Methodology, Ethics and Practice of Data Privacy

实验部分

Lan Zhang

School of Computer Science and Technology

University of Science and Technology of China

Spring 2021

Part 1

K-Anonymity

K-Anonymity简介

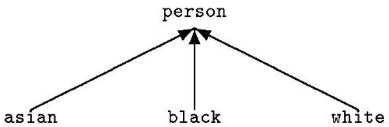
Every QI-cluster contains k or more tuples. (k=4)

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	20-29	Any	130**	Asian	Heart disease
2	Bruce	20-29	Any	130**	Asian	Heart disease
3	Cary	20-29	Any	130**	Asian	Viral infection
4	Dick	20-29	Any	130**	Asian	Viral infection
5	Eshwar	40-59	Any	14***	Asian	Cancer
6	Fox	40-59	Any	14***	Asian	Flu
7	Gary	40-59	Any	14***	Asian	Heart disease
8	Helen	40-59	Any	14***	Asian	Flu
9	Igor	30-39	Any	1322*	American	Cancer
10	Jean	30-39	Any	1322*	American	Cancer
11	Ken	30-39	Any	1322*	American	Cancer
12	Lewis	30-39	Any	1322*	American	Cancer

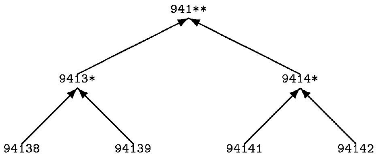
Identifier attributes **Quasi-identifiers, QI** **Sensitive attributes**

K-Anonymity 算法1 -- Samarati算法

- 技术：
  - Generalization: 泛化;
  - Suppression: 不发布/删除。
- 单个(categorical) Attribute：
  - 预先定义泛化层数，设可以删除的最大记录数$MaxSup$;
  - 先泛化到某一层，再删除记录数小于k的QI-cluster使得满足K-Anonymity（Full domain generalization）；



VGH_{R_0}



VGH_{Z_0}

[1] Samarati P. Protecting respondents identities in microdata release[J]. IEEE transactions on Knowledge and Data Engineering, 2001, 13(6): 1010-1027.

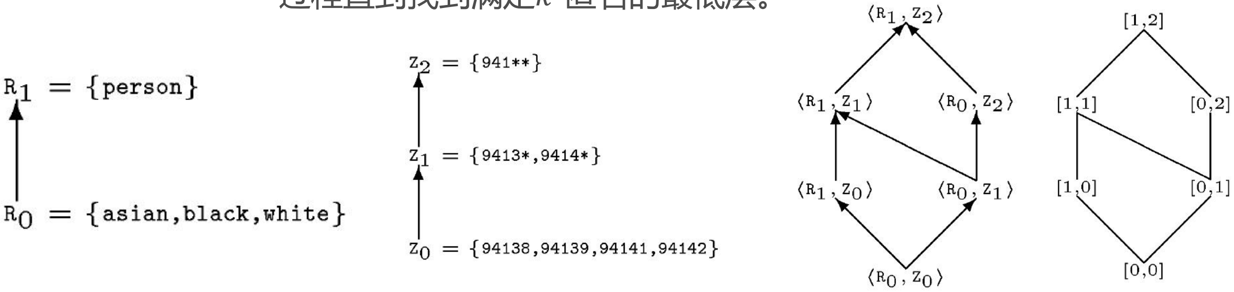
K-Anonymity 算法1 -- Samarati算法

多个(categorical) Attributes：

- 预定义泛化层数，构建lattice，如右下角的图；
- 例子：泛化到$< R_1, Z_1 >$对应的距离向量为[1, 1]；
- 要求泛化后的表格在满足K-Anonymity、删除的记录数不超过$MaxSup$的条件下，距离向量的元素之和尽可能得小。

基本过程（二分）：

- 结构高度为h(下面的例子$h = 3$)，检查高为$h/2$的节点能否满足k-匿名，满足则继续检查$h/4$高度的结点；否则检查$3h/4$高度的结点。重复这一过程直到找到满足k-匿名的最低层。



