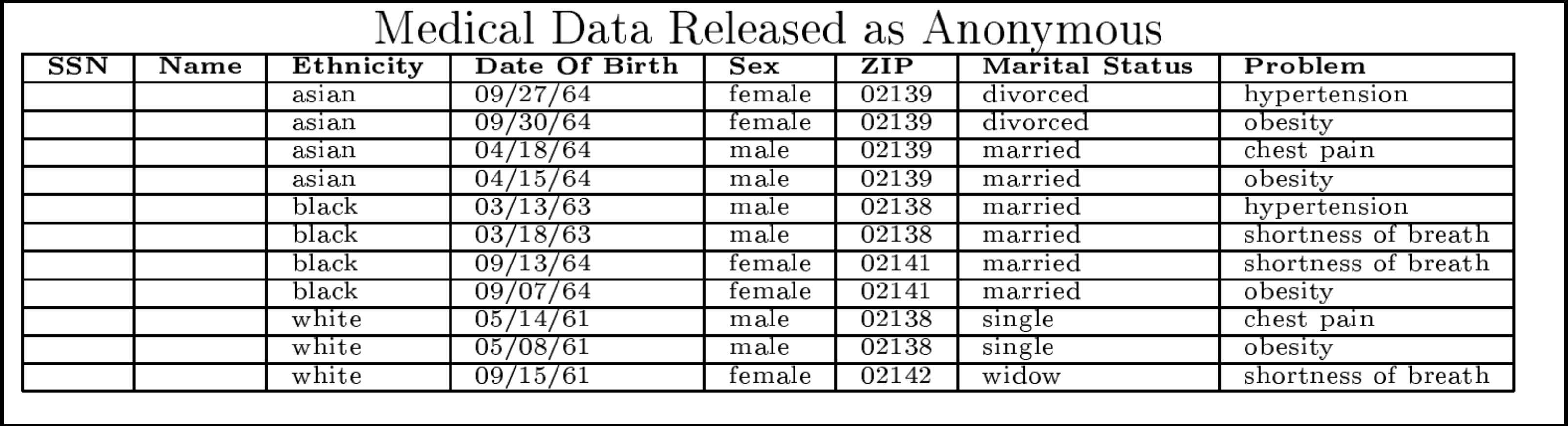
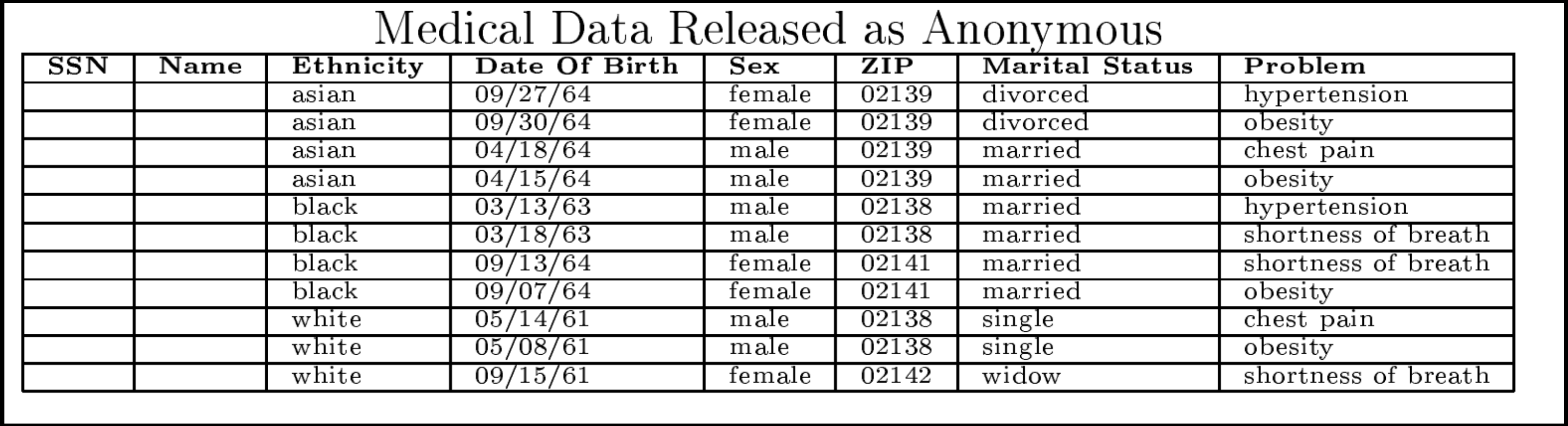
Disclosure risk



Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Terms

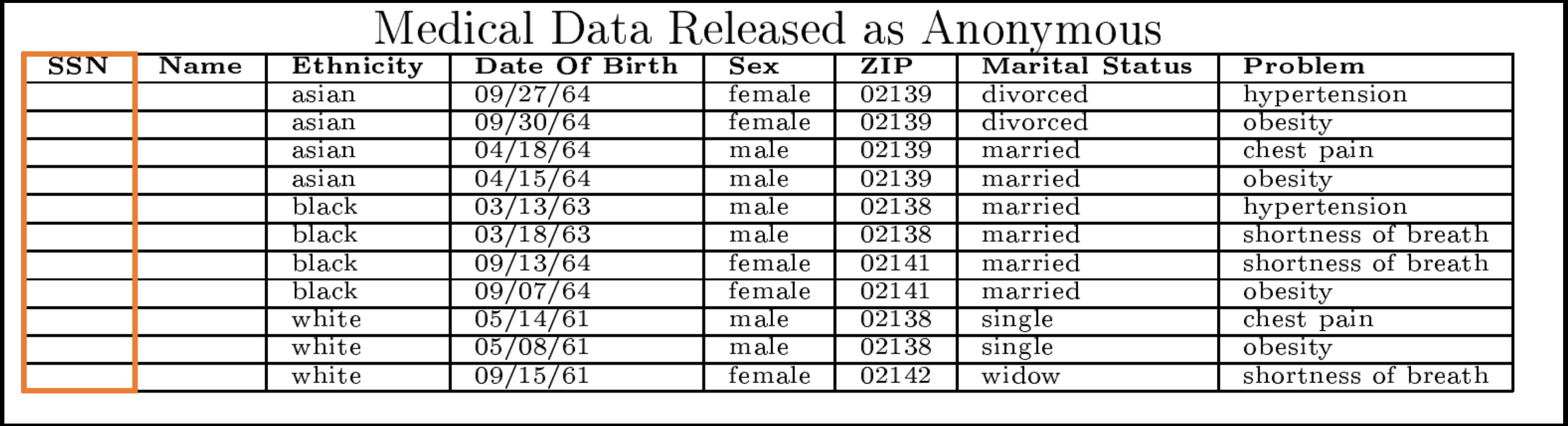


Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Attributes: Let T is a table with a finite number of tuples. The finite set of attributes of T are {A1,…,An}

Identifiers

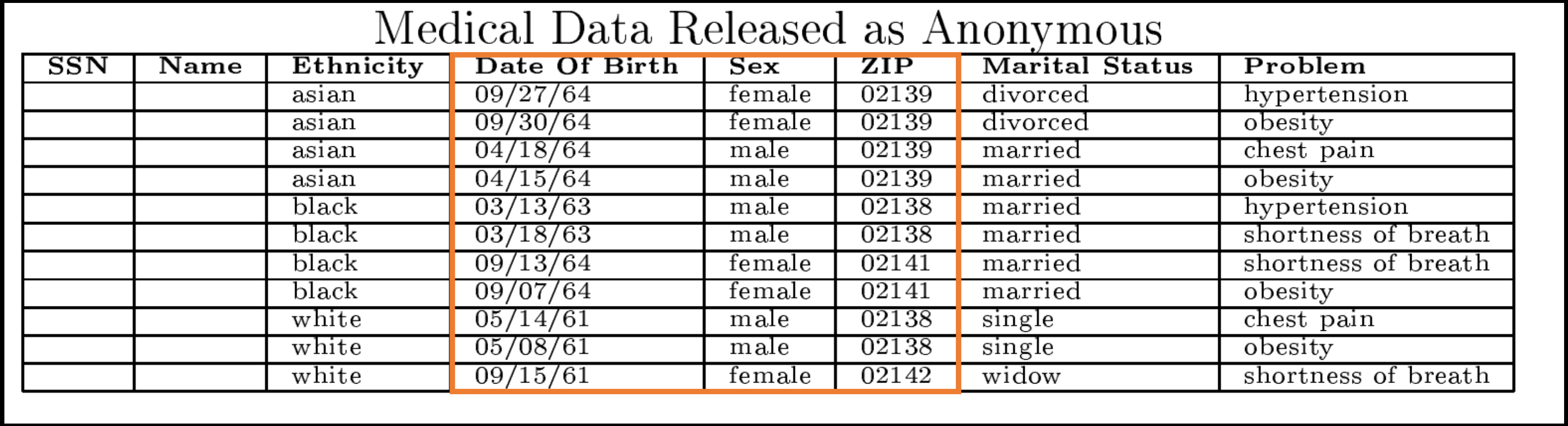


Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Given a population of entities U, an entity-specific table T(A1,…,An), $f_c: U \rightarrow T$ and $f_g: T \rightarrow U'$, where $U \subseteq U'$. A set of attributes $\mathcal{A}$ in Table $\mathcal{T}$ is an identifier if $\mathcal{T}[\mathcal{A}]$ can uniquely map an entity.

Quasi-identifiers



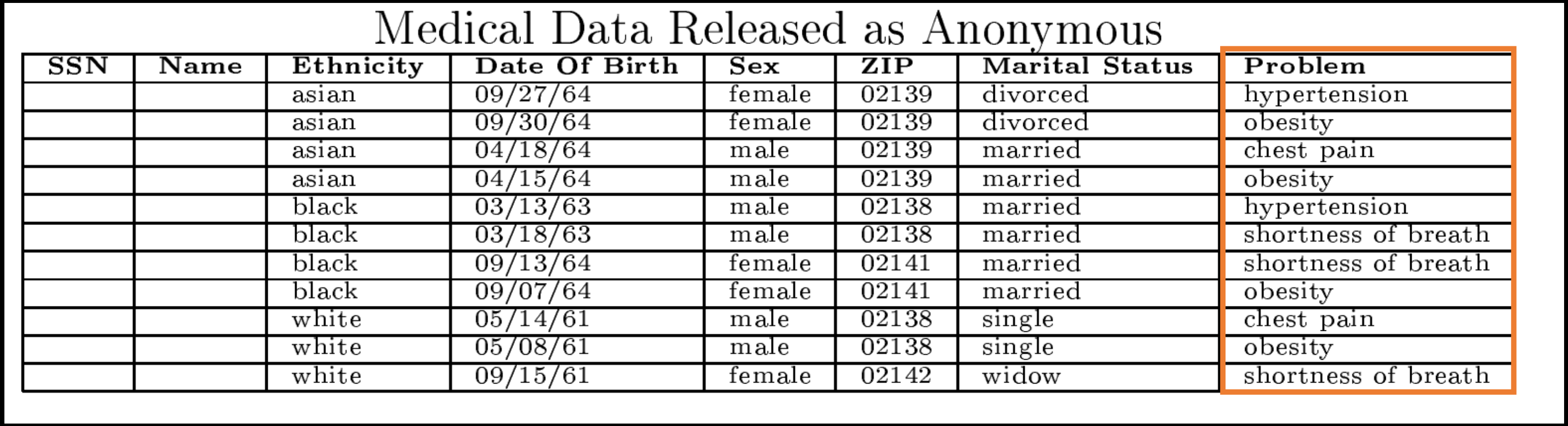
Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Quasi identifier: Given a population of entities U, an entity-specific table T(A1,…,An), f_c: $U \rightarrow T$ and f_g: $T \rightarrow U'$, where $U \subseteq U'$.

A quasi-identifier of T, written Q_T, is a set of attributes $\{Ai,…,Aj\} \subseteq \{A1,…,An\}$ where: $\exists p_i \in U$ such that $f_g(f_c(pi)[QT]) = p_i$.

