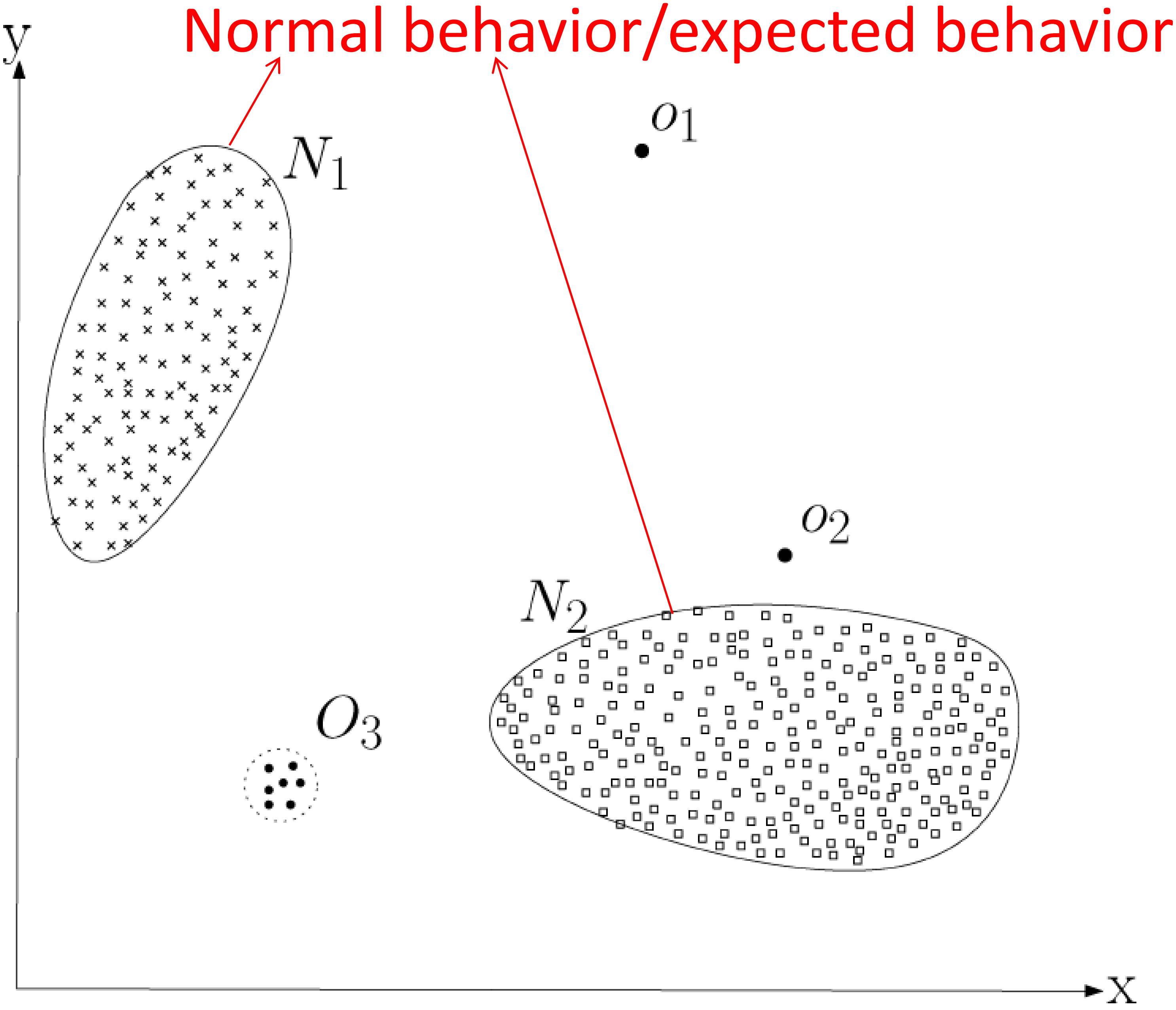
Anomaly Detection

You Chen

Two questions:
(1) What is Anomaly Detection?
(2) What are Anomalies?

- Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior

- Anomalies are patterns in data that do not conform to a well defined notion of normal behavior



Normal behavior/expected behavior

What are Challenges of Anomaly Detection?

- Defining a normal region that encompasses every possible normal behavior is very difficult
- When anomalies are the result of malicious actions, the malicious adversaries often adapt themselves to make the anomalous observations appear normal
- Normal behavior keeps evolving and a current notion of normal behavior might not be sufficiently representative in the future

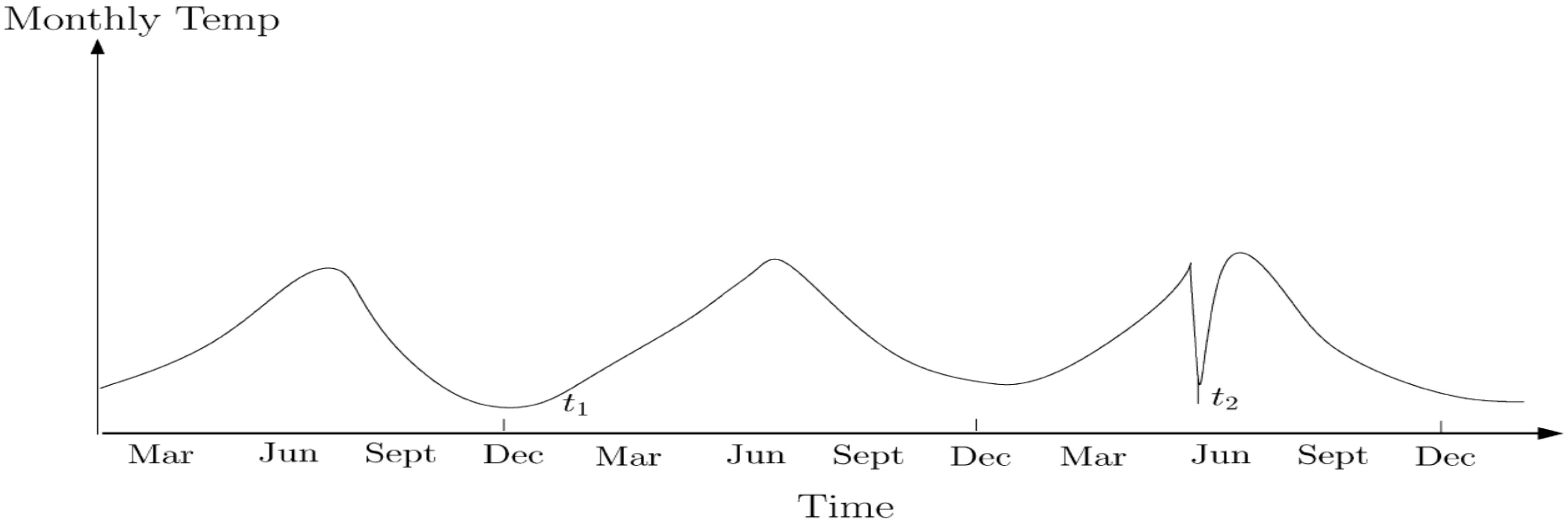
Challenges

- The exact notion of an anomaly is different for different application domains

- Availability of labeled data for training/validation of models and noise existing in the data

Type of Anomaly

- Point Anomalies
- Contextual Anomalies



Type of Anomaly

- Collective Anomalies

  . . . http-web, buffer-overflow, http-web, http-web, smtp-mail, ftp, http-web, ssh, smtp-mail,

  http-web, ssh, buffer-overflow, ftp, http-web, ftp, smtp-mail,http-web . . .

Three Anomaly Detection Models

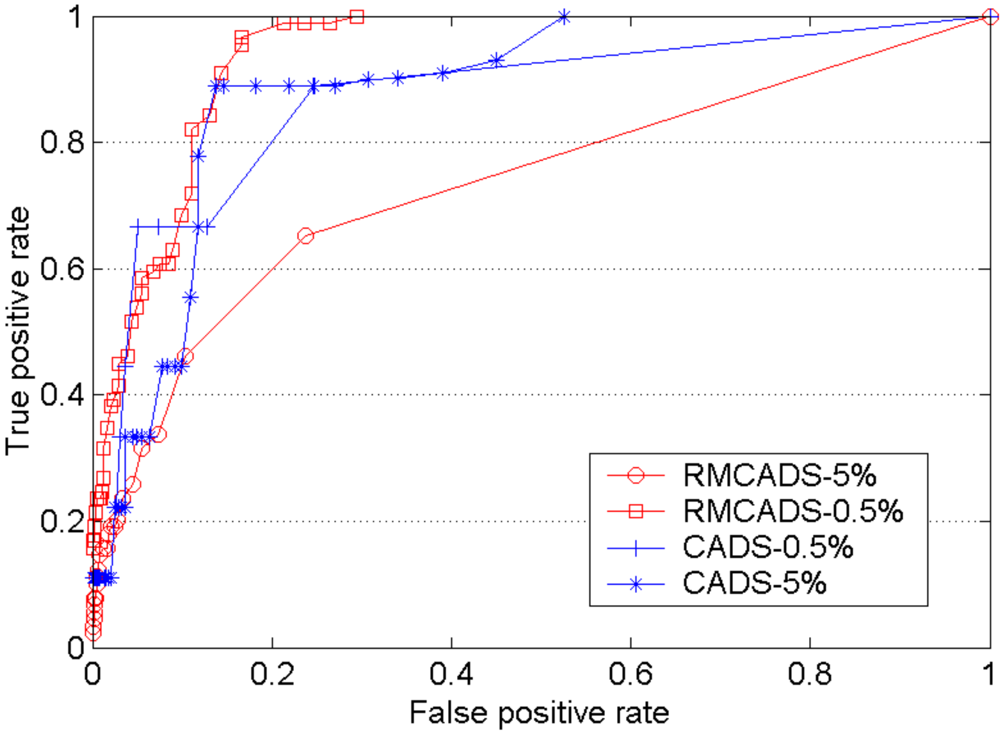
- Supervised Anomaly Detection
- Semi-supervised Anomaly Detection
- Unsupervised Anomaly Detection

