K- and $K^m$-Anonymity Examples

Leon S.L. Wang
Department of Information Management
National University of Kaohsiung
Kaohsiung, Taiwan 81148
Outline

1. Relational data + DGH
   - K-anonymity, Samarati TKDE ’01, bottom-up
2. Relational data, no DGH
   - K-anonymity, Aggarwal SIGMOD ‘06, clustering
   - K-anonymity, Park SIGMOD ‘07, minimize suppression
3. Set data + DGH
   - $K^m$-anonymity, Terrovitis VLDB ‘08, bottom-up
   - K-anonymity, He VLDB ‘09, top-down
4. Set data, no DGH
   - K-anonymity, Motwani, arXiv ‘08, suppression + flipping
   - $K^m$-anonymity, Xu KDD, ICDM ’08, $(h,k,p)$ – coherence, delete moles
Outline

- Graph data + DGH
  - K-anonymity, Campan PinKDD ’08, SaNreeA
- Graph data, no DGH
  - K-degree, Liu SIGMOD ’08
  - K-automorphism, VLDB ’09
  - K-isomorphism, Cheng SIGMOD ‘10
- Graph data, edge weight
  - K-anonymous weight, Liu ICIS’10
Relational data + DGH

- K-anonymity, re-identification

Medical Data Released as Anonymous

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

Voter List

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>City</th>
<th>ZIP</th>
<th>DOB</th>
<th>Sex</th>
<th>Party</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sue J. Carlson</td>
<td>900 Market St.</td>
<td>San Francisco</td>
<td>94142</td>
<td>9/15/61</td>
<td>female</td>
<td>democrat</td>
<td></td>
</tr>
</tbody>
</table>
Relational data + DGH

- Domain Generalization Hierarchy

\[ R_1 = \{ \text{person} \} \]

\[ R_0 = \{ \text{asian, black, white} \} \]

\[ \text{DGH}_{R_0} \]

\[ Z_2 = \{ 941** \} \]

\[ Z_1 = \{ 9413*, 9414* \} \]

\[ Z_0 = \{ 94138, 94139, 94141, 94142 \} \]

\[ \text{DGH}_{Z_0} \]

\[ \text{VGH}_{R_0} \]

\[ 941** \]

\[ 9413* \]

\[ 94138 \]

\[ 94139 \]

\[ \text{VGH}_{Z_0} \]

\[ 94141 \]

\[ 94142 \]

\[ 9414* \]
Relational data + DGH

- Minimal generalization, 2 QI, \(k = 2\), bottom-up

<table>
<thead>
<tr>
<th>Race: (R_0)</th>
<th>ZIP: (Z_0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>asian</td>
<td>94138</td>
</tr>
<tr>
<td>asian</td>
<td>94138</td>
</tr>
<tr>
<td>asian</td>
<td>94142</td>
</tr>
<tr>
<td>asian</td>
<td>94142</td>
</tr>
<tr>
<td>black</td>
<td>94138</td>
</tr>
<tr>
<td>black</td>
<td>94141</td>
</tr>
<tr>
<td>black</td>
<td>94142</td>
</tr>
<tr>
<td>white</td>
<td>94138</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: (R_1)</th>
<th>ZIP: (Z_0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>94138</td>
</tr>
<tr>
<td>person</td>
<td>94142</td>
</tr>
<tr>
<td>person</td>
<td>94138</td>
</tr>
<tr>
<td>person</td>
<td>94142</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: (R_0)</th>
<th>ZIP: (Z_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>asian</td>
<td>9413*</td>
</tr>
<tr>
<td>asian</td>
<td>9413*</td>
</tr>
<tr>
<td>asian</td>
<td>9414*</td>
</tr>
<tr>
<td>asian</td>
<td>9414*</td>
</tr>
<tr>
<td>black</td>
<td>9413*</td>
</tr>
<tr>
<td>black</td>
<td>9414*</td>
</tr>
<tr>
<td>black</td>
<td>9414*</td>
</tr>
<tr>
<td>white</td>
<td>9413*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: (R_0)</th>
<th>ZIP: (Z_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>asian</td>
<td>941**</td>
</tr>
<tr>
<td>asian</td>
<td>941**</td>
</tr>
<tr>
<td>asian</td>
<td>941**</td>
</tr>
<tr>
<td>asian</td>
<td>941**</td>
</tr>
<tr>
<td>black</td>
<td>941**</td>
</tr>
<tr>
<td>black</td>
<td>941**</td>
</tr>
<tr>
<td>black</td>
<td>941**</td>
</tr>
<tr>
<td>white</td>
<td>941**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: (R_1)</th>
<th>ZIP: (Z_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>9413*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: (R_1)</th>
<th>ZIP: (Z_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>9413*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
</tbody>
</table>