Sensitive attributes

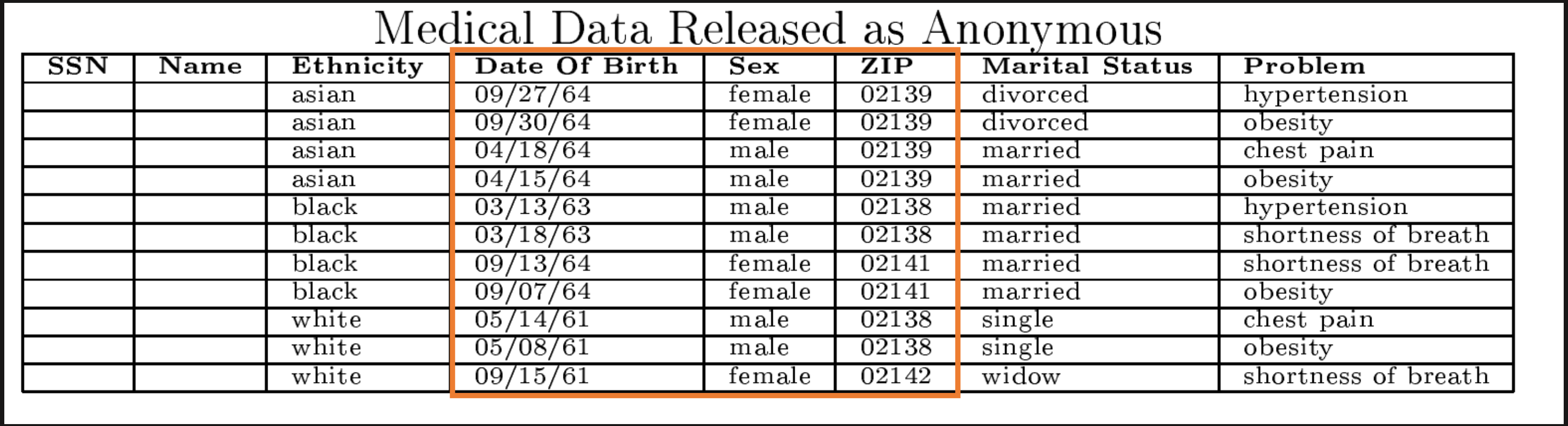


Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

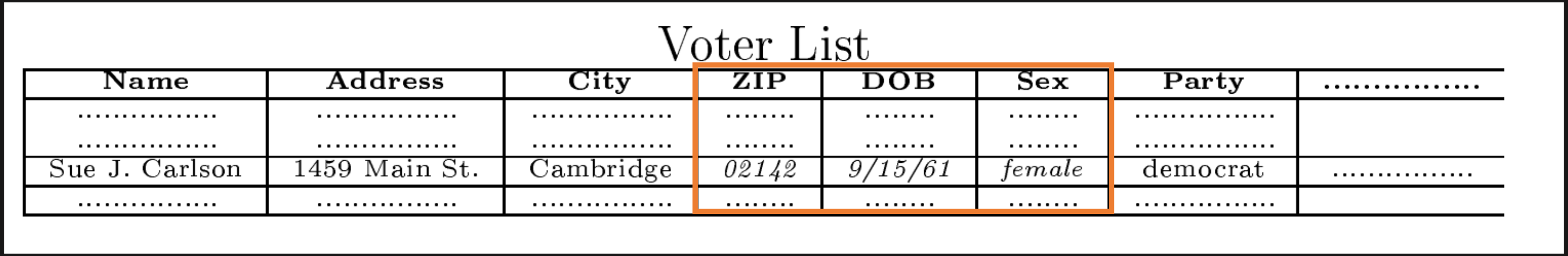
A *sensitive attribute* is an attribute whose value for any particular individual must be kept secret from people who have no direct access to the original data.

Linking attack with Quasi identifiers



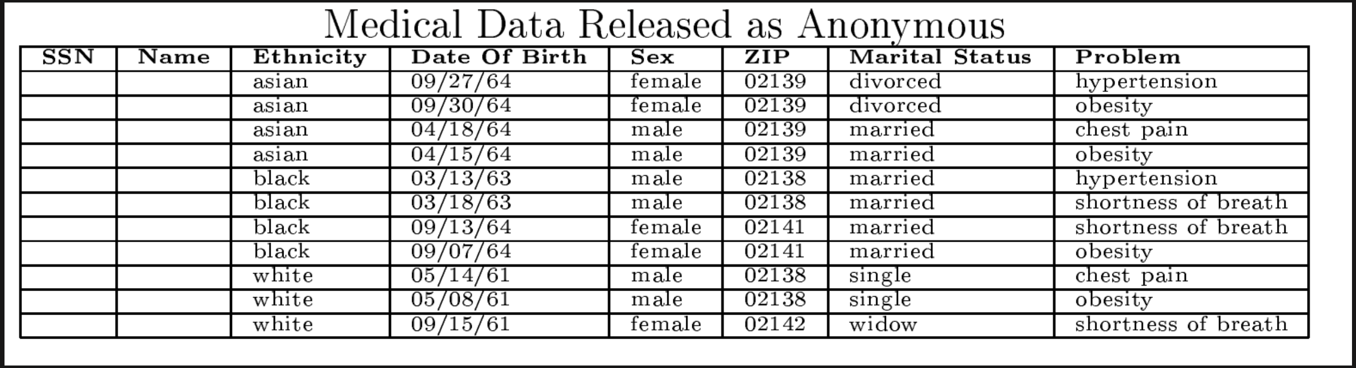
Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath



Voter List

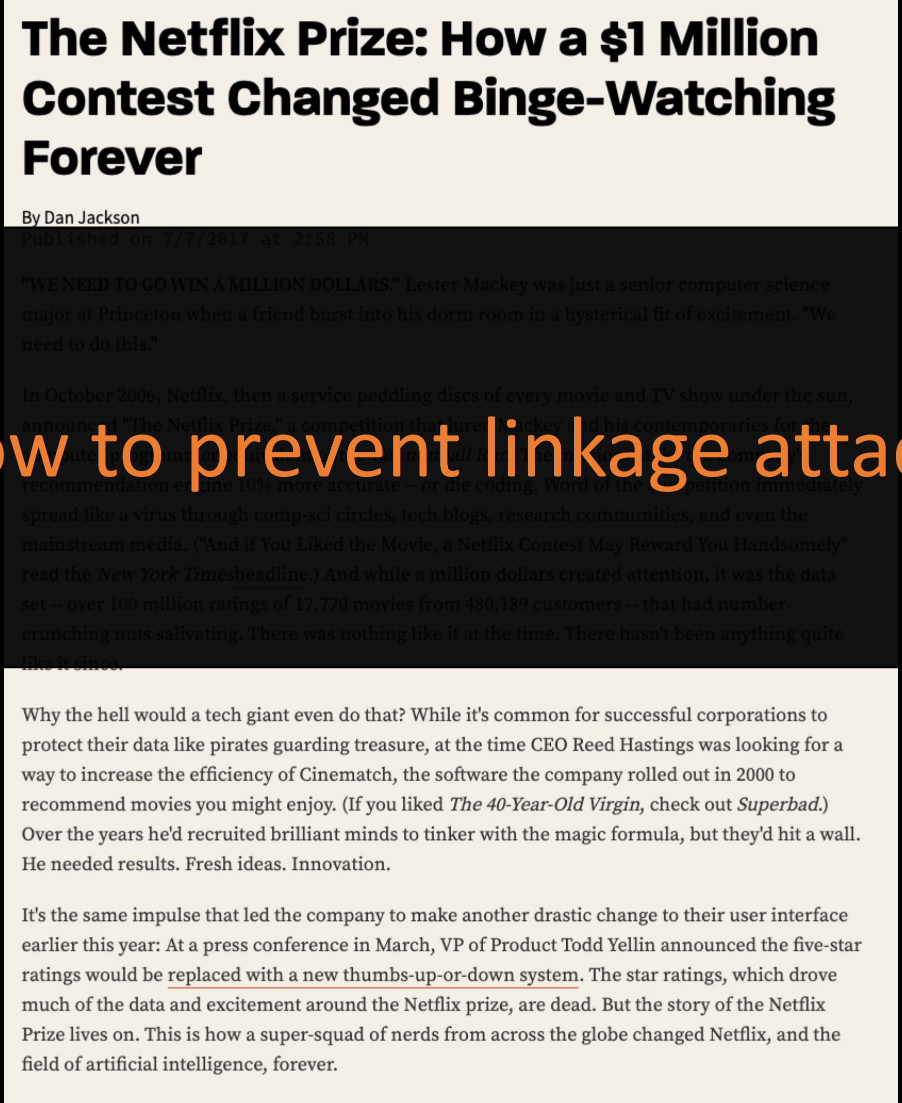
Name	Address	City	ZIP	DOB	Sex	Party	....
....	....	....	....	....	....	....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	....
....	....	....	....	....	....	....	



Medical Data Released as Anonymous

SSN	**Name**	**Ethnicity**	**Date Of Birth**	**Sex**	**ZIP**	**Marital Status**	**Problem**
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

How to prevent linkage attacks?



The Netflix Prize: How a $1 Million Contest Changed Binge-Watching Forever

By Dan Jackson

Published on 7/7/2017 at 2:58 PM

"WE NEED TO GO WIN A MILLION DOLLARS." Lester Mackey was just a senior computer science major at Princeton when a friend burst into his dorm room in a hysterical fit of excitement. "We need to do this."

In October 2006, Netflix, then a service peddling discs of every movie and TV show under the sun, announced "The Netflix Prize," a competition that lured Mackey and his contemporaries for the computer programmer equivalent of the *Cannonball Run*. The mission: Make the company's recommendation engine 10% more accurate -- or die coding. Word of the competition immediately spread like a virus through comp-sci circles, tech blogs, research communities, and even the mainstream media. ("And if You Liked the Movie, a Netflix Contest May Reward You Handsomely" read the *New York Times*headline.) And while a million dollars created attention, it was the data set -- over 100 million ratings of 17,770 movies from 480,189 customers -- that had number-crunching nuts salivating. There was nothing like it at the time. There hasn't been anything quite like it since.

Why the hell would a tech giant even do that? While it's common for successful corporations to protect their data like pirates guarding treasure, at the time CEO Reed Hastings was looking for a way to increase the efficiency of Cinematch, the software the company rolled out in 2000 to recommend movies you might enjoy. (If you liked *The 40-Year-Old Virgin*, check out *Superbad*.) Over the years he'd recruited brilliant minds to tinker with the magic formula, but they'd hit a wall. He needed results. Fresh ideas. Innovation.

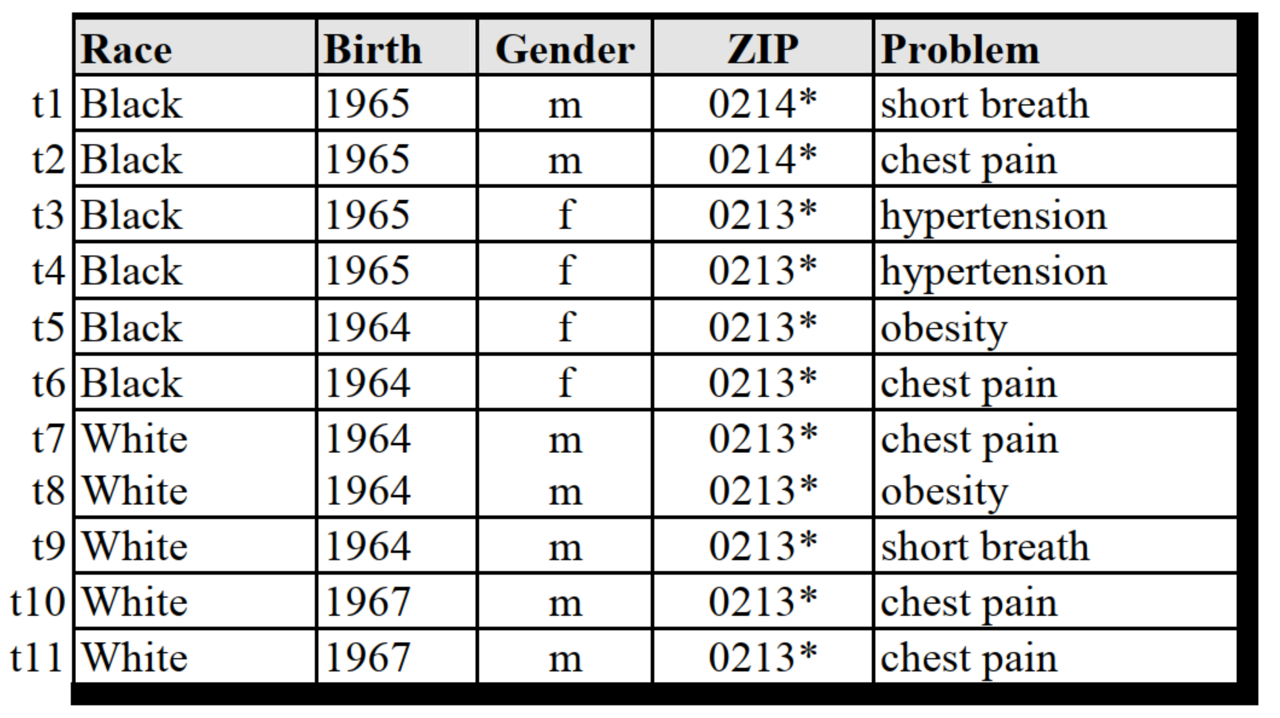
It's the same impulse that led the company to make another drastic change to their user interface earlier this year: At a press conference in March, VP of Product Todd Yellin announced the five-star ratings would be replaced with a new thumbs-up-or-down system. The star ratings, which drove much of the data and excitement around the Netflix prize, are dead. But the story of the Netflix Prize lives on. This is how a super-squad of nerds from across the globe changed Netflix, and the field of artificial intelligence, forever.