Output of Anomaly Detection

- Scores
  - Scoring techniques assign an anomaly score to each instance in the test data depending on the degree to which that instance is considered an anomaly
- Labels
  - Techniques in this category assign a label (normal or anomalous) to each test instance

Performances measurement

- ROC curve and AUC

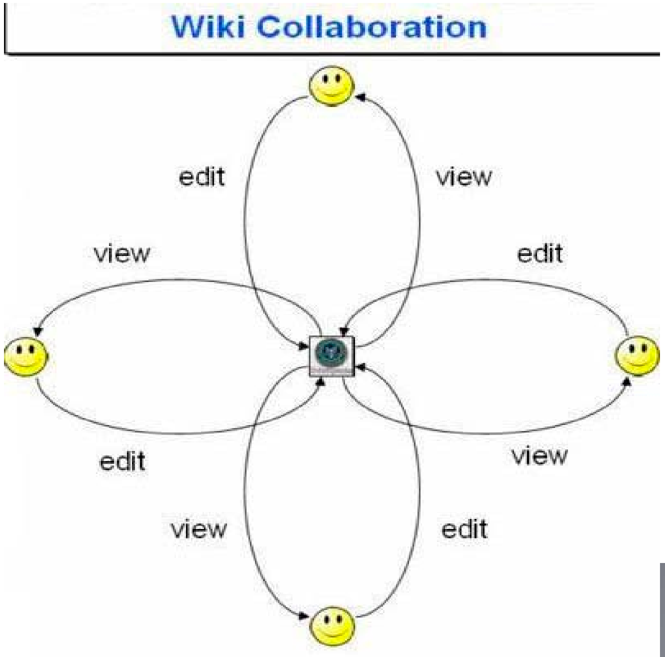


Applications

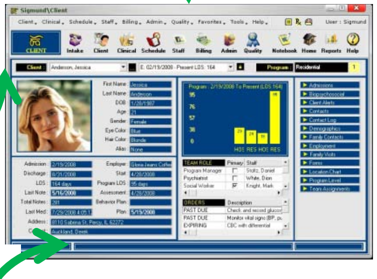
- Intrusion Detection
  - Host-Based Intrusion Detection Systems
  - Network Intrusion Detection Systems
- Medical and Public Health Anomaly Detection
  - The data typically consists of records that may have several different types of features, such as patient age, blood group, and weight
    - Aim at detecting anomalous records (point anomalies)
  - Another form of data handled by anomaly detection techniques in this domain is time-series data, such as Electrocardiograms
    - Collective anomaly detection techniques have been applied to detect anomalies in such data
  - Access Logs of Electronic Health Records Systems (EHR) points or collective anomalies.

Detection of Insider Threats in Collaborative Information System via Relational Analysis of Access Logs

Motivation



Wiki Collaboration



Client : Anderson, Jessica
E: 02/19/2008 - Present LOS: 164
Program : Residential 1

First Name : Jessica
Last Name : Anderson
DOB : 1/28/1987
Age : 21
Gender : Female
Eye Color : Blue
Hair Color : Blonde
Alias : None

Program : 2/19/2008 To Present (LOS 164)

Admission : 2/19/2008
Discharge : 8/31/2008
LOS : 164 days
Last Note : 5/16/2008
Total Notes : 281
Last Med : 7/29/2008 4:05:1
Address : 8110 Sabena St. Percy, IL 62272
Contact : Auckland, Derek

Employer : Gloria Jeans Coffee
Start : 4/28/2008
Program LOS : 95 days
Assessment : 4/28/2008
Behavior Plan :
Plan : 5/19/2008

TEAM ROLE	Primary	Staff
Program Manager	□	Stoltz, Daniel
Psychiatrist	□	White, Dion
Social Worker	☑	Knight, Mark

ORDERS	Description
PAST DUE	Check and record glucos
PAST DUE	Monitor vital signs (BP, pu
EXPIRING	CBC with differential

Motivation

Collaborative information system facilitate flexible participation and coordination between disparate users over common tasks

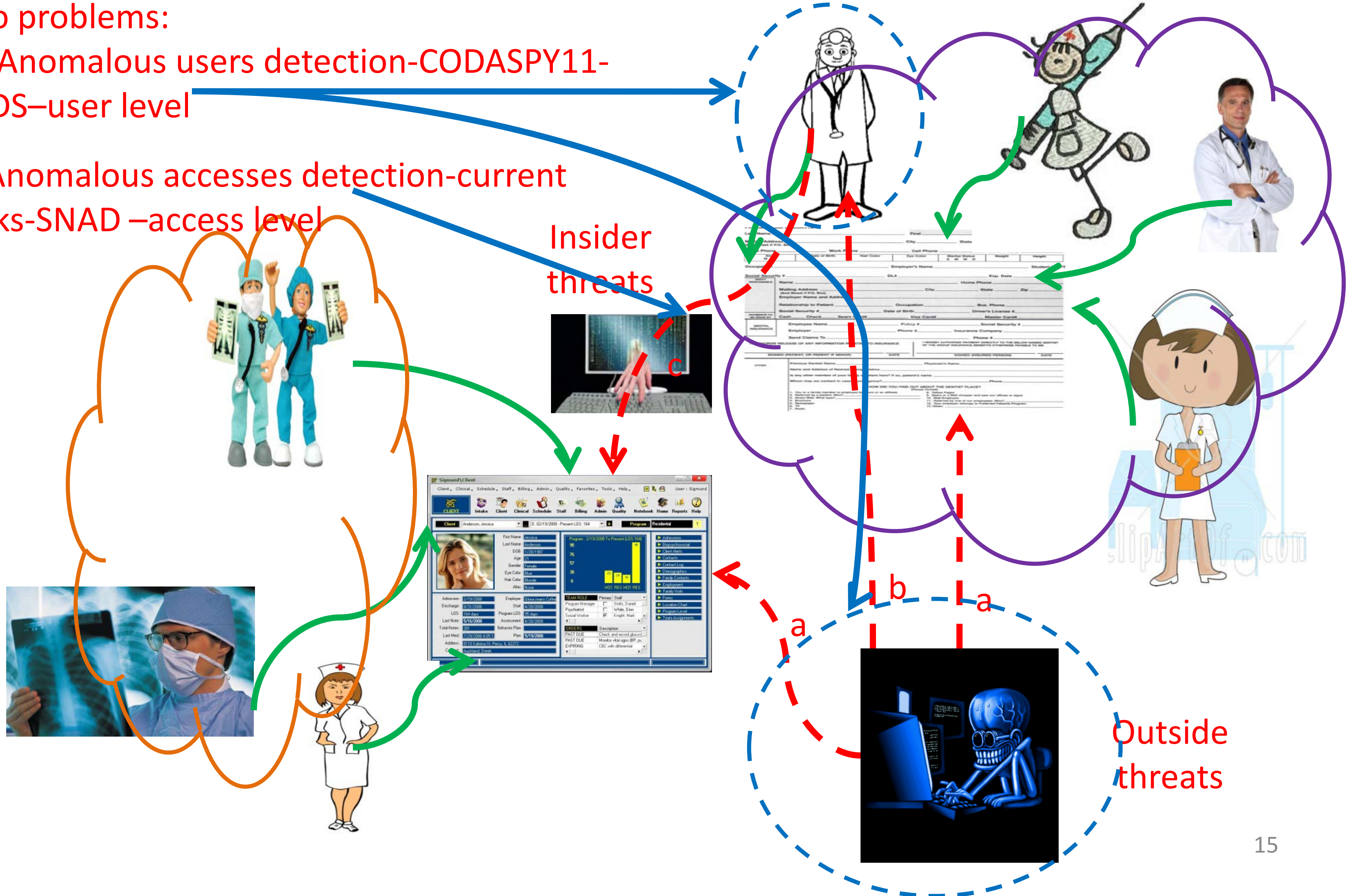
- Benefits
  - Increasing organizational efficiency
  - Shaving administrative costs
  - Facilitating social engagement
- Weakness
  - Misuse and abuse which lead to the exploitation of private information
    - Collaborative information System (CIS) are dynamic and complex
    - They consists of large number of users, permissions and ad hoc relationships between users and data elements

Problems

Two problems:

(1) Anomalous users detection-CODASPY11-CADS–user level

(2) Anomalous accesses detection-current works-SNAD –access level



Background

In a CIS, users' roles and permissions are dynamic and change over time.

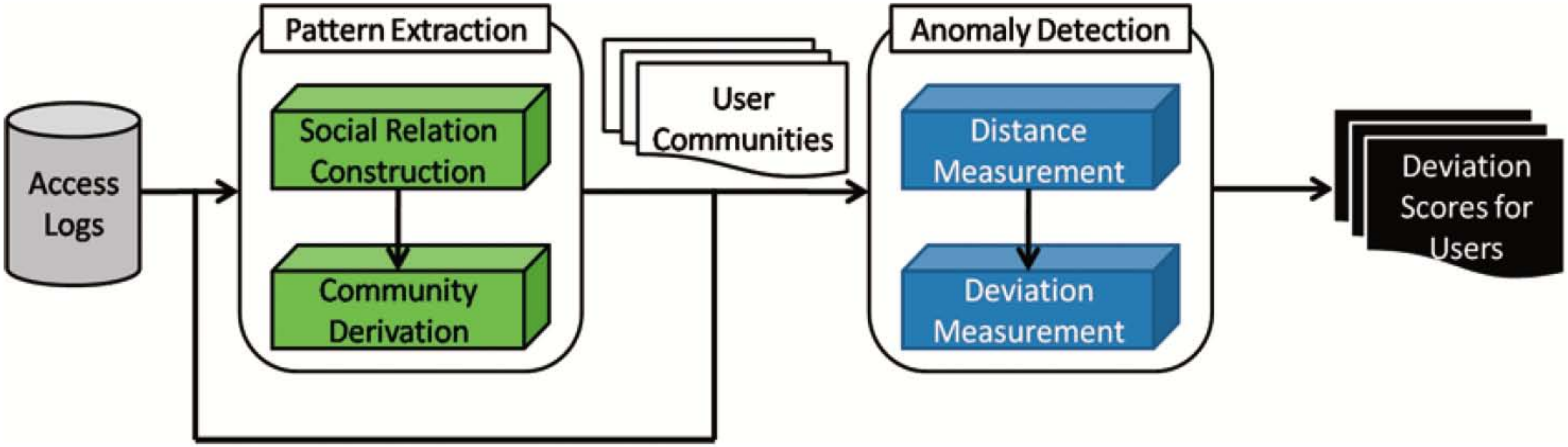
- Access Control
  - Access Matrix Model [37]
  - Role Mining [20, 27, 41]
  - Task-based Access Control and Team-based Access Control [29, 40, 11]
- Behavioral Modeling
  - Graph-based anomaly detection [7, 19, 24]
  - Community based anomaly detection [6, 28, 32]
  - Nearest neighbor based anomaly detection [18, 26, 33]