**One-step generalization**

**Two-step generalization**
Relational data + DGH

- Minimal generalization, 2 QI, for different $k$

<table>
<thead>
<tr>
<th>Race: $R_0$</th>
<th>ZIP: $Z_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>asian</td>
<td>94138</td>
</tr>
<tr>
<td>asian</td>
<td>94139</td>
</tr>
<tr>
<td>asian</td>
<td>94141</td>
</tr>
<tr>
<td>asian</td>
<td>94142</td>
</tr>
<tr>
<td>black</td>
<td>94138</td>
</tr>
<tr>
<td>black</td>
<td>94139</td>
</tr>
<tr>
<td>black</td>
<td>94141</td>
</tr>
<tr>
<td>black</td>
<td>94142</td>
</tr>
<tr>
<td>white</td>
<td>94138</td>
</tr>
<tr>
<td>white</td>
<td>94139</td>
</tr>
<tr>
<td>white</td>
<td>94141</td>
</tr>
<tr>
<td>white</td>
<td>94142</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: $R_1$</th>
<th>ZIP: $Z_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>94138</td>
</tr>
<tr>
<td>person</td>
<td>94139</td>
</tr>
<tr>
<td>person</td>
<td>94141</td>
</tr>
<tr>
<td>person</td>
<td>94142</td>
</tr>
<tr>
<td>person</td>
<td>94138</td>
</tr>
<tr>
<td>person</td>
<td>94139</td>
</tr>
<tr>
<td>person</td>
<td>94141</td>
</tr>
<tr>
<td>person</td>
<td>94142</td>
</tr>
<tr>
<td>person</td>
<td>94138</td>
</tr>
<tr>
<td>person</td>
<td>94139</td>
</tr>
<tr>
<td>person</td>
<td>94141</td>
</tr>
<tr>
<td>person</td>
<td>94142</td>
</tr>
<tr>
<td>person</td>
<td>94138</td>
</tr>
<tr>
<td>person</td>
<td>94139</td>
</tr>
<tr>
<td>person</td>
<td>94141</td>
</tr>
<tr>
<td>person</td>
<td>94142</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: $R_0$</th>
<th>ZIP: $Z_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>asian</td>
<td>9413*</td>
</tr>
<tr>
<td>asian</td>
<td>9413*</td>
</tr>
<tr>
<td>asian</td>
<td>9414*</td>
</tr>
<tr>
<td>asian</td>
<td>9414*</td>
</tr>
<tr>
<td>black</td>
<td>9413*</td>
</tr>
<tr>
<td>black</td>
<td>9413*</td>
</tr>
<tr>
<td>black</td>
<td>9414*</td>
</tr>
<tr>
<td>black</td>
<td>9414*</td>
</tr>
<tr>
<td>white</td>
<td>9413*</td>
</tr>
<tr>
<td>white</td>
<td>9413*</td>
</tr>
<tr>
<td>white</td>
<td>9414*</td>
</tr>
<tr>
<td>white</td>
<td>9414*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: $R_1$</th>
<th>ZIP: $Z_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>9413*</td>
</tr>
<tr>
<td>person</td>
<td>9413*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
<tr>
<td>person</td>
<td>9413*</td>
</tr>
<tr>
<td>person</td>
<td>9413*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: $R_0$</th>
<th>ZIP: $Z_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>asian</td>
<td>941**</td>
</tr>
<tr>
<td>asian</td>
<td>941**</td>
</tr>
<tr>
<td>asian</td>
<td>941**</td>
</tr>
<tr>
<td>black</td>
<td>941**</td>
</tr>
<tr>
<td>black</td>
<td>941**</td>
</tr>
<tr>
<td>black</td>
<td>941**</td>
</tr>
<tr>
<td>white</td>
<td>941**</td>
</tr>
<tr>
<td>white</td>
<td>941**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: $R_1$</th>
<th>ZIP: $Z_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>941**</td>
</tr>
<tr>
<td>person</td>
<td>941**</td>
</tr>
<tr>
<td>person</td>
<td>941**</td>
</tr>
<tr>
<td>person</td>
<td>941**</td>
</tr>
<tr>
<td>person</td>
<td>941**</td>
</tr>
<tr>
<td>person</td>
<td>941**</td>
</tr>
<tr>
<td>person</td>
<td>941**</td>
</tr>
<tr>
<td>person</td>
<td>941**</td>
</tr>
</tbody>
</table>
Relational data + DGH