K-anonymity or *“hide in the crowd”*

Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals.

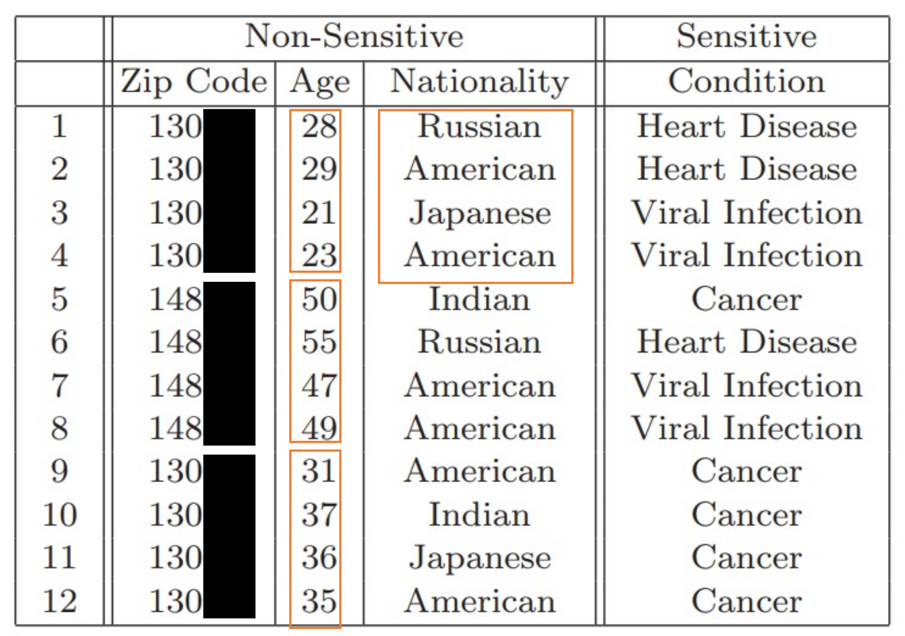
Let T(A1,...,An) be a table and QI_T be the quasi-identifier associated with it. T is said to satisfy k-anonymity if and only if each sequence of values in $T[QI_T]$ appears with at least k occurrences in $T[QI_T]$.

2-anonymous table

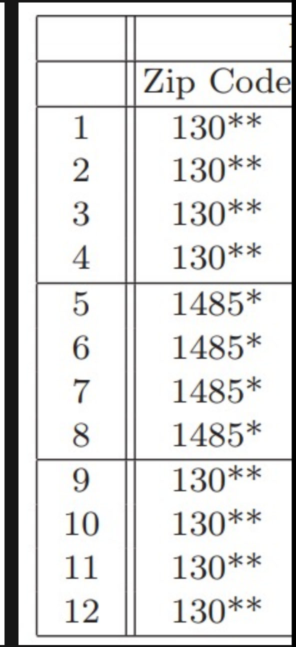


	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

4-anonymous table



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130[redacted]	28	Russian	Heart Disease
2	130[redacted]	29	American	Heart Disease
3	130[redacted]	21	Japanese	Viral Infection
4	130[redacted]	23	American	Viral Infection
5	148[redacted]	50	Indian	Cancer
6	148[redacted]	55	Russian	Heart Disease
7	148[redacted]	47	American	Viral Infection
8	148[redacted]	49	American	Viral Infection
9	130[redacted]	31	American	Cancer
10	130[redacted]	37	Indian	Cancer
11	130[redacted]	36	Japanese	Cancer
12	130[redacted]	35	American	Cancer



	Zip Code
1	130**
2	130**
3	130**
4	130**
5	1485*
6	1485*
7	1485*
8	1485*
9	130**
10	130**
11	130**
12	130**

[redacted]

Generalization process

$$origin \in \{US, UK, Germany, China, Korea, \dots\}$$
$$age \in \{18, 19, \dots, 99\}$$
$$zip \in \{85281, 47408, \dots\}$$

$$origin \in \{NA, Europe, Asia, \dots\}$$
$$age \in \{[18-29], [30-45], [46-65], [66-99]\}$$
$$zip \in \{8528*, 4740*\}$$

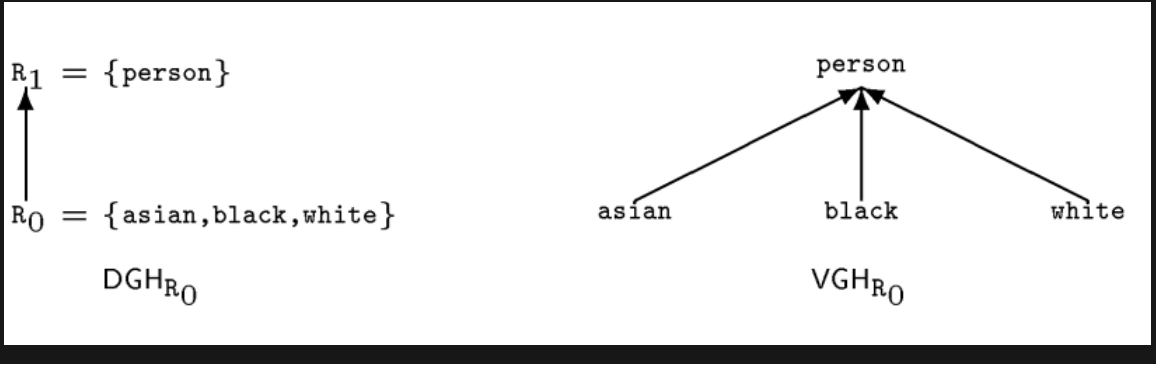
Domain generalization

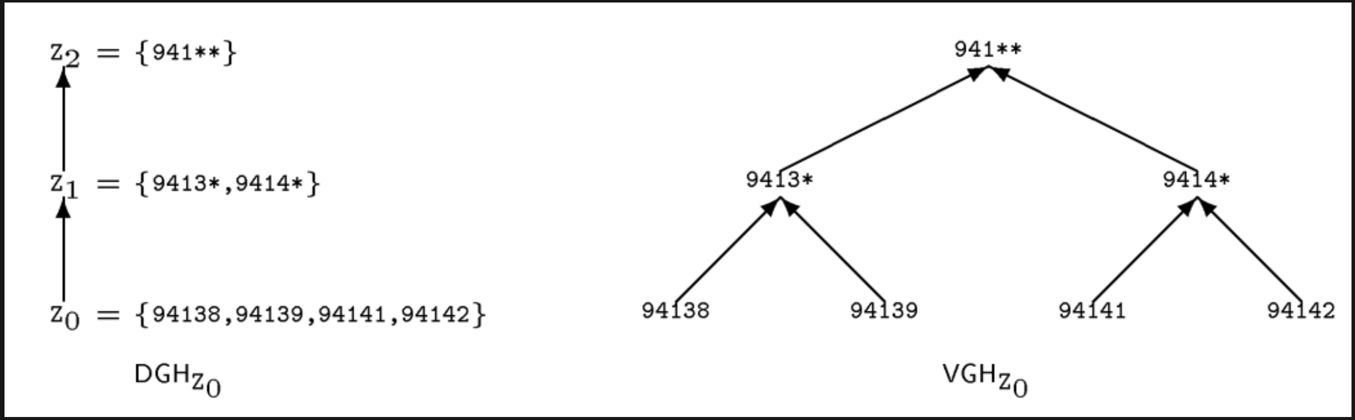
Given two domains D_i and D_j, relationship $D_i \leq D_j$ describes the fact that values in domain D_j are generalization of values in domain D_i.

1. $\forall D_i, D_j, D_k \in Dom: D_i \leq D_j, D_i \leq D_k \Rightarrow D_j \leq D_k \ or \ D_k \leq D_j$
2. All maximal element of *Dom* are singleton.

$zip \in \{8528 *, 4740 *\} \Rightarrow \{852 **, 474 **\} \Rightarrow .. \Rightarrow \{*****, *****\}$

Domain Generalization Hierarchy (DGH)





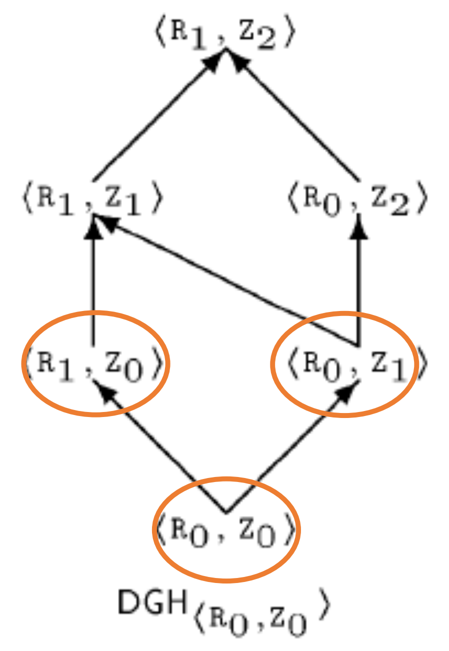

<person, 941**>

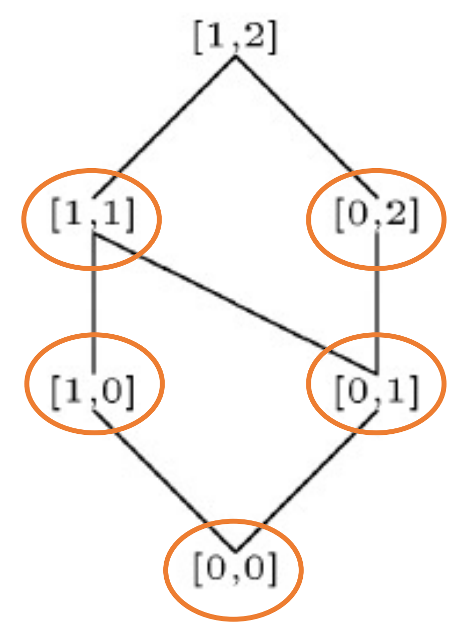
<person, 9414*> <Asian, 941**>

<Asian, 9414*> <person, 94141>

<Asian, 94141>

Domain Generalization Hierarchy (DGH)

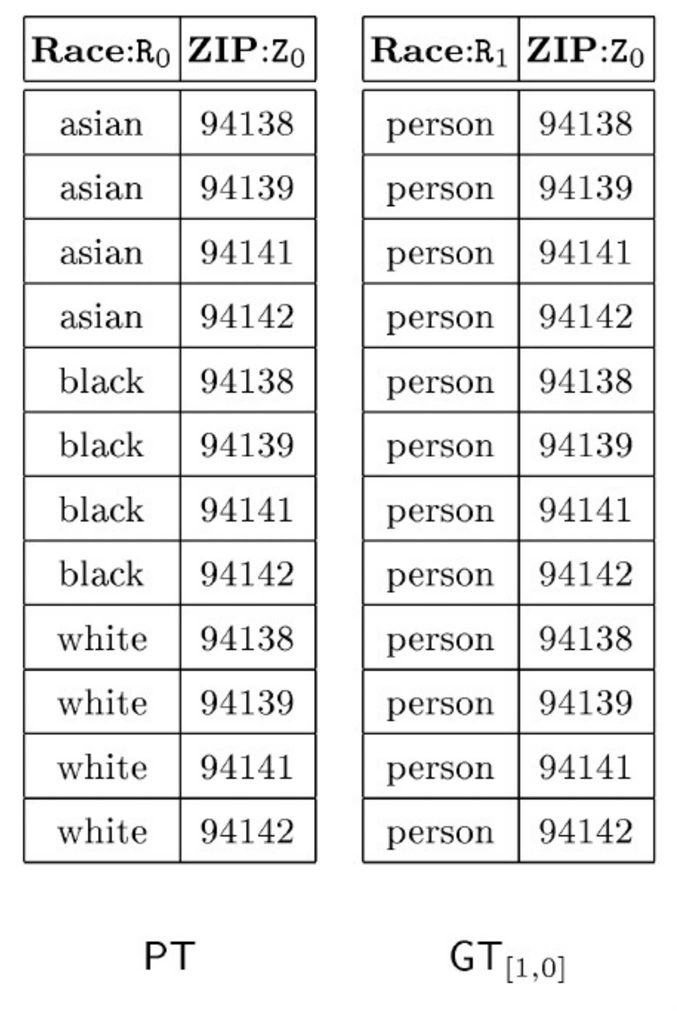




Generalized table

$\mathrm{T_i}$ is a generalization of table $\mathrm{T_j}$, defined on the same set of attributes, iff

1. $|T_i| = |T_j|$
2. The domain of each attribute in $\mathrm{T_j}$is equal to or is a generalization of, the domain corresponding attribute in $\mathrm{T_j}$
3. A bijective functions maps each tuple $\mathrm{t_i} \in T_i$ to one tuple in $t_j \in T_j$ where attribute values in t_j is equal or more generic than attribute values in t_i