K-Anonymity 算法1 -- Samarati算法

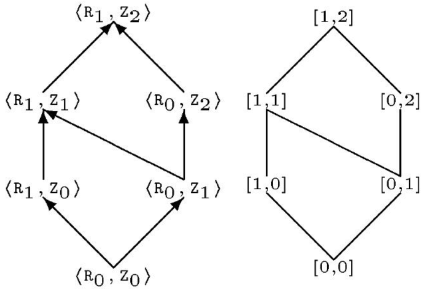
Find_vector

INPUT: Table T_i = $\mathsf{PT}[QI]$ to be generalized, anonymity requirement k, suppression threshold MaxSup, lattice VL_{DT} of the distance vectors corresponding to the domain generalization hierarchy DGH_{DT}, where DT is the tuples of the domains of the quasi-identifier attributes.

OUTPUT: The distance vector sol of a generalized table GT_{sol} that is a k-minimal generalization of $\mathsf{PT}[QI]$ according to Definition 4.3.

METHOD: Executes a binary search on VL_{DT} based on height of vectors in VL_{DT}.

```
1. $low$ :=0; $high$:=$height(\top, \mathsf{VL}_{DT})$; $sol$ := $\top$
2. while $low < high$
    2.1 $try := \lfloor \frac{low+high}{2} \rfloor$
    2.2 $Vectors := \{vec \mid height(vec, \mathsf{VL}_{DT}) = try\}$
    2.3 $reach\_k$:= false
    2.4 while $Vectors \neq \emptyset \wedge reach\_k \neq$ true do
            Select and remove a vector $vec$ from $Vectors$
            if satisfies($vec$,$k$,$T_i$,MaxSup) then $sol$:= $vec$; $reach\_k$:= true
    2.5 if $reach\_k$ = true then $high$:= $try$ else $low$:= $try + 1$
3. Return $sol$
```



- ⊤ 表示完全generalization的Table；
- VL_{DT}表示右下角的图；
- 2.2 Vectors:表示元素之和为try的距离向量的集合；

K-Anonymity 算法2 -- Mondrian算法

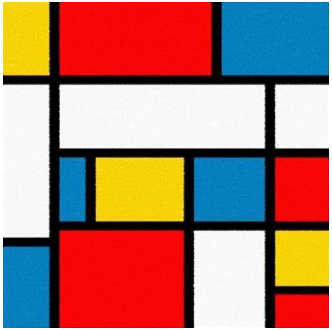
- 技术：
  - Generalization: 泛化；
- 单个（数值型）Attribute：
  - 以所有记录在该属性取值的中位数将记录划分为两部分，然后每一部分继续以中位数划分为两个区间（有两种方式）。
  - 重复这个过程，直到每个区间包含的记录数>=k，且不能再划分。此时每个区间都是一个等价类，记录泛化为对应范围。
- 中位数划分的两种方式（后面介绍第一种）：
  - 如$k = 2$， $\text{dataset} = [1, 2, 3, 3, 4, 5]$；
  - 第一种划分：[1, 2, 3, 3], [4, 5]; (strict partitioning)
  - 第二种划分: [1, 2, 3], [3, 4, 5]。

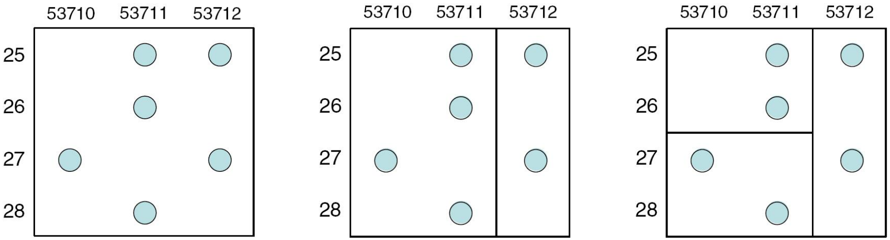
LeFevre K, DeWitt D J, Ramakrishnan R. Mondrian multidimensional k-anonymity[C]//22nd International conference on data engineering (ICDE'06). IEEE, 2006: 25-25.

