COMP7370 Advanced Computer and Network Security

Generalizing Data to Provide Anonymity when Disclosing Information (2)

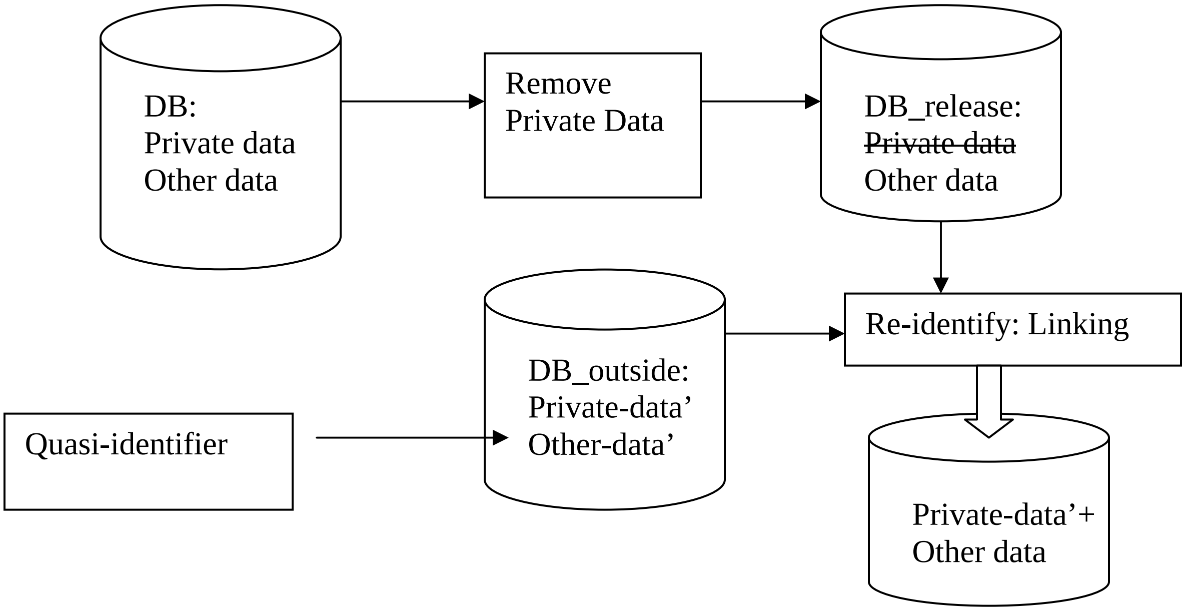
- Comments on homework 2.

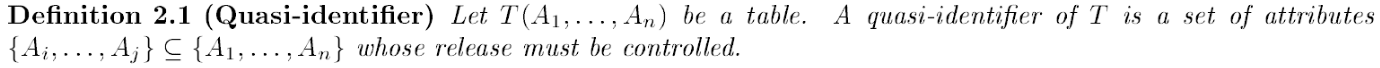
Topics:

1. Problem description
    a. How to define a term formally?

Review: Problem Description

- Motivation:
    - Protect individual-specific (private) data
        - e.g., name, address, phone number, SSN
    - Limitation: (see slide09b, p2)
        - **Re-identifying** anonymous data
        - Link to outside data (public data)
        - e.g., voting list - use DOB (12%); DOB+gender (29%); DOB+Zip (69%)





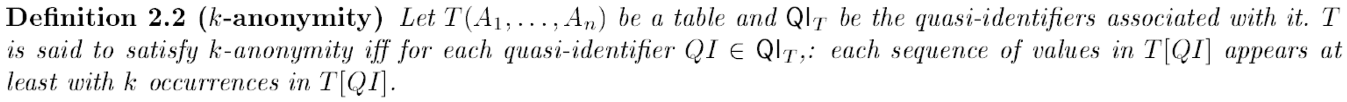
Definition 2.1 (Quasi-identifier) *Let $T(A_1, \ldots, A_n)$ be a table. A quasi-identifier of T is a set of attributes $\{A_i, \ldots, A_j\} \subseteq \{A_1, \ldots, A_n\}$ whose release must be controlled.*

- PT (or Private Table) = DB_release
- QI_PT = (u1, …, u_m) in PT (released DB)
- Goal: how to control QI_PT? or what info should be released to public?

  e.g.: QI_PT1= (DOB, gender, zip, marital status, problem) or
  QI_PT2= (gender, zip, marital status, problem) or
  QI_PT3= (DOB, marital status, problem)

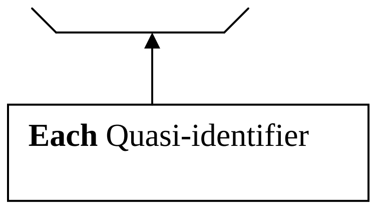
  To match at least *k* individuals

- Anonymity constraint: (see Lec11a, p4)



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 $QI \in \mathsf{QI}_T$,: each sequence of values in $T[QI]$ appears at least with k occurrences in $T[QI]$.*

DB_release + DB_outside -> at least k individual



Each Quasi-identifier

Question: Why only consider one quasi-identifier? (see Lec11a.ppt, p5)

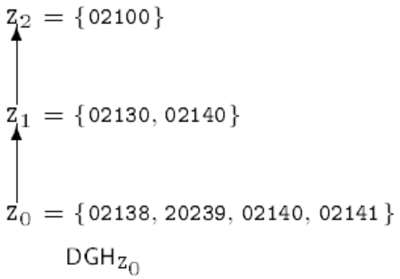
Answer: If each quasi-identifier in the released data satisfy k-anonymity, then the combination of released data to external sources cannot match lower than k individuals

Topic 2: Generalization of PT (private table, DB_release)

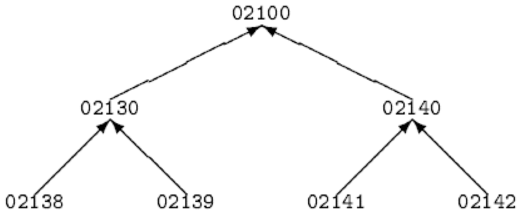
- Related K-anonymity ideas:
  - change unusual information to typical values, e.g. 4/1/1969 -> 1/1/1969
  - insert complementary records
  - swapping entries
  - scrambling records
- New idea: re-coding values -> make more general
  - e.g.: zip code 02139 -> 02130 (last digit is replaced by ‘0’, less informative)

- Domain:
  - e.g.: zip code domain, number domain, string domain.

Every attribute is in the ground domain



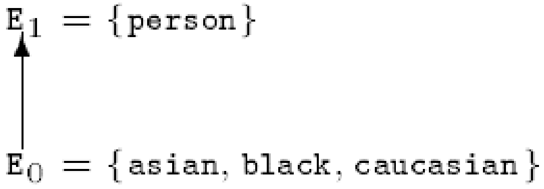
Key: Less informative
domain generalization hierarchy



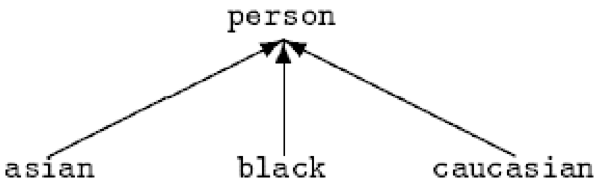
VGH_{Z_0}

value generalization hierarchy

Question: give Eo, can you provide E1?



DGH_{E_0}



VGH_{E_0}