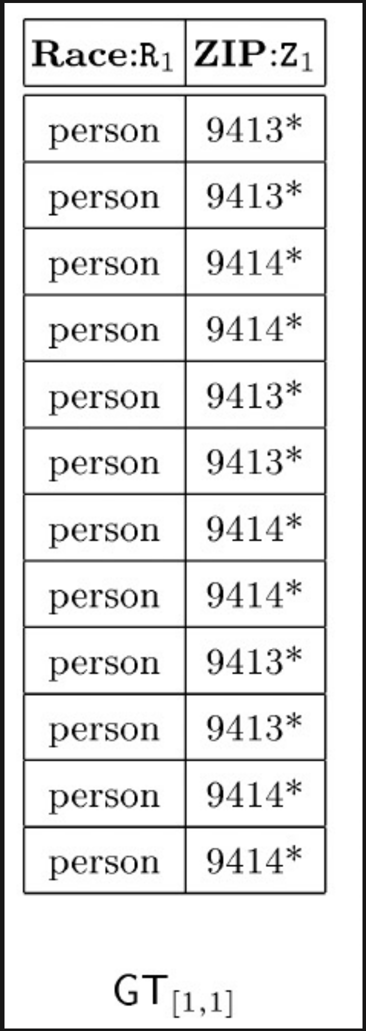
Race:R_0	**ZIP**:Z_0
asian	94138
asian	94139
asian	94141
asian	94142
black	94138
black	94139
black	94141
black	94142
white	94138
white	94139
white	94141
white	94142

PT

Race:R_1	**ZIP**:Z_0
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142

$GT_{[1,0]}$

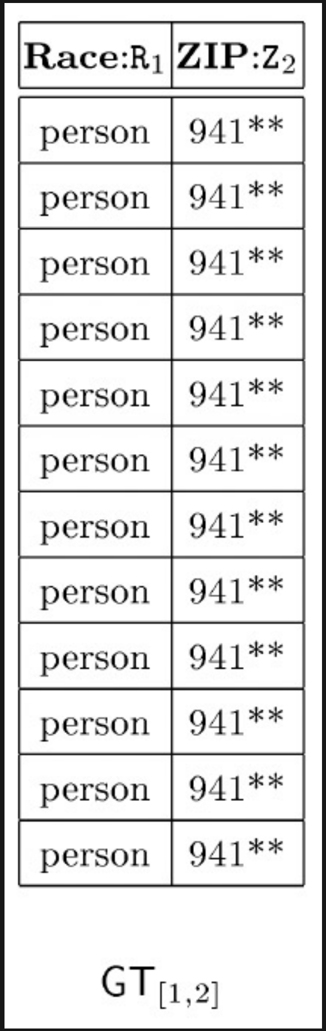
k= 1, 2



Race:R_1	**ZIP**:Z_1
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9413*
person	9414*
person	9414*

$GT_{[1,1]}$

k= 1, 2, …, 6



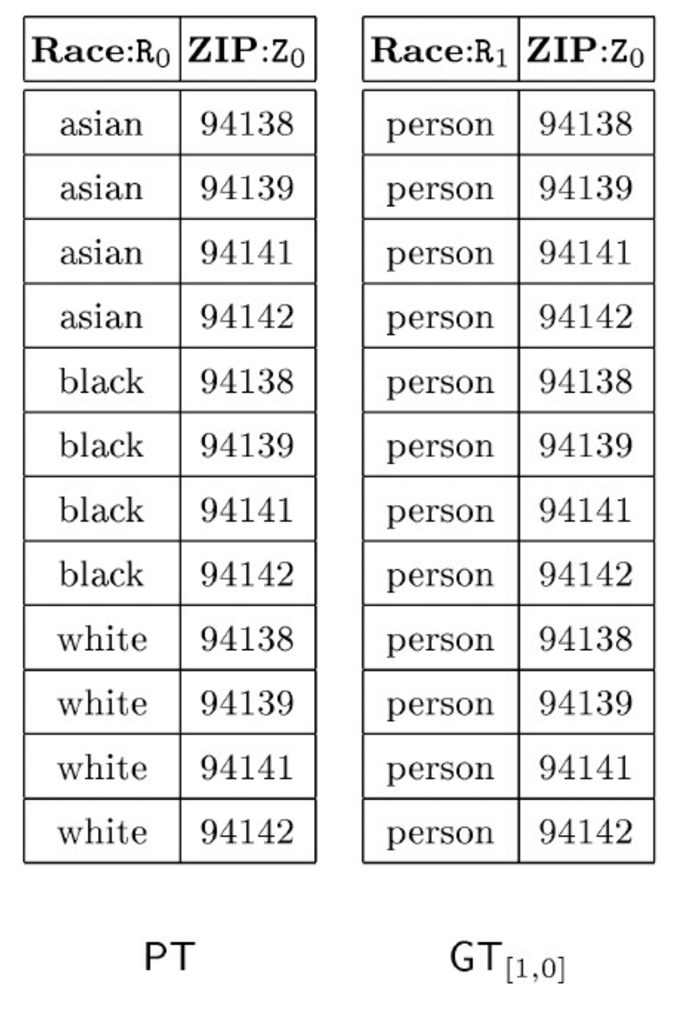
Race:R_1	**ZIP**:Z_2
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**

$GT_{[1,2]}$

k= *

Distance vector

Let T_i and T_j be two tables such that $T_i \leq T_j$. The distance vector of T_j from T_i is the vector $DV = [d_1, \ldots, d_n]$, where each d_i is length of unique path between corresponding domains in the domain generalization hierarchy.



Race:R_0	**ZIP**:Z_0
asian	94138
asian	94139
asian	94141
asian	94142
black	94138
black	94139
black	94141
black	94142
white	94138
white	94139
white	94141
white	94142

PT

Race:R_1	**ZIP**:Z_0
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142

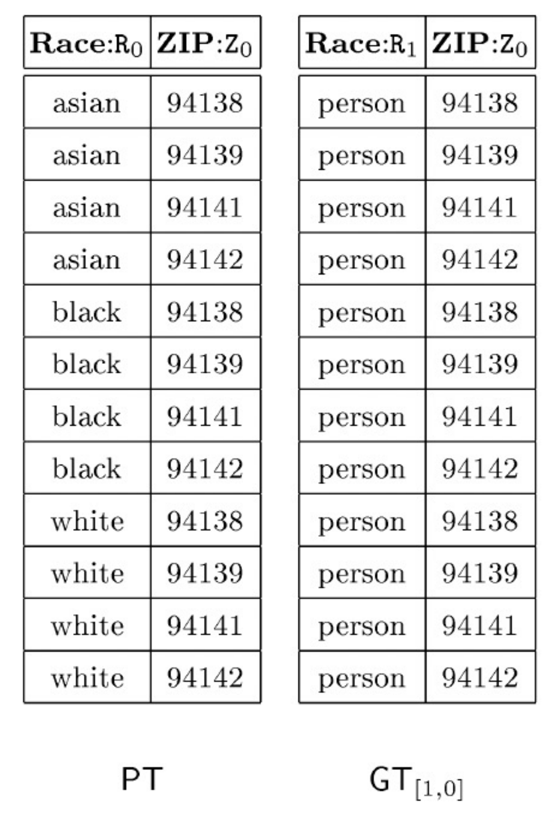
$GT_{[1,0]}$

k minimal generalization

Let T_i and T_j be two tables such that $T_i \leq T_j$. T_j is a k-minimal generalization of T_i iff

1. T_j satisfies k-anonymity
2. $\forall T_k : T_i \leq T_j, T_k$ satisfies k anonymity implies ! $(DV_{\{i,k\}} \leq DV_{\{i,j\}})$

Example

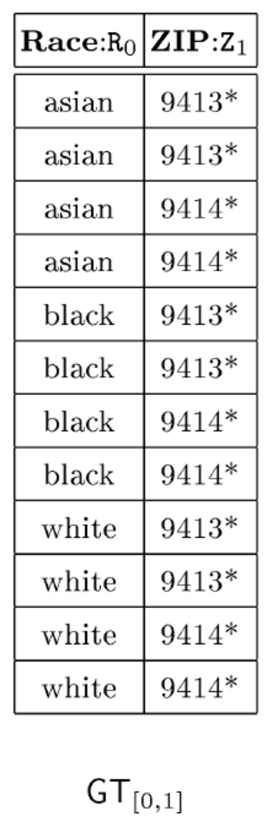


Race:R_0	**ZIP**:Z_0
asian	94138
asian	94139
asian	94141
asian	94142
black	94138
black	94139
black	94141
black	94142
white	94138
white	94139
white	94141
white	94142

PT

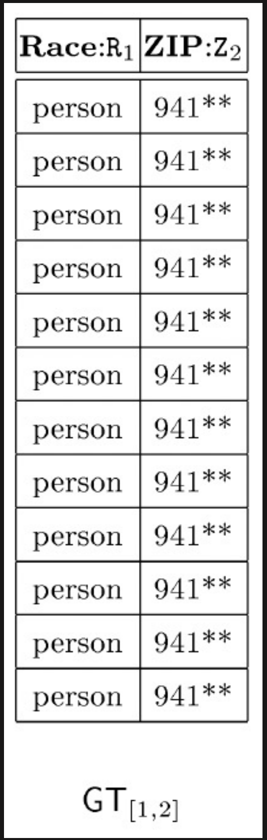
Race:R_1	**ZIP**:Z_0
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142

$GT_{[1,0]}$



Race:R_0	**ZIP**:Z_1
asian	9413*
asian	9413*
asian	9414*
asian	9414*
black	9413*
black	9413*
black	9414*
black	9414*
white	9413*
white	9413*
white	9414*
white	9414*