K-Anonymity 算法2 -- Mondrian算法

» 多个（数值型）Attributes：

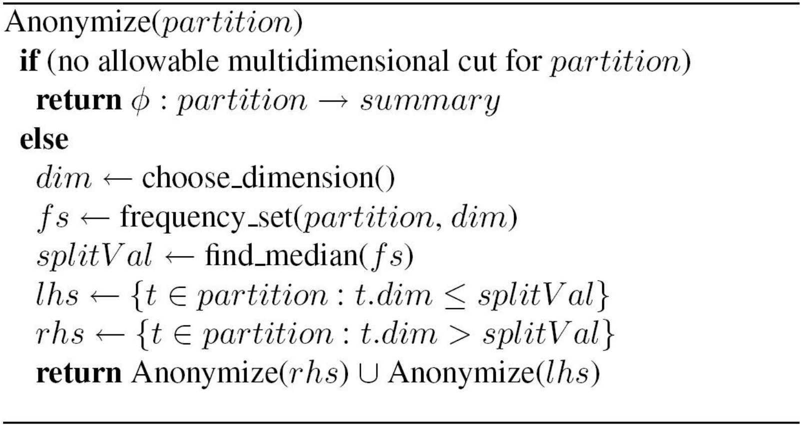
- 每个Partition单独选择一个属性，可以选择范围最大的属性，或者随机选；
- 找到属性的中位数，对Partition划分；
- 重复上述过程，直到不能划分为止。





K-Anonymity 算法2 -- Mondrian算法

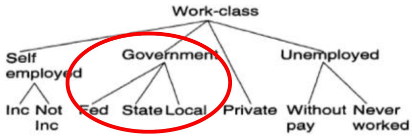
- 算法（strict multidimensional partitioning）：



```
Anonymize($partition$)
 if (no allowable multidimensional cut for $partition$)
   return $\phi : partition \rightarrow summary$
 else
   $dim \leftarrow$ choose_dimension()
   $fs \leftarrow$ frequency_set($partition, dim$)
   $splitVal \leftarrow$ find_median($fs$)
   $lhs \leftarrow \{t \in partition : t.dim \leq splitVal\}$
   $rhs \leftarrow \{t \in partition : t.dim > splitVal\}$
   return Anonymize($rhs$) $\cup$ Anonymize($lhs$)
```

评价指标 Loss Metric (LM)

- LM[1] is defined in terms of a normalized loss for each attribute of every tuple.

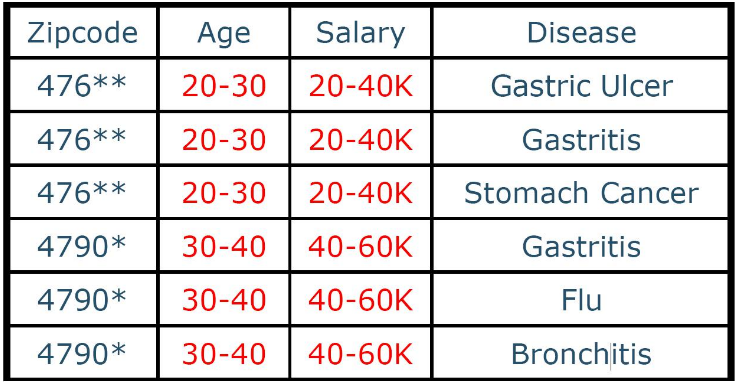


- Quantify the loss when a leaf node value cannot be disambiguated from another value due to generation.
- **Categorical attribute A:** For a tuple t, suppose the value of **t[A] has been generalized to x**. Letting |A| represent the total number of leaf nodes in the tree; Letting M represent the number of leaf nodes in the subtree rooted at x, then the **loss for t[A] is (M – 1)/(|A| – 1).**
- **What is the loss for “State”? 2/7**
- The loss for attribute A is the average of the loss for all tuples t. The LM for the entire data set is the sum of the losses for each attribute.

