Administrative

• HW3 released
  – Due October 21
  – Question 2 requires installation of AdFisher
    • Start this question early!

• In-class discussion of privacy practices of organizations next Monday
  – Details on piazza
18734 Recitation

Distance Metrics

October 7, 2016
## Figure 1. Inpatient Microdata

<table>
<thead>
<tr>
<th>Zip Code</th>
<th>Age</th>
<th>Nationality</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13053</td>
<td>28</td>
<td>Russian</td>
</tr>
<tr>
<td>2</td>
<td>13068</td>
<td>29</td>
<td>American</td>
</tr>
<tr>
<td>3</td>
<td>13068</td>
<td>21</td>
<td>Japanese</td>
</tr>
<tr>
<td>4</td>
<td>13053</td>
<td>23</td>
<td>American</td>
</tr>
<tr>
<td>5</td>
<td>14853</td>
<td>50</td>
<td>Indian</td>
</tr>
<tr>
<td>6</td>
<td>14853</td>
<td>55</td>
<td>Russian</td>
</tr>
<tr>
<td>7</td>
<td>14850</td>
<td>47</td>
<td>American</td>
</tr>
<tr>
<td>8</td>
<td>14850</td>
<td>49</td>
<td>American</td>
</tr>
<tr>
<td>9</td>
<td>13053</td>
<td>31</td>
<td>American</td>
</tr>
<tr>
<td>10</td>
<td>13053</td>
<td>37</td>
<td>Indian</td>
</tr>
<tr>
<td>11</td>
<td>13068</td>
<td>36</td>
<td>Japanese</td>
</tr>
<tr>
<td>12</td>
<td>13068</td>
<td>35</td>
<td>American</td>
</tr>
</tbody>
</table>

## Figure 2. 4-anonymous Inpatient Microdata

<table>
<thead>
<tr>
<th>Zip Code</th>
<th>Age</th>
<th>Nationality</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>130**</td>
<td>&lt; 30</td>
<td>American</td>
</tr>
<tr>
<td>2</td>
<td>130**</td>
<td>&lt; 30</td>
<td>American</td>
</tr>
<tr>
<td>3</td>
<td>130**</td>
<td>&lt; 30</td>
<td>Japanese</td>
</tr>
<tr>
<td>4</td>
<td>130**</td>
<td>&lt; 30</td>
<td>American</td>
</tr>
<tr>
<td>5</td>
<td>1485*</td>
<td>≥ 40</td>
<td>American</td>
</tr>
<tr>
<td>6</td>
<td>1485*</td>
<td>≥ 40</td>
<td>Russian</td>
</tr>
<tr>
<td>7</td>
<td>1485*</td>
<td>≥ 40</td>
<td>American</td>
</tr>
<tr>
<td>8</td>
<td>1485*</td>
<td>≥ 40</td>
<td>American</td>
</tr>
<tr>
<td>9</td>
<td>130**</td>
<td>3*</td>
<td>American</td>
</tr>
<tr>
<td>10</td>
<td>130**</td>
<td>3*</td>
<td>American</td>
</tr>
<tr>
<td>11</td>
<td>130**</td>
<td>3*</td>
<td>Japanese</td>
</tr>
<tr>
<td>12</td>
<td>130**</td>
<td>3*</td>
<td>American</td>
</tr>
</tbody>
</table>
Goal of Statistical Disclosure Control

Reveal accurate statistics about a population while preserving the privacy of individuals
Stephanie Sun is one inch shorter than the average Russian woman.
Differential Privacy

Control

Experimental

Does smoking increase your chances of getting cancer?

Control

Experimental

Does smoking increase your chances of getting cancer?

