COMP7370 Advanced Computer and Network Security

Generalizing Data to Provide Anonymity when Disclosing Information (3)

- Comments on homework 2.

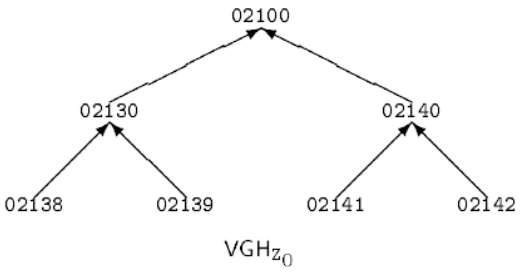
Review: **K-anonymity**

Definition 2.2 (k-anonymity) *Let $T(A_1, \ldots, A_n)$ be a table and QI_T be the quasi-identifiers associated with it. T is said to satisfy k-anonymity iff for each quasi-identifier $QI \in QI_T$,: each sequence of values in $T[QI]$ appears at least with k occurrences in $T[QI]$.*

T[Q1]: T[ethnicity] : each sequence; 2 sequences; both 5 "black" and 6 "Caucasian" >= k (say k is 4)
T[Q2]: T[sex]: 2 sequences: both 7 "male" and 4 "female" >= 4

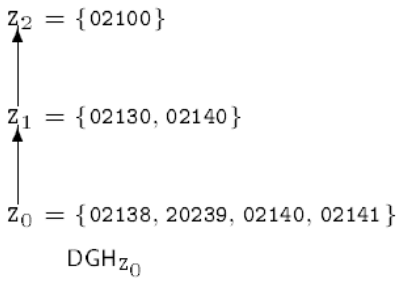
Key Approach: Hide partial information by generalizing values

How to present a process of hiding partial information?
Value Generalization Hierarchies



value generalization hierarchy

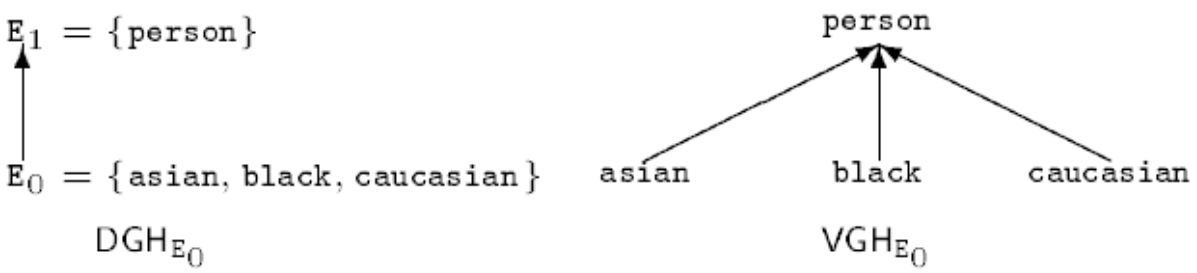
Question: Why value generalization hierarchies are not enough to present the idea of hiding information?



Key: Less informative
domain generalization hierarchy

- Domain:
  - e.g.: zip code domain, number domain, string domain.
  - Every attribute is in the ground domain

Question: give Eo, can you provide E1?



Definition 3.1 (k-minimal generalization – wrt a quasi-identifier) *Let T_i and T_j be two tables such that $T_i \leq T_j$. T_j is said to be a* k-minimal *generalization of a table T_i* wrt to a quasi-identifier *QI iff:*

1. *T_j satisfies k-anonymity wrt QI*
2. *$\forall T_z : T_i \leq T_z, T_z \leq T_j$, T_z satisfies k-anonymity wrt $QI \Rightarrow T_z[QI] = T_j[QI]$.*

Use slide 7 to explain: T_i = PT, T_j = GT[0,1];
PT --- transformed -> GT[0,1]
(1) k-anonymity
(2) Minimal
Question: (why minimal matters?)

Eth:E_0	ZIP:Z_0
a	38
a	39
a	41
a	42
b	38
b	39
b	41
b	42
c	38
c	39
c	41
c	42

PT

Eth:E_0	ZIP:Z_1
a	30
a	30
a	40
a	40
b	30
b	30
b	40
b	40
c	30
c	30
c	40
c	40

$GT_{[0,1]}$