Anonymity

Anonymity

- Is the state of being not observable within a set of subjects, the anonymity set
  - e.g. a particular person among all possible persons
  - e.g. a particular voter among all possible voters
  - e.g. a particular address among all possible addresses
- The anonymity set is the set of all possible subjects
  - From the attacker's point of view
- Is a context-dependent concept
  - The context defines the anonymity set regarding a particular action

Microdata privacy issues

- Microdata
  - Information at the level of individual respondents
- Privacy issues
  - Microdata is often used for several studies
  - How can we share microdata among companies without exposing its source?
    - The identity of the persons that provided it

Microdata privacy enhancing: Removal of potentially unique IDs

- Basic strategy
  - By removing potentially unique IDs we cannot link microdata items from several databases
- Candidate IDs
  - Name
  - National IDs (passport, identity card, etc.)
  - Social Security ID, Tax ID, etc.
  - Phone numbers
  - Car plate numbers
- Not enough!
  - A study in the States proved that 87% of its the population could be identified using a link attack using 3 non-unique attributes
    - 5-digit ZIP code, gender and birthday

Microdata privacy enhancing: Noise

- Basic strategy
  - Add noise to stored data or to the result of queries
- Issues
  - Privacy is achieved at the cost of integrity

Microdata privacy enhancing: K-anonymity

L. Sweeney, "K-anonymity: A Model for Protecting Privacy", Int. Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 2002

- Definition
  - No query can deliver an anonymity set with less than k entries
- Privacy-critical attributes
  - (Unique) identifiers
  - Quasi-identifiers
    - When combined can produce unique tuples
  - Sensitive attributes
    - Potentially unique per subject
    - Disease, salary, crime committed

K-anonimity: Implementation approaches

- Suppression of quasi-identifiers
  - Simple to perform
  - Information loss
- Generalization of quasi-identifiers
  - Transformation of quasi-identifiers in other ones less specific
    - e.g. 7-digit ZIP → 4-digit ZIP
    - e.g. ages w/ 1 year granularity → 5 or 10 year granularity
  - There is not a complete loss of information
    - But the generalization should not potentiate wrong data interpretations
  - We must ensure that there are at least k entries with equal generalized quasi-identifiers

Example

Name	Age	Sex	Zip Code	Illness
Sam	29	M	43102	Diabetes
gloria	38	F	43102	Breat cancer
Adam	51	M	43102	Colon cancer
Eric	29	M	43102	Diabetes
Tanisha	34	F	43102	HIV
Don	51	M	43102	Heart disease

Example: Identifiers

Name	Age	Sex	Zip Code	Illness
Sam	29	M	43102	Diabetes
gloria	38	F	43102	Breat cancer
Adam	51	M	43102	Colon cancer
Eric	29	M	43102	Diabetes
Tanisha	34	F	43102	HIV
Don	51	M	43102	Heart disease

Example: Quasi identifiers

Name	Age	Sex	Zip Code	Illness
Sam	29	M	43102	Diabetes
gloria	38	F	43102	Breat cancer
Adam	51	M	43102	Colon cancer
Eric	29	M	43102	Diabetes
Tanisha	34	F	43102	HIV
Don	51	M	43102	Heart disease

Example:
Sensitive attributes

Name	Age	Sex	Zip Code	Illness
Sam	29	M	43102	Diabetes
gloria	38	F	43102	Breat cancer
Adam	51	M	43102	Colon cancer
Eric	29	M	43102	Diabetes
Tanisha	34	F	43102	HIV
Don	51	M	43102	Heart disease

K-anonymity 1st step:
Remove unique identifiers

Age	Sex	Zip Code	Illness
29	M	43102	Diabetes
38	F	43102	Breat cancer
51	M	43102	Colon cancer
29	M	43102	Diabetes
34	F	43102	HIV
51	M	43102	Heart disease

K-anonymity 2nd step: Generalization

Age	Sex	Zip Code	Illness
30	M	43102	Diabetes
40	F	43102	Breat cancer
50	M	43102	Colon cancer
30	M	43102	Diabetes
30	F	43102	HIV
50	M	43102	Heart disease

K-anonymity : 2-anonymity possible results

Age	Sex	Zip Code	Illness
30	M	43102	Diabetes
40	F	43102	Breat cancer
50	M	43102	Colon cancer
30	M	43102	Diabetes
30	F	43102	HIV
50	M	43102	Heart disease

Issue:
Sensitive attribute disclosure

Age	Sex	Zip Code	Illness
30	M	43102	Diabetes
40	F	43102	Breat cancer
50	M	43102	Colon cancer
30	M	43102	Diabetes
30	F	43102	HIV
50	M	43102	Heart disease

L-Diversity

Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data (TKDD), 1.1, 2007.

- K-anonymity is not enough!
- Homogeneity attack
  - The attacker knows the generalized QIs of a target
  - A query reveals the exact same sensitive attributes
  - The attacker gets the sensitive attribute of the target
  - Issue: lack of diversity in the results
- Background knowledge attack
  - The attacker can filter out query results using known information

Solution:
l-diverse k-anonymity

- Results from a k-anonymity result of a query must contain l different values for each sensitive attribute

l-diversity :
2-anonymity 1-diversity results

Age	Sex	Zip Code	Illness
30	M	43102	Diabetes
40	F	43102	Breat cancer
50	M	43102	Colon cancer
30	M	43102	Diabetes
30	F	43102	HIV
50	M	43102	Heart disease

l-diversity :
2-anonymity 2-diversity results

Age	Sex	Zip Code	Illness
30	M	43102	Diabetes
40	F	43102	Breat cancer
50	M	43102	Colon cancer
30	M	43102	Diabetes
30	F	43102	HIV
50	M	43102	Heart disease