# An Efficient Local Recoding Anonymization in Data without Hierarchical Taxonomies
# 階層的分類を伴わない効果的な匿名化のための ローカルリコーディング

Mohammad Rasool Sarrafi Aghdam　　Professor Noboru Sonehara

## どんな研究？

個人情報を保護するために，データを公開する際にプライバシー保護のために広く用いられているモデルとしてk-匿名性（k-anonymity）がある。
k-匿名性とは，個人情報から個人を特定することを難しくするための技術である。しかしながら，k-匿名性を用いたデータは情報が喪失されるため正確さに欠ける。現存するk-匿名化アプローチは大規模な情報喪失という点において問題がある。そこで我々はattributes hierarchical taxonomies から独立した数値属性とカテゴリー属性を含むデータ間の距離の計算にもとづく新しいモデルとSpatialDistance (SD) ヒューリスティックアルゴリズムを提唱する。我々の研究では，SDは既存のよく知られているアルゴリズムに比べて有意に情報喪失を減少させることが示されている。


## Abstract

To protect the privacy of individuals, a model that is widely used for privacy preservation in publishing micro-data, is **k-anonymity**. It reduces the linking confidence between sensitive information and specific individual. However, k-anonymous dataset loses its accuracy due to the **information loss**. Most of the existing k-anonymization approaches suffer from huge information loss. We propose a new model and **SpatialDistance (SD)** heuristic algorithm based on **distance calculation** between tuples including **numerical** and **categorical** attributes which is independent of attributes hierarchical taxonomies and **enhances** the **data utility** significantly.


## 状況設定

**Successful Linking Attack**

| Tuple | Age | Gender | ZIP | Disease |
|---|---|---|---|---|
| t1 | 25 | Female | 22370 | HIV |
| t2 | 40 | Male | 55490 | Malaria |
| t3 | 35 | Female | 22410 | Asthma |
| t4 | 45 | Male | 55410 | Influenza |

**Patient Data** Released by Hospital

- The **disease** of **Anna** is disclosed and her privacy is violated.
- **High Data Utility** but **NO Privacy** in Original dataset

Privacy Violation

**Unsuccessful Linking Attack**

| Name | Age | Gender | ZIP |
|---|---|---|---|
| Mia | 35 | Female | 22915 |
| Jack | 40 | Male | 75942 |
| Anna | 25 | Female | 22370 |
| Sara | 35 | Female | 65784 |

**External Data**

**Clustering + Generalization = k-anonymity**

| Tuple | Age | Gender | ZIP | Disease |
|---|---|---|---|---|
| t1 = t1&t3 | 25 ~ 35 | Female | 22*** | HIV |
| t2 = t2&t4 | 40 ~ 45 | Male | 554** | Malaria |
| t3 = t1&t3 | 25 ~ 35 | Female | 22*** | Asthma |
| t4 = t2&t4 | 40 ~ 45 | Male | 554** | Influenza |

**2-Anonymous** Patient Dataset

- **Linking confidence** is reduced by ***1/k*** ratio
- The **privacy** of individuals is **protected** to some extent. However, the data is **NOT** as **accurate** as the data in original dataset

## 研究内容

**SD algorithm** suggests to **cluster** the tuples for **generalization** through **local recoding** based on **distance calculation** to **Enhance** the **Utility** of Data in k-anonymization.

**Total Distance ($t_i$, $t_j$) = D(Numerical($t_i$, $t_j$) ) + D(Categorical($t_i$, $t_j$) )**

- **Numerical Attribute:** $D(t_i, t_j)_{Att_x} = \frac{|x_i - x_j|}{R(Att_x)}$
- **Categorical Attribute:**
  - Cardinality ($Att_x$) = 1 then Distance (a, a) = 0
  - Cardinality($Att_x$) = 2 then Distance (a, b) = 1
  - Cardinality($Att_x$) > 2
    1. Measuring similarity based on observation probability
    2. Defining distance based on measured similarities

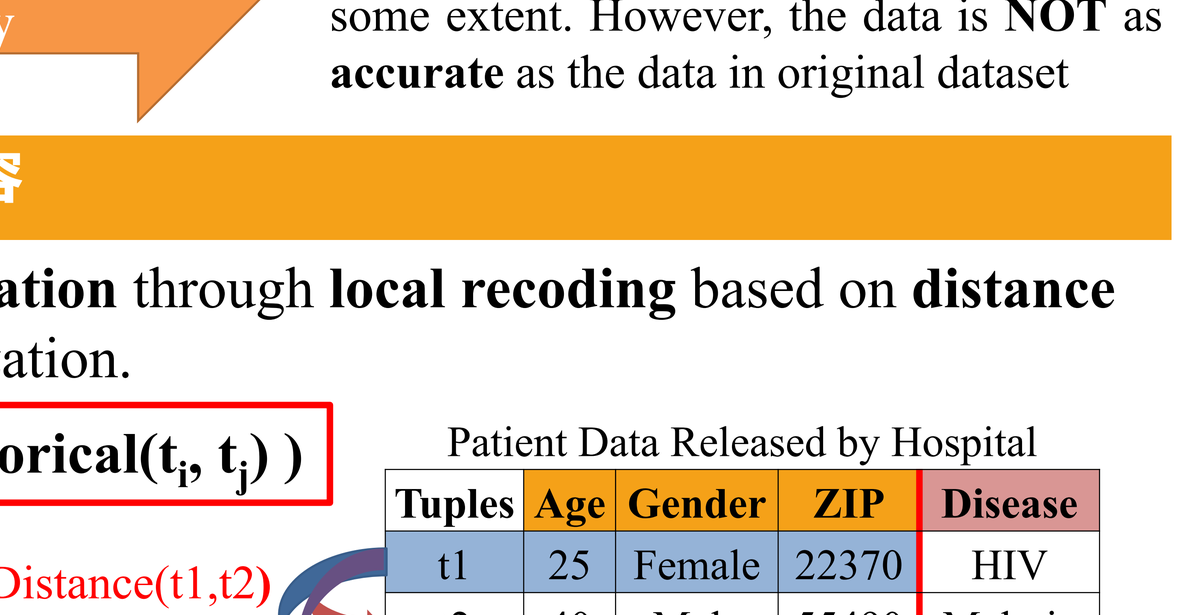
| | Japan | USA | China |
|---|---|---|---|
| Male | 5 | 5 | 1 |
| Female | 3 | 0 | 6 |

- Assuming, values in tuple 1 are **Male** (Gender), **Japan** (Nationality)
- **Japan** is **closer** to **USA** than **China** based on the contingency table
  - Because the probability of being **Male from Japan** is **closer** to **Male from USA**

Patient Data Released by Hospital

| Tuples | Age | Gender | ZIP | Disease |
|---|---|---|---|---|
| t1 | 25 | Female | 22370 | HIV |
| t2 | 40 | Male | 55490 | Malaria |
| t3 | 35 | Female | 22410 | Asthma |
| t4 | 45 | Male | 55410 | Influenza |

Distance(t1,t2)
**Distance(t1,t3)**
Distance(t1,t4)




連絡先：Mohammad Rasool Sarrafi Aghdam ／ 総合研究大学院大学　複合科学研究科情報学専攻
Email : rasool@nii.ac.jp