[1]V. S. Iyengar , “Transforming data to satisfy privacy constraints,” in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

评价指标 Loss Metric (LM)

- LM is defined in terms of a normalized loss for each attribute of every tuple.
- **Numerical information:** For a tuple t, suppose the value of t[A] has been generalized to an interval $[L_i, U_i]$. Letting the lower and upper bounds in the table for A be L and U. The normalized loss for this entry is given by $\boldsymbol{(U_i - L_i)/(U - L)}$**.**

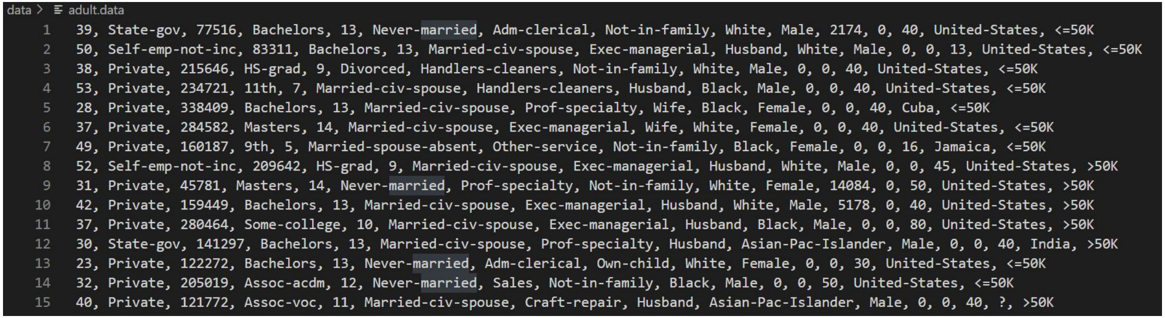


Zipcode	Age	Salary	Disease
476**	20-30	20-40K	Gastric Ulcer
476**	20-30	20-40K	Gastritis
476**	20-30	20-40K	Stomach Cancer
4790*	30-40	40-60K	Gastritis
4790*	30-40	40-60K	Flu
4790*	30-40	40-60K	Bronchitis

- The loss for age [20-30] is (30-20)/(40-20)

Adult数据集介绍（文件夹中有提供）

- 下载链接：https://archive.ics.uci.edu/ml/datasets/adult
- 有32561条数据，删除空的或有?的行后，剩余**30162**条。
- 15个attributes ['age', 'work_class', 'final_weight', 'education', 'education_num', 'marital_status', 'occupation', 'relationship', 'race', 'sex', 'capital_gain', 'capital_loss', 'hours_per_week', 'native_country', 'class']



```
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, <=50K
53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K
28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K
37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K
49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female, 0, 0, 16, Jamaica, <=50K
52, Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 45, United-States, >50K
31, Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, White, Female, 14084, 0, 50, United-States, >50K
42, Private, 159449, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 5178, 0, 40, United-States, >50K
37, Private, 280464, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Black, Male, 0, 0, 80, United-States, >50K
30, State-gov, 141297, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, Asian-Pac-Islander, Male, 0, 0, 40, India, >50K
23, Private, 122272, Bachelors, 13, Never-married, Adm-clerical, Own-child, White, Female, 0, 0, 30, United-States, <=50K
32, Private, 205019, Assoc-acdm, 12, Never-married, Sales, Not-in-family, Black, Male, 0, 0, 50, United-States, <=50K
40, Private, 121772, Assoc-voc, 11, Married-civ-spouse, Craft-repair, Husband, Asian-Pac-Islander, Male, 0, 0, 40, ?, >50K
```

实验要求

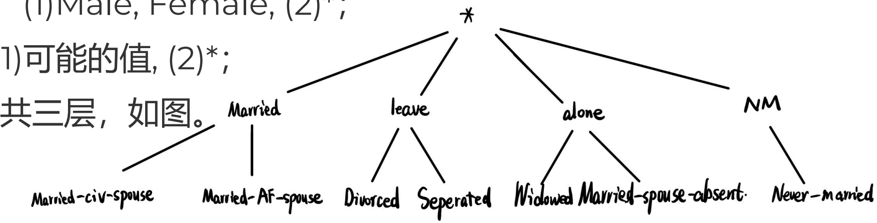
- 必做部分80：
  - 代码正确：50 （每个算法各25)
  - 代码清晰有注释：10
  - 实验报告，测试分析结果和讨论：20
- 选做部分20：
  - Samarati算法可能会有很多解满足要求，调研并探究如何选择输出保证结果的可用性尽可能大，说说你的启发，(e.g.:选用合适的评价指标评价不同的输出) 15
  - Mondrian算法处理categorical（如Gender） 5
- 代码抄袭0，会有查重