- Minimal generalization/suppression, $k=2$

<table>
<thead>
<tr>
<th>Race: $R_0$</th>
<th>ZIP: $Z_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>asian</td>
<td>94138</td>
</tr>
<tr>
<td>asian</td>
<td>94138</td>
</tr>
<tr>
<td>asian</td>
<td>94142</td>
</tr>
<tr>
<td>asian</td>
<td>94142</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: $R_1$</th>
<th>ZIP: $Z_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>94138</td>
</tr>
<tr>
<td>person</td>
<td>94138</td>
</tr>
<tr>
<td>person</td>
<td>94142</td>
</tr>
<tr>
<td>person</td>
<td>94142</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: $R_0$</th>
<th>ZIP: $Z_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>asian</td>
<td>9413*</td>
</tr>
<tr>
<td>asian</td>
<td>9414*</td>
</tr>
<tr>
<td>black</td>
<td>9414*</td>
</tr>
<tr>
<td>black</td>
<td>9414*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: $R_1$</th>
<th>ZIP: $Z_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>9413*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
<tr>
<td>person</td>
<td>9414*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: $R_0$</th>
<th>ZIP: $Z_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>asian</td>
<td>941**</td>
</tr>
<tr>
<td>asian</td>
<td>941**</td>
</tr>
<tr>
<td>black</td>
<td>941**</td>
</tr>
<tr>
<td>black</td>
<td>941**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race: $R_1$</th>
<th>ZIP: $Z_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>941**</td>
</tr>
<tr>
<td>person</td>
<td>941**</td>
</tr>
<tr>
<td>person</td>
<td>941**</td>
</tr>
<tr>
<td>person</td>
<td>941**</td>
</tr>
</tbody>
</table>

Optimal, $\text{maxSupp}=0$

Optimal, $\text{maxSupp}=1$

Suppression is delete
Relational data - no DGH

- Aggarwal PODS ‘06, clustering
- 2-Anonymity with suppression
- *All attributes suppressed* !!!

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amy</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>Brian</td>
<td>27</td>
<td>60</td>
</tr>
<tr>
<td>Carol</td>
<td>29</td>
<td>100</td>
</tr>
<tr>
<td>David</td>
<td>35</td>
<td>110</td>
</tr>
<tr>
<td>Evelyn</td>
<td>39</td>
<td>120</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amy</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Brian</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Carol</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>David</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Evelyn</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
Relational data - no DGH

- Aggarwal PODS ‘06, clustering
- 2-Anonymity with generalization, *Pre-specified ranges*
- *NP-hard, O*(k) *approximation for k-anonymity*

