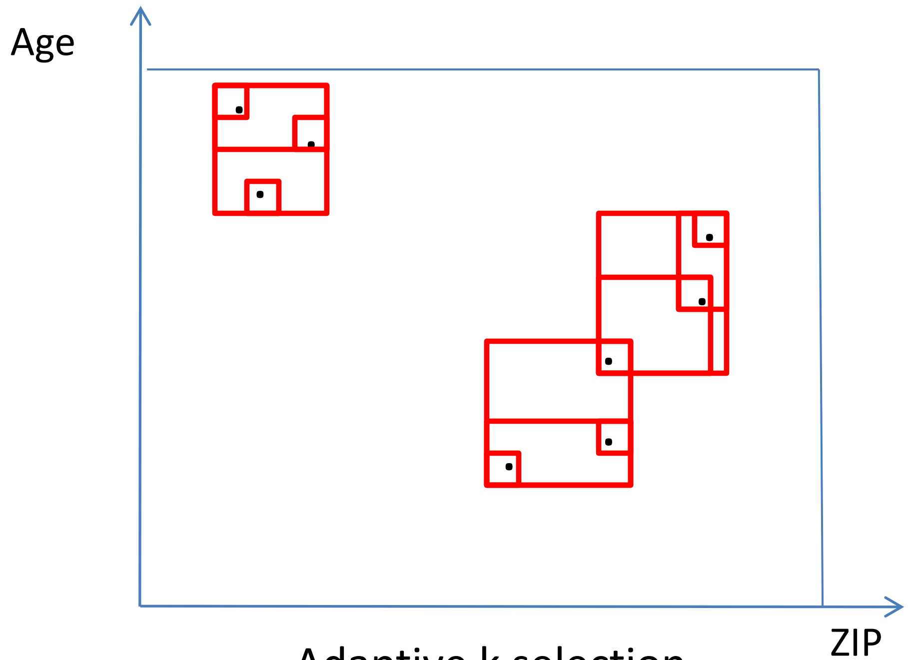
Move E to E

}

}

Generalize all equivalent classes in set E and output the table *T'*

Simple Example



Adaptive k selection

Metrics for Evaluation

- Normalized Information Loss Metric

$$IL_{E_i} = \frac{\prod_n (Max^{A_n}_{E_i} - Min^{A_n}_{E_i})}{|T|}$$

- Privacy Gain Metric

$$PG_{E_i} = 1 - \frac{1}{|E_i|}$$

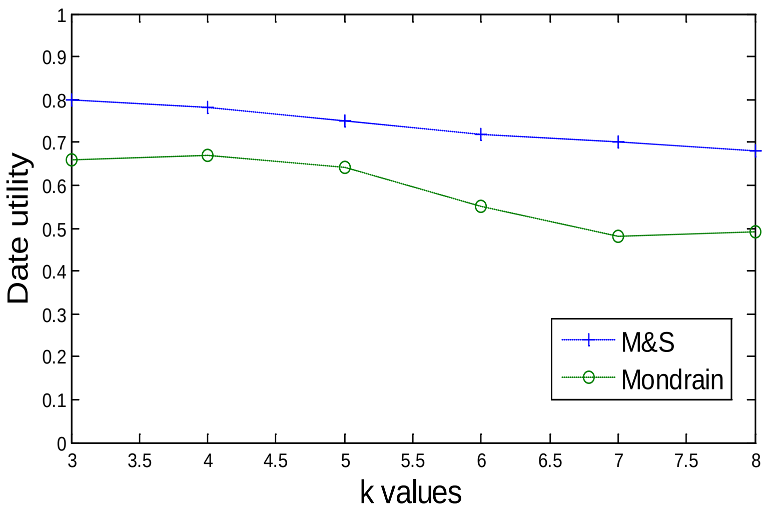
- Anonymization Quality Metric

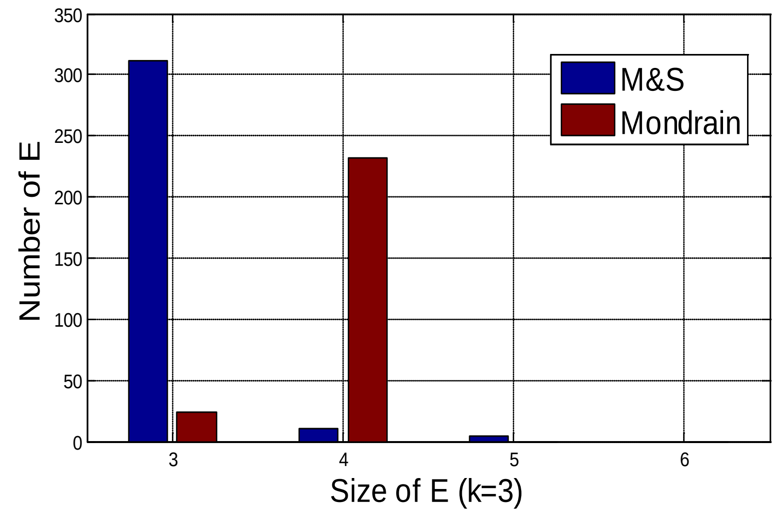
$$IPG_{E_i} = (1 - IL_{E_i}) \times PG_{E_i}$$

E_i : Equivalent class i:
A_n : Attribute n:
T : Orginal Dataset:

Preliminary Results

Experiment Setup:
1000 record, with 2 quasi-attributes, uniform distribution in each attribute





1. Our algorithm achieve higher utility since all the size of anonymized group are more close to k.
2. For non-uniform dataset, our proposed algorithm can provide flexible anonymized group size.

Conclusions

- Proposed algorithm achieves better quality of output dataset in terms of utility and same efficient as Mondrian
- Also suitable for real-world non-uniform dataset
- Compatible with extra improvement of *k*-anonymity such as *l*-diversity
- Future work is to test more real-world datasets, and add more features as *l*-diversity

Thank you!

Any questions?