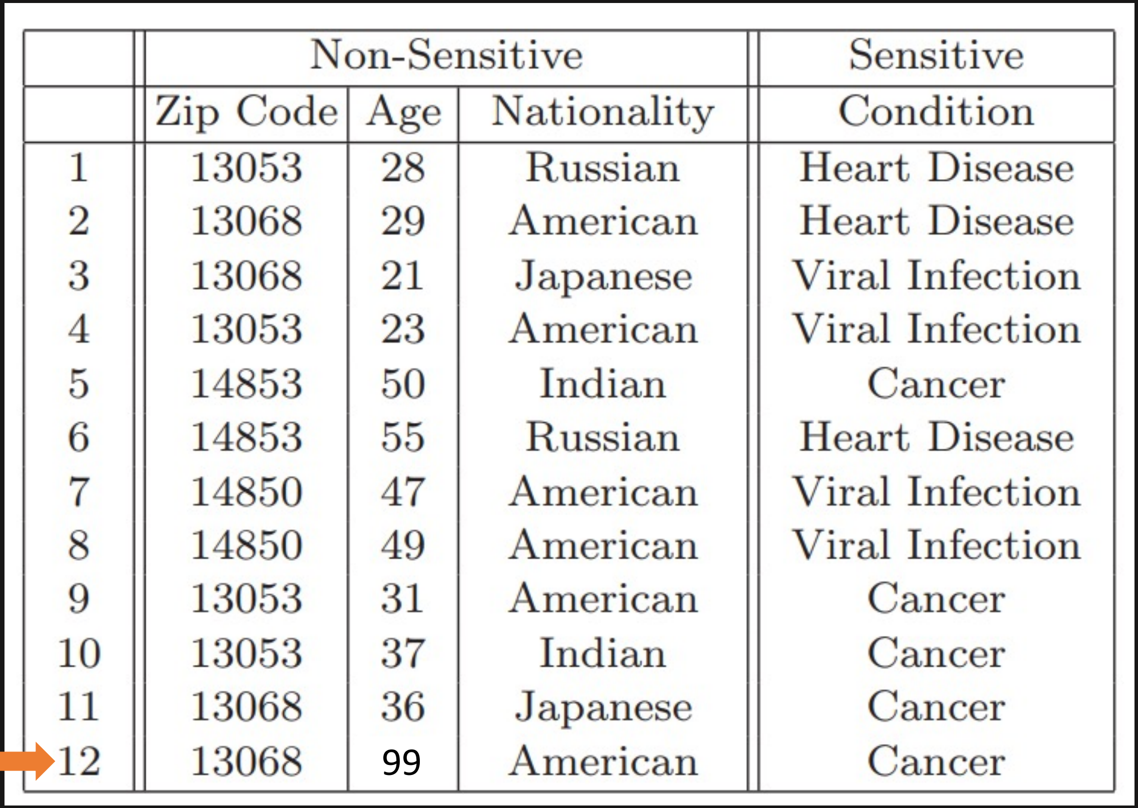
$GT_{[0,1]}$



Race:R_1	**ZIP**:Z_2
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**

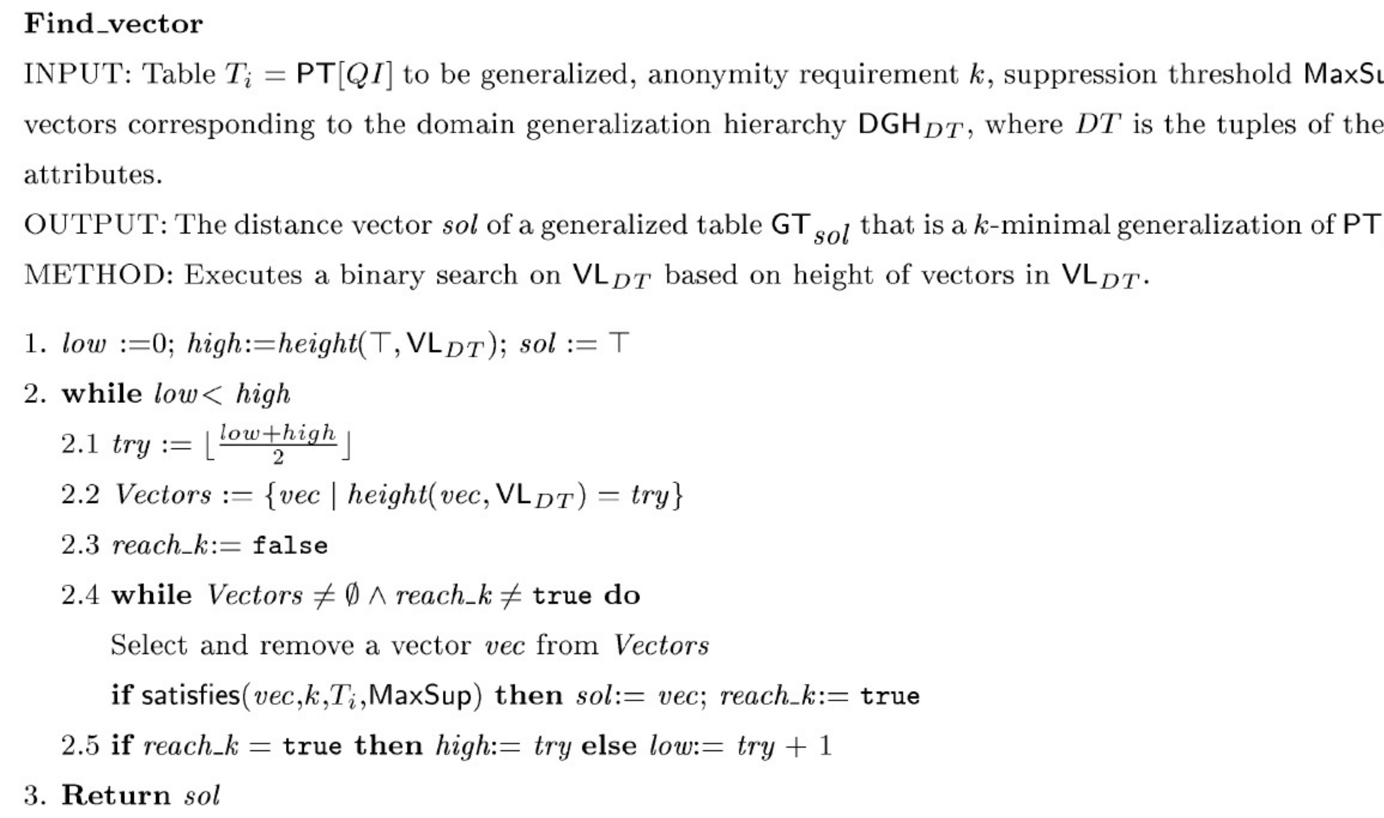
$GT_{[1,2]}$

Suppression



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	99	American	Cancer

Finding k-anonymous solution



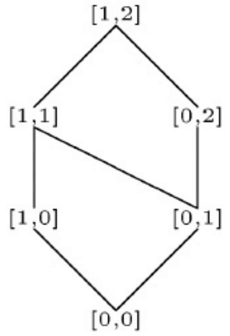
Find_vector

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 $sol$
```



heights: {0, 1 , 2 ,3}

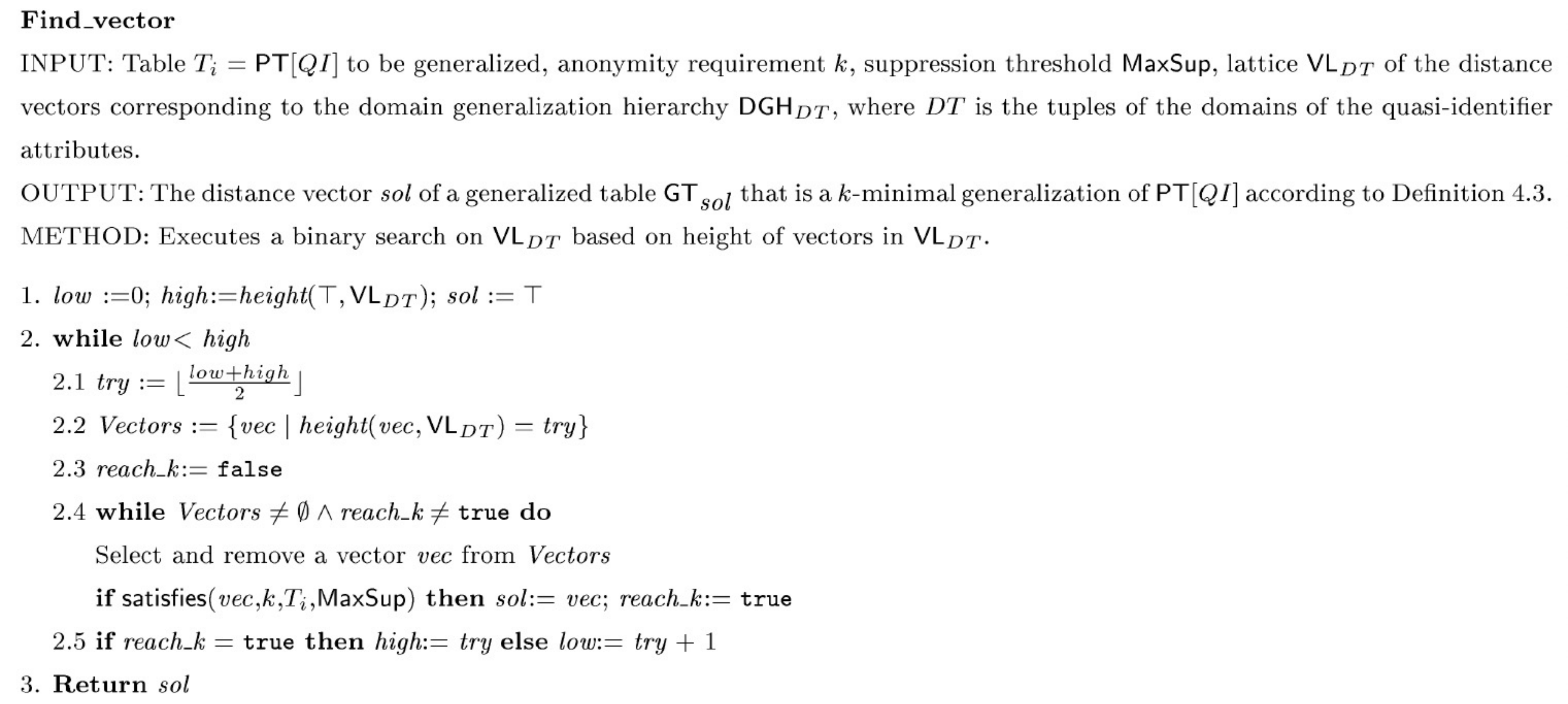
Resources

Hundepol et al. Handbook on statistical data disclosure

Samarati Protecting Respondents' Identities in Microdata Release

LeFevre et al. Incognito: Efficient Full Domain K-Anonymity

Finding k-anonymous solution



Find_vector

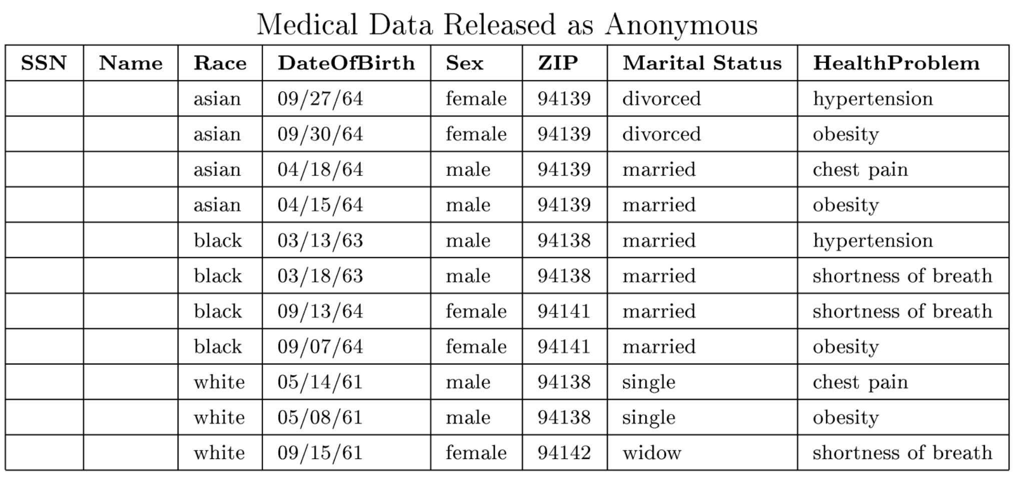
INPUT: Table $T_i = \mathsf{PT}[QI]$ to be generalized, anonymity requirement k, suppression threshold MaxSup, lattice VL_{DT} of the distance vectors corresponding to the domain generalization hierarchy DGH_{DT}, where DT is the tuples of the domains of the quasi-identifier attributes.

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}.

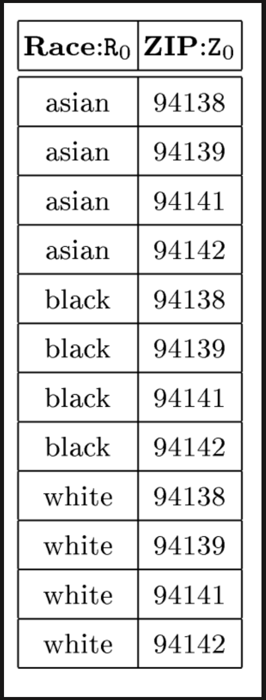
```
1. $low := 0$; $high := height(\top, \mathsf{VL}_{DT})$; $sol := \top$
2. while $low < high$
   2.1 $try := \lfloor \frac{low+high}{2} \rfloor$
   2.2 $Vectors := \{vec \mid height(vec, \mathsf{VL}_{DT}) = try\}$
   2.3 $reach\_k :=$ false
   2.4 while $Vectors \neq \emptyset \wedge reach\_k \neq$ 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 $sol$
```

Example input

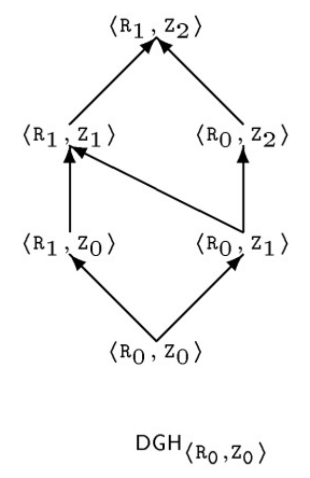


Medical Data Released as Anonymous

SSN	**Name**	**Race**	**DateOfBirth**	**Sex**	**ZIP**	**Marital Status**	**HealthProblem**
		asian	09/27/64	female	94139	divorced	hypertension
		asian	09/30/64	female	94139	divorced	obesity
		asian	04/18/64	male	94139	married	chest pain
		asian	04/15/64	male	94139	married	obesity
		black	03/13/63	male	94138	married	hypertension
		black	03/18/63	male	94138	married	shortness of breath
		black	09/13/64	female	94141	married	shortness of breath
		black	09/07/64	female	94141	married	obesity
		white	05/14/61	male	94138	single	chest pain
		white	05/08/61	male	94138	single	obesity
		white	09/15/61	female	94142	widow	shortness of breath

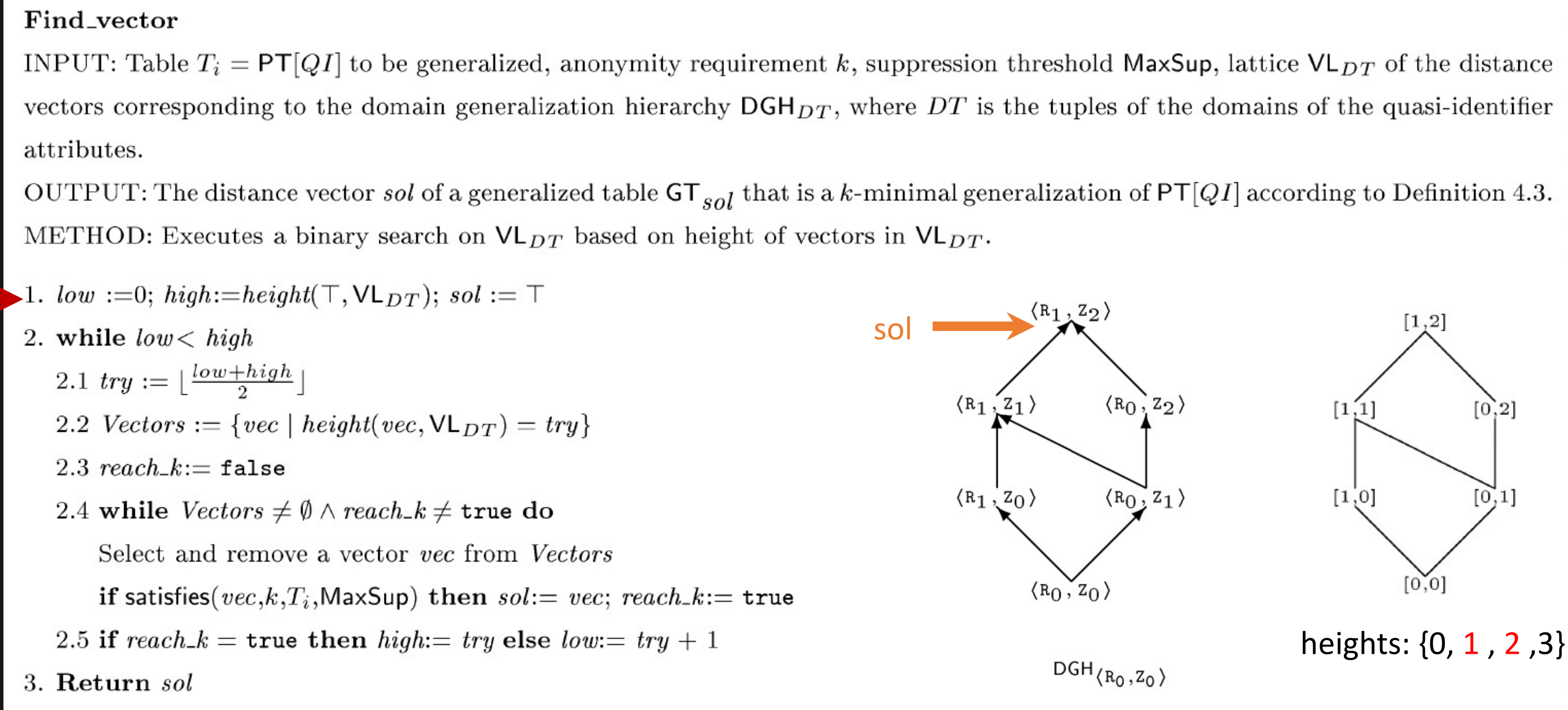


Race:R_0	**ZIP**:Z_0
asian	94138
asian	94139
asian	94141
asian	94142
black	94138
black	94139
black	94141
black	94142
white	94138
white	94139
white	94141
white	94142



Maxsup=2, k=2

Execution



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.

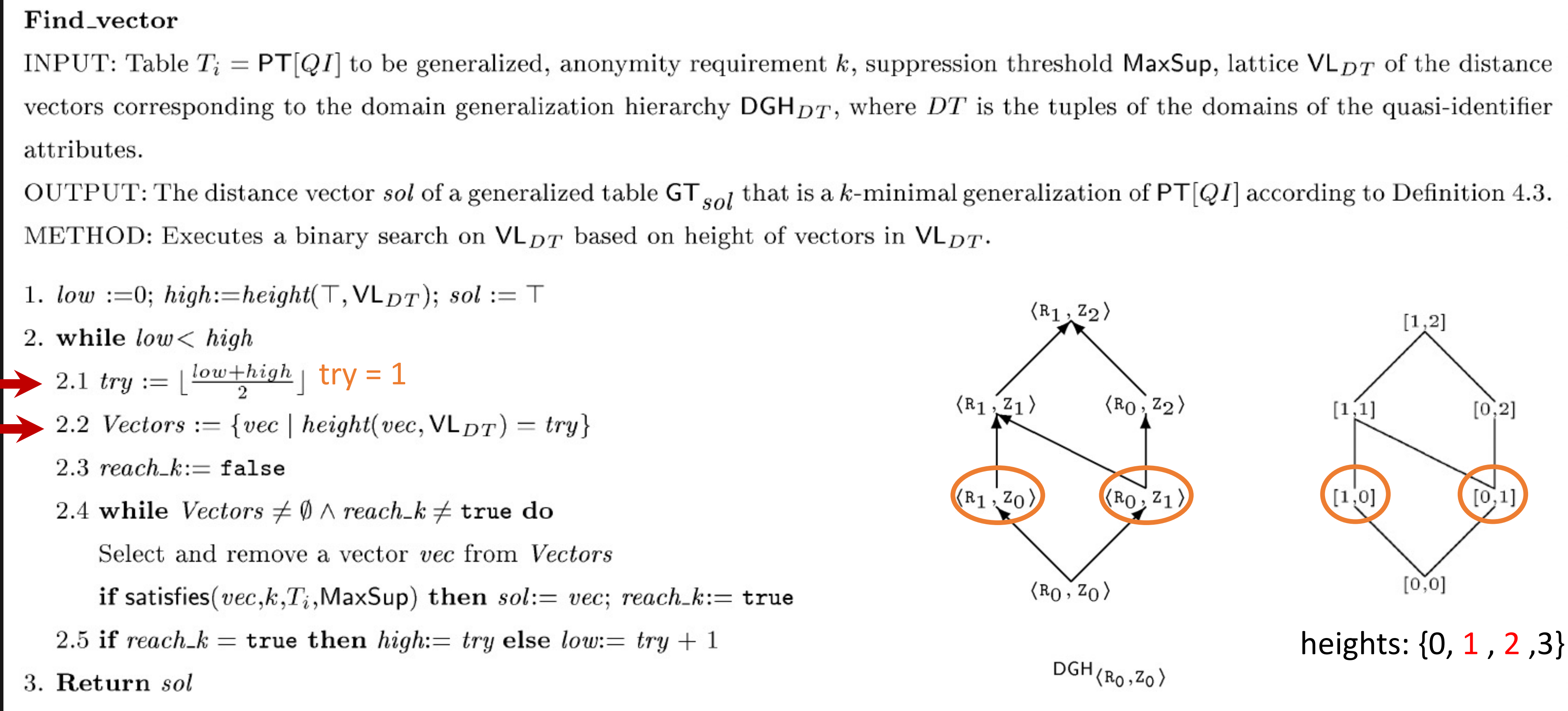
OUTPUT: The distance vector *sol* of a generalized table GT_{sol} that is a k-minimal generalization of $\mathsf{PT}[QI]$ according to Definition 4.3.

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 $sol$
```