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amy</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>Brian</td>
<td>27</td>
<td>60</td>
</tr>
<tr>
<td>Carol</td>
<td>29</td>
<td>100</td>
</tr>
<tr>
<td>David</td>
<td>35</td>
<td>110</td>
</tr>
<tr>
<td>Evelyn</td>
<td>39</td>
<td>120</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amy</td>
<td>20-30</td>
<td>50-100</td>
</tr>
<tr>
<td>Brian</td>
<td>20-30</td>
<td>50-100</td>
</tr>
<tr>
<td>Carol</td>
<td>20-30</td>
<td>50-100</td>
</tr>
<tr>
<td>David</td>
<td>30-40</td>
<td>100-150</td>
</tr>
<tr>
<td>Evelyn</td>
<td>30-40</td>
<td>100-150</td>
</tr>
</tbody>
</table>
### 2-Anonymity with Clustering

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amy</td>
<td>[25-29]</td>
<td>[50-100]</td>
</tr>
<tr>
<td>Brian</td>
<td>[25-29]</td>
<td>[50-100]</td>
</tr>
<tr>
<td>Carol</td>
<td>[25-29]</td>
<td>[50-100]</td>
</tr>
<tr>
<td>David</td>
<td>[35-39]</td>
<td>[110-120]</td>
</tr>
<tr>
<td>Evelyn</td>
<td>[35-39]</td>
<td>[110-120]</td>
</tr>
</tbody>
</table>

#### Cluster centers and cluster sizes published

- Cluster centers: $27 = (25+27+29)/3$
- Cluster sizes: $70 = (50+60+100)/3$
- $37 = (35+39)/2$
- $115 = (110+120)/2$

#### Constant factor approximation algorithms
Relational data - no DGH

- Convert quasi-identifiers into points in a **metric space**
  - Converting (gender, zip code, DOB) into points in a metric space not easy.
  - Define distance function on each attribute.
  - E.g. on Zip code:
    - $D(\text{Zip1}, \text{Zip2}) =$ physical distance between locations Zip1 and Zip2.
  - Weight attributes, weighted sum of attribute distances gives metric.
Relational data - no DGH

• Cluster Quasi-identifiers so that each cluster has at least $r$ members for anonymity.

• Publish cluster centers for anonymity with number of point and radius

• Tight clusters $\Rightarrow$ Usefulness of data for mining

• Large number of points per cluster $\Rightarrow$ Anonymity

• Propose two clustering algorithms
  • $r$-Gather Clustering
  • $r$-Cellular Clustering
r-Gather Clustering

10 points, radius 5

20 points, radius 10

50 points, radius 20

Minimize the maximum radius: 20
Cellular Clustering Metric

10 points, radius 5

20 points, radius 10

50 points, radius 20

Minimize Cellular Clustering Metric: \(10 \times 5 + 20 \times 10 + 50 \times 20\)

\[= 50 + 200 + 1000 = 1250\]
Relational data - no DGH

- Park, SIGMOD ‘07, approximation algorithms
  - Minimize the number of suppression
    - Partition table into subsets with size \([k, 2k-1]\)
    - K-anonymize each subset

<table>
<thead>
<tr>
<th>Age</th>
<th>Marital status</th>
<th>Home country</th>
<th>Gender</th>
<th>Age</th>
<th>Marital status</th>
<th>Home country</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>20~29</td>
<td>Single</td>
<td>USA</td>
<td>Male</td>
<td>20~29</td>
<td>Single</td>
<td>USA</td>
<td>*</td>
</tr>
<tr>
<td>30~39</td>
<td>Divorce</td>
<td>China</td>
<td>Female</td>
<td>30~39</td>
<td>*</td>
<td>*</td>
<td>Female</td>
</tr>
<tr>
<td>20~29</td>
<td>Single</td>
<td>USA</td>
<td>Female</td>
<td>20~29</td>
<td>Single</td>
<td>USA</td>
<td>*</td>
</tr>
<tr>
<td>30~39</td>
<td>Separation</td>
<td>Korea</td>
<td>Female</td>
<td>30~39</td>
<td>*</td>
<td>*</td>
<td>Female</td>
</tr>
</tbody>
</table>