Yes

Yes
Input perturbation

Noise injection
L1 Distance

• Between two points
  \(- (x_1, x_2, ..., x_n) \text{ and } (y_1, y_2, ..., y_n)\)

• \(\sum_i |x_i - y_i|\)
Distance between functions

• Between two discrete functions
  – $m_1(x), m_2(x)$
  – $x \in \{x_1, x_2, \ldots, x_n\}$
  – $\sum_i |m_1(x_i) - m_2(x_i)|$

• Between two continuous functions
  – $n_1(y), n_2(y)$
  – $y \in [y_1, y_2]$
  – $\int_{y_1}^{y_2} |n_1(y) - n_2(y)| \, dy$

$d_1(f, g)$ is the shaded area
Distance between probability distributions

• Between two discrete distributions
  – PMFs $p_1(x), p_2(x)$
  – $x \in \{x_1, x_2, ..., x_n\}$
  – $\sum_i |p_1(x_i) - p_2(x_i)|$

• Between two continuous distributions
  – PDFs $f_1(y), f_2(y)$
  – $y \in [y_1, y_2]$}
  – $\int_{y_1}^{y_2} |f_1(y) - f_2(y)| dy$
Exercise

• Find L1 distance between the following continuous distributions:
  
  – \( f_1(x) = \frac{x}{12} \quad x \in [1, 5] \)
  
  – \( f_2(x) = \frac{1}{4} \quad x \in [1, 5] \)
Solution: $\frac{1}{3}$
Individual Fairness

Treat *similar* individuals *similarly*

- Similar for the purpose of the classification task
- Similar distribution over outcomes
Fairness through Awareness

Metric $d : V \times V \rightarrow \mathbb{R}$

Lipschitz condition $\|M(x) - M(y)\| \leq d(x, y)$

$M : V \rightarrow \Delta(O)$

$V$: Individuals

$O$: outcomes
Statistical Distance

• P, Q are probability measures on a finite domain A.

• Statistical distance between P and Q is:

\[ D(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)| \]

• where M(x)=P, M(y)=Q, O=A
• P, Q are probability measures on a finite domain A.

• Statistical distance between P and Q is:
  \[ D(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)| \]

  • where M(x)=P, M(y)=Q, O=A

**Example:** High D

A= \{0,1\}

P(0) = 1, P(1) = 0
Q(0) = 0, Q(1) = 1

D(P, Q) = 1
Statistical Distance

• P, Q are probability measures on a finite domain A.

• Statistical distance between P and Q is:

$$D(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)|$$

• where M(x)=P, M(y)=Q, O=A

Example: Low D
A= \{0,1\}
P(0) = 1, P(1) = 0
Q(0) = 1, Q(1) = 0
D(P, Q) = 0
Statistical Distance

• $P, Q$ are probability measures on a finite domain $A$.

• Statistical distance between $P$ and $Q$ is:

$$D(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)|$$

• where $M(x) = P, M(y) = Q, \Omega = A$

Example: Mid D

$A = \{0, 1\}$

$P(0) = P(1) = \frac{1}{2}$

$Q(0) = \frac{3}{4}, Q(1) = \frac{1}{4}$

$D(P, Q) = \frac{1}{4}$
Installing and Running AdFisher
Setting up the environment

• AdFisher has been tested on Ubuntu 16.04 with Firefox 45.
• Use a VM if you are running Windows or Mac
  – https://www.virtualbox.org/wiki/Downloads
• Ubuntu
  – https://www.ubuntu.com/download/download/desktop
Downgrade Firefox to Version 45

```
firefox --version Mozilla Firefox 47.0
apt-get remove firefox
wget
tar -xjf firefox-45.0.tar.bz2
mv firefox /opt/firefox45
ln -s /opt/firefox45/firefox /usr/bin/firefox
firefox --version Mozilla Firefox 45.0
```

Installing the AdFisher

• Clone the git repository
  – https://github.com/tadatitam/info-flow-experiments

• Follow the instructions to install the python packages AdFisher uses:
  – https://github.com/tadatitam/info-flow-experiments/tree/master/AdFisher
Testing AdFisher

• Cd into AdFisher/examples
• Run python demo_exp.py