heights: {0, 1 , 2 ,3}

Execution



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.

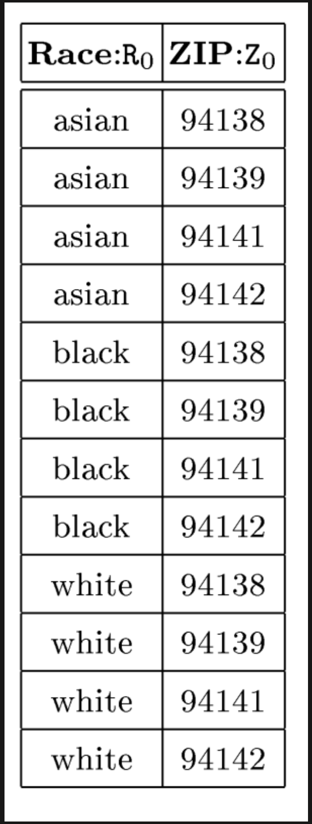
METHOD: Executes a binary search on VL_{DT} based on height of vectors in VL_{DT}.

```
1. $low$ :=0; $high$:=$height(\top, \mathsf{VL}_{DT})$; $sol$ := $\top$
2. while $low < high$
    2.1 $try$ := $\lfloor \frac{low+high}{2} \rfloor$   try = 1
    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 $sol$
```

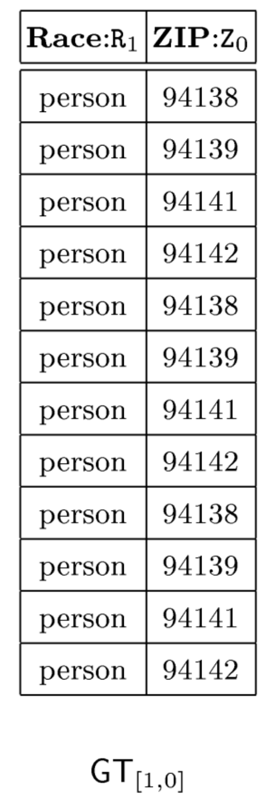
$\mathsf{DGH}_{\langle R_0, Z_0 \rangle}$

heights: {0, 1 , 2 ,3}

Applying the vectors



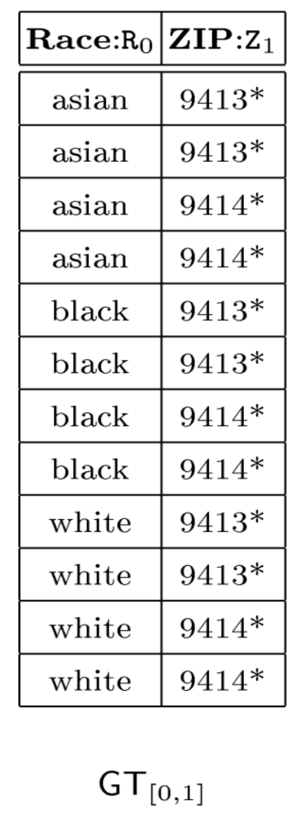
Race:R_0	**ZIP**:Z_0
asian	94138
asian	94139
asian	94141
asian	94142
black	94138
black	94139
black	94141
black	94142
white	94138
white	94139
white	94141
white	94142



Race:R_1	**ZIP**:Z_0
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142

GT$_{[1,0]}$

<R1,Z0>

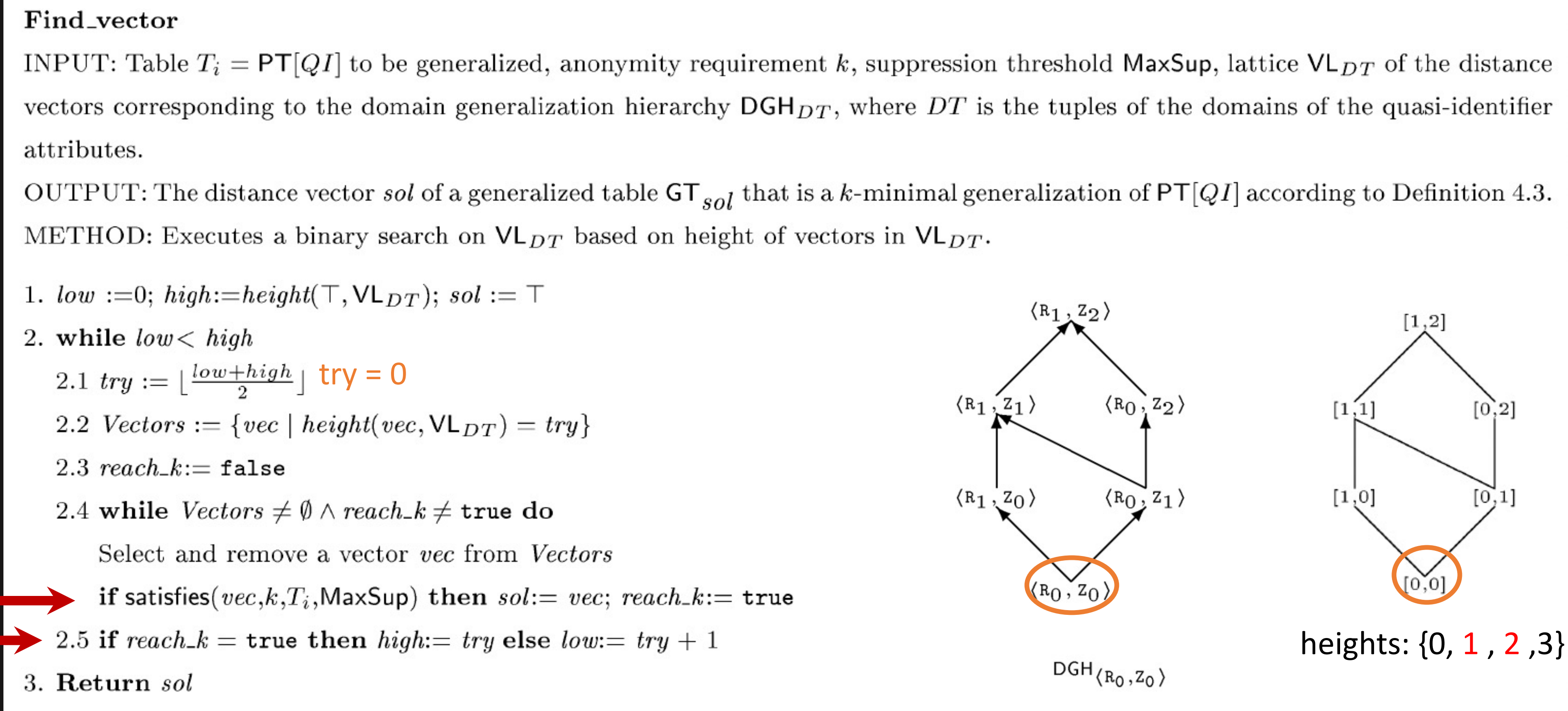


Race:R_0	**ZIP**:Z_1
asian	9413*
asian	9413*
asian	9414*
asian	9414*
black	9413*
black	9413*
black	9414*
black	9414*
white	9413*
white	9413*
white	9414*
white	9414*

GT$_{[0,1]}$

<R0,Z1>

Execution



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.