(a) An employee table  
(b) A 2-anonymized table
Relational data - no DGH

- Park, SIGMOD ‘07, approximation algorithms
- Original table, $k = 4$

<table>
<thead>
<tr>
<th>Age</th>
<th>Marital status</th>
<th>Home country</th>
<th>Gender</th>
<th>Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>20~29</td>
<td>Single</td>
<td>USA</td>
<td>Female</td>
<td>Master</td>
</tr>
<tr>
<td>20~29</td>
<td>Single</td>
<td>USA</td>
<td>Female</td>
<td>Doctor</td>
</tr>
<tr>
<td>20~29</td>
<td>Single</td>
<td>China</td>
<td>Male</td>
<td>Master</td>
</tr>
<tr>
<td>20~29</td>
<td>Single</td>
<td>China</td>
<td>Male</td>
<td>Doctor</td>
</tr>
<tr>
<td>20~29</td>
<td>Divorce</td>
<td>USA</td>
<td>Male</td>
<td>Master</td>
</tr>
<tr>
<td>20~29</td>
<td>Divorce</td>
<td>USA</td>
<td>Male</td>
<td>Doctor</td>
</tr>
<tr>
<td>30~39</td>
<td>Single</td>
<td>USA</td>
<td>Male</td>
<td>Master</td>
</tr>
<tr>
<td>30~39</td>
<td>Single</td>
<td>USA</td>
<td>Male</td>
<td>Doctor</td>
</tr>
</tbody>
</table>

(a) An employee table
Relational data - no DGH

- Park, SIGMOD ‘07, approximation algorithms
- Partition table into subsets with size [4, 7]

<table>
<thead>
<tr>
<th>Age</th>
<th>Marital status</th>
<th>Home country</th>
<th>Gender</th>
<th>Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>20~29</td>
<td>Single</td>
<td>USA</td>
<td>Female</td>
<td>Master</td>
</tr>
<tr>
<td>20~29</td>
<td>Single</td>
<td>USA</td>
<td>Female</td>
<td>Doctor</td>
</tr>
<tr>
<td>20~29</td>
<td>Single</td>
<td>China</td>
<td>Male</td>
<td>Master</td>
</tr>
<tr>
<td>20~29</td>
<td>Single</td>
<td>China</td>
<td>Male</td>
<td>Doctor</td>
</tr>
<tr>
<td>20~29</td>
<td>Divorce</td>
<td>USA</td>
<td>Male</td>
<td>Master</td>
</tr>
<tr>
<td>20~29</td>
<td>Divorce</td>
<td>USA</td>
<td>Male</td>
<td>Doctor</td>
</tr>
<tr>
<td>30~39</td>
<td>Single</td>
<td>USA</td>
<td>Male</td>
<td>Master</td>
</tr>
<tr>
<td>30~39</td>
<td>Single</td>
<td>USA</td>
<td>Male</td>
<td>Doctor</td>
</tr>
</tbody>
</table>
Relational data - no DGH

- Park, SIGMOD ‘07, approximation algorithms
- Optimal 4-anonymity

<table>
<thead>
<tr>
<th>Age</th>
<th>Marital status</th>
<th>Home country</th>
<th>Gender</th>
<th>Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>20~29</td>
<td>Single</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>20~29</td>
<td>Single</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>20~29</td>
<td>Single</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>20~29</td>
<td>Single</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>*</td>
<td>*</td>
<td>USA</td>
<td>Male</td>
<td>*</td>
</tr>
<tr>
<td>*</td>
<td>*</td>
<td>USA</td>
<td>Male</td>
<td>*</td>
</tr>
<tr>
<td>*</td>
<td>*</td>
<td>USA</td>
<td>Male</td>
<td>*</td>
</tr>
<tr>
<td>*</td>
<td>*</td>
<td>USA</td>
<td>Male</td>
<td>*</td>
</tr>
</tbody>
</table>
Set data + DGH

- $K^m$-anonymity, Terrovitis VLDB ‘08, bottom-up
  - The support count of every $m$-itemset $\geq k$

- For example, $k = 2$
Set data + DGH

- $K^m$-anonymity, expanded DB, count-tree and bottom-up
- $K = 2$
Set data + DGH

- K-anonymity, He VLDB ‘09, top-down
- The support count of every transaction $\geq k$
- Quality metric, Normalized Certainty Penalty (NCP)

