Insider Thread Detection in Electronic Medical Record Systems

You Chen

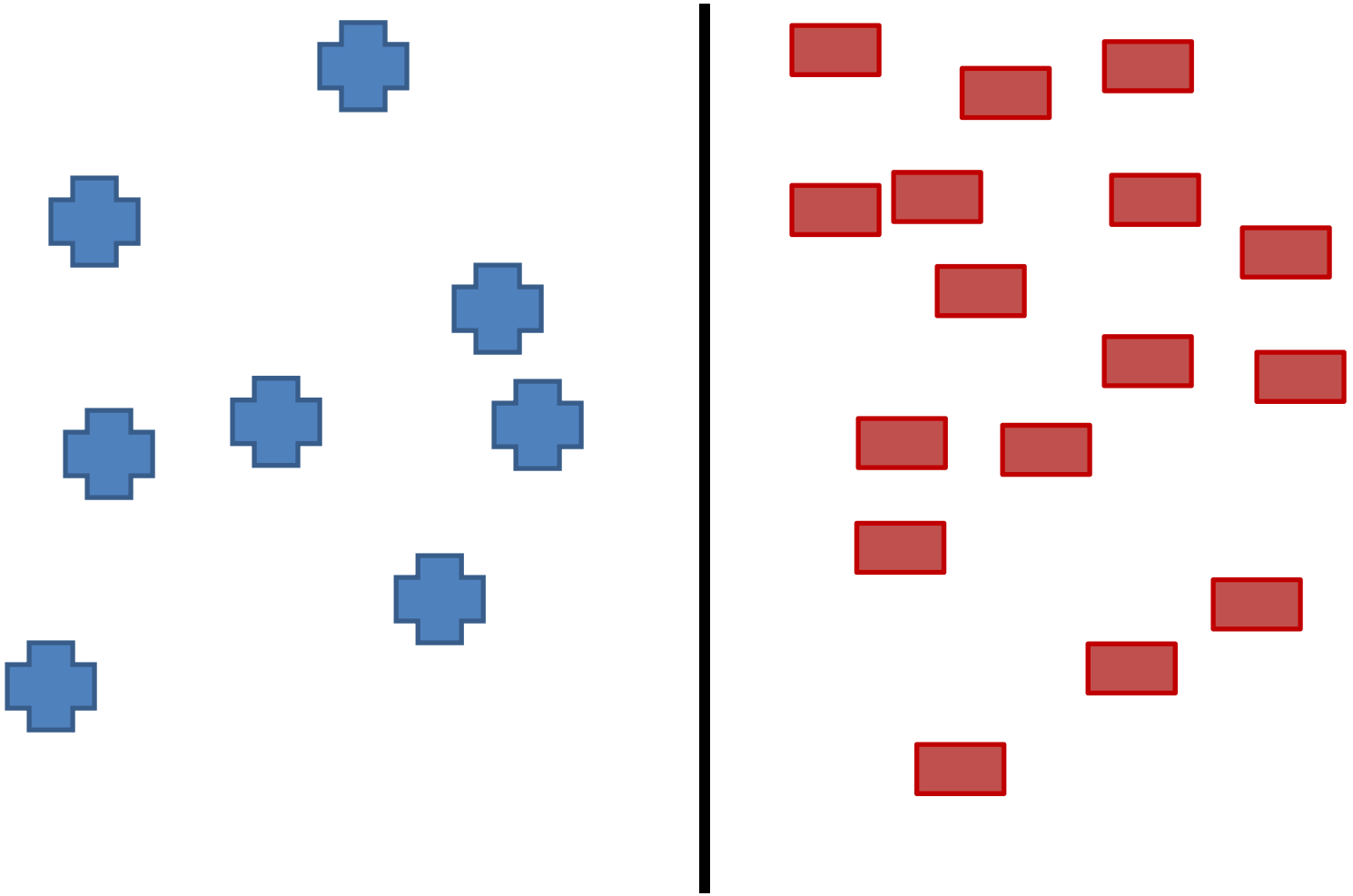
Feb, 04, 2015

You.chen@vanderbilt.edu

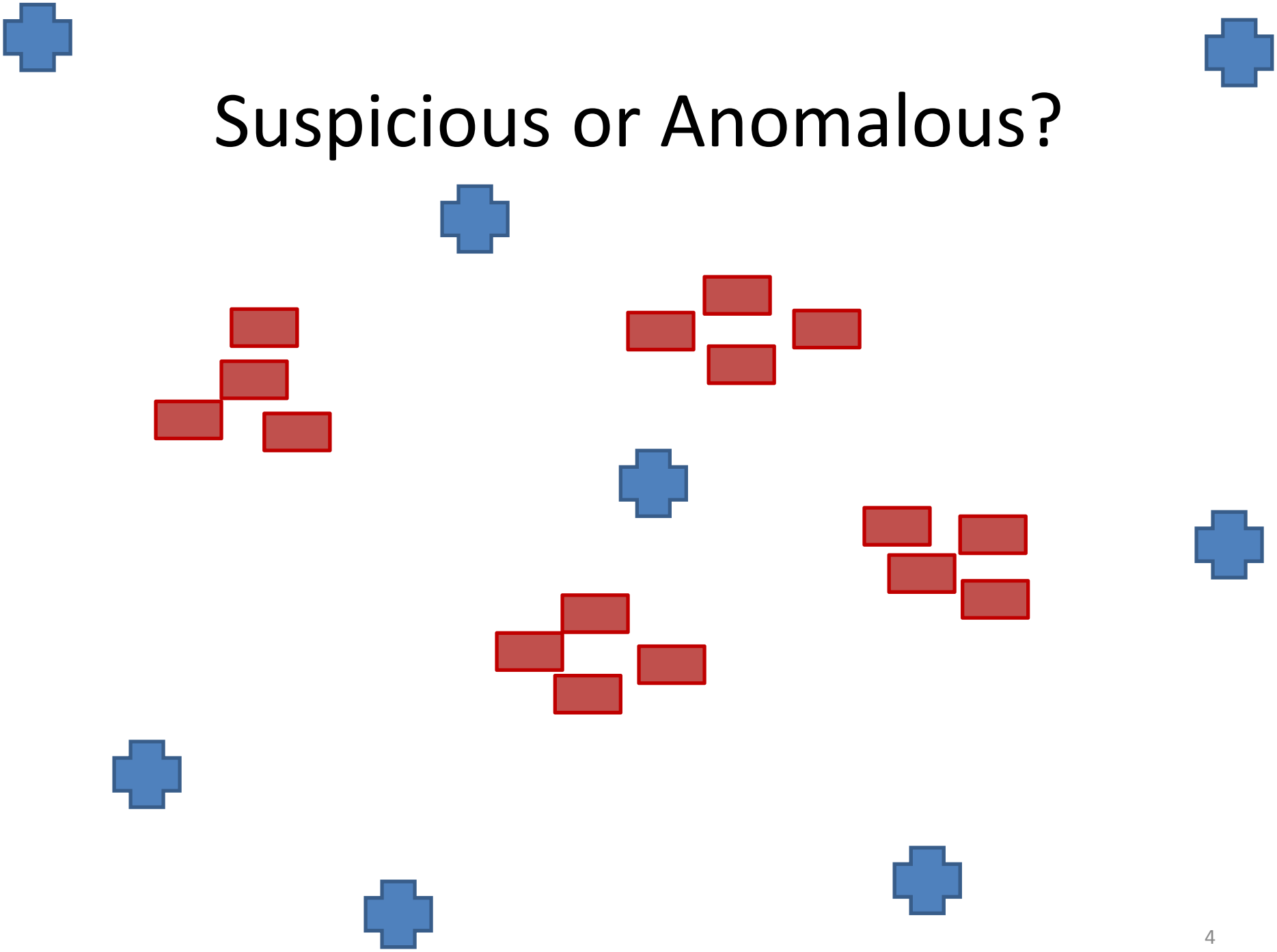
What Makes Sense?

- Dr. Smith's access of Peggy Johnson's medical record was strange
- Dr. Smith's access was 10 standard deviations away from normal behavior in his hospital
- Dr. Smith's access was strange because he is a neonatologist and he accessed the record of a 100 year-old woman who, for the past year, has only been treated by gerontologists

Suspicious or Anomalous?



Suspicious or Anomalous?



How Did We Get Here?

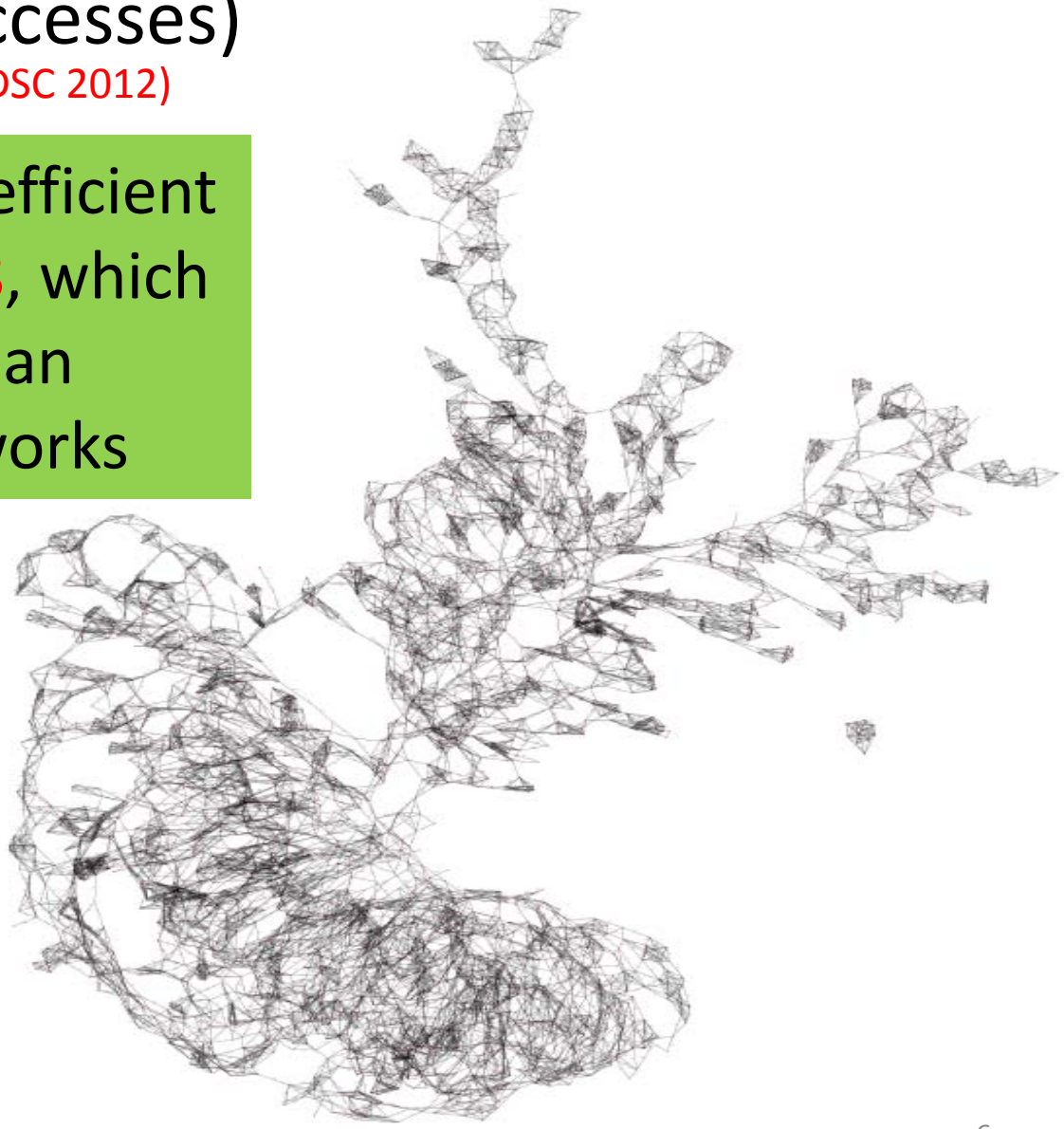
- Collaborative systems are about social phenomena
- People *should* form communities
- We should be able to measure deviation from community structure

6-Nearest Neighbor Network-Vanderbilt Medical Center (1 day of accesses)

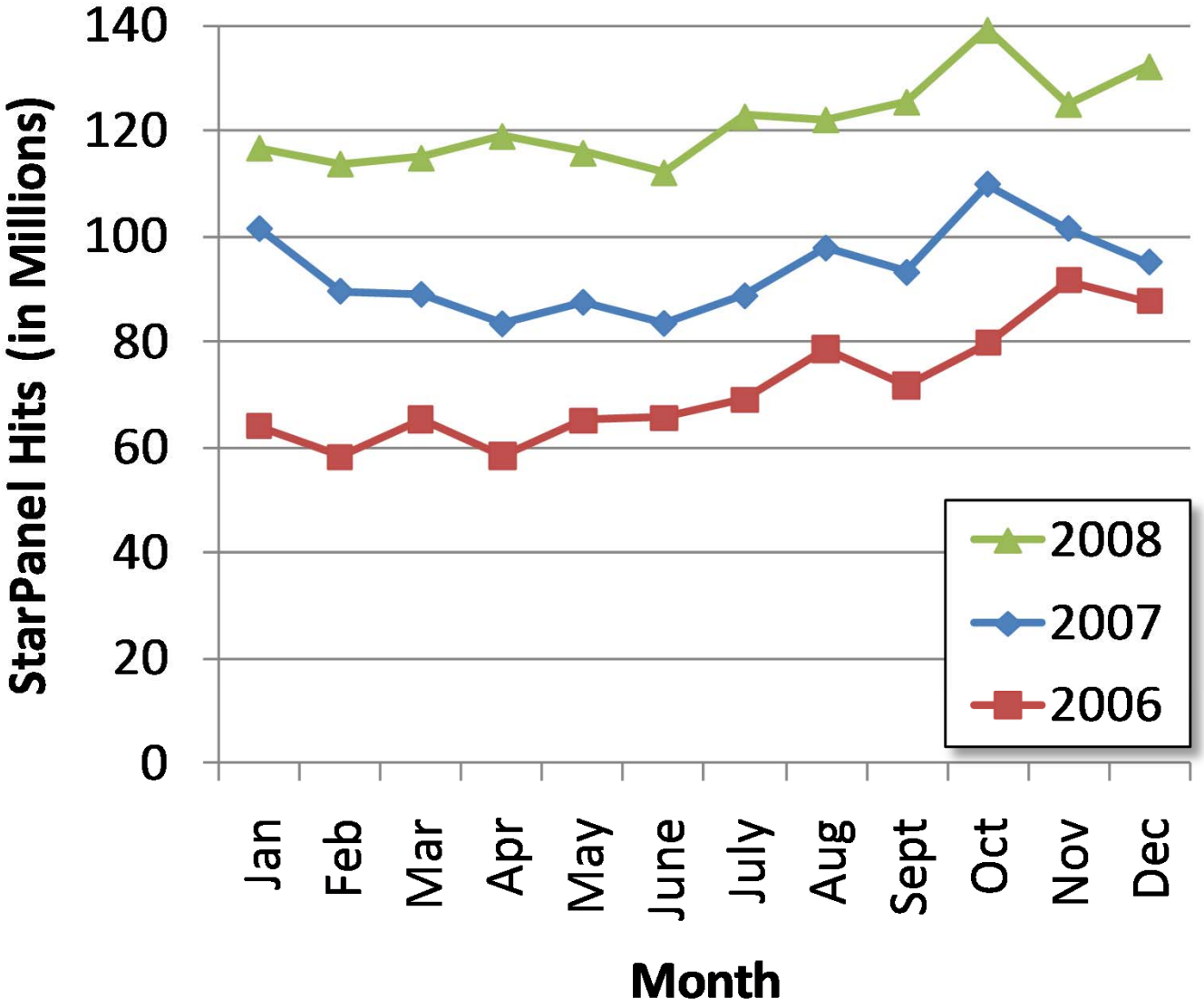
(Chen, Nyemba, & Malin – IEEE TDSC 2012)

The average cluster coefficient for this network is 0.48, which is significantly larger than 0.001 for random networks

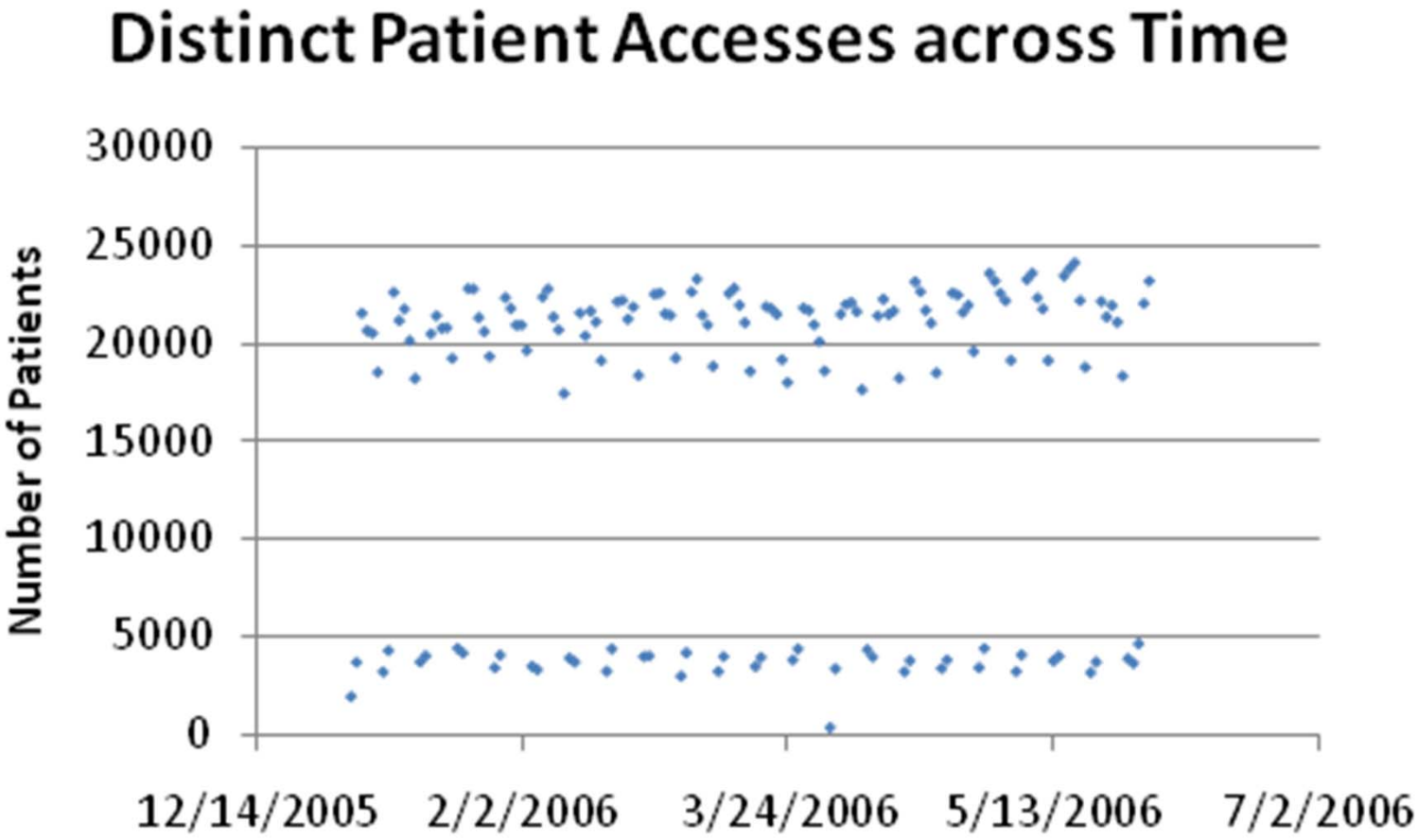
Users exhibit collaborative behavior in the Vanderbilt StarPanel System



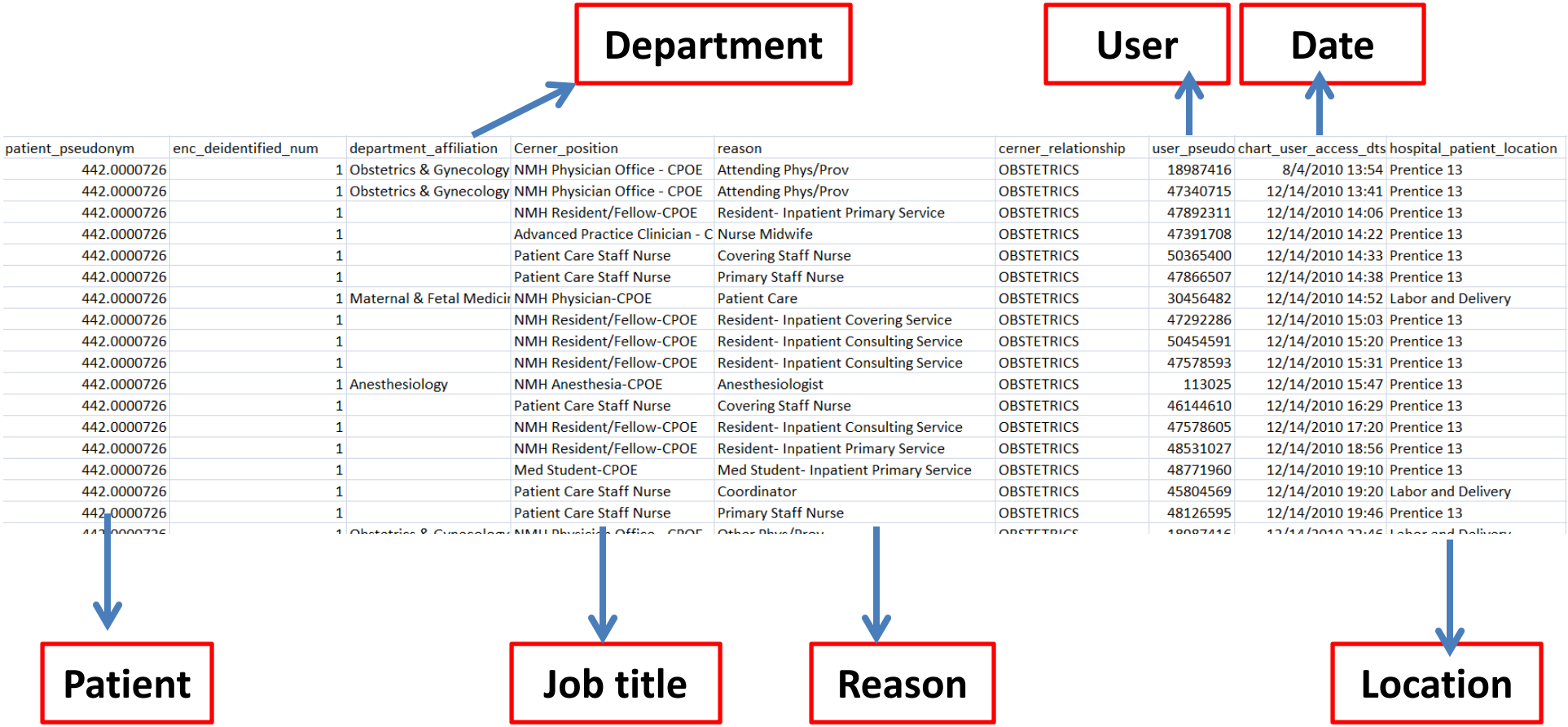
What type of data we have?
StarPanel system Growth in Use



Periodicity!-Week Day VS. Week End



Examples of Accesses



patient_pseudonym	enc_deidentified_num	department_affiliation	Cerner_position	reason	cerner_relationship	user_pseudo	chart_user_access_dts	hospital_patient_location
442.0000726	1	Obstetrics & Gynecology	NMH Physician Office - CPOE	Attending Phys/Prov	OBSTETRICS	18987416	8/4/2010 13:54	Prentice 13
442.0000726	1	Obstetrics & Gynecology	NMH Physician Office - CPOE	Attending Phys/Prov	OBSTETRICS	47340715	12/14/2010 13:41	Prentice 13
442.0000726	1		NMH Resident/Fellow-CPOE	Resident- Inpatient Primary Service	OBSTETRICS	47892311	12/14/2010 14:06	Prentice 13
442.0000726	1		Advanced Practice Clinician - C	Nurse Midwife	OBSTETRICS	47391708	12/14/2010 14:22	Prentice 13
442.0000726	1		Patient Care Staff Nurse	Covering Staff Nurse	OBSTETRICS	50365400	12/14/2010 14:33	Prentice 13
442.0000726	1		Patient Care Staff Nurse	Primary Staff Nurse	OBSTETRICS	47866507	12/14/2010 14:38	Prentice 13
442.0000726	1	Maternal & Fetal Medici	NMH Physician-CPOE	Patient Care	OBSTETRICS	30456482	12/14/2010 14:52	Labor and Delivery
442.0000726	1		NMH Resident/Fellow-CPOE	Resident- Inpatient Covering Service	OBSTETRICS	47292286	12/14/2010 15:03	Prentice 13
442.0000726	1		NMH Resident/Fellow-CPOE	Resident- Inpatient Consulting Service	OBSTETRICS	50454591	12/14/2010 15:20	Prentice 13
442.0000726	1		NMH Resident/Fellow-CPOE	Resident- Inpatient Consulting Service	OBSTETRICS	47578593	12/14/2010 15:31	Prentice 13
442.0000726	1	Anesthesiology	NMH Anesthesia-CPOE	Anesthesiologist	OBSTETRICS	113025	12/14/2010 15:47	Prentice 13
442.0000726	1		Patient Care Staff Nurse	Covering Staff Nurse	OBSTETRICS	46144610	12/14/2010 16:29	Prentice 13
442.0000726	1		NMH Resident/Fellow-CPOE	Resident- Inpatient Consulting Service	OBSTETRICS	47578605	12/14/2010 17:20	Prentice 13
442.0000726	1		NMH Resident/Fellow-CPOE	Resident- Inpatient Primary Service	OBSTETRICS	48531027	12/14/2010 18:56	Prentice 13
442.0000726	1		Med Student-CPOE	Med Student- Inpatient Primary Service	OBSTETRICS	48771960	12/14/2010 19:10	Prentice 13
442.0000726	1		Patient Care Staff Nurse	Coordinator	OBSTETRICS	45804569	12/14/2010 19:20	Labor and Delivery
442.0000726	1		Patient Care Staff Nurse	Primary Staff Nurse	OBSTETRICS	48126595	12/14/2010 19:46	Prentice 13

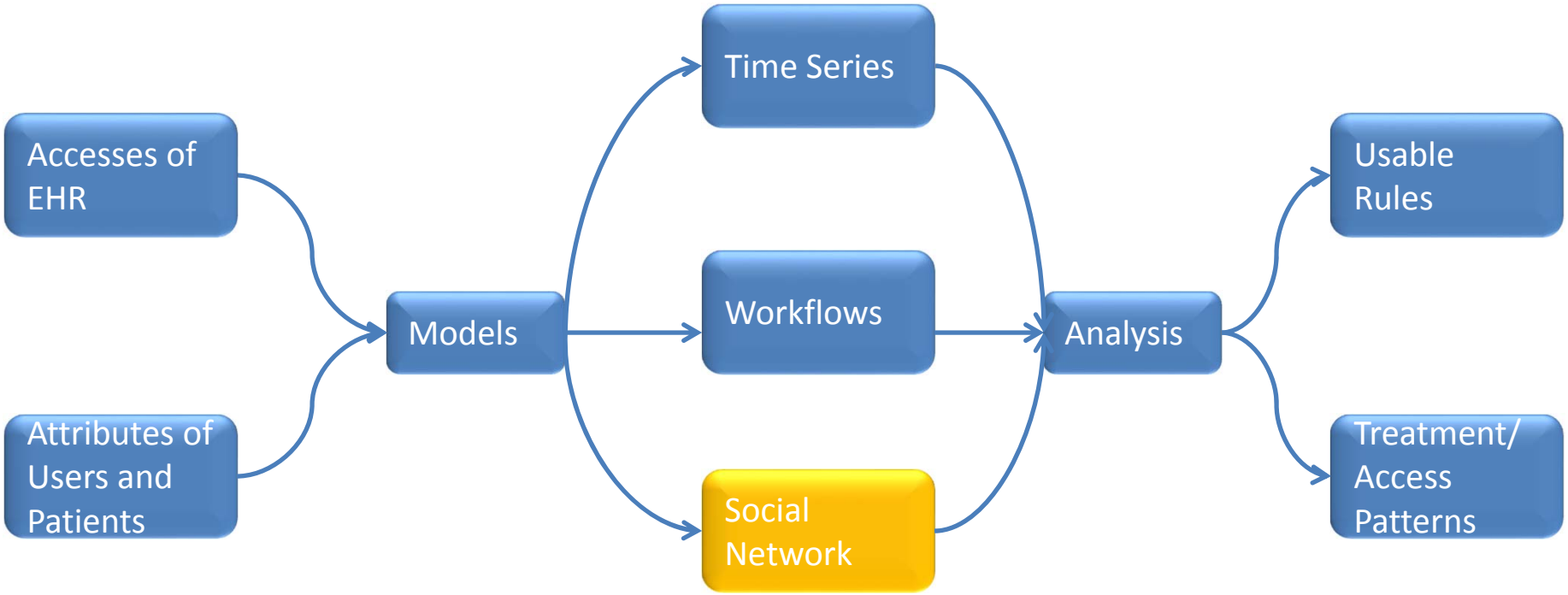
Examples of Patient Diagnosis Codes

Diagnosis codes

patient_study_id	enc_deiden	DX_codes
442.0000726	1	2165 , 65921, 65951, 66401, 66481, V270
442.0001714	1	V053 , V3000
442.0002396	1	4019 , 4111 , 41401, 4142 , 4739 , 49390
442.0002775	2	1122 , 20300, 25000, 27651, 40390, 5845 , 5859 , 591 , 5933 , V1005, V1046
442.0002775	1	1534 , 185 , 1962 , 1974 , 20300, 25000, 2809 , 40390, 56089, 5849 , 5859 , 59080, 591 , 78791, 7907
442.0003301	1	76408, 76529, V053 , V3100
442.0004873	1	V270 , V8535, 27800, 64911, 64971, 65841, 66401
442.0005024	1	4019 , 72252, V1582
442.0005968	1	5990 , 2724 , 311 , 4019 , 44022
442.0006352	1	65971, V270
442.0007008	1	25000, 6144 , 99859, V1042
442.0007371	1	V707 , 2859 , 33394, 4019 , 71690, 74190, V420
442.0007707	1	30000, 49121, 51889, 60000, 7850
442.0007707	2	78052, V1083, 30001, 496 , 0549 , 1120 , 2768 , 30000
442.0008016	1	V053 , V3001
442.0008405	1	2449 , 25080, 4019 , 41400, 42731, 4280 , 60000, V4581, V5861
442.0009617	1	V053 , V3000

Patient

Various Ways of Access Logs Auditing



Automatic Detection of Insider Threats through Social Network Analysis

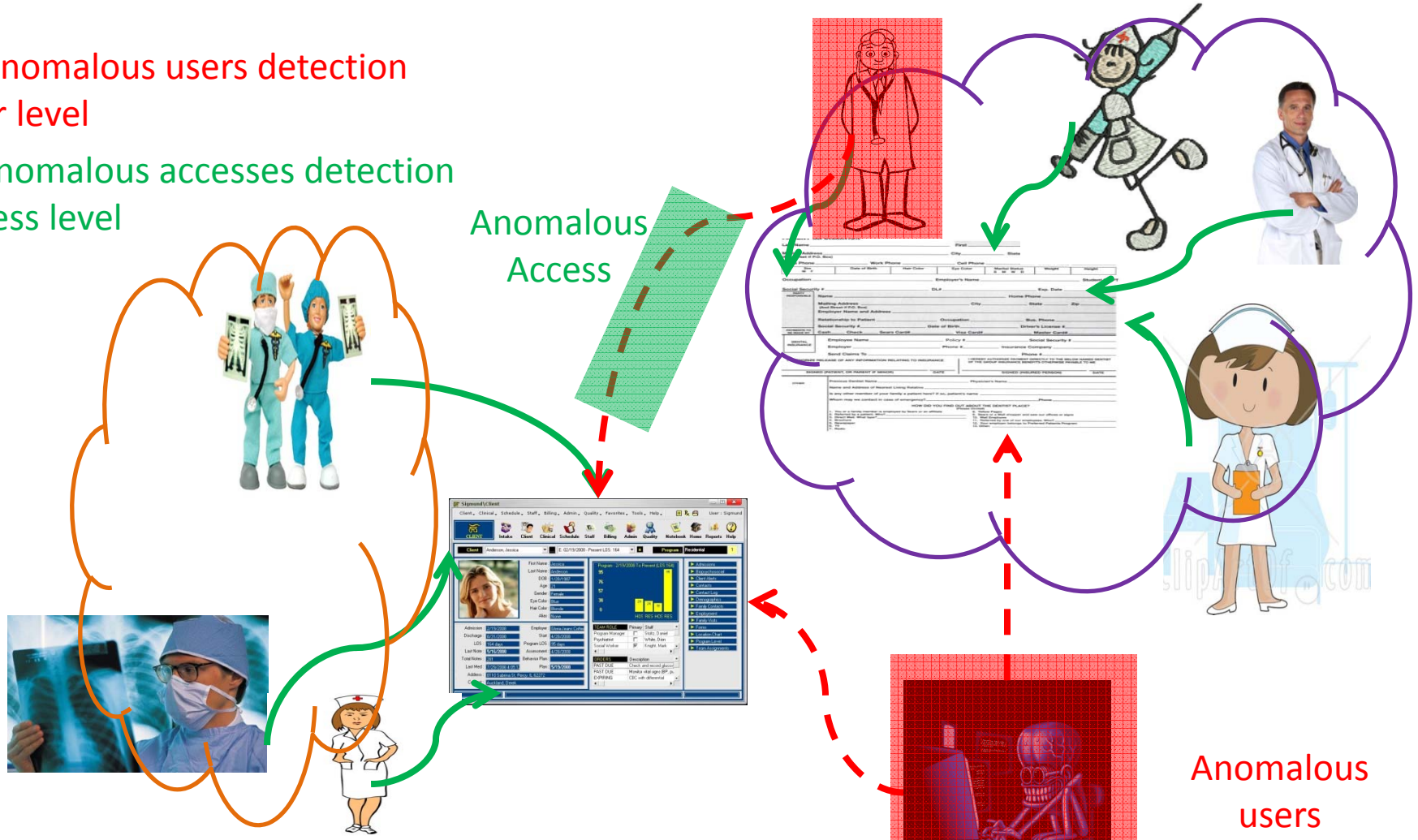
- User Level
  - Anomalous users detection
- Access Level
  - Anomalous insider actions detection
  - Specific actions of anomalous users

Two Typical Attacks

Intruders have complete knowledge of the system and its policies

(1) Anomalous users detection –user level

(2) Anomalous accesses detection –access level



Where are We Going?

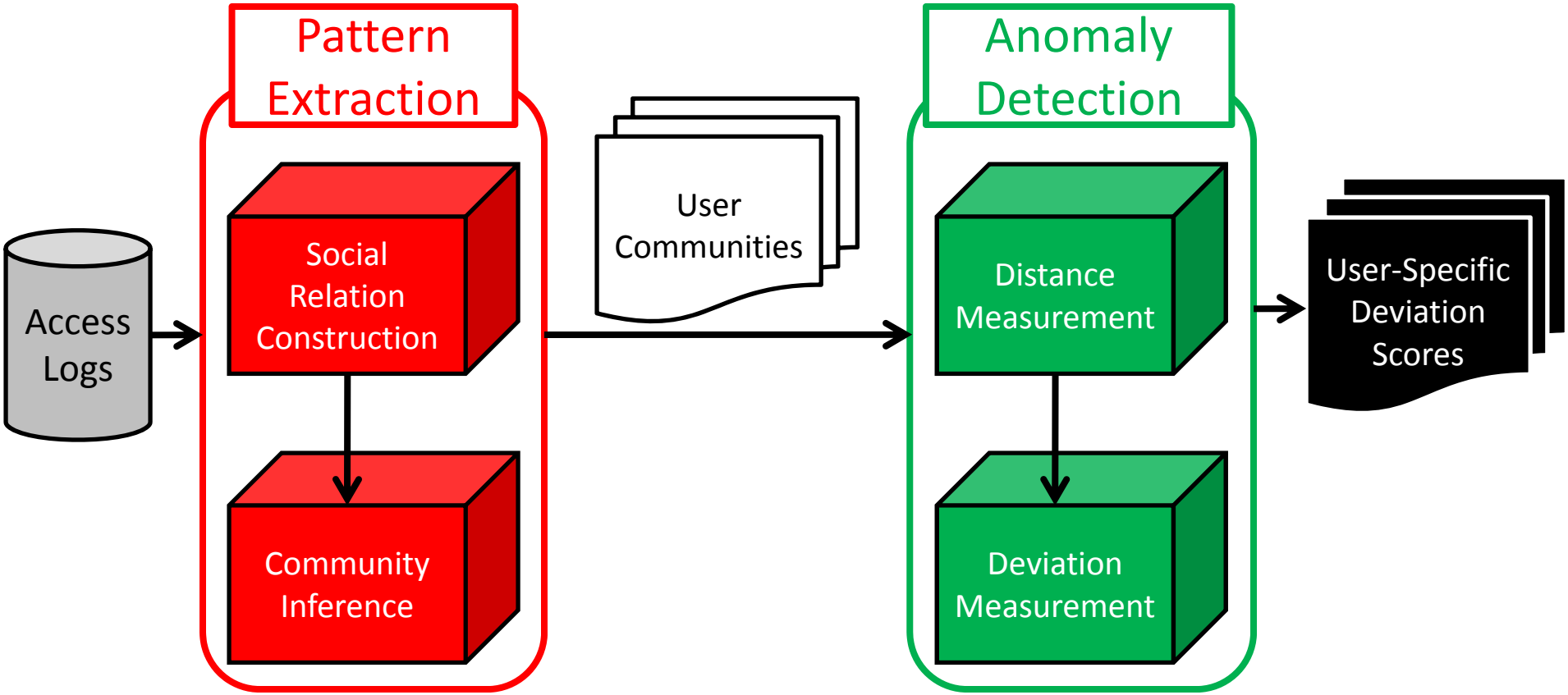
User Level Anomaly Detection

Community Anomaly Detection System (CADS) and its extension MetaCADS

Chen et al. IEEE TDSC: You Chen, Steve Nyemba and Bradley Malin. Detecting Anomalous Insiders in Collaborative Information Systems. IEEE Transaction on Dependable and Secure Computing. Vol.9.No 3, p332-344.

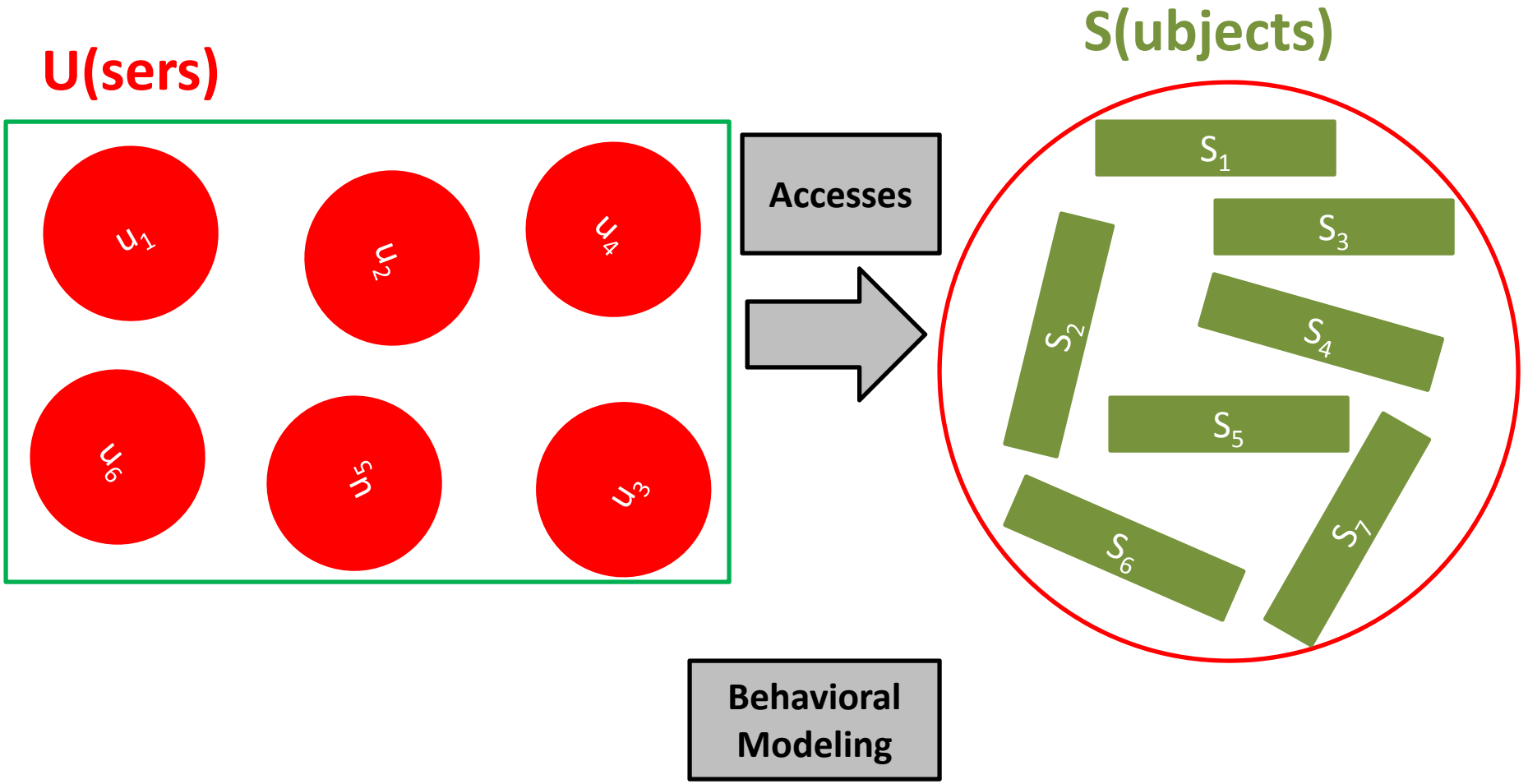
Chen & Malin – ACM CODASPY 2011: You Chen and Bradley Malin. Detection of Anomalous Insiders in Collaborative Environments via Relational Analysis of Access Logs. Proceedings of ACM Conference on Data and Application Security and Privacy. 2011, p63-74

Community-Based Anomaly Detection (CADS)

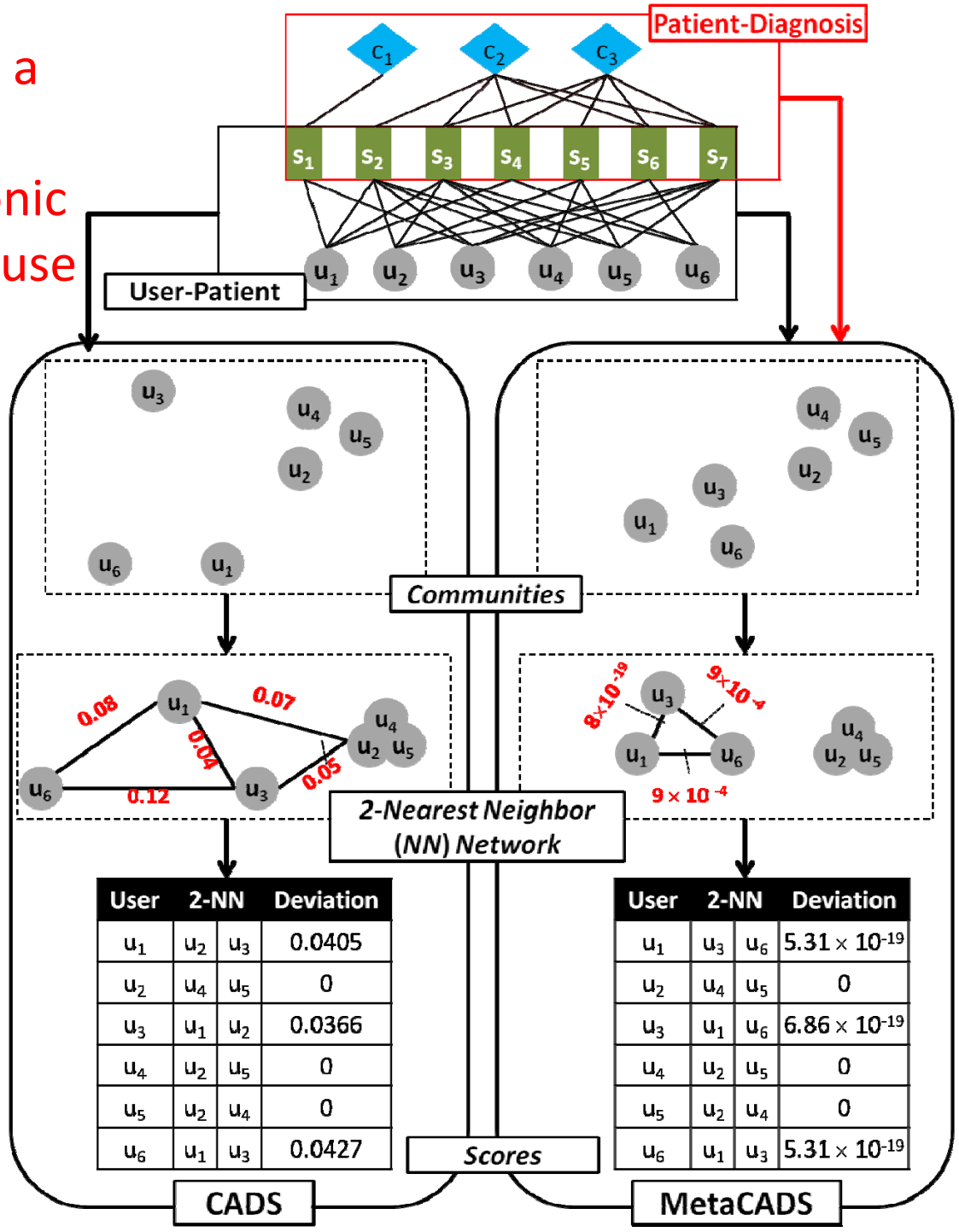


(Chen & Malin – ACM CODASPY 2011)

Two general objects of health information system

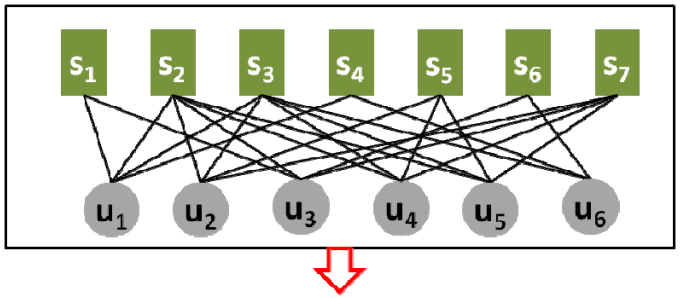


Social Networks are a Novel Approach to Discovery of Electronic Medical Record Misuse

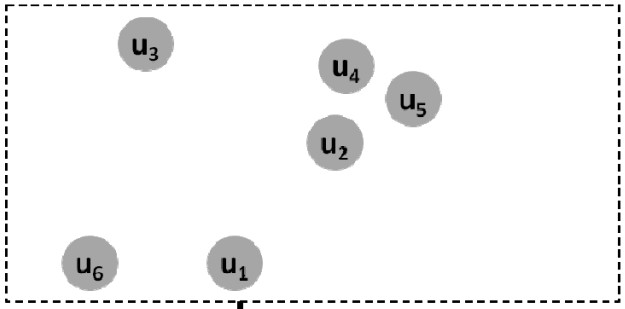


User	2-NN		Deviation
u_1	u_2	u_3	0.0405
u_2	u_4	u_5	0
u_3	u_1	u_2	0.0366
u_4	u_2	u_5	0
u_5	u_2	u_4	0
u_6	u_1	u_3	0.0427

User	2-NN		Deviation
u_1	u_3	u_6	5.31×10^{-19}
u_2	u_4	u_5	0
u_3	u_1	u_6	6.86×10^{-19}
u_4	u_2	u_5	0
u_5	u_2	u_4	0
u_6	u_1	u_3	5.31×10^{-19}

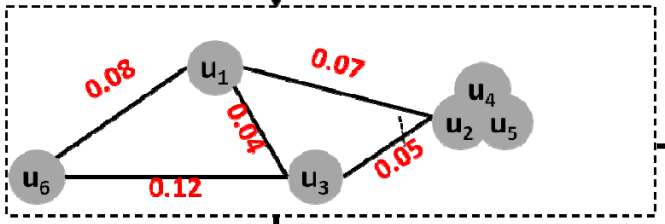


Bipartite Graph->Access Network of Users



Communities via Singular Value Decomposition

Distance via Weighted Euclidean Distance



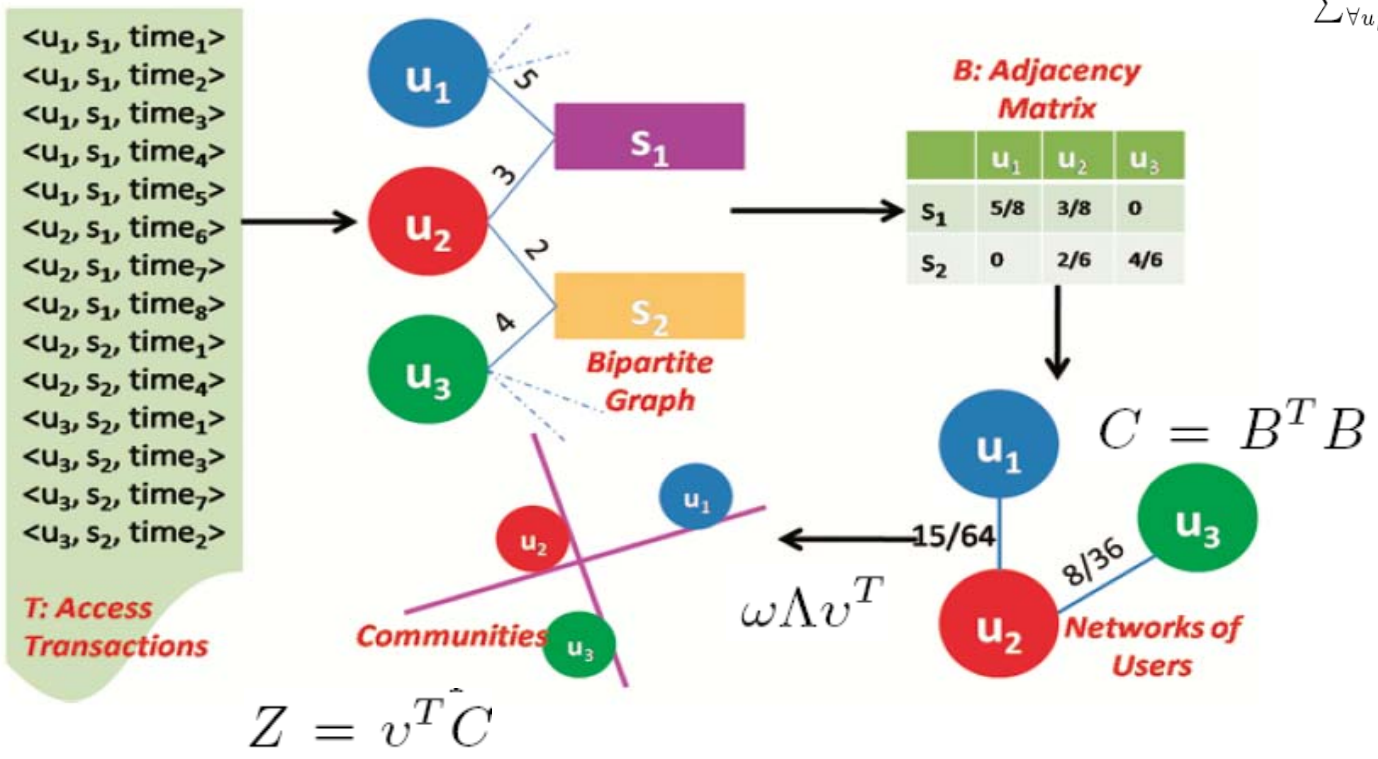
Nearest Neighbor Network

User	2-NN		Deviation
u_1	u_2	u_3	0.0405
u_2	u_4	u_5	0
u_3	u_1	u_2	0.0366
u_4	u_2	u_5	0
u_5	u_2	u_4	0
u_6	u_1	u_3	0.0427

Deviation Scores Calculation

Community Pattern Extraction

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



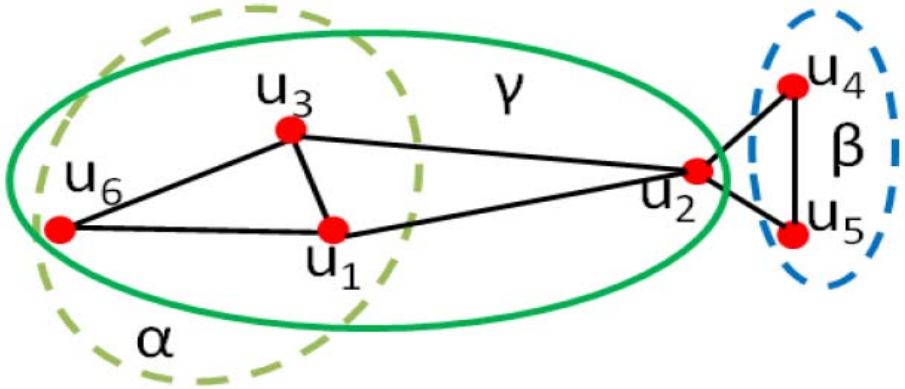
Distance measurement of pairs of users

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

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

How Do We Set “k”-NN?

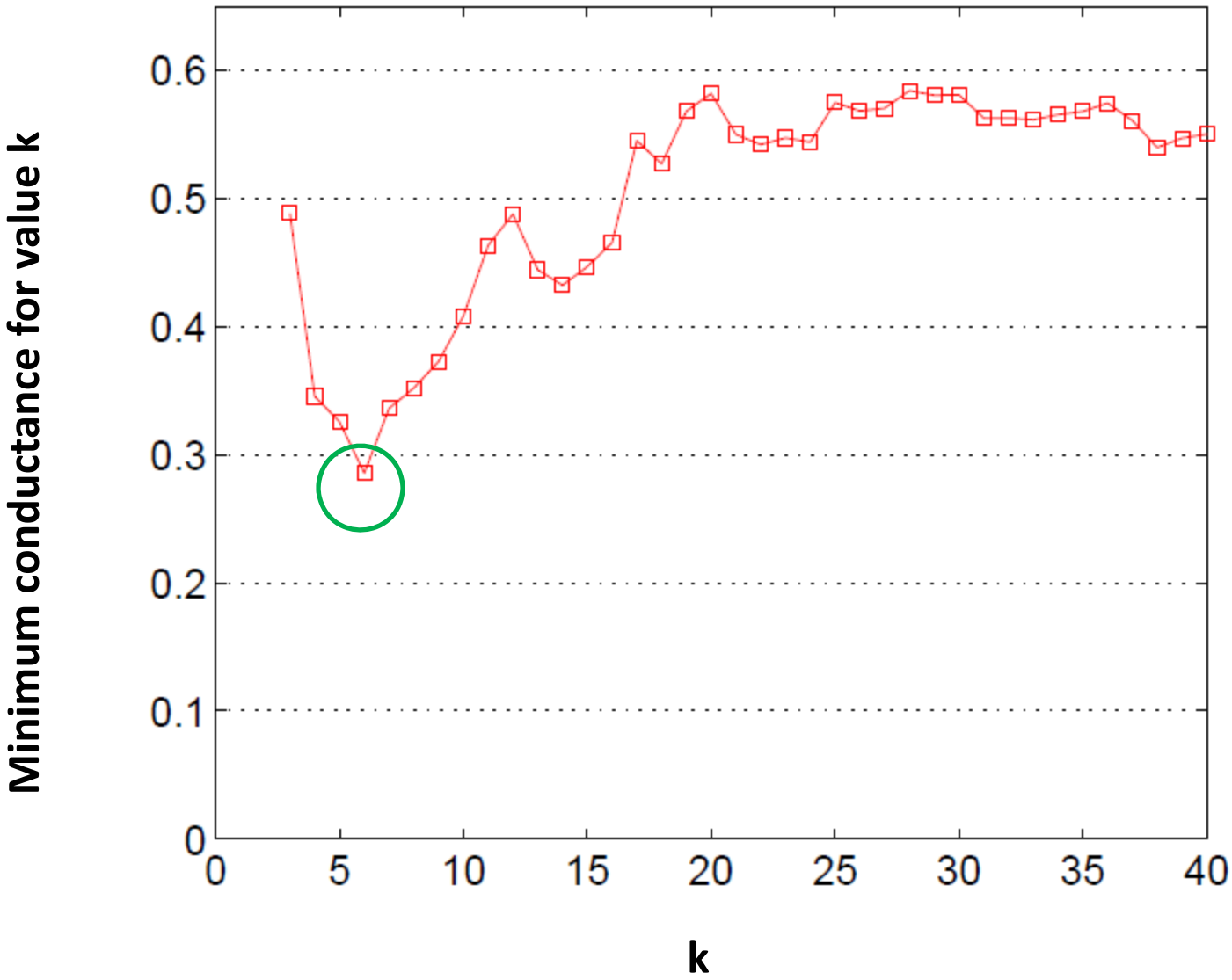
- Conductance- a measure of community quality



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

$$\psi(\alpha) < \psi(\beta) = \psi(\gamma)$$

Minimum conductance at k=6



6-Nearest Neighbor Network-Vanderbilt Medical Center (1 day of accesses)

Measuring Deviation from k-NN

- Every user is assigned a radius *r*:
  - the distance to his k[th] nearest neighbor
- Smaller the radius → higher density in user's network

$$Dev(u_i) = \sqrt{\frac{\sum_{u_j \in knni} (r_j - \bar{r})^2}{k}}$$

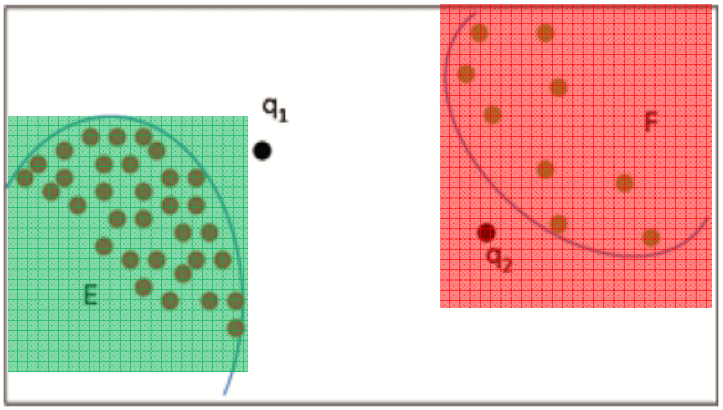
$$\bar{r} = \frac{\sum_{u_j \in knni} r_j}{k}$$

Radius for points in the green area are 1, and for q_1 is 3

$$\bar{r} = \frac{2+2+2+2+3}{5} = 2.2$$

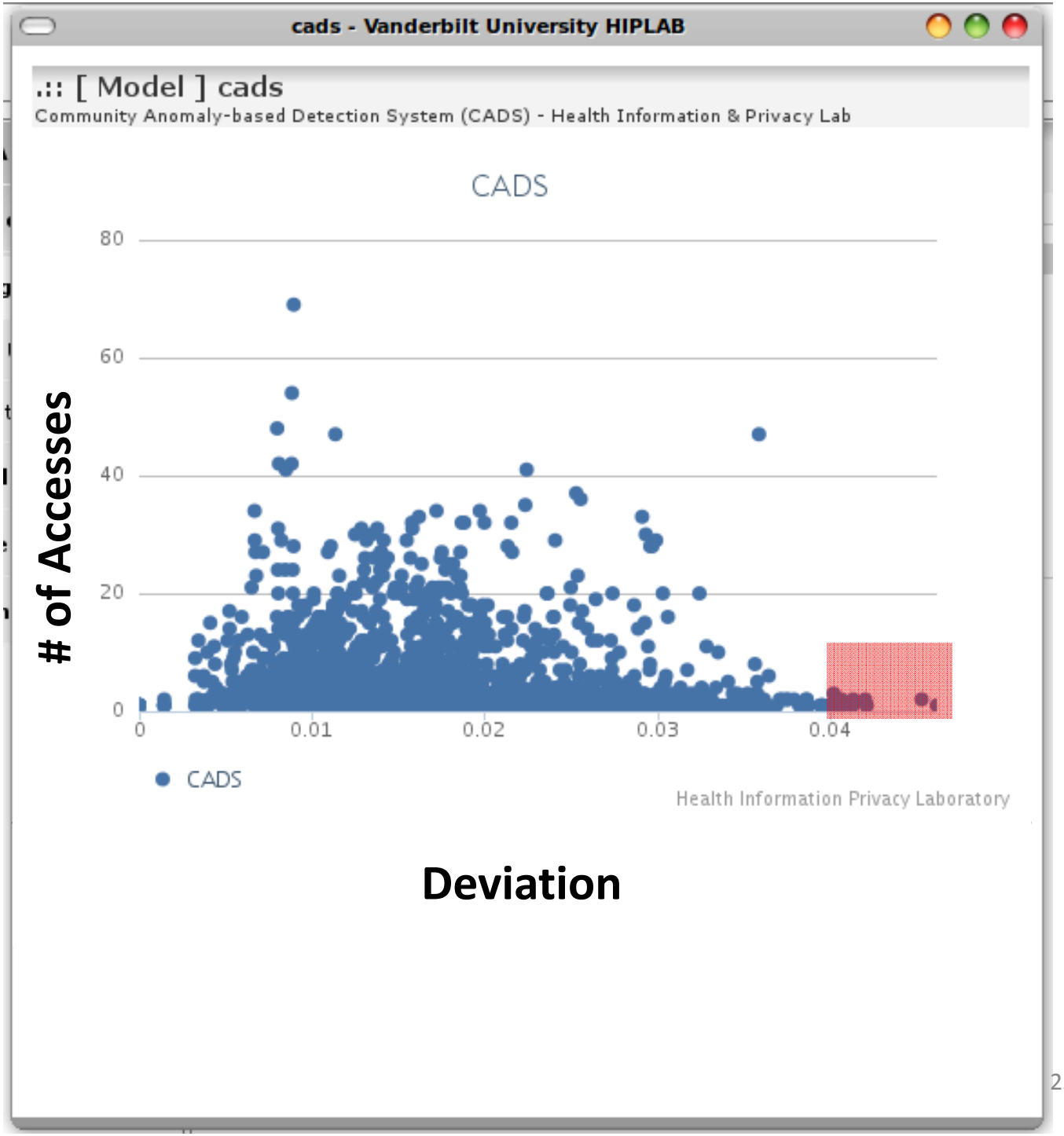
$$Dev(q_1) = \sqrt{\frac{(2-2.2)^2 \times 4 + (3-2.2)^2}{5}} = 0.42$$

5 nearest

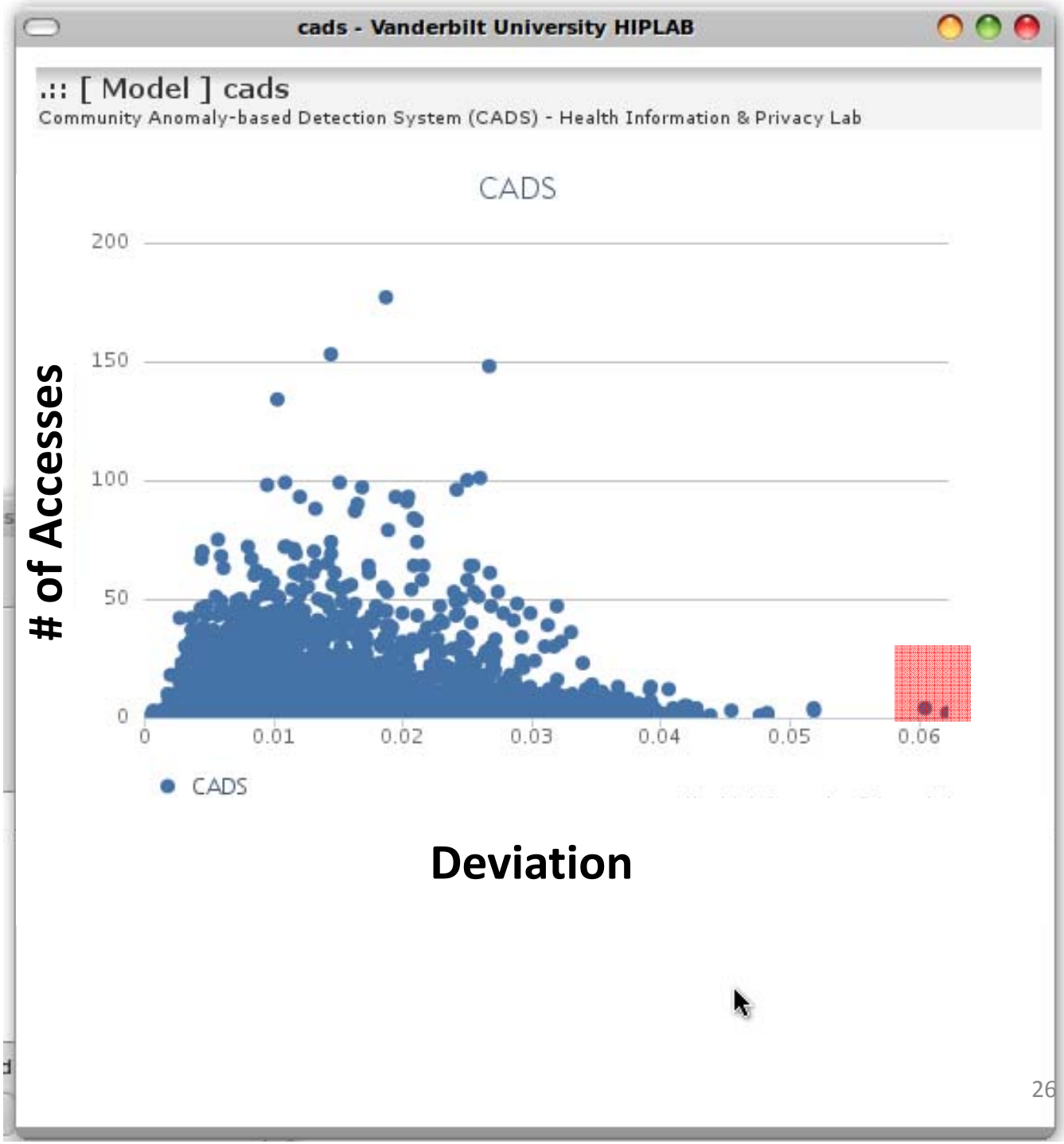


Radius for points are larger than 10,

CADS on Vanderbilt Dataset



CADS on Northwestern Dataset



Example Environments

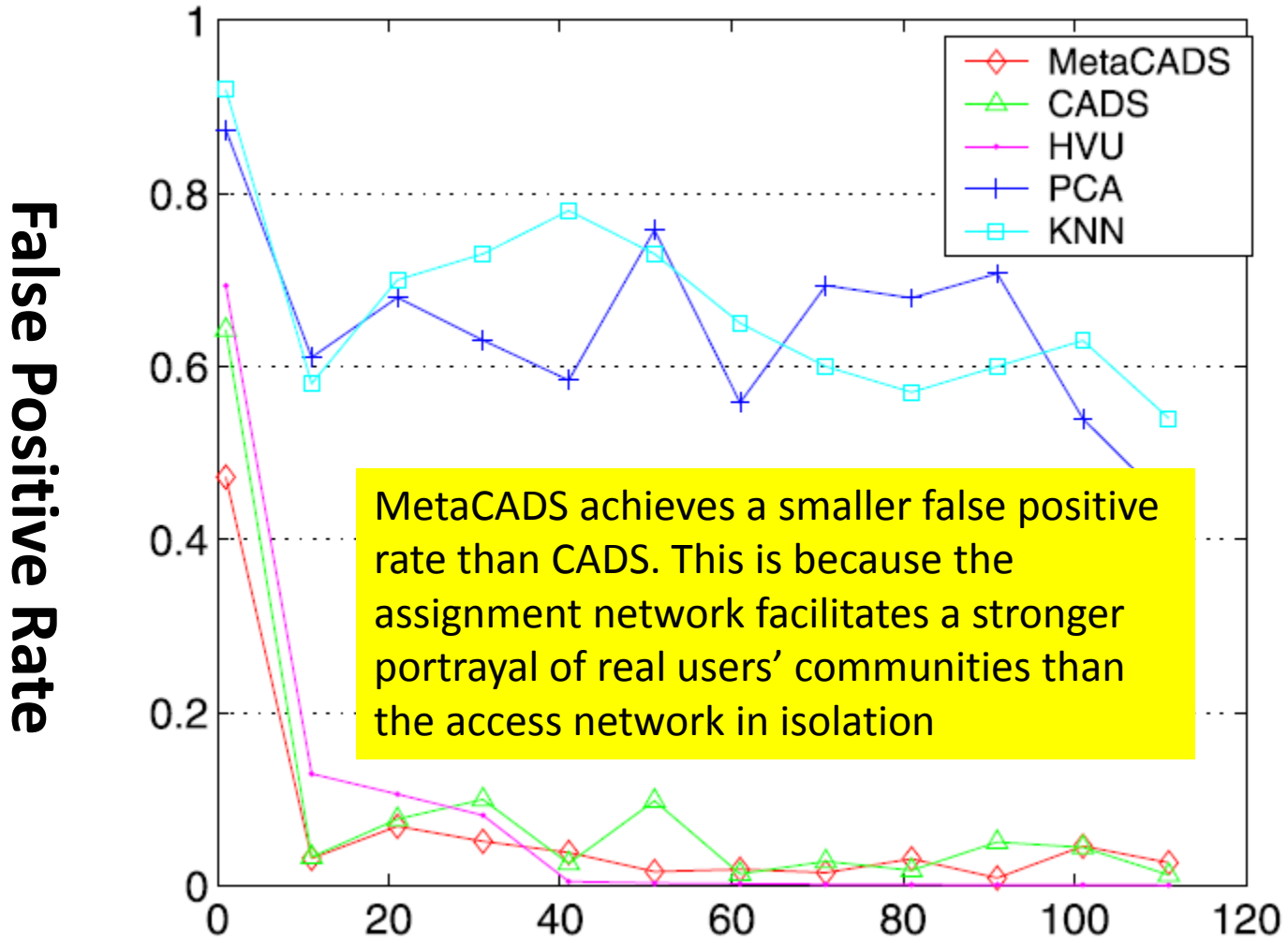
Electronic Health Records (EHR)

- Vanderbilt University Medical Center “StarPanel” Logs
- 3 months in 2010
- Arbitrary Day
  - ≈ 4,208 users
  - ≈ 1,006 patients
  - ≈ 1,482 diagnoses
  - ≈ 22,014 accesses of subjects
  - ≈ 4,609 assignments of diagnoses

Experimental Design

- Datasets are not annotated for illicit behavior
- We simulated users in several settings to test:
  - Sensitivity to number of patients accessed of a specific users
    - Range from 1 to 120
  - 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(0.5%~5%) and records accessed (1~150)

Exp1: False Positive Rate Decreases, when the Number of Subjects Accessed Increases



Exp2: Detection Rate With Various Mix Rates of Real and Simulated Users

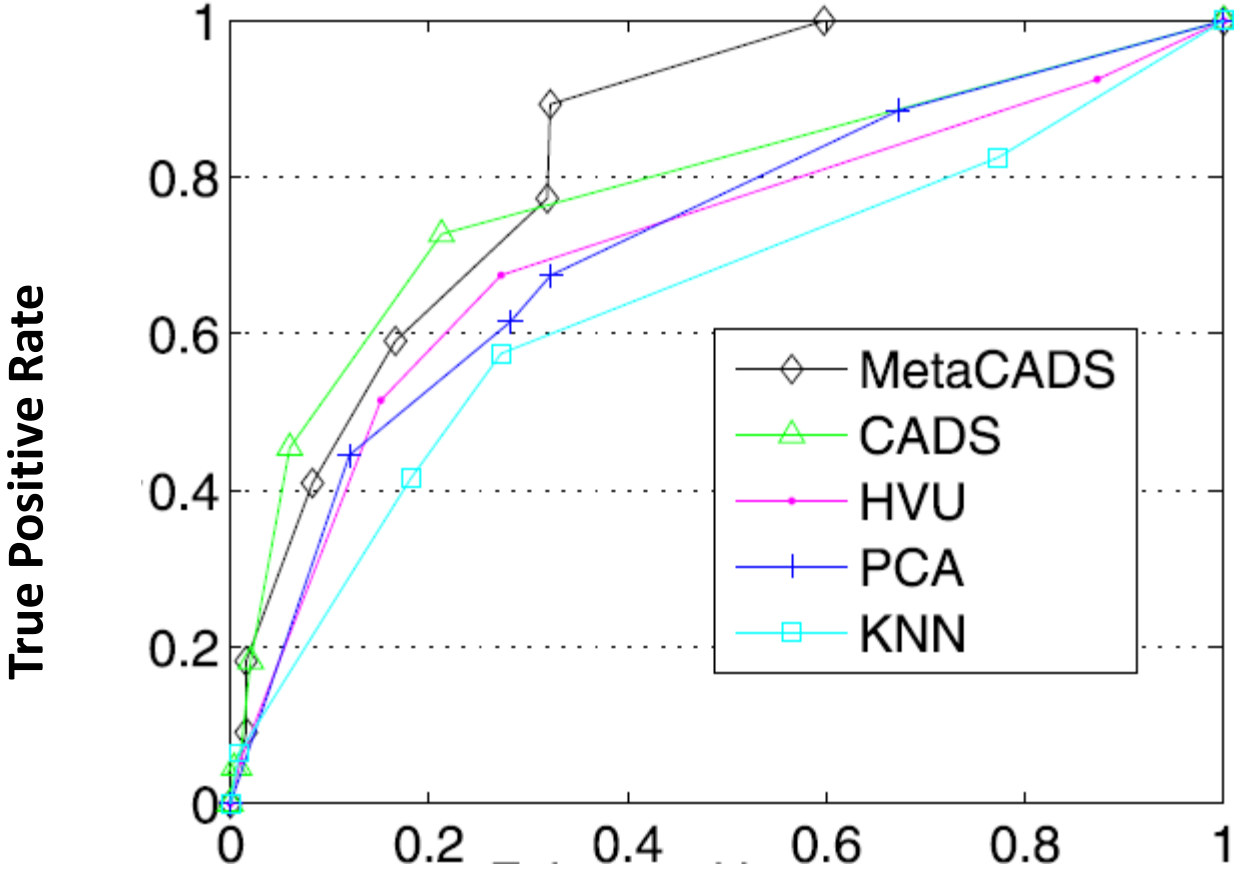
	MIX RATE		
MODEL	0.5%	2%	5%
MetaCADS	**0.92±0.02**	0.90±0.01	0.87±0.03
CADS	0.91±0.01	**0.94±0.02**	**0.94±0.01**
KNN	0.75±0.02	0.73±0.03	0.72±0.04
PCA	0.72±0.03	0.74±0.02	0.75±0.03
HVU	0.68±0.03	0.68±0.03	0.68±0.03

when the number of simulated users is low (i.e., 0.5 percent), MetaCADS yields a slightly higher AUC than CADS (0.92 versus 0.91)

As the number of simulated users increases, CADS clearly dominates MetaCADS. The performance rate of CADS increases from 0.91 to 0.94, while MetaCADS decreases from 0.92 to 0.87.

Because when the number of simulated users increases, they have more frequent categories in common. In turn, these categories enable simulated users to form more communities than those based on patients alone, thus lowering their deviation scores.

Exp3: MetaCADS dominates when the mix rate is low (mix rate = 0.5%)



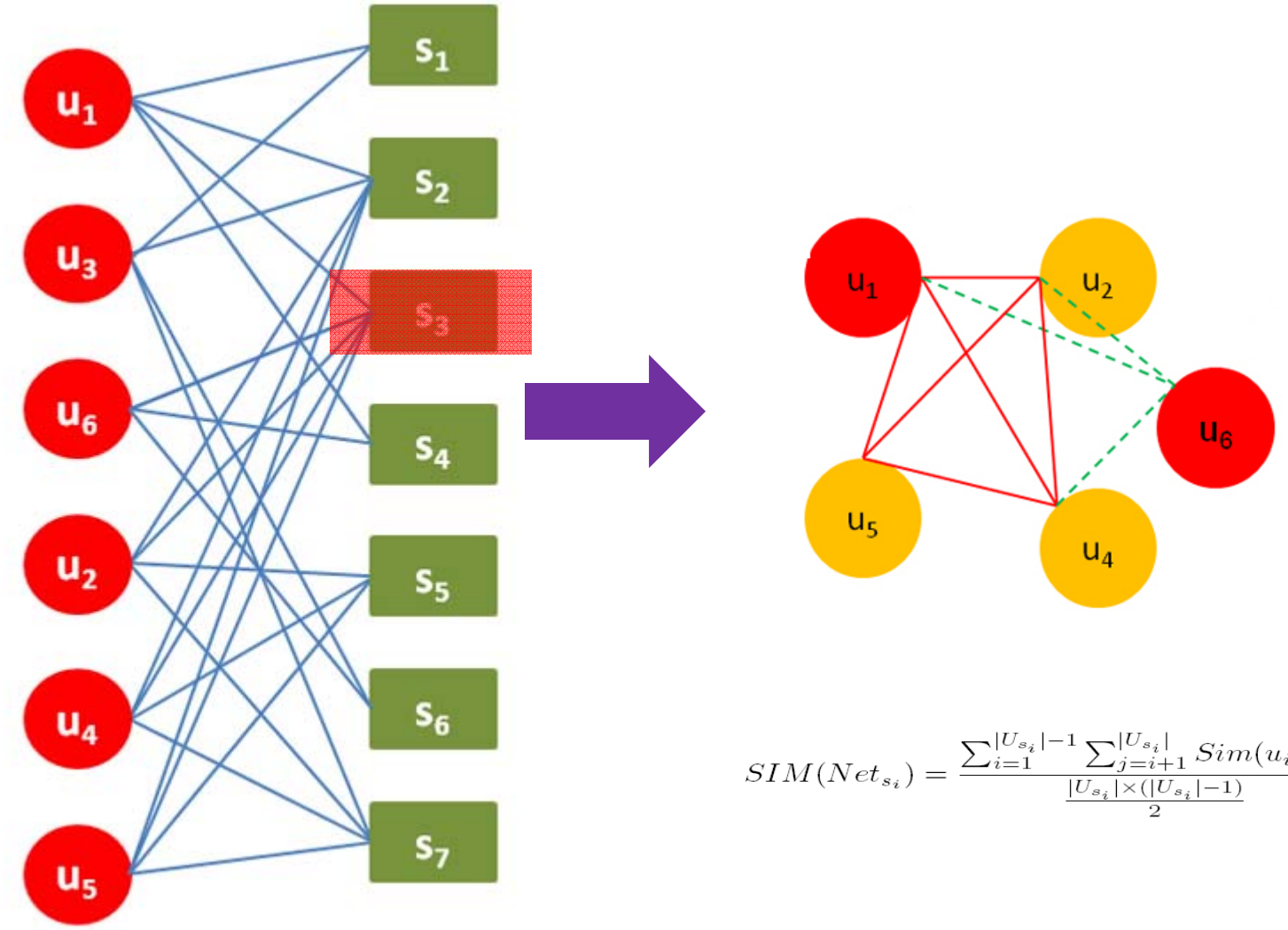
Where are We Going?

Access Level Anomaly Detection

Specialized Network Anomaly Detection (SNAD)

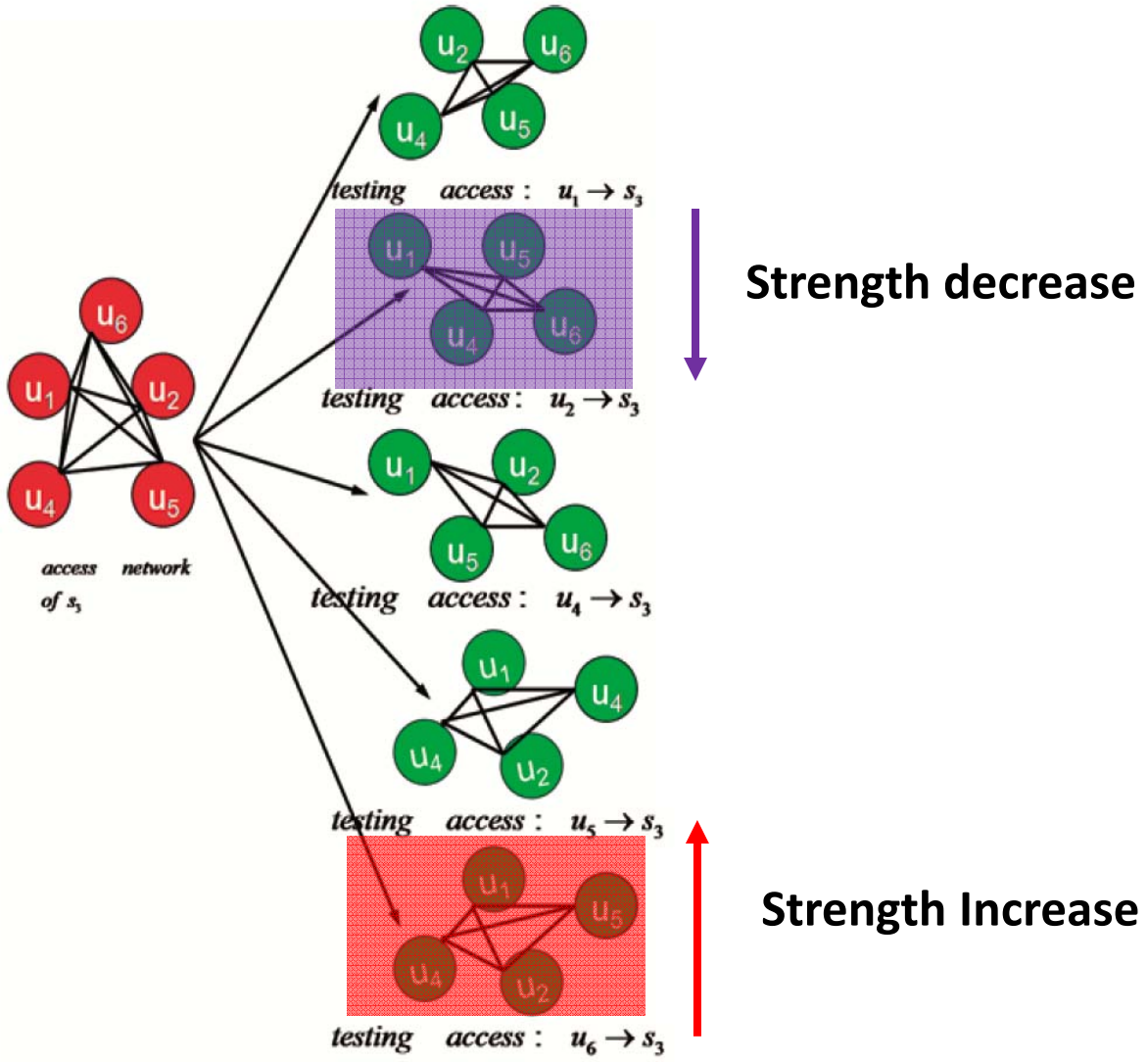
Chen et al. Security Informatics: You Chen, Steve Nyemba, Wen Zhang and Bradley Malin. Specializing Network Analysis to Detect Anomalous Insider Actions. Security Informatics. **1**:5, 2012, p1-24.
Chen et al. ISI: You Chen, Steve Nyemba, Wen Zhang and Bradley Malin. Leveraging Social Networks to Detect Anomalous Insider Actions in Collaborative Environments. IEEE Intelligence and Security Informatics (ISI). 2011.p119-124.

Local Access Network Construction

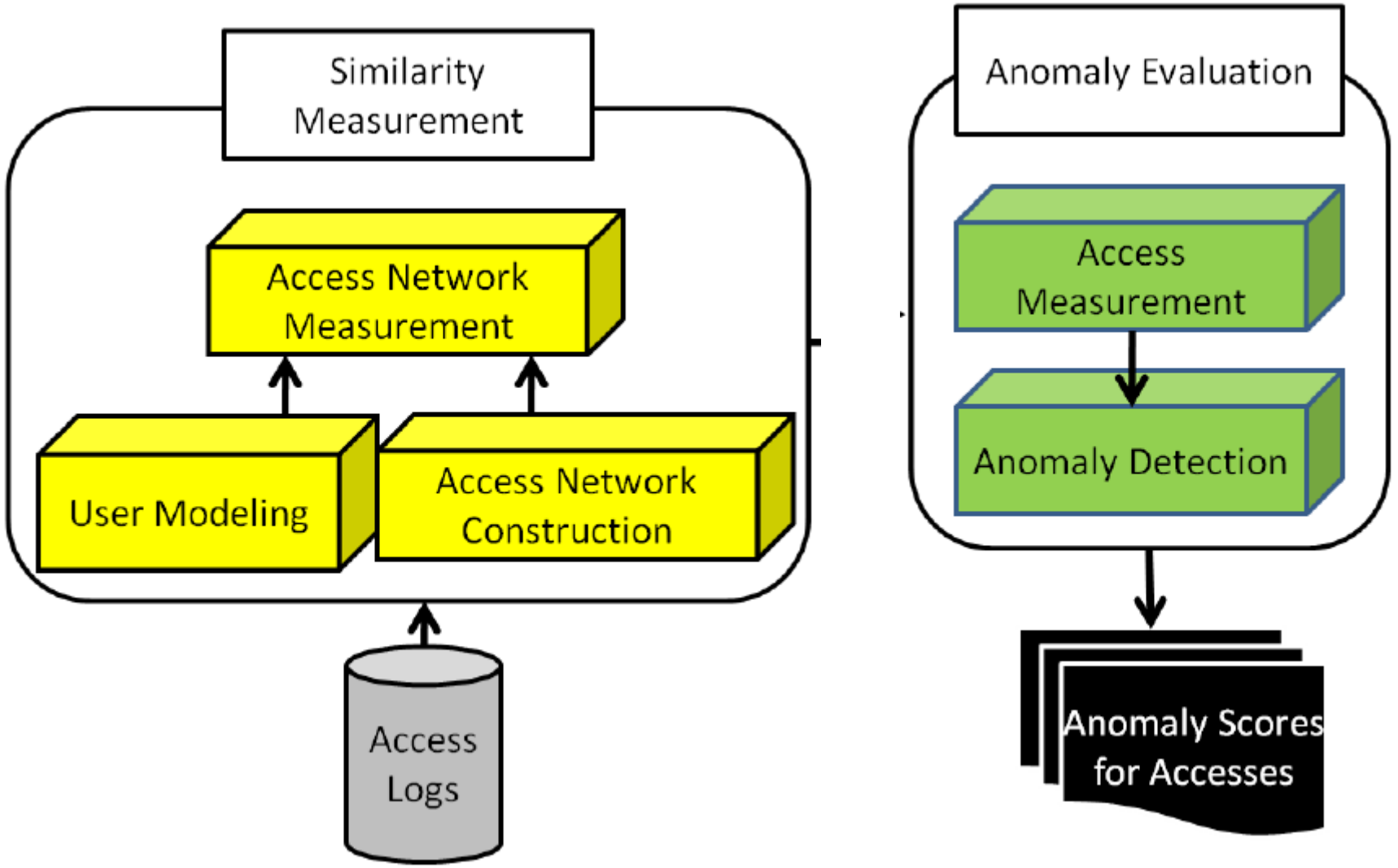


$$SIM(Net_{s_i}) = \frac{\sum_{i=1}^{|U_{s_i}|-1} \sum_{j=i+1}^{|U_{s_i}|} Sim(u_i, u_j)}{\frac{|U_{s_i}| \times (|U_{s_i}|-1)}{2}}$$

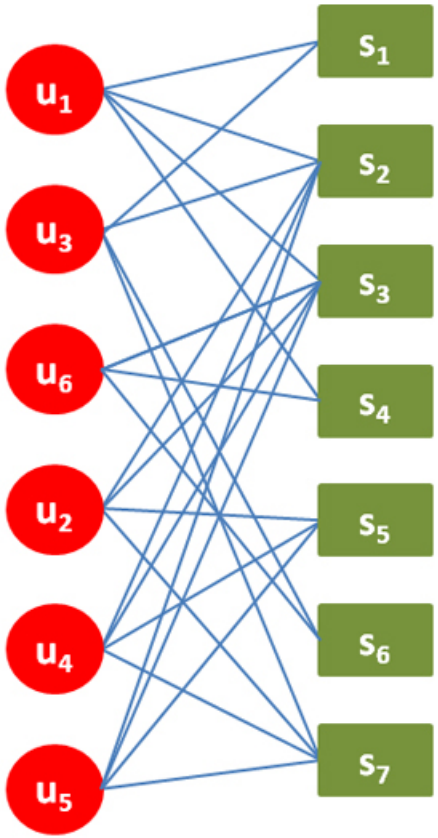
Changes of relation strength of a local network could be leveraged for detection of anomalous accesses



SNAD Framework



User Modeling



	u_1	u_2	u_3	u_4	u_5	u_6
s_1	1	0	1	0	0	0
s_2	1	1	1	1	1	0
s_3	1	1	0	1	1	1
s_4	1	0	0	0	0	1
s_5	0	1	0	1	1	0
s_6	0	0	1	0	0	1
s_7	0	1	1	1	1	0

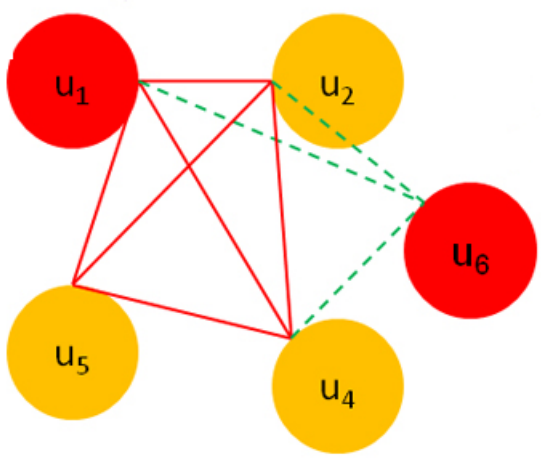
Relationship Measurement

	u_1	u_2	u_3	u_4	u_5	u_6
s_1	1	0	1	0	0	0
s_2	1	1	1	1	1	0
s_3	1	1	0	1	1	1
s_4	1	0	0	0	0	1
s_5	0	1	0	1	1	0
s_6	0	0	1	0	0	1
s_7	0	1	1	1	1	0

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

	u_1	u_2	u_4	u_5	u_6
u_1	1.00				
u_2	0.50	1.00			
u_4	0.50	1.00	1.00		
u_5	0.50	1.00	1.00	1.00	
u_6	0.58	0.29	0.29	0.29	1.00

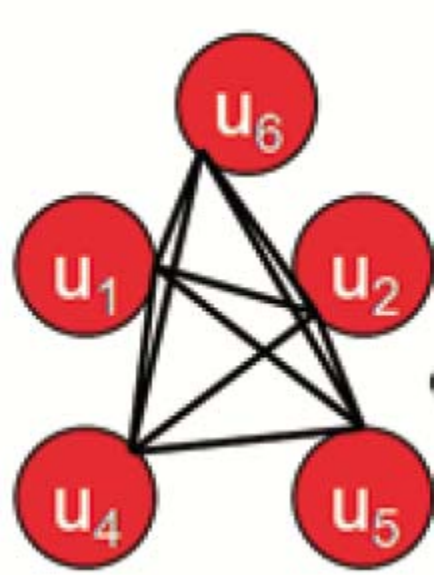
Relation Strength of Local Access Network



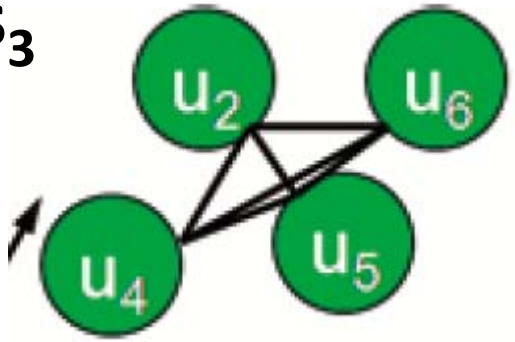
	u_1	u_2	u_4	u_5	u_6
u_1	1.00				
u_2	0.50	1.00			
u_4	0.50	1.00	1.00		
u_5	0.50	1.00	1.00	1.00	
u_6	0.58	0.29	0.29	0.29	1.00

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

Measuring Accesses through Changes of Network Similarity



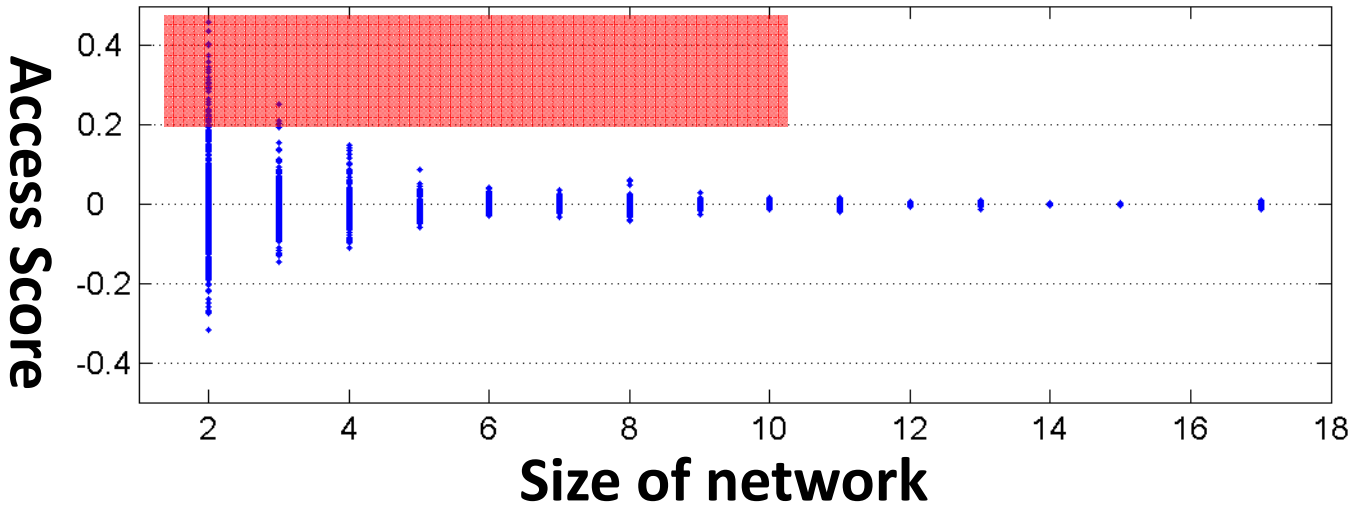
Access: u_1->s_3

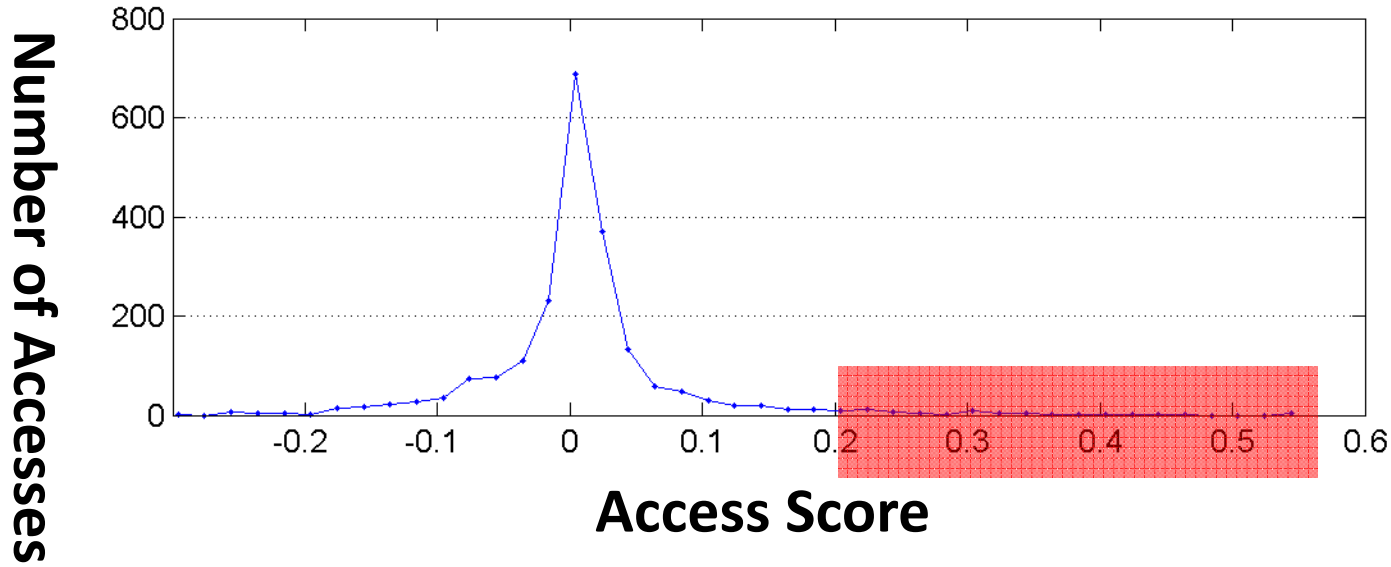


Network	Similarity	Size
u_1,u_2,u_4,u_5,u_6	0.59	5
u_2,u_4,u_5,u_6	0.64	4

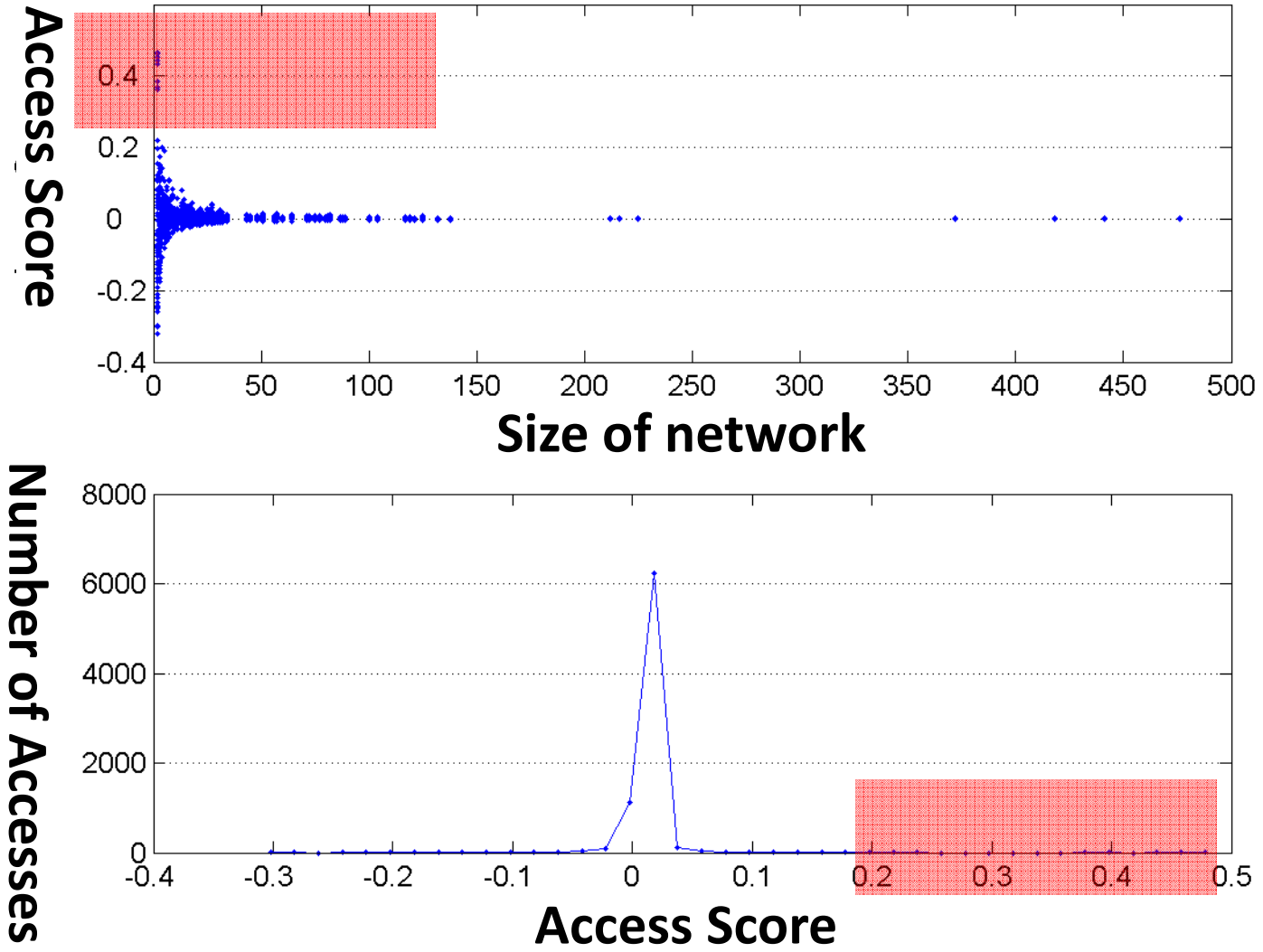
Access	Score	Size
u1-s3	0.05	4

In EHR System-one week





In Wiki-one week



Evaluation