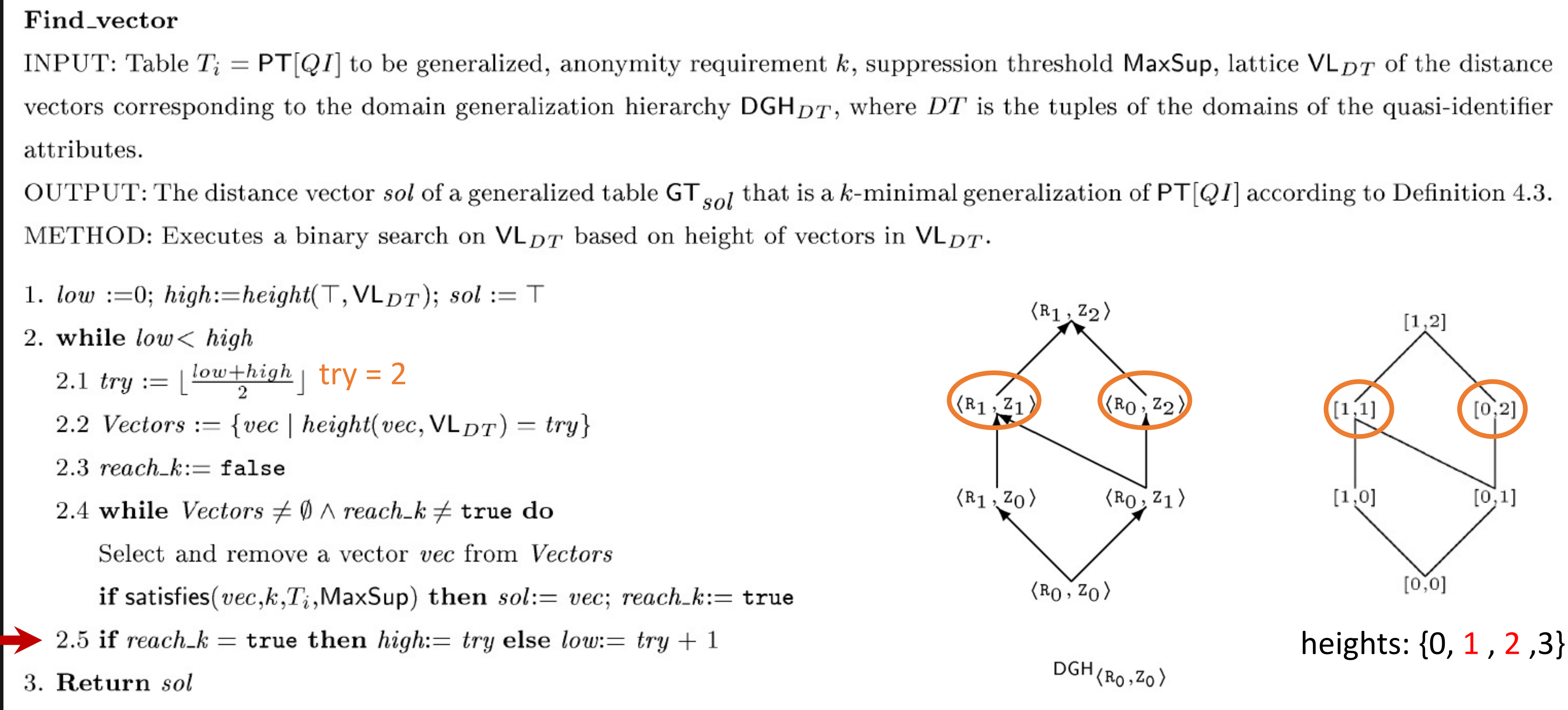
OUTPUT: The distance vector sol of a generalized table GT_{sol} that is a k-minimal generalization of $\mathsf{PT}[QI]$ according to Definition 4.3.