<table>
<thead>
<tr>
<th>TID</th>
<th>Original Data</th>
<th>Local Recoding (2-anonymity)</th>
<th>Global Recoding (2$^2$-anonymity)</th>
<th>Global Recoding (2-anonymity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>${a_1}$</td>
<td>${A}$</td>
<td>${a_1}$</td>
<td>${A}$</td>
</tr>
<tr>
<td>$T_2$</td>
<td>${a_1, a_2}$</td>
<td>${A}$</td>
<td>${a_1, a_2}$</td>
<td>${A}$</td>
</tr>
<tr>
<td>$T_3$</td>
<td>${b_1, b_2}$</td>
<td>${b_1, b_2}$</td>
<td>${B}$</td>
<td>${B}$</td>
</tr>
<tr>
<td>$T_4$</td>
<td>${b_1, b_2}$</td>
<td>${b_1, b_2}$</td>
<td>${B}$</td>
<td>${B}$</td>
</tr>
<tr>
<td>$T_5$</td>
<td>${a_1, a_2, b_2}$</td>
<td>${a_1, a_2, B}$</td>
<td>${a_1, a_2, B}$</td>
<td>${A, B}$</td>
</tr>
<tr>
<td>$T_6$</td>
<td>${a_1, a_2, b_2}$</td>
<td>${a_1, a_2, B}$</td>
<td>${a_1, a_2, B}$</td>
<td>${A, B}$</td>
</tr>
<tr>
<td>$T_7$</td>
<td>${a_1, a_2, b_1, b_2}$</td>
<td>${a_1, a_2, B}$</td>
<td>${a_1, a_2, B}$</td>
<td>${A, B}$</td>
</tr>
</tbody>
</table>
Set data - no DGH

- K-anonymity, Motwani, arXive ’08
- Approximation and flipping

<table>
<thead>
<tr>
<th>ID</th>
<th>e₁</th>
<th>e₂</th>
<th>e₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>S₁</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>S₂</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S₃</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S₄</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S₅</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S₆</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

(a) Original dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>e₁</th>
<th>e₂</th>
<th>e₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>S₁</td>
<td>1</td>
<td>*</td>
<td>0</td>
</tr>
<tr>
<td>S₂</td>
<td>*</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S₃</td>
<td>*</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S₄</td>
<td>1</td>
<td>*</td>
<td>0</td>
</tr>
<tr>
<td>S₅</td>
<td>*</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S₆</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(b) \( S_* \)

<table>
<thead>
<tr>
<th>ID</th>
<th>e₁</th>
<th>e₂</th>
<th>e₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>S₁</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S₂</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S₃</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S₄</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S₅</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S₆</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(c) \( S_{flip} \)
Set data - no DGH

- $K^m$-anonymity, Xu KDD, ICDM ’08, $(h,k,p)$ – coherence
  - Every $p$-item subset must appear in at least $k$ transactions
  - The probability of linking an individual to a private item is limited to $h$
  - For $k=2$, $p=2$, $h=80%$; $ab$ and $bf$ are moles
  - $|ab| = 1$; $|bf| = 2$, $T_2, T_3$, but can infer SA is Hepatitis 100%

<table>
<thead>
<tr>
<th>TID</th>
<th>Activities</th>
<th>Medical History</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>a, c, d, f, g</td>
<td>Diabetes</td>
</tr>
<tr>
<td>$T_2$</td>
<td>a, b, c, f</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>$T_3$</td>
<td>b, d, f, ☒</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>$T_4$</td>
<td>b, c, g, ☒, ☒</td>
<td>HIV</td>
</tr>
<tr>
<td>$T_5$</td>
<td>a, c, f, g</td>
<td>HIV</td>
</tr>
</tbody>
</table>
Set data - no DGH

- $K^m$-anonymity (for simplicity, consider only $k, p$),
  - For $k=2, p=2$;
  - $ab, ad, bd, bg, cd, dg$ are moles
  - All other frequent itemsets are nuggets
  - Strategy: to delete all minimal moles but maintain maximal nuggets