实验要求

- 实现k-Anonymity两种算法（后面两页有具体要求）：
  - Samarati算法（categorical 型）
  - Mondrian算法（数值型）
- 实验报告：问题描述、程序使用指南、实验结果分析、讨论与总结。

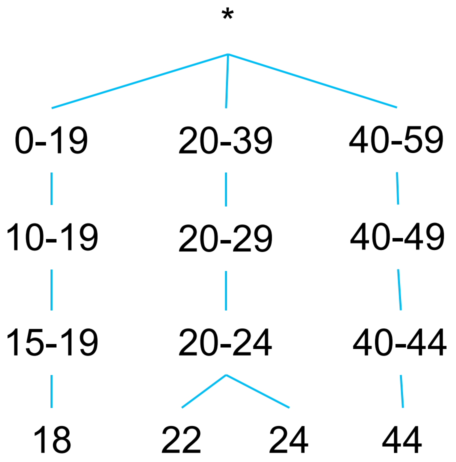
实验要求(Samarati算法)

- Samarati算法：
  - 使用Adult数据集；
  - QI={age, gender, race, marital _status} （ categorical 型），S = {occupation}
  - 输入：data, k, maxSup (data是数据集，k是K-Anonymity的参数，maxSup表示最大suppression的个数)；
  - 输出：匿名后的数据集。
  - 评价指标： 运行时间和LM 。
- 可取K=10,maxSup=20。并测试不同的$k, maxSup$对实验结果的影响。
- Age: 共五层，(1)原始值，(2)range-5, (3)range-10, (4)range-20, (5)*；
- Gender:共两层，(1)Male, Female, (2)*；
- Race: 共两层，(1)可能的值, (2)*；
- Marital_status:共三层，如图。



实验要求(Samarati算法)

- Age的层次类似左下图，共五层，(1)原始值，(2)range-5, (3)range-10, (4)range-20, (5)*;
- Gender、marital_status、race的层次以右下文件形式给出：
  - 子节点,父节点



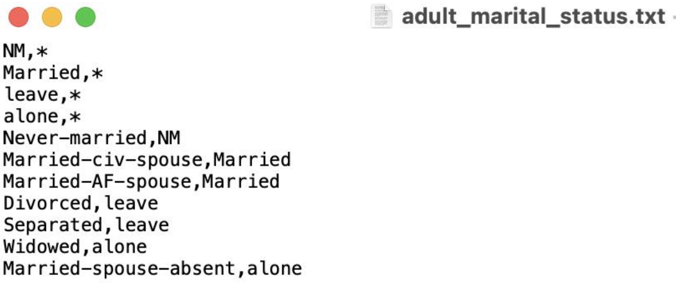
*

0-19 20-39 40-59

10-19 20-29 40-49

15-19 20-24 40-44

18 22 24 44



```
NM,*
Married,*
leave,*
alone,*
Never-married,NM
Married-civ-spouse,Married
Married-AF-spouse,Married
Divorced,leave
Separated,leave
Widowed,alone
Married-spouse-absent,alone
```

实验要求(Mondrian算法)

- Mondrian算法:
  - 使用Adult数据集;
  - QI={age, education_num}（数值型）, S = {occupation};
  - 输入：data, k (data是数据集，k是K-Anonymity的参数）；
  - 输出：匿名后的数据集。
  - 评价指标：运行时间和LM。
- 可取k=10 。并测试不同的k 对实验结果的影响。

参考资料

- Samarati P. Protecting respondents identities in microdata release[J]. IEEE transactions on Knowledge and Data Engineering, 2001.
- LeFevre K, DeWitt D J, Ramakrishnan R. Mondrian multidimensional k-anonymity[C].(ICDE'06). IEEE, 2006.
- V. S. Iyengar , “Transforming data to satisfy privacy constraints,” in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.
- https://blog.csdn.net/xff1994/article/details/83149116

THANKS!

Any questions?