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 $try := \lfloor \frac{low+high}{2} \rfloor$   try = 0
    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 $sol$
```

heights: {0, 1 , 2 ,3}

Execution



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}.

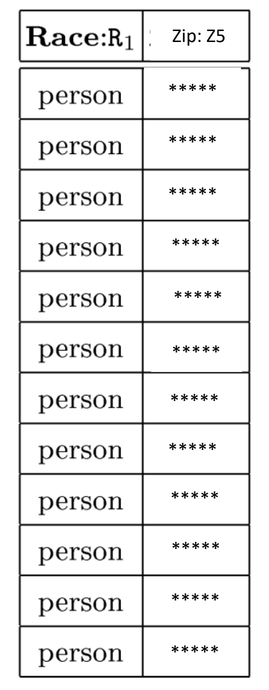
```
1. $low := 0$; $high := height(\top, \mathsf{VL}_{DT})$; $sol := \top$
2. while $low < high$
    2.1 $try := \lfloor \frac{low+high}{2} \rfloor$
    2.2 $Vectors := \{vec \mid height(vec, \mathsf{VL}_{DT}) = try\}$
    2.3 $reach\_k :=$ false
    2.4 while $Vectors \neq \emptyset \land reach\_k \neq$ 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 $sol$
```

try = 2

$\mathsf{DGH}_{\langle R_0, Z_0\rangle}$

heights: {0, 1 , 2 ,3}

The most general table

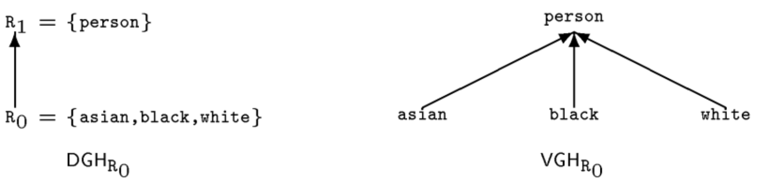


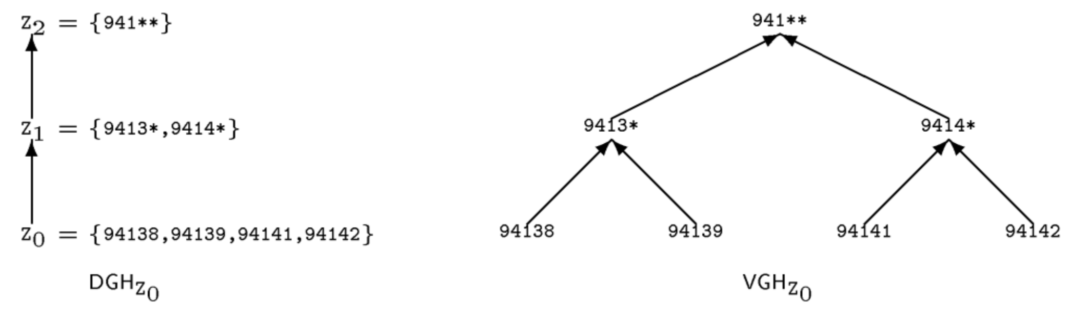
Race:R_1	Zip: Z5
person	*****
person	*****
person	*****
person	*****
person	*****
person	*****
person	*****
person	*****
person	*****
person	*****
person	*****
person	*****

Which *minimal* generalization to pick?

Minimum absolute distance prefers the *smallest total number of generalization steps.*

***Minimum relative distance** prefers the generalization(s) that has the smallest relative distance.*





D(<R1,Z1>) = 1+1/2 = 1.5 D(<R0,Z2>) = 0+2/2=1

Which *minimal* generalization to pick?

Minimum absolute distance prefers the *smallest total number of generalization steps.*

Minimum relative distance *prefers the generalization(s) that has the smallest relative distance.*

Maximum distribution *prefers the generalization(s) that contains the greatest number of distinct tuples.*

Minimum suppression *prefers the generalization(s) that suppresses less tuples.*

Utility metrics

Loss of information

Use of sanitized data

The number of generalization/suppression steps

Generalizing zip code

$$\{85281, 47408, \dots\} \Rightarrow \{8528*, 4740*, \dots\} \Rightarrow \{852**, 474**, \dots\}$$
$$\Rightarrow .. \Rightarrow \{*****,*****\}$$

$$\text{<Race, Zip>} \xrightarrow[\text{operation}]{\text{k-anon}} [1,1]$$

$$\text{(Asian, 85281)} \xrightarrow[\text{operation}]{\text{k-anon}} \text{(person, 8528*)}$$

Loss = height of the distance vector ([1,1]) = 2

The number of generalization/suppression steps

Generalizing zip code

$$\{85281, 47408, \ldots\} \Rightarrow \{8528*, 4740*, \ldots\} \Rightarrow \{852**, 474**, \ldots\}$$
$$\Rightarrow .. \Rightarrow \{*****,*****\}$$

85281
8528*
852**
85***
*****

Information loss is not actually same for each subsequent step.

One idea is to double the loss at each subsequent step.

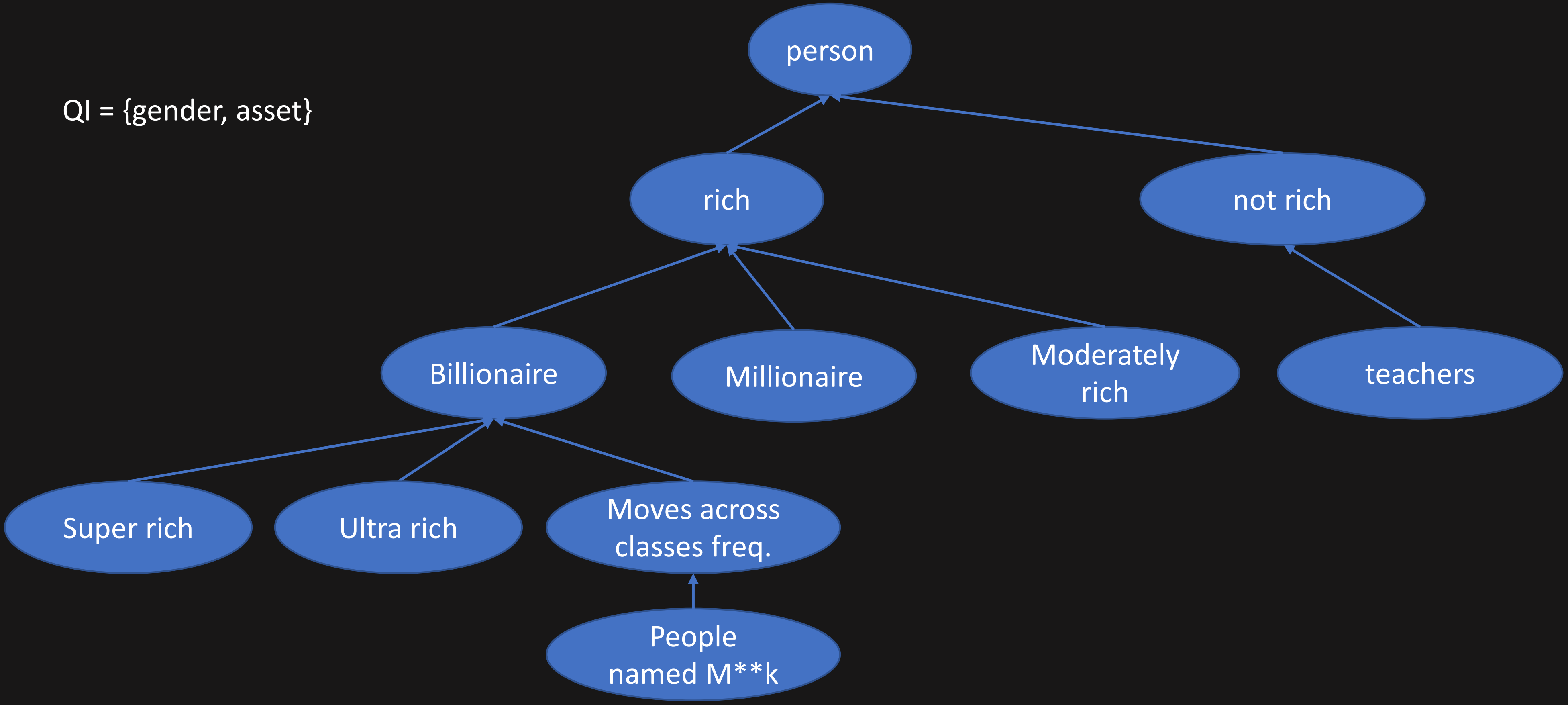
The number of generalization/suppression steps

$$\text{(Asian, 85281)} \xrightarrow[\text{operation}]{\text{k-anon}} \text{(person, 8528*)}$$

The same number of steps can result in unequal information loss for different attributes.

Loss metric

QI = {gender, asset}

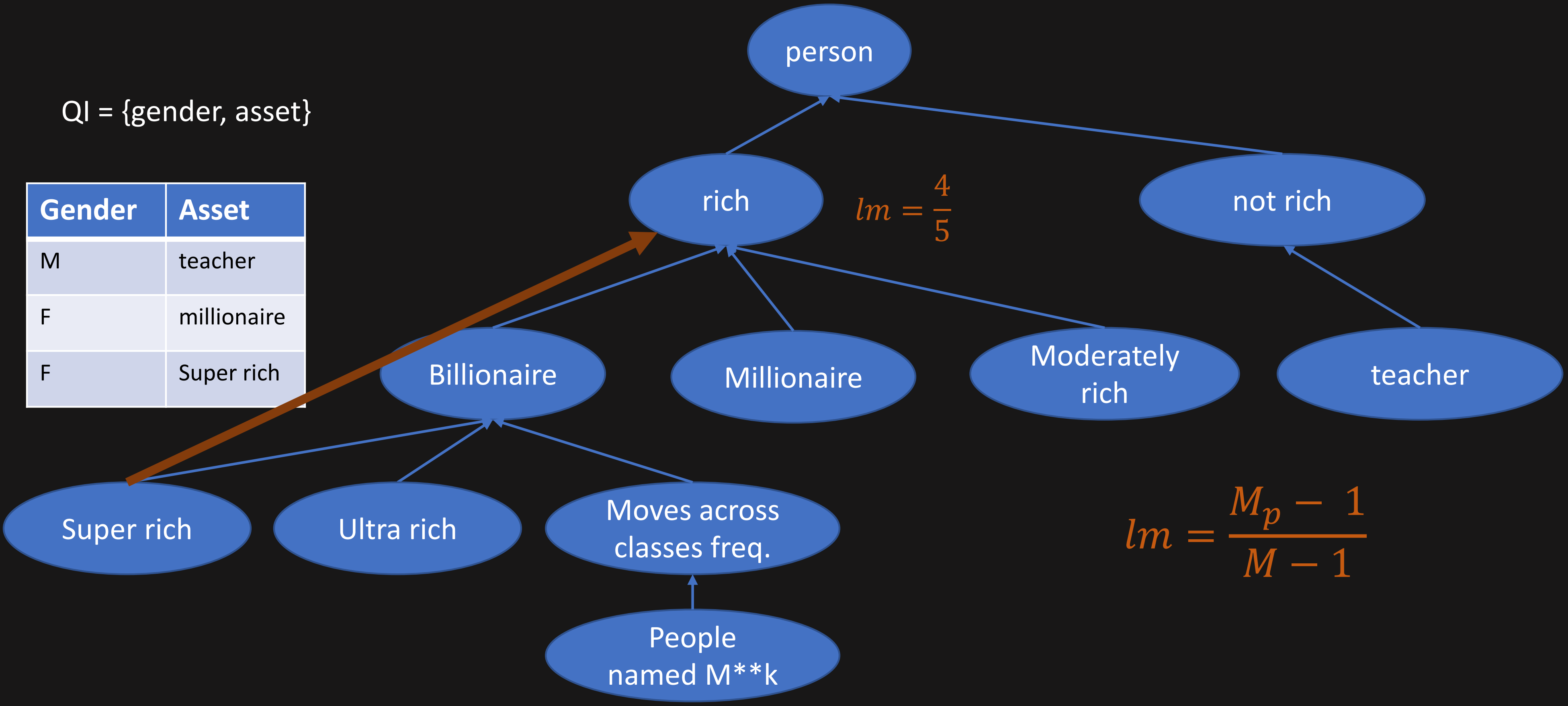


Iyengar, *Transforming data to satisfy privacy constraints*

Loss metric

QI = {gender, asset}

Gender	Asset
M	teacher
F	millionaire
F	Super rich



$lm = \frac{4}{5}$

$$lm = \frac{M_p - 1}{M - 1}$$

Iyengar, *Transforming data to satisfy privacy constraints*

Discernibility metric (DM)

Zip	Race	Disease
85281	Asian	Migraine
85282	Asian	Migraine
85283	Black	Heart
47408	Black	Heart
47401	Black	Migraine
47403	Black	Heart

Zip	Race	Disease
8528*	Asian	Migraine
8528*	Asian	Migraine
85283	Black	Heart ✗
4740*	Black	Heart
4740*	Black	Heart
47403	Black	Heart

Equivalence class

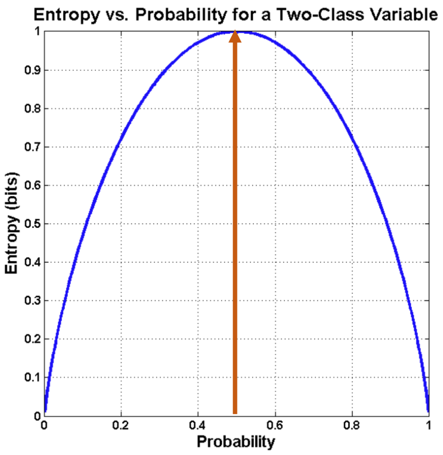
Equivalence class

$$DM = n * S + \sum_{i=1}^{NEQ} |EQ_i|^2$$

n= |T|
S= number of suppressed tuples
NEQ= number of equivalence class
|EQi| = size of EQi

Bayardo et al. *Data privacy through optimal k-anonymity*

Information theoretic measures



Entropy of a random variable X:

$$H(X) = -\sum_i p(x_i)\log_2(p(x_i))$$

Coin toss: $X = \{head, tail\}$

$$H(X) = -p(head)\log_2 p(head)) \ - p(tail)\log_2(p(tail))$$

$$= -\frac{1}{2}(-1) - \frac{1}{2}(-1) = 1$$

Information theoretic measures

Entropy of a random variable X:

$$H(X) = -\sum_{i} p(x_i) \log_2(p(x_i))$$

Dice: $X = \{1,2,3,4,5,6\}$

$$H(X) = -6 \ * \frac{1}{6} * \log_2 \frac{1}{6} = \sim 2.585$$

1. When all values of X are equally likely, entropy is the highest.

2. The larger the domain of X, the higher the entropy.

How the concept of entropy can be used as a utility measure?

Distortion

Gender	Race
M	Black
M	Asian
F	Asian
M	White
F	White
F	Black

QI = {Gender, Race} = X

$$X = \{\{M, Black\}, \{M, Asian\}, \dots, \{F, Black\}\}$$

$$H(X) = -6 \ * \frac{1}{6} * \log_2 \frac{1}{6} = \sim 2.585$$

Gender	Race
*	*
*	*
*	*
*	*
*	*
*	*

$$X = \{\{*,*\}, \{*,*\}, \dots .\}$$

$$H(X) = -6 \ * 1 * \log_2 1 = 0$$

$$Distortion = \frac{\left(H\left(QI_{pre}\right) \ - H\left(QI_{post}\right)\right)}{\log_2(\#r)} = \frac{2.585}{\log_2 6} = 1$$

Distortion

Gender	**Race**
M	Black
M	Asian
F	Asian
M	White
F	White
F	Black

QI = {Gender, Race} = X

$$X = \{\{M, Black\}, \{M, Asian\}, \ldots, \{F, Black\}\}$$

$$H(X) = -6 \ * \frac{1}{6} * \log_2 \frac{1}{6} = {\sim}2.585$$

Gender	**Race**
*	Black
*	Asian
*	Asian
*	White
*	White
*	Black

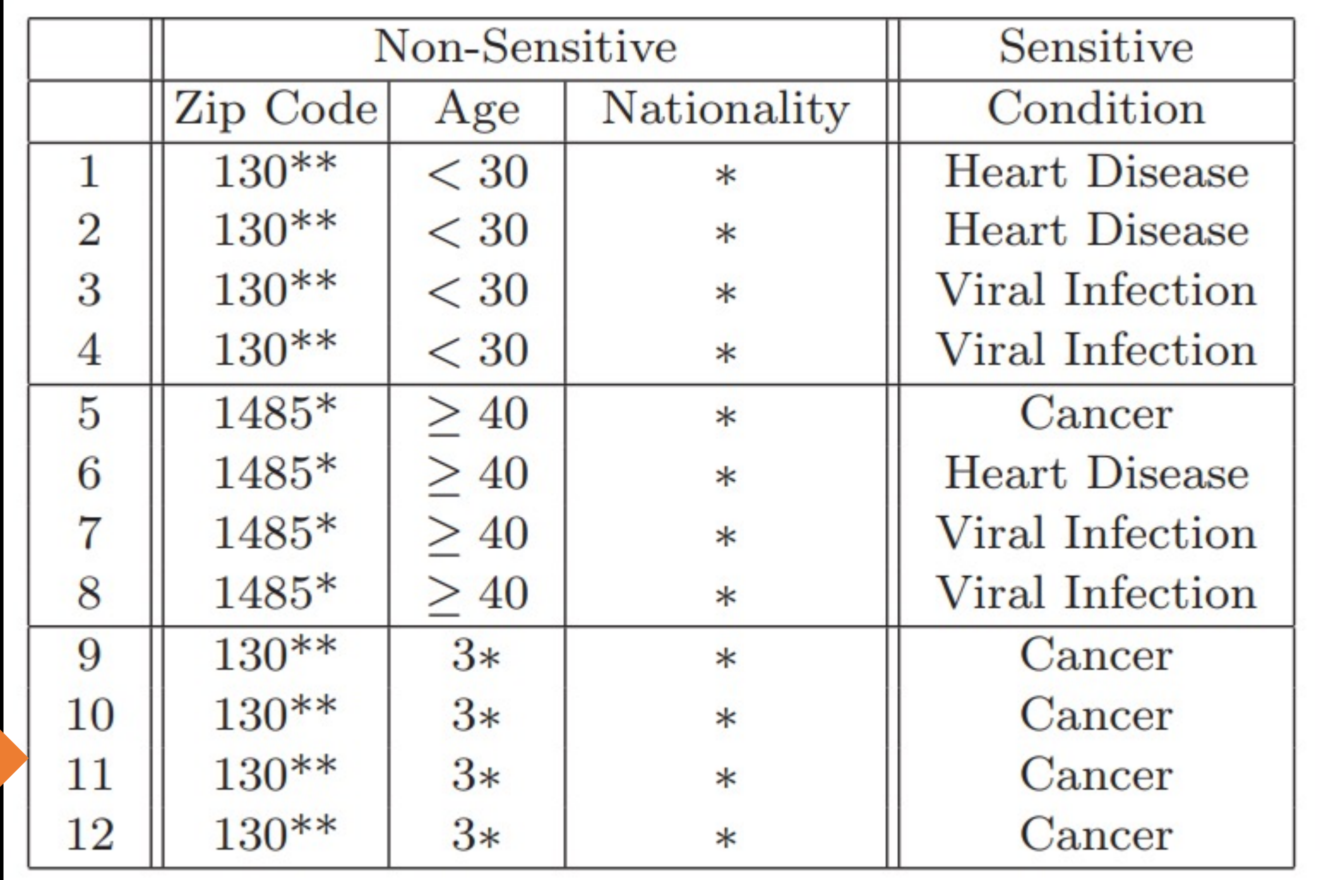
$$X = \{\{*, Black\}, \{*, Asian\}, \ldots.\}$$

$$H(X) = -3 \ * \frac{1}{3} * \log_2 \frac{1}{3} = 1.585$$

$$Distortion = \frac{\left(H(QI_{pre}) \ - H(QI_{post})\right)}{\log_2(\#r)} = \frac{1}{\log_2 6} < 1$$

Xu and Ye, *Risk & distortion-based k-anonymity*

How good is k-anonymity?



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

Resources

Utility metrics: Ch. 3 of Chen at al. *Privacy-Preserving Data Publishing*

link: https://www.researchgate.net/publication/220626610_Privacy-Preserving_Data_Publishing

https://www.youtube.com/watch?v=yQq1-_ujXrM