They are designed to detect if the user is behaving in an anomalous manner, they do not model if any particular action executed by a user is anomalous

Methods-unsupervised-points anomalies

- Detection of anomalous insiders in collaborative environments via relational analysis of access logs-CADS
- Leveraging social networks to detects of anomalous insider actions in collaborative environments-SNAD

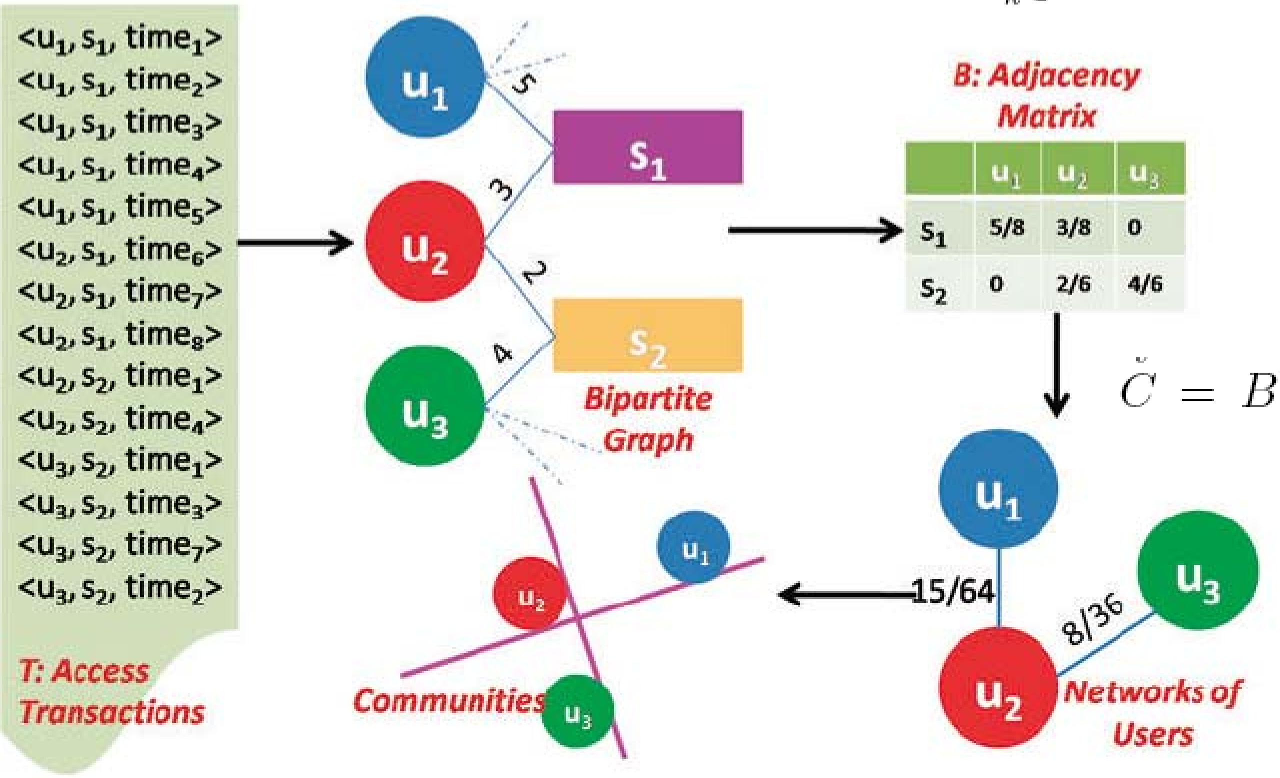
Detection of anomalous insiders in collaborative environments via relational analysis of access logs



Community Pattern Extraction

- Access Networks of Users and Subjects
- Relational Networks of Users
- Community Inference via Spectral Analysis

$$B(i,j) = \frac{count(\langle u_j, s_i, time\rangle)}{\sum_{\forall u_k \in U} count(\langle u_k, s_i, time\rangle)}$$



<u1, s1, time1>
<u1, s1, time2>
<u1, s1, time3>
<u1, s1, time4>
<u1, s1, time5>
<u2, s1, time6>
<u2, s1, time7>
<u2, s1, time8>
<u2, s2, time1>
<u2, s2, time4>
<u3, s2, time1>
<u3, s2, time3>
<u3, s2, time7>
<u3, s2, time2>

T: Access Transactions

B: Adjacency Matrix

	u1	u2	u3
s1	5/8	3/8	0
s2	0	2/6	4/6

$$\check{C} = B^T B$$

$$\omega \Lambda v^T$$

$$Z = v^T C$$

$$Z_i = [Z_{i1}, Z_{i2}, ..., Z_{in}]$$

Community-based Anomaly Detection

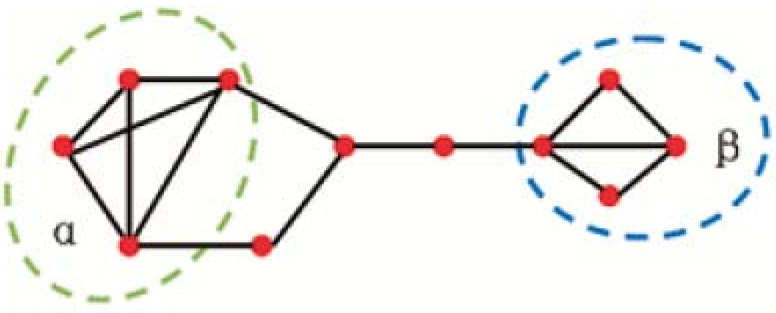
- Finding K Nearest Neighbors
- Measuring Deviation from K Nearest Neighbor Networks

$$Dis(u_i, u_j) = \sqrt{\sum_{q=1}^{l}((Z_{qi} - Z_{qj})^2 \times \lambda_q / \lambda_{total})}$$

$$\sum_{i=1}^{l} \lambda_i / \sum_{j=1}^{n} \lambda_j (l \prec n) \qquad \lambda_{total} = \sum_{j=1}^{l} \lambda_j$$

- How to determine K?
  - Network community profile which is a measure of community quality [21,22] $\phi(k) = min_{|A|=k}\psi(A)$