<table>
<thead>
<tr>
<th>TID</th>
<th>Activities</th>
<th>Medical History</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>a, c, d, f, g</td>
<td>Diabetes</td>
</tr>
<tr>
<td>$T_2$</td>
<td>a, b, c, f</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>$T_3$</td>
<td>b, d, f, $\n$</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>$T_4$</td>
<td>b, c, g, $\n$</td>
<td>HIV</td>
</tr>
<tr>
<td>$T_5$</td>
<td>a, c, f, g</td>
<td>HIV</td>
</tr>
</tbody>
</table>
Set data - no DGH

- **Strategy**: to delete all moles but maintain nuggets
- **ab, ad, bd, bg, cd, dg** are moles
- 5 candidate items to be deleted: a, b, c, d, g
- **Greedy**: delete the item that can affect most moles
- **Criteria**: $\text{MME}(e) / \text{IL}(e)$, # of moles / # of items deleted
- **Delete a** => $2/3$
- **Delete b** => $3/2$
- **Delete c** => $1/4$
- **Delete d** => $4/2$
- **Delete g** => $2/3$

<table>
<thead>
<tr>
<th>TID</th>
<th>Activities</th>
<th>Medical History</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>a, c, d, f, g</td>
<td>Diabetes</td>
</tr>
<tr>
<td>$T_2$</td>
<td>a, b, c, f</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>$T_3$</td>
<td>b, d, f, X</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>$T_4$</td>
<td>b, c, g, Y, X</td>
<td>HIV</td>
</tr>
<tr>
<td>$T_5$</td>
<td>a, c, f, g</td>
<td>HIV</td>
</tr>
</tbody>
</table>
Set data - no DGH

- **Strategy:** to delete all moles but maintain nuggets
- **Two moles left:** ab, bg
- **3 candidate items to be deleted:** a, b, g
- **Greedy:** delete the item that can affect most moles
- **Criteria:** $MME(e) / IL(e)$, # of moles / # of items deleted

- *Delete a* $\Rightarrow$ 1/3
- *Delete b* $\Rightarrow$ 2/3
- *Delete g* $\Rightarrow$ 1/3

<table>
<thead>
<tr>
<th>TID</th>
<th>Activities</th>
<th>Medical History</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>a, c, d, f, g</td>
<td>Diabetes</td>
</tr>
<tr>
<td>$T_2$</td>
<td>a, b, c, f</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>$T_3$</td>
<td>b, d, f</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>$T_4$</td>
<td>b, c, g</td>
<td>HIV</td>
</tr>
<tr>
<td>$T_5$</td>
<td>a, c, f, g</td>
<td>HIV</td>
</tr>
</tbody>
</table>
### K- anonymity and K\(^m\)- anonymity

<table>
<thead>
<tr>
<th>Data type</th>
<th>K-anonymity</th>
<th>K(^m)- anonymity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relational + DGH</td>
<td>Samarati TKDE ’01, Sweeney IJUFKS ’02; l-diversity, Machanavajjhala, TKDD ’07; t-closeness, Li, ICDE ’07;</td>
<td>Mohammed KDD ‘09, LKC-privacy, top-down;</td>
</tr>
<tr>
<td>Relational, no DGH</td>
<td>Aggarwal PODS’06, clustering</td>
<td></td>
</tr>
<tr>
<td>Set + DGH</td>
<td>He VLDB ‘09, top-down</td>
<td>Terrovitis VLDB ‘08, bottom-up</td>
</tr>
<tr>
<td>Set, no DGH</td>
<td>Motwani arXiv ‘08, Relational clustering</td>
<td>Xu KDD, ICDM ’08, (h,k,p) – coherence, delete moles</td>
</tr>
<tr>
<td>Graph + DGH</td>
<td>Campan PinKDD08, SaNGreeA;</td>
<td></td>
</tr>
<tr>
<td>Graph, no DGH</td>
<td>Chang VLDB ‘09, Predictive anonymity; Zon VLDB ‘09, K-automorphism for multiple structural attacks; Bhagat VLDB ‘09, class-based anonymity;</td>
<td></td>
</tr>
</tbody>
</table>