Uncovering Anomalous Usage of Medical Records via Social Network Analysis

You Chen, Ph.D.
Biomedical Informatics Dept., School of Medicine
EECS Dept., School of Engineering
November 2, 2011

(Joint work with Bradley Malin, Steve Nyemba, and Wen Zhang)
Fact Sheet 8a: HIPAA Basics: Medical Privacy in the Electronic Age

[Also see our FAQ on medical privacy.]

1. Introduction
2. HIPAA Privacy Rule: Benefits and Shortcomings
4. Medical Information: What Does HIPAA Cover?
5. Control of Your Medical Information: "Consent" and "Authorization"
6. More About Your Right to Access Your Medical Records
7. Your Health Records and Your Employer
8. Your Health Records and the Government
9. Your Health Information and Your Credit Report
10. HIPAA and Your Daily Routine
11. Complaints and Penalties for Violations
12. The HIPAA Security Rule
13. Electronic Health Records (EHRs)

However, HIPAA's shortcomings and lack of clarity have fed the public's concern about the potential risks to privacy associated with having the most personal data imaginable stored in electronic format. Add to this, the nearly constant barrage of news stories about health data being accessed by hackers, lost with laptop computers, or simply read by curious employees, and it is little wonder consumers are concerned about privacy.
California Hospitals Fined $675,000 For Privacy Violations: Jun 11, 2010

What Every Risk Manager Needs To Know About Copy Machines: May 8, 2010 CBS exposes the risks to healthcare and other entities for possible data breach violations, financial privacy laws violations, HIPAA violations, and makes them vulnerable to being targeted by criminals for theft or other crimes. The video also shows an example of how they can be exploited by terrorists. Of course, that is in addition to the risks posed by improper use and distribution of copies themselves. View Here.

Be Prepared To Deal With Exploding Medical ID Theft And Privacy Issues In Healthcare: May 3, 2010 Medical ID Theft is hitting the headlines as organized crime and ID thieves grab millions in false claims and leave innocent patients and healthcare providers with the bills. By Stephen A. Frew JD.

Hospital Employee Gets Jail Time For HIPAA Violation: Apr 29, 2010 Hospital employee sentenced to federal prison for 3-week long medical records spree.

UCLA Employee Indicted For Celebrity Privacy Violations: May 8, 2008 Hospital employee sells celebrity medical info to tabloids.
Two Typical Attacks

(1) Anomalous users detection – user level
(2) Anomalous accesses detection – access level

Intruders have little knowledge of the system and the anticipated behavior

Intruders have complete knowledge of the system and its policies
Related Research

Access Control Models

- Role-based
- Situation-based
- Role Mining

Do not capture the dynamic relationships among users in collaborative information systems
Does not offer stability of access control model over time

Auditing Models

- Community Anomaly Detection
- Specialized Anomaly Detection
- K-Nearest Neighbors
- Spectral Analysis
- High-Volume Users
- Machine Learning
- Case-Based Reasoning

Related Research

Anomalous Usage
Two general objects of health information system

U(sers)

S(subjects)

Accesses

Behavioral Modeling
Where are We Going?

User Level Anomaly Detection
Community Anomaly Detection System (CADS)
(ACM CODASPY’11)

Access Level Anomaly Detection
Specialized Network Anomaly Detection (SNAD)
(IEEE ISI’11)
Social Networks are a Novel Approach to Discovery of Electronic Medical Record Misuse

CADS: Leverages a **global** view of the network

SNAD: A **local** view of the network

Anomalous Usage
Example Environments

Electronic Health Records (EHR)

- Vanderbilt University Medical Center
  “StarPanel” Logs
- 6 months in 2006
- Arbitrary Week
  \( \approx 2,300 \) users
  \( \approx 35,000 \) patient records
  \( \approx 66,000 \) accesses
Where are We Going?

• User Level: Community Anomaly Detection System (CADS) (ACM CODASPY’11)
  – Framework of CADS
  – An Example of CADS
  – Experimental Evaluation
  – Limitation

• Access Level: Specialized Network Anomaly Detection (SNAD) (IEEE ISI’11)
Community-Based Anomaly Detection (CADS)

Access Logs

Pattern Extraction
- Social Relation Construction
- Community Derivation

Communities of users

Anomaly Detection
- Distance Measurement
- Deviation Measurement

Deviation Scores for Users

Nearest neighbor networks

<user, subject, time>
Where are We Going?

• User Level: Community Anomaly Detection System (CADS) (ACM CODASPY’11)
  – Framework of CADS
  – An Example of CADS
  – Experimental Evaluation
  – Limitation

• Access Level: Specialized Network Anomaly Detection (SNAD) (IEEE ISI’11)
Bipartite graph
Communities Derivation
Communities
Distance Measurement
Nearest Neighbor Network
Deviation Measurement
Deviation Scores

<table>
<thead>
<tr>
<th>User</th>
<th>2-NN</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>u₁</td>
<td>u₂</td>
<td>u₃</td>
</tr>
<tr>
<td>u₂</td>
<td>u₄</td>
<td>u₅</td>
</tr>
<tr>
<td>u₃</td>
<td>u₁</td>
<td>u₂</td>
</tr>
<tr>
<td>u₄</td>
<td>u₂</td>
<td>u₅</td>
</tr>
<tr>
<td>u₅</td>
<td>u₂</td>
<td>u₄</td>
</tr>
<tr>
<td>u₆</td>
<td>u₁</td>
<td>u₃</td>
</tr>
</tbody>
</table>
How Do We Set “k”-NN?

- Conductance- a measure of community quality (Kannan et al)

\[ \psi(\beta) = \frac{2}{4}, \psi(\alpha) = \frac{2}{8}, \psi(\gamma) = \frac{2}{\min\{4,12\}} \]

\[ \psi(\alpha) < \psi(\beta) = \psi(\gamma) \]
Minimum conductance at $k=6$
The average cluster coefficient for this network is 0.48, which is significantly larger than 0.001 for random networks.

Users exhibit collaborative behavior in the health information system.
Measuring Deviation from k-NN

• Every user is assigned a radius \( d \):
  – the distance to his \( k^{th} \) nearest neighbor

• Smaller the radius \( \rightarrow \) higher density in user’s network
Where are We Going?

• User Level: Community Anomaly Detection System (CADS) (ACM CODASPY’11)
  – Framework of CADS
  – An Example of CADS
  – Experimental Evaluation
  – Limitation

• Access Level: Specialized Network Anomaly Detection (SNAD) (IEEE ISI’11)
Experimental Design

- Datasets are not annotated for illicit behavior
- We simulated users in several settings to test:
  - Sensitivity to number of records accessed
    - Range from 1 to 1,000
  - Sensitivity to number of anomalous users
    - simulated users correspond to 0.5% to 5% of total users
    - Number of records accessed fixed to 5
  - Sensitivity to diversity
    - Random number of users and records accessed
Deviation and Detection Rate Increases with Number of Subjects Accessed

![Graph showing deviation and false positive rate against patients accessed.](image)

- Deviation vs. Patients Accessed
- False Positive Rate vs. Patients Accessed

Anomalous Usage

© You Chen, 2011
Detection Rate With Various Mix Rates of Real and Simulated Users

![Graph showing detection rate with various mix rates of real and simulated users. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. The graph includes lines for different mix rates: 0.5%, 1%, 2%, and 5%.]
CADS Outperforms Competitors
(mix rate = 0.5%)

Anomalous Usage
Where are We Going?

• User Level: Community Anomaly Detection System (CADS) (ACM CODASPY’11)
  – Framework of CADS
  – An Example of CADS
  – Experimental Evaluation
  – Limitation

• Access Level: Specialized Network Anomaly Detection (SNAD) (IEEE ISI’11)
Some Limitations

• Simulated users are indicative of misuse of the system...
  ...but actual illicit behavior may be more directed.

• “False positives” are not necessarily false!
  (Adjudication by EHR privacy experts under way)

• Need to specialize tool to account for semantics of users and subjects
  – User: {Role, Department, Residence}
  – Patient: {Diagnosis, Procedure, Demographics, Residence}

• Anomalous users... not anomalous accesses
  – Need to account for insiders that deviate by only a couple of actions
  – Work underway (about to be submitted), but it’s detection is “local”, not “global”
Where are We Going?

- User Level: Community Anomaly Detection System (CADS) (ACM CODASPY’11)
- Access Level: Specialized Network Anomaly Detection (SNAD) (IEEE ISI’11)
  - Framework of SNAD
  - An Example of CADS
  - Experimental Evaluation
  - Limitation
Where are We Going?

- User Level: Community Anomaly Detection System (CADS) (ACM CODASPY’11)
- Access Level: Specialized Network Anomaly Detection (SNAD) (IEEE ISI’11)
  - Framework of SNAD
  - An Example of SNAD
  - Experimental Evaluation
  - Limitation
User Modeling

\[ IDF(u_i) = \log \frac{|S|}{1 + |\{s_j, \text{ where } SU(j, i) > 0\}|} \]
Access Network Construction

Similarity Measurement

Access Network Measurement

User Modeling

Access Logs

Access Network Construction

Graph representation with nodes and edges indicating user and access network relationships.
Access Network Measurement

\[
Sim(u_i, u_j) = \frac{U_i \cdot U_j}{||U_i|| \times ||U_j||}
\]

Anomalous Usage

© You Chen, 2011
Measuring Accesses for Changes in Network Similarity

Access: $u_1 \rightarrow s_3$

<table>
<thead>
<tr>
<th>Network</th>
<th>Similarity</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1, u_2, u_4, u_5, u_6$</td>
<td>0.59</td>
<td>5</td>
</tr>
<tr>
<td>$u_2, u_4, u_5, u_6$</td>
<td>0.64</td>
<td>4</td>
</tr>
</tbody>
</table>

Access | Score | Size |
-------|-------|------|
$u_1$-$s_3$ | 0.05  | 4    |
Where are We Going?

• User Level: Community Anomaly Detection System (CADS) (ACM CODASPY’11)

• Access Level: Specialized Network Anomaly Detection (SNAD) (IEEE ISI’11)
  – Framework of SNAD
  – An Example of SNAD
  – **Experimental Evaluation**
  – Limitation
Experimental Design

• Datasets are not annotated for illicit behavior

• We simulated users in several settings to test:
  – Sensitivity to number of subjects accessed
    • Range from 1 to 1,00
  – Sensitivity to number of anomalous users
    • Range from 2 to 20
    • Number of subjects accessed fixed to 5
  – Sensitivity to diversity
    • Random number of users and subjects accessed
SNAD: Deviation Rate Increase with Number of Subjects Accessed

Number of Subjects the Intruder Accesses

AUC

SNAD
Spectral model on IDF matrix
Spectral model on binary matrix
SNAD: Deviation Rate Increases with Number of Intruders

![Graph showing the relationship between AUC and the number of intruders.](image)
SNAD Outperforms Competitors When the Number of Intruders & Accessed Subjects is Random
Where are We Going?

- User Level: Community Anomaly Detection System (CADS) (ACM CODASPY’11)
- Access Level: Specialized Network Anomaly Detection (SNAD) (IEEE ISI’11)
  - Framework of SNAD
  - An Example of SNAD
  - Experimental Evaluation
  - Limitation
Limitations

• SNAD has high performance in Vanderbilt’s EHR system because
  – organization is collaborative
  – access networks have high network similarity

• SNAD may not be appropriate for large access network with low network similarity
  – Absence of a user has little influence on the similarity.
Conclusions

• It is an effective way by using social network analysis to detect anomalous usages of electronic health records, such as CADS and SNAD

• Adding semantic information of users and subjects will make social network analysis be more understandable
References

Acknowledgements

Vanderbilt
- Erik Boczko, Ph.D., Ph.D.
- Josh Denny, M.D.
- Dario Giuse, Dr. Ing
- Bradley Malin, Ph.D.
- Steve Nyemba, M.S.
- John Paulett, M.S.
- Jian Tian
- Wen Zhang, M.S.

UIUC
- Carl Gunter, Ph.D.
- Igor Svecs

Northwestern
- David Liebovitz, M.D.
- Sanjay Mehotra, Ph.D.

Funding
- National Science Foundation
  - CCF-0424422 (TRUST)
  - CNS-0964063
- National Institutes of Health
  - R01LM010207
Questions? Comments?

you.chen@vanderbilt.edu

Health Information Privacy Lab:
http://www.hiplab.org/
SNAD assumes that access scores are approximately distributed around a well-centered mean.
The correlation coefficient between real and Laplace distribution is 0.886.