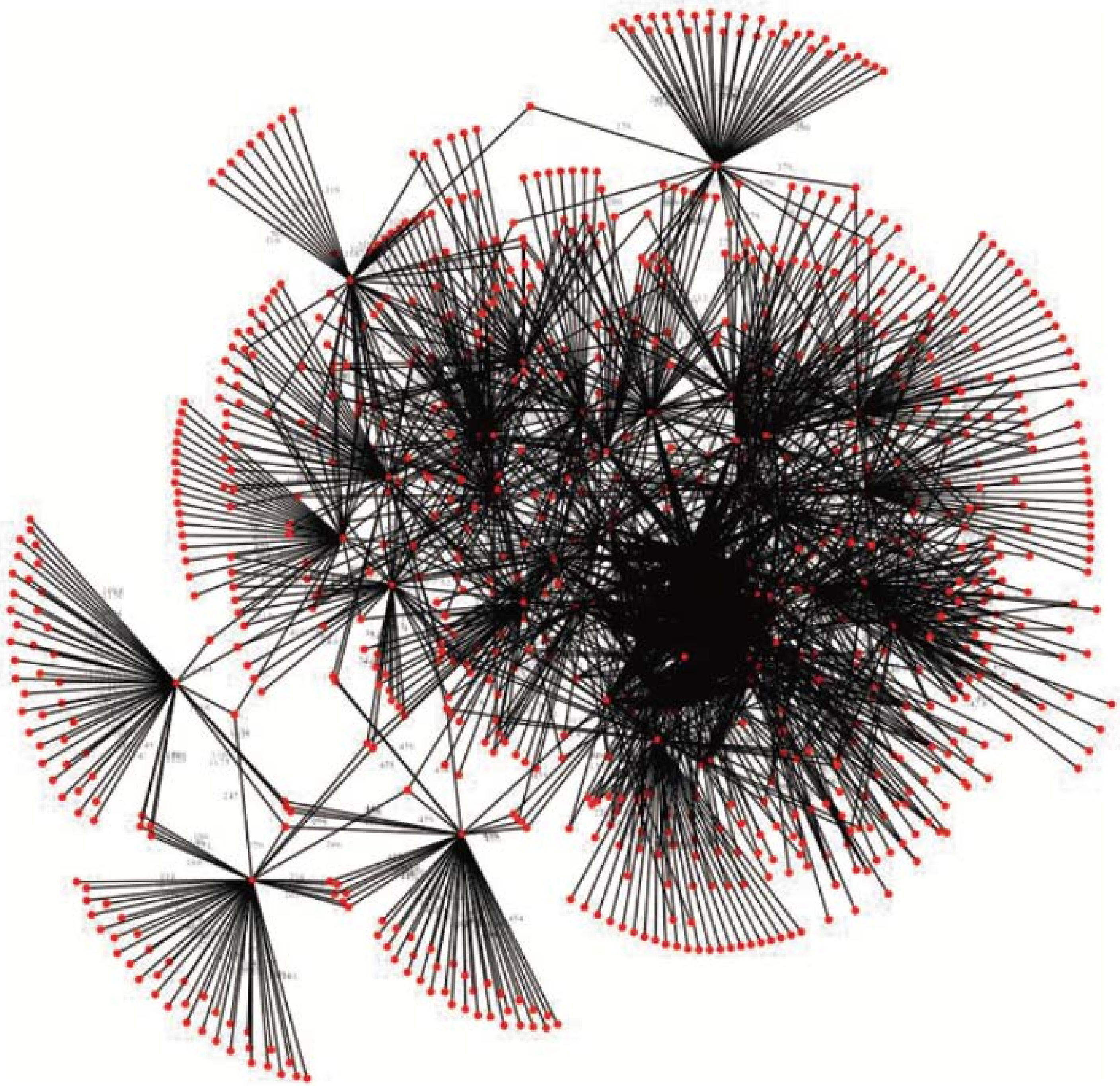
2/14



1/11

$$\psi(A) = N_A / min(Vol(A), Vol(V \setminus A))$$

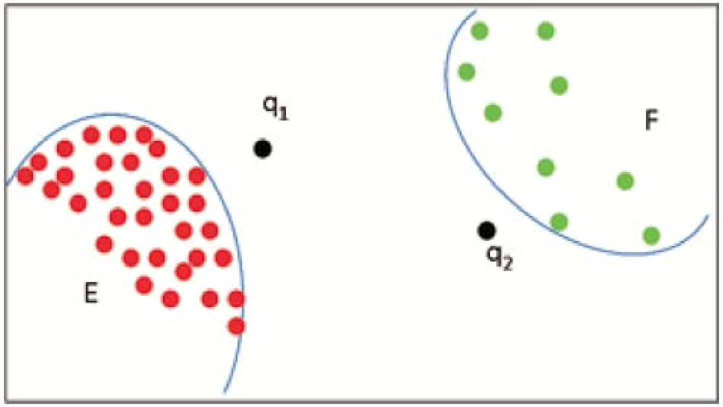
$N_A = |(g,h) : g \in A, h \notin A|$ $Vol(A) = \sum_{g \in A} d(g)$



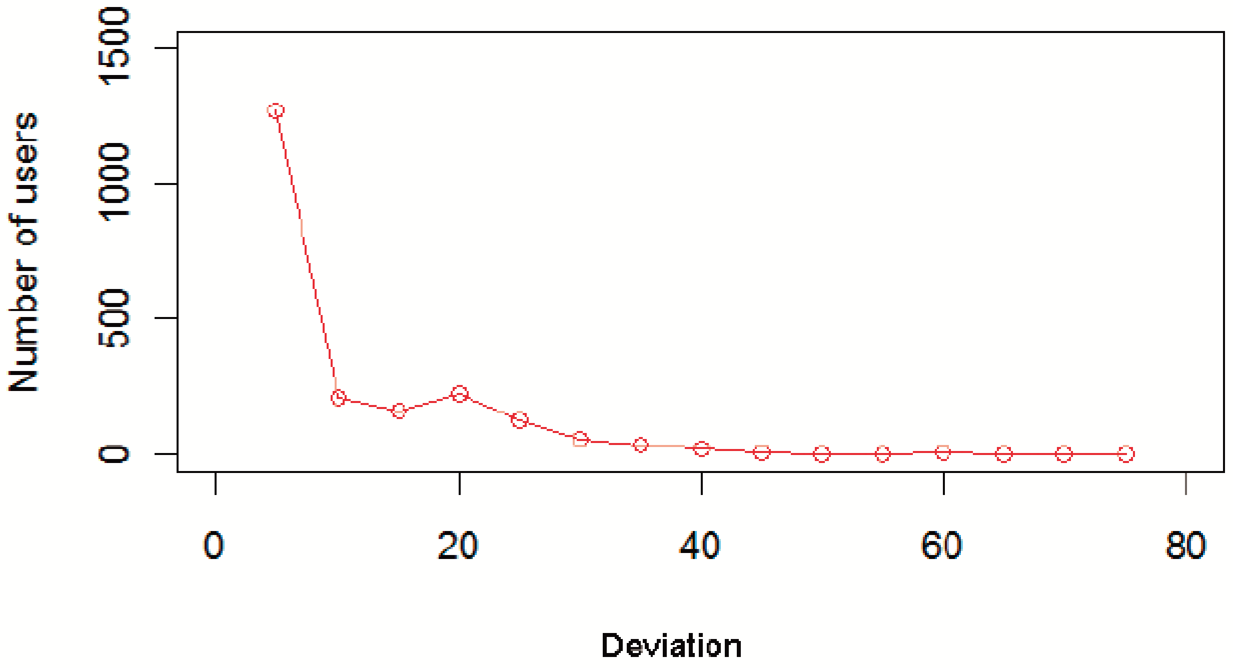
- How to measure deviation from nearest neighbors?
  - Every user can be assigned a radius value *d* by recording the distance to his k^th nearest neighbor
  - The smaller the radius, the higher density of the user's network

$$Dev(u_i) = \sqrt{\sum_{u_j \in knn_i} (d_j - \bar{d})^2 / k}$$

$$\bar{d} = \sum_{u_j \in knn_i} d_j / (k+1)$$



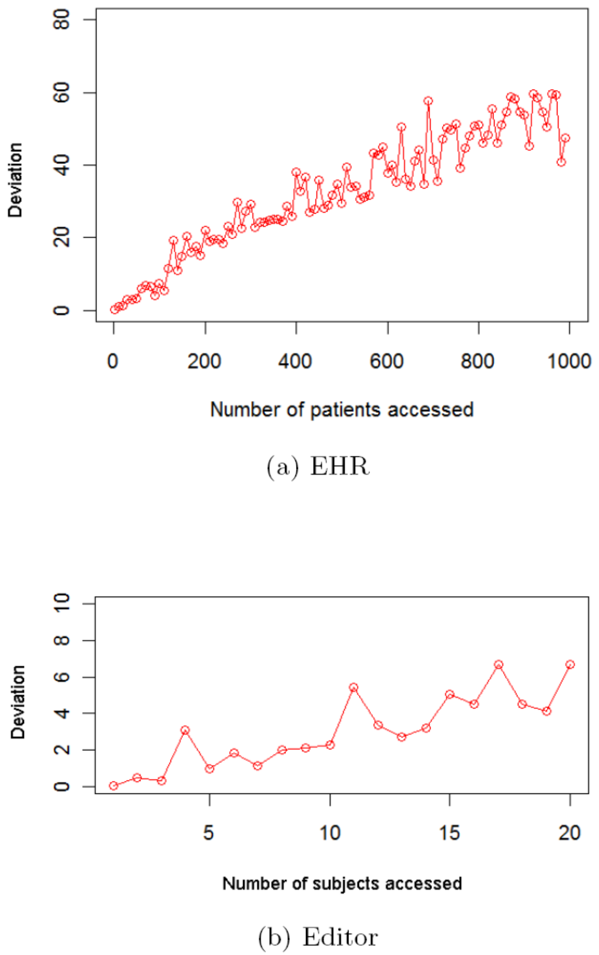
- Distribution of user deviations on a real EHR data set



- Experiments
  - EHR access log dataset
    - Analyzing 6 months of access logs from the year 2006
    - The access matrix of users and patients is very sparse
    - In an arbitrary week, there are 35,531 patients, 2,377 users and 66,441 access transactions; In other words, only 0.07% of the possible user-patient edges were observed
  - Public data set
    - The editorial board memberships for a set of journals in biomedical informatics over the years of 2000 to 2005
    - It contains 1,245 editors and 49 journals

- Simulation of users
  - Sensitivity to number of records accessed
    - Ranging from 1 to 1,000 in EHR data set
    - 1 to 20 in Editor data set
  - Sensitivity to number of anomalous users
    - The simulated users from 0.5% to 5% of the total number of users
  - Sensitivity to diversity

