Special Topics in Security and Privacy of Medical Information

Sujata Garera

Privacy
- HIPAA
- Anonymity
- Hippocratic databases

HIPAA
- Health Insurance
- Portability and Accountability
- Act of 1996
HIPAA

- What is HIPAA?
  - Legislation to protect workers
  - Protect integrity and confidentiality of health information

One of the security standards states the following:
Physical safeguards to guard data integrity, confidentiality, and availability. These safeguards protect physical computer systems, medical documents and related buildings and equipment from intrusion. The use of locks, keys, and administrative measures used to control access to computer systems and facilities are also included.

HIPAA rules

- Privacy rule
  - Regulations for use and disclosure of Protected Health Information
- Security rule
  - Deals specifically with Electronic PHI
  - Three types of safeguards
    - Administrative, physical and technical

HIPAA

- Why do we need such rules?
HIPAA

- Protect PHI
  - Protected health information may be transmitted in any form
  - Limit sharing of PHI without appropriate authorization
  - Sharing of PHI may be permitted for clinical purposes

Who do the rules apply to?

- Covered Entities
- Business Associates

Penalties

- Civil monetary penalties
  - Unintentional
    - $100 per violation
  - Intentional
    - Upto a maximum of $250,000 and/or imprisonment for ten years if offense includes intent
Stimulus bills dramatically modifies certain rules

- Enforcement and penalties
  - Damages can now be sought!
- New business associate requirements
  - Business associate contracts may have to be revised
- Breach notifications
  - Notifications must be made with respect to breach of unsecured PHI
- Accounting of certain PHI disclosures
  - Right to obtain accounting of disclosure

From the HIPAA blog

- Actually, early on in the life of the HIPAA Security Rule, many IT guru types jumped onto the encryption bandwagon with both feet, saying things like "encryption is industry standard and failure to encrypt is per se an unreasonable violation of the Security Rule," or "sending email over the internet in clear text (i.e., unencrypted) is a violation of the Security Rule." Well, the Security Rule has always listed encryption as an adoptable standard, not a required one; that means any covered entity must review its operations, practices, capabilities and finances and determine whether it should encrypt, but that it may reasonably determine that encryption is not necessary for structural, organizational, operational, or financial reasons. I have consistently advised people that you have to take an honest look, but if you determine that encryption isn't necessary, you don't have to do it.
- That reasoning holds true today; you are still Security Rule compliant if you've made this determination. HOWEVER, under the new Data Breach Rules, your obligations upon a data breach are dramatically higher if you do not encrypt. Encryption, done properly, will be a "get out of jail free" card if you have a data breach.

De-identifying patient records

- More attention is being placed on anonymizing patient records
- De-identifying data escapes the requirements of HIPAA and breach notification
- When properly implemented privacy of individuals is protected in healthcare datasets and
- De-identified data useful for research purposes as well
SCRUB System

- Replacing personally identifying information in medical records, the Scrub System
  - Latanya Sweeny [1996]
  - Utilize many detection techniques to label personally identifying information
    - E.g. John Smith and A.W. Gross are both full names
      - Identifying a full name is not the same as recognizing its independent parts

Personal information formats

<table>
<thead>
<tr>
<th>Names</th>
<th>Phone numbers</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frank Graves</td>
<td>255-1423</td>
<td>March 1, 1991</td>
</tr>
<tr>
<td>F. R. Graves, MD</td>
<td>(304) 255-1423</td>
<td>3/1/91</td>
</tr>
<tr>
<td>Dr. Graves</td>
<td>304/255-1423</td>
<td>first of March</td>
</tr>
<tr>
<td>Frank Red Graves</td>
<td>255-1000 ext 1423</td>
<td>1-MAR-91</td>
</tr>
<tr>
<td>&quot;Red&quot; Graves</td>
<td>phone: 255-1423</td>
<td>03-01-91</td>
</tr>
</tbody>
</table>
| frank red graves  | extension 1423 | March 1st     

Table 1. Samples of personal information.

Entities recognized by Scrub

<table>
<thead>
<tr>
<th>Scrub</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Identification block</td>
</tr>
<tr>
<td>2.</td>
<td>Name</td>
</tr>
<tr>
<td>3.</td>
<td>Address block</td>
</tr>
<tr>
<td>4.</td>
<td>Full name</td>
</tr>
<tr>
<td>5.</td>
<td>Address block</td>
</tr>
<tr>
<td>6.</td>
<td>Phone block</td>
</tr>
<tr>
<td>7.</td>
<td>Email</td>
</tr>
<tr>
<td>8.</td>
<td>Contact information</td>
</tr>
<tr>
<td>9.</td>
<td>Address block</td>
</tr>
<tr>
<td>10.</td>
<td>Phone number</td>
</tr>
<tr>
<td>11.</td>
<td>First name</td>
</tr>
<tr>
<td>12.</td>
<td>Last name</td>
</tr>
<tr>
<td>13.</td>
<td>Date</td>
</tr>
<tr>
<td>14.</td>
<td>Reference number</td>
</tr>
<tr>
<td>15.</td>
<td>Medical area</td>
</tr>
</tbody>
</table>

Table 1. List of the entities recognized by Scrub are listed above in relative order of precedence.
Prior instances found

- Knowing instances found is useful in reducing ambiguity
  - Virgina P. Weston
  - “After seeing Virgina this time I feel much better”
  - Second reference now disambiguated to be a persons name rather than a state due to first reference

Templates and probabilities

<table>
<thead>
<tr>
<th>Phone numbers</th>
<th>Template</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>255-1423</td>
<td>400-5555</td>
<td>80</td>
</tr>
<tr>
<td>555-1423</td>
<td>400-5555</td>
<td>85</td>
</tr>
<tr>
<td>904/255-1423</td>
<td>.444/.444/.444</td>
<td>50</td>
</tr>
<tr>
<td>555-1000 ext 1423</td>
<td>.444-.444-.444-444</td>
<td>70</td>
</tr>
<tr>
<td>extension 1423</td>
<td>ext*#*</td>
<td>40</td>
</tr>
<tr>
<td>phone 255-1423</td>
<td>(int, int*)</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>.444/.444</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Samples of templates and their probabilities. The d is a digit, the asterisk matches any wild character and the set notation denotes possibilities.

Replacement

- With each detection algorithm there is a replacement algorithm
  - If detection was of a date replacement may involve lumping days to the first of the nearest month or some other grouping
Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Other documents</th>
<th>Letters only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Straight search</td>
<td>32%</td>
<td>37%</td>
</tr>
<tr>
<td>Search with cues</td>
<td>32%</td>
<td>64%</td>
</tr>
<tr>
<td>Scrub (threshold 0.7)</td>
<td>99%</td>
<td>99%</td>
</tr>
<tr>
<td>Scrub (threshold 0.5, false positive reduction)</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 4. Comparisons of Scrub to standard techniques.

Anonymity

- Does SCRUB provide anonymity?
  - Can you be guaranteed that the data cannot be manipulated, matched or linked to any individual?

- Scrub merely de-identifies data
- Anonymity is still not provided with the scrub system
  - Possible to employ reverse scrubbing techniques to identify and link to an individual
Anonymity

- Can the following data be considered as anonymous?

<table>
<thead>
<tr>
<th>ZIP Code</th>
<th>Birthdate</th>
<th>Gender</th>
<th>Ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>33171</td>
<td>7/1/71</td>
<td>m</td>
<td>Caucasian</td>
</tr>
<tr>
<td>02137</td>
<td>2/18/73</td>
<td>f</td>
<td>Black</td>
</tr>
<tr>
<td>20612</td>
<td>3/7/73</td>
<td>m</td>
<td>Asian</td>
</tr>
</tbody>
</table>

Table 2: Do-identified data that is not anonymous

Anonymous data challenges

- 1997 voting list for Cambridge MA contains demographics on 54,805 voters.
  - Birth date alone can uniquely identify the name and address of 12% of the voters
  - 29% by just birth date and gender
  - 69% by birth date and a 5 digit ZIP code
  - 97% when a full postal code and birthdate

Anonymous data challenges

- Unique and unusual information may appear in data sets
  - Medical records of a pediatric hospital where only one patient is above 45 years of age
  - Hospitals maternity ward contains only one patient who gave birth to octuplets
    - E.g. Octomom
Anonymous data challenges

- Measuring the degree of anonymity is also challenging [1997]
  - SSAs general rule for data: Any subset of the data that can be defined in terms of combinations of characteristics must contain at least 5 individuals
    - Minimal bin size
    - How is the bin size relate to anonymity?

Anonymous data challenges

- Medical databases
  - Minimum bin size should be much larger than the SSA guidelines
  - Determining the optimal bin size to ensure anonymity is tricky
    - Map 10 people to one bin
    - Map 100 people to one bin
    - Map 1000 people to one bin

Datafly
User also provides a profile of the person who is the recipient of the data.
- Specifies whether the recipient could have or would use information external to database for linking records
- 0 represents full trust and 1 full distrust
- Profile thresholds help to restore the effective bin size by forcing these fields to adhere to bin sizes larger than overall anonymity level.

Used a database of pediatric medical record
- 300 records

Table 3. The slide bar highlights a 0.7 anonymity level for the birth date field of the Cambridge voters data.

Table 4. Results as the overall anonymity level increases. Each cell reports the minimum anonymity level that achieved the required bin size.
Datafly

- Maintaining anonymity as per preferred user levels
  - Interestingly:
    - IRB required for researchers to use data but not administrators
    - Datafly can be used to make uniform policies based on trust levels assigned

Suggested readings

- Scrub and Datafly systems available online on the website