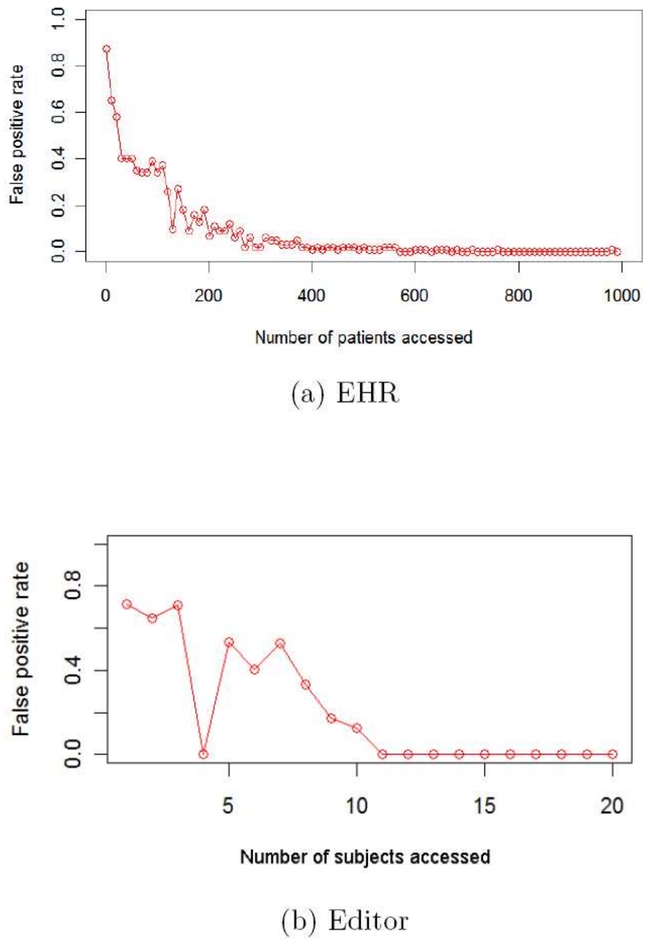
- Results



(a) EHR

(b) Editor

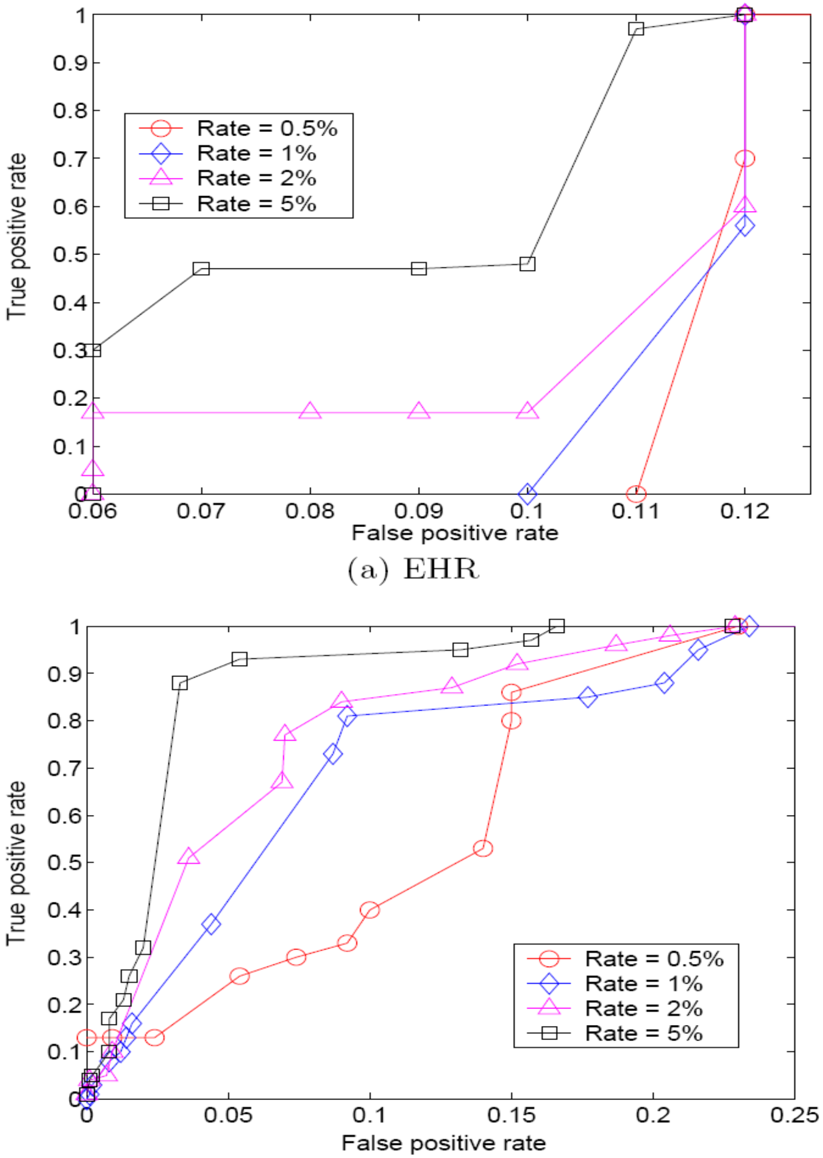
Figure 8: CADS deviation score of the simulated user as a function of number of subjects accessed.



(a) EHR

(b) Editor

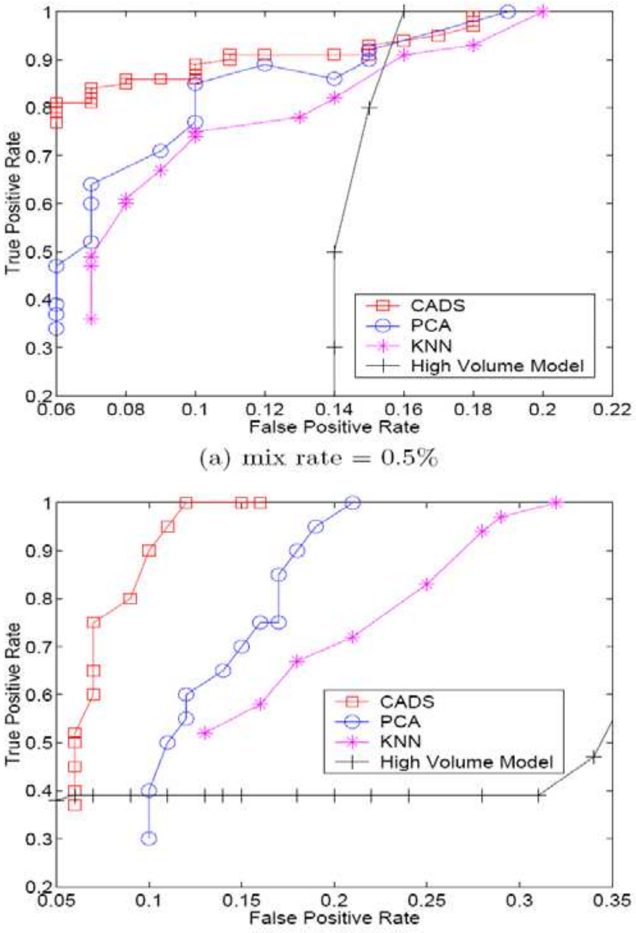
Figure 9: Rate of detection of the simulated user via the largest CADS deviation score as a function of the number of patients accessed.



(a) EHR

(b) Editor

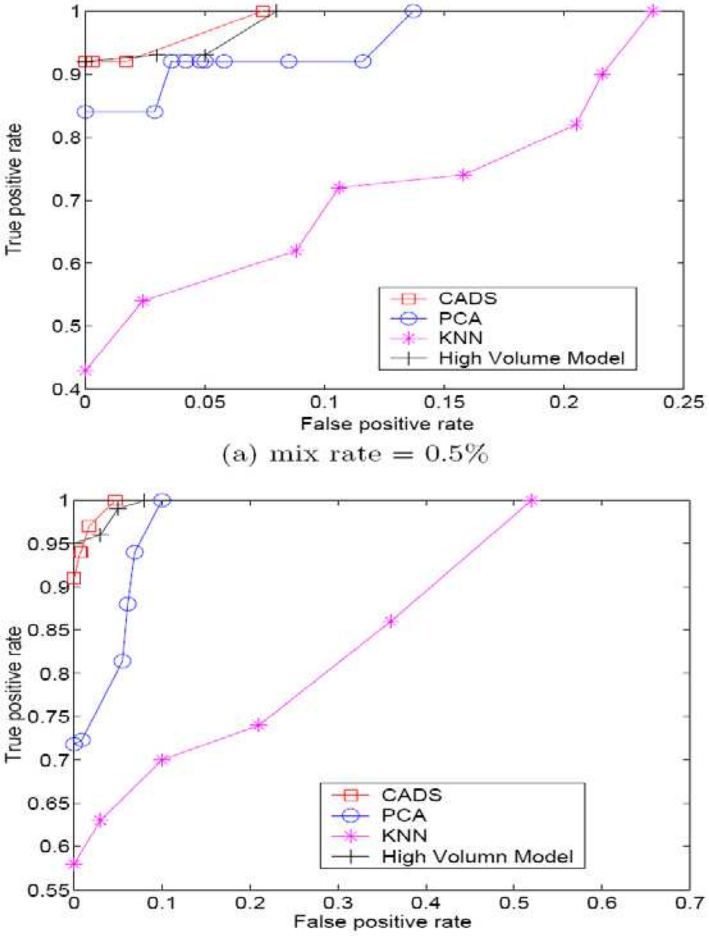
Figure 10: CADS performance with various mix rates of simulated and real users.



(a) mix rate = 0.5%

(b) mix rate = 5%

Figure 11: Comparison of different anomaly detection methods on the EHR dataset. The number of accessed subjects for simulated user is random.



(a) mix rate = 0.5%

(b) mix rate = 5%

Figure 12: Comparison of different anomaly detection methods on the Editor dataset. The number of accessed subjects for simulated user is random.

- Conclusions
  - A Community-based Anomaly Detection Model (CADS) to predict which users are anomalous
  - CADS calculates deviation of users based on their nearest neighbor's networks
  - We compared CADS with other three models: PCA, KNN and high volume users. The experimental findings suggest that
    - when the number of users and complexity of the social networks in the CIS are low, very simple models of anomaly detection, such as high volume user detection, may be sufficient.
    - As the complexity of the system grows tools that model complex behavior, tools such as CADS, are certainly necessary.

- Limitations
  - This work did not incorporate additional semantics that are often associated with users and subjects that could be useful in constructing more meaningful patterns
  - CADS model aims to detect anomalous insiders, but this is only one type of anomalous insiders; Models could be developed to search for anomalies at the level of individual accesses or sequences of events

- Leveraging social networks to detects of anomalous insider actions in collaborative environments

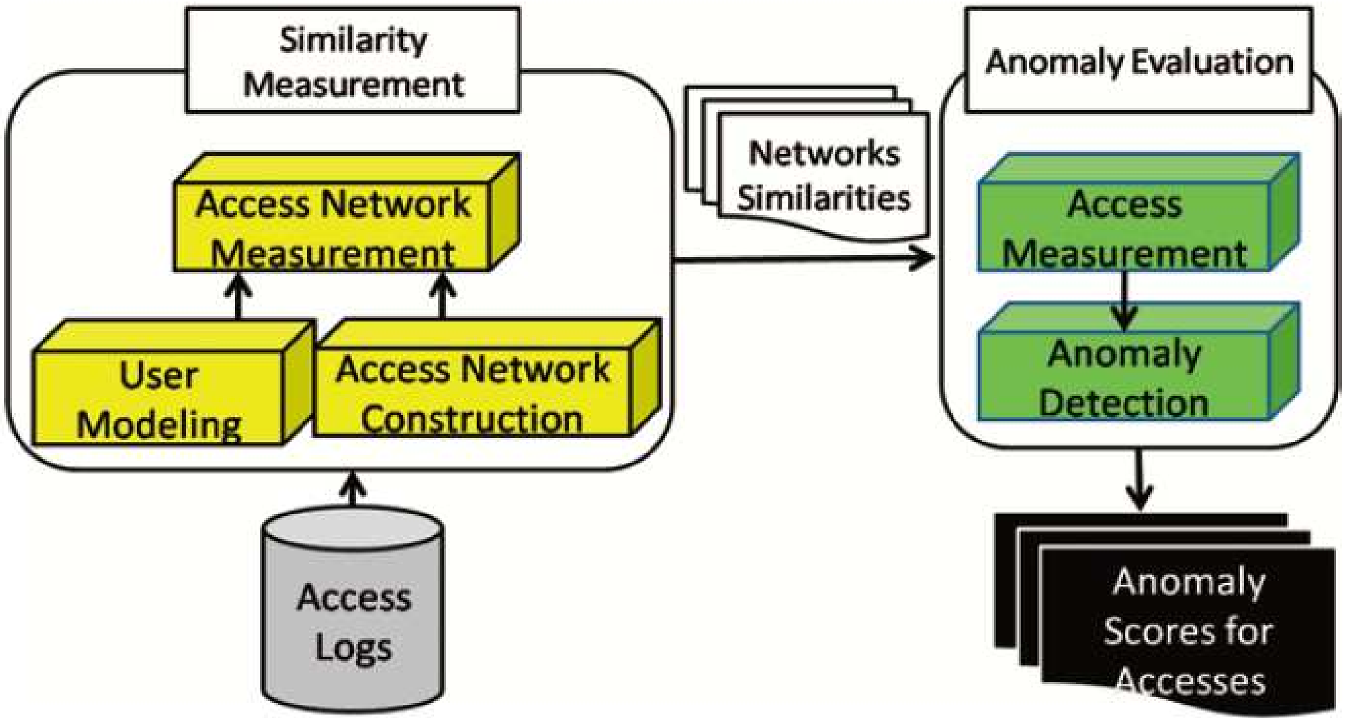
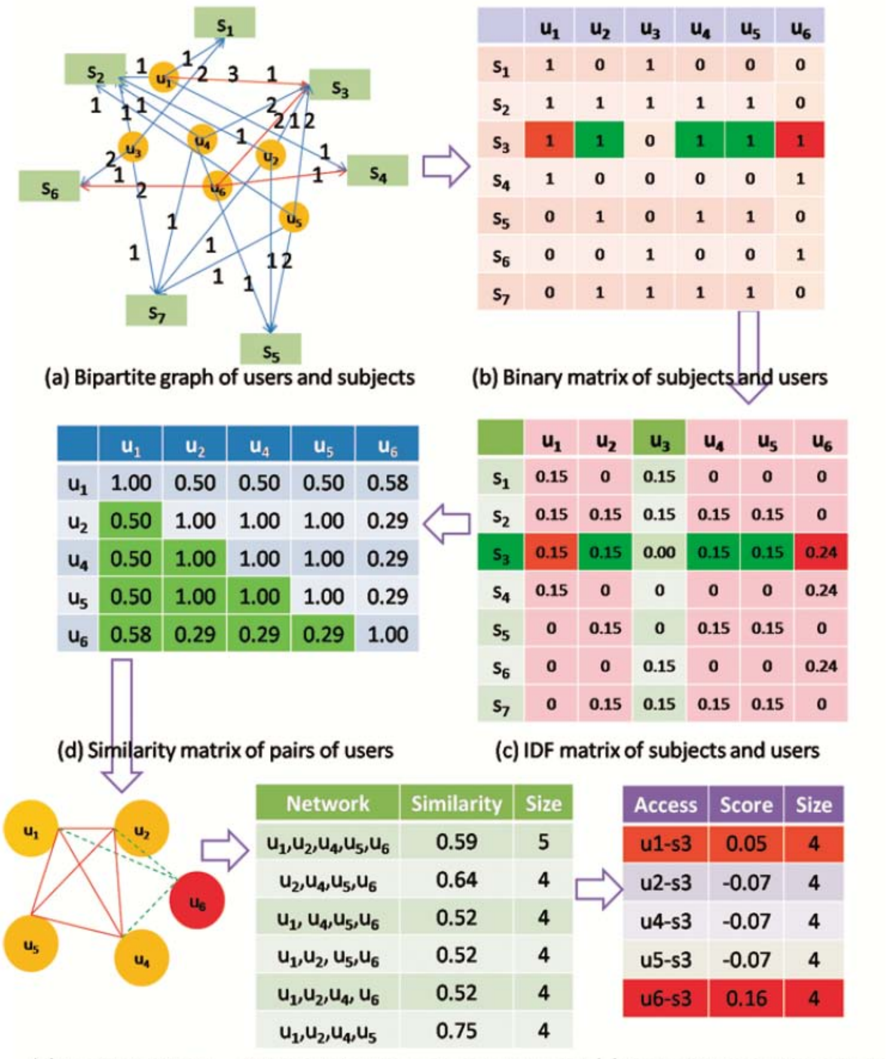


Figure 1: Overall framework of specialized network anomaly detection (SNAD).

- An example to illustrate the model



(a) Bipartite graph of users and subjects

	u1	u2	u3	u4	u5	u6
s1	1	0	1	0	0	0
s2	1	1	1	1	1	0
s3	1	1	0	1	1	1
s4	1	0	0	0	0	1
s5	0	1	0	1	1	0
s6	0	0	1	0	0	1
s7	0	1	1	1	1	0

(b) Binary matrix of subjects and users

	u1	u2	u3	u4	u5	u6
s1	0.15	0	0.15	0	0	0
s2	0.15	0.15	0.15	0.15	0.15	0
s3	0.15	0.15	0.00	0.15	0.15	0.24
s4	0.15	0	0	0	0	0.24
s5	0	0.15	0	0.15	0.15	0
s6	0	0	0.15	0	0	0.24
s7	0	0.15	0.15	0.15	0.15	0

(c) IDF matrix of subjects and users

	u1	u2	u4	u5	u6
u1	1.00	0.50	0.50	0.50	0.58
u2	0.50	1.00	1.00	1.00	0.29
u4	0.50	1.00	1.00	1.00	0.29
u5	0.50	1.00	1.00	1.00	0.29
u6	0.58	0.29	0.29	0.29	1.00

(d) Similarity matrix of pairs of users

(e) Access network

Network	Similarity	Size
u1,u2,u4,u5,u6	0.59	5
u2,u4,u5,u6	0.64	4
u1, u4,u5,u6	0.52	4
u1,u2, u5,u6	0.52	4
u1,u2,u4, u6	0.52	4
u1,u2,u4,u5	0.75	4

(f) Similarities of access network and

Access	Score	Size
u1-s3	0.05	4
u2-s3	-0.07	4
u4-s3	-0.07	4
u5-s3	-0.07	4
u6-s3	0.16	4

(g) Anomaly scores of accesses

$$S = \{s_1, \ldots, s_m\}$$

$$U = \{u_1, \ldots, u_n\}$$

We say U_{s_i} is the set of users that accessed s_i

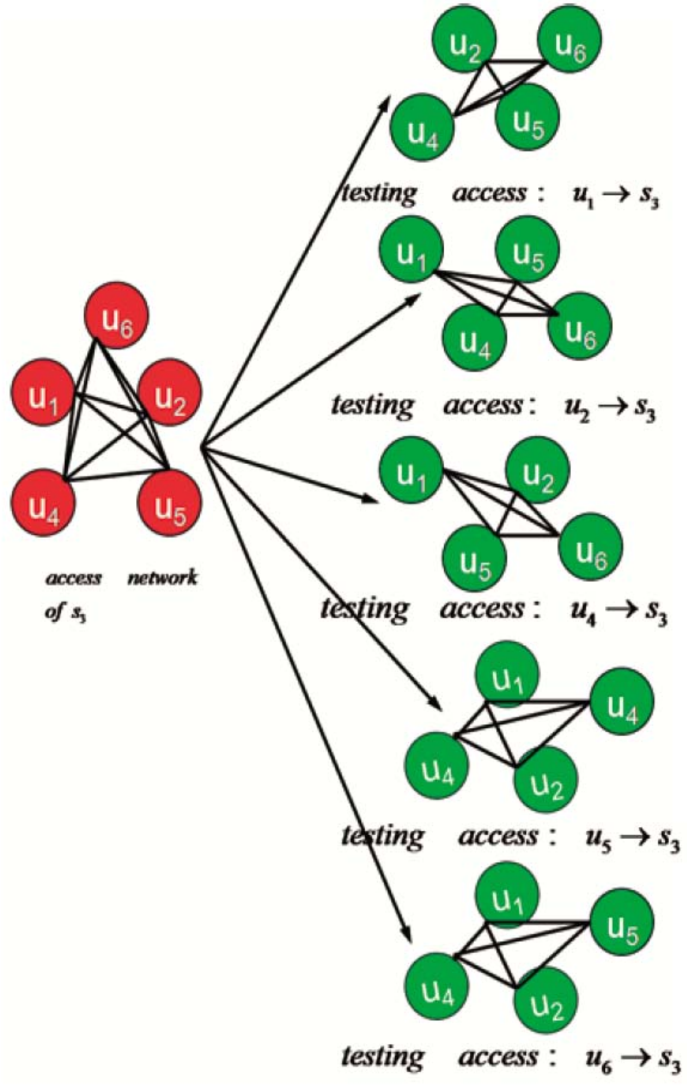
We say Net_{s_i} is a full-connected graph of U_{s_i}

$$IDF(u_i) = log\frac{|S|}{1+|\{s_j, \quad where \quad SU(j,i)>0\}|}$$

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

$$SIM(Net_{s_i}) = \frac{\sum_{i=1}^{|U_{s_i}|-1} \sum_{j=i+1}^{|U_{s_i}|} Sim(u_i, u_j)}{\frac{|U_{s_i}| \times (|U_{s_i}|-1)}{2}}$$

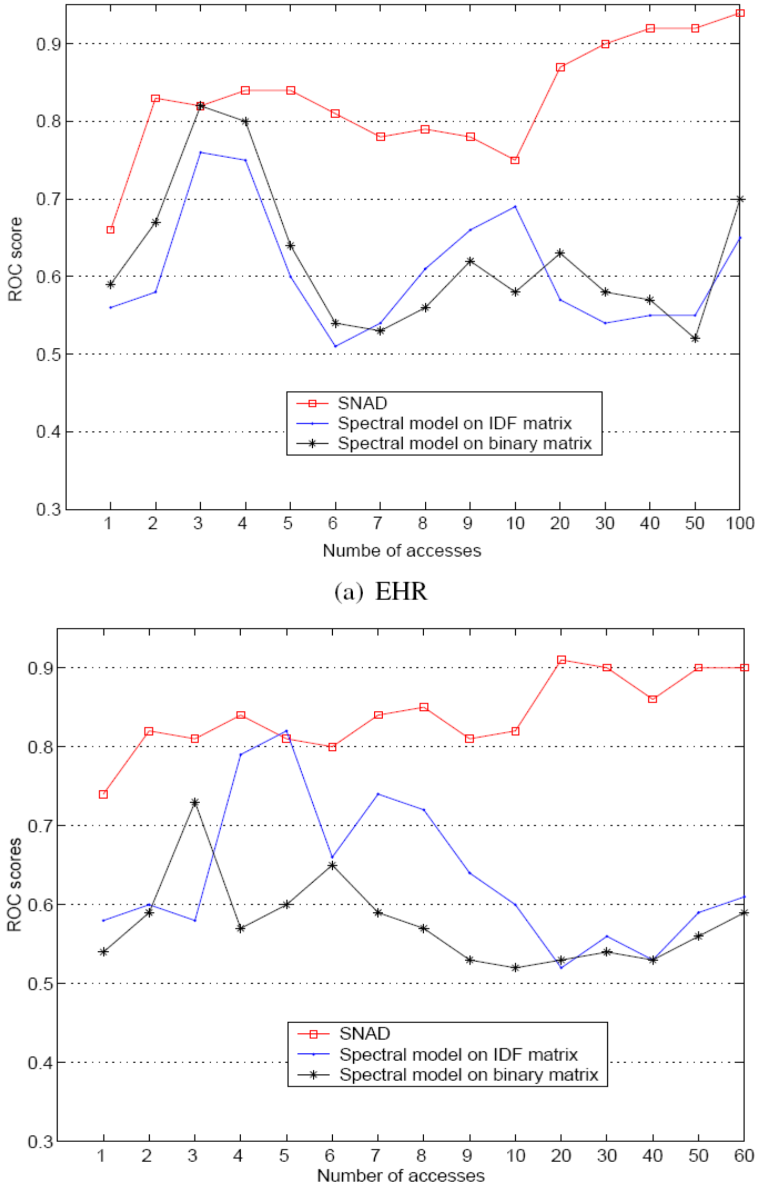
- Access Measurement


u_6
u_1
u_2
u_4
u_5
access network of s_3
u_2 u_6 u_4 u_5
testing access : $u_1 \rightarrow s_3$
u_1 u_5 u_4 u_6
testing access : $u_2 \rightarrow s_3$
u_1 u_2 u_5 u_6
testing access : $u_4 \rightarrow s_3$
u_1 u_4 u_4 u_2
testing access : $u_5 \rightarrow s_3$
u_1 u_5 u_4 u_2
testing access : $u_6 \rightarrow s_3$

$$Score(u_j \rightarrow s_3) = SIM(Net_{s_{3j}}) - SIM(Net_{s_3})$$

- Experiments
  - EHR data set
    - 30 weeks of 2006
    - 2,281 users, 13,148 patients and 44,250 accesses
  - Wiki edit data set
    - 2007-2008
    - 28,186 revisions, 240 articles , and 3,952 users

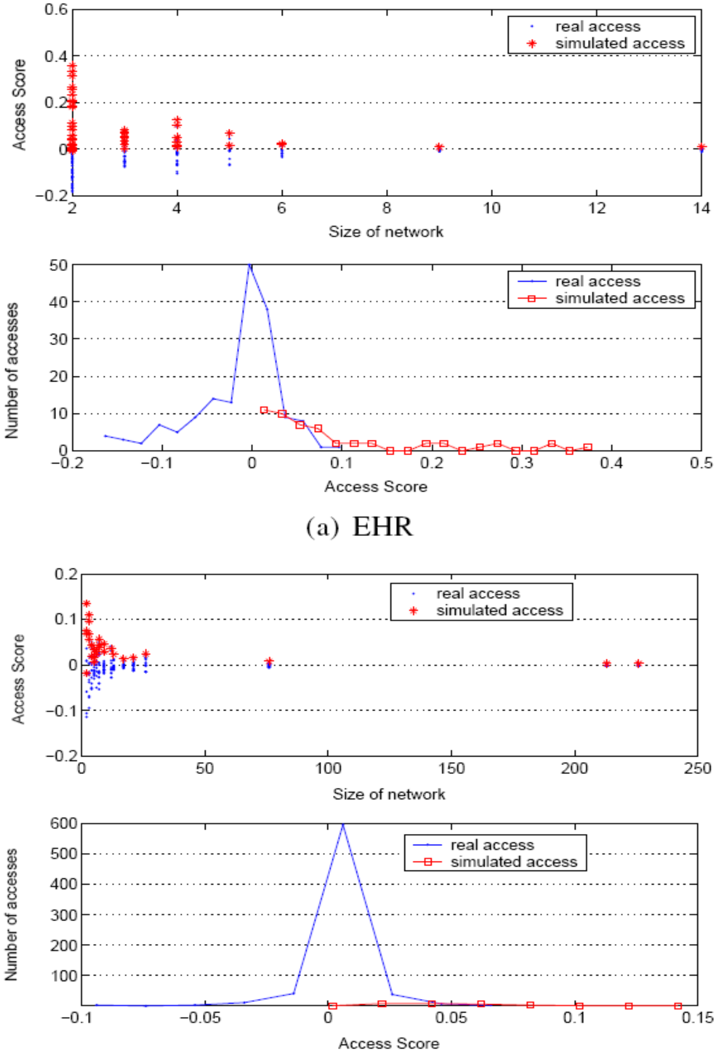
- Results and analysis
  - Random number of simulated accesses
  - Random number of anomalous insiders
  - Diversity



(a) EHR

(b) Wiki

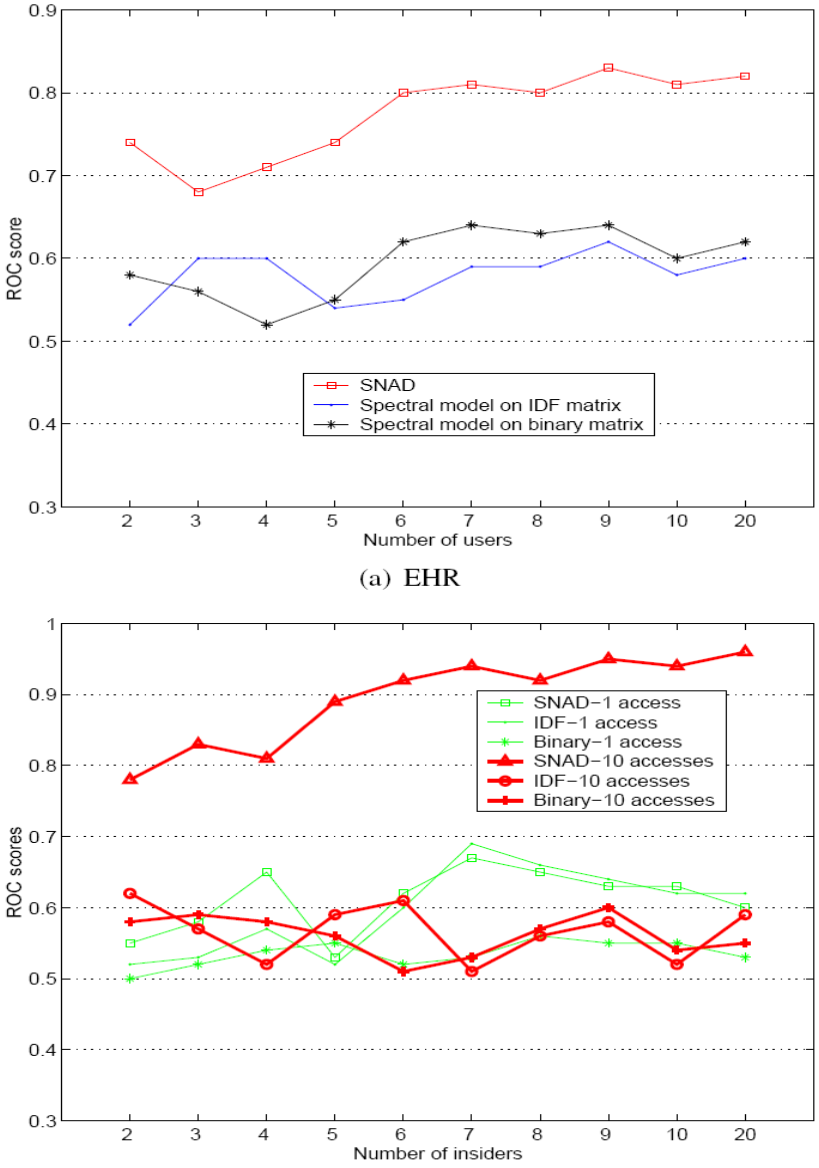
Figure 6: Average roc scores of three models on different number of simulated accesses



(a) EHR

(b) Wiki

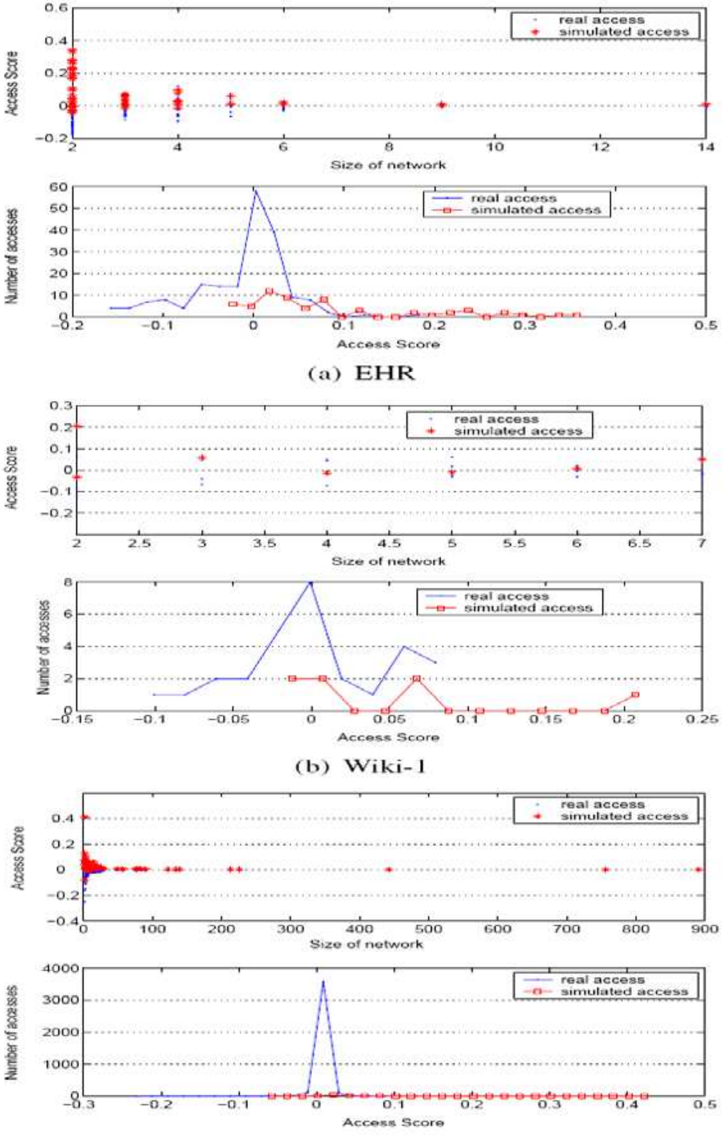
Figure 7: Distribution of real access scores and simulated access scores. The scores are evaluated by SNAD on a random week, and the number of simulated accesses is 50 in EHR data set and 30 in Wiki data set



(a) EHR

(b) Wiki

Figure 10: ROC scores of three models on different number of insiders, each insider has 3 accesses in EHR data set, and have 1 access and 10 accesses in Wiki data set

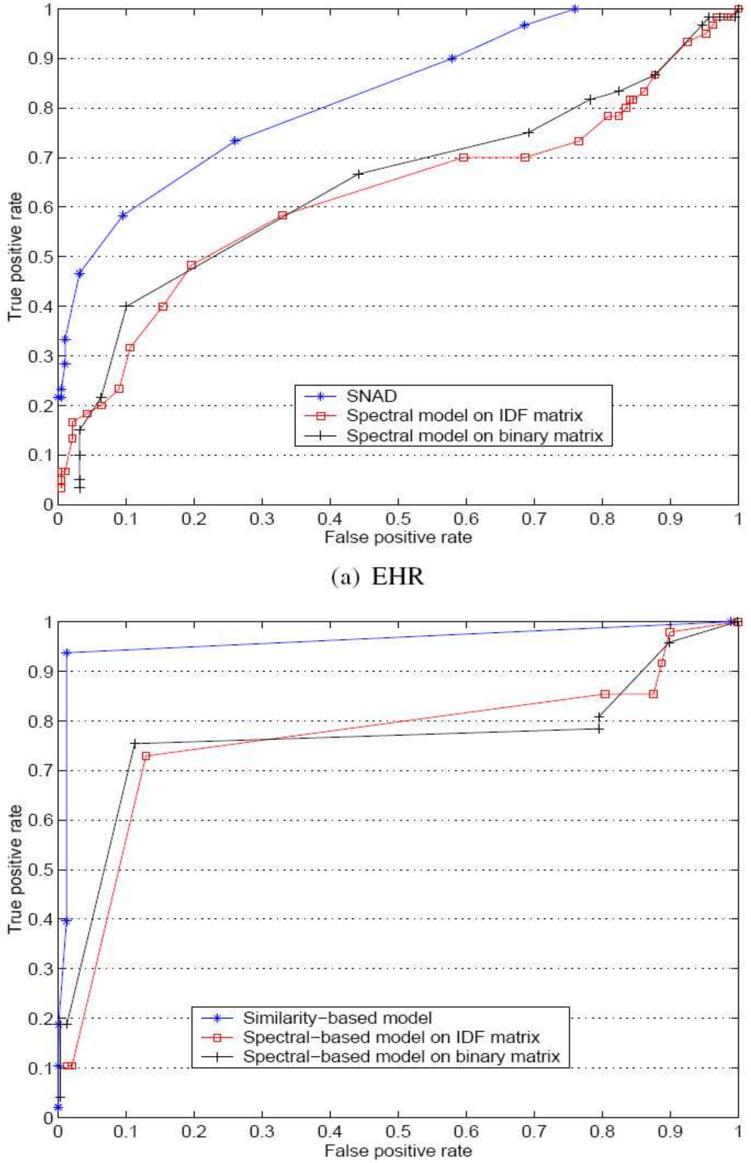


(a) EHR

(b) Wiki-1

(c) Wiki-10

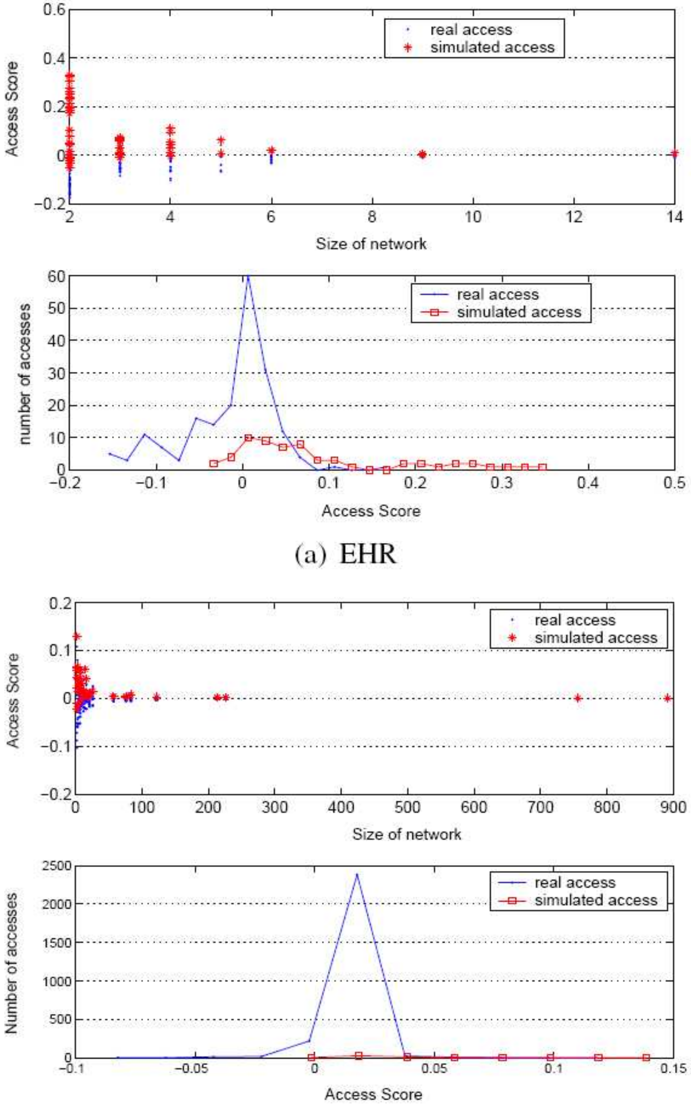
Figure 11: Distribution of real access scores and simulated access scores. The scores are evaluated by SNAD on a random week, the number of insiders is 20 in EHR data set and 7 in Wiki data set



(a) EHR

(b) Wiki

Figure 13: Comparisons of ROC curves for three models on EHR data set and Wiki data set.



(a) EHR

(b) Wiki

Figure 14: Distribution of real access scores and simulated access scores. The scores are evaluated by SNAD on a random week, the number of insiders and simulated accesses are both random

- Conclusions
  - We present a specialized network anomaly detection model (SNAD) to prevent anomalous insider accesses
  - In the experiments, we mixed simulated accesses with into systems of real accesses and evaluated the anomaly detection models on two types of access logs: 1) a real electronic health record system (EHR) and 2) a publicly-available set of Wikipedia revisions. Our results illustrate that SNAD exhibits the highest performance at detecting simulated internal accesses
  - Since SNAD is an unsupervised learning system, we believe it may be implemented in real time environments without training
  - However, and in particular, we intend to validate and improve our system with adjudication through real human experts.

Future works

- The goal of the current work was to determine how the basic information in the access logs could assist in anomaly detection
- In a CIS, an access log contains a wealth of information regarding how users interact with the data stored in the system
  - The access logs in an electronic health record (EHR) system records information about when a user accesses a patient’s record, from what computer the access was made, and which section of the record was accessed.
  - Moreover, the healthcare organization often retains meta information about the user, such as the job title and department in which the user works
  - EHR itself contains a great deal of knowledge about the patient whose information was accessed, such as demographics, diagnoses received, and procedures performed

- Next steps are
  - to specialize the models to work with the semantics of the healthcare domain
  - to use the ”role” or ”departmental affiliation” of the EHR users to construct more specific models about the users
  - To use the “diagnoses” or “treatments performed” for the patients to determine if clinically-related groups of patients are accessed in similar ways

References

- VARUN CHANDOLA, ARINDAM BANERJEE, and VIPIN KUMAR. Anomaly Detection. Journal, ACM Computing Survey. Volume 41 Issue 3, July 2009
- You Chen and Bradley Malin. Detection of Anomalous Insiders in Collaborative Environments via Relational Analysis of Access Logs. ACM CODASPY11, pp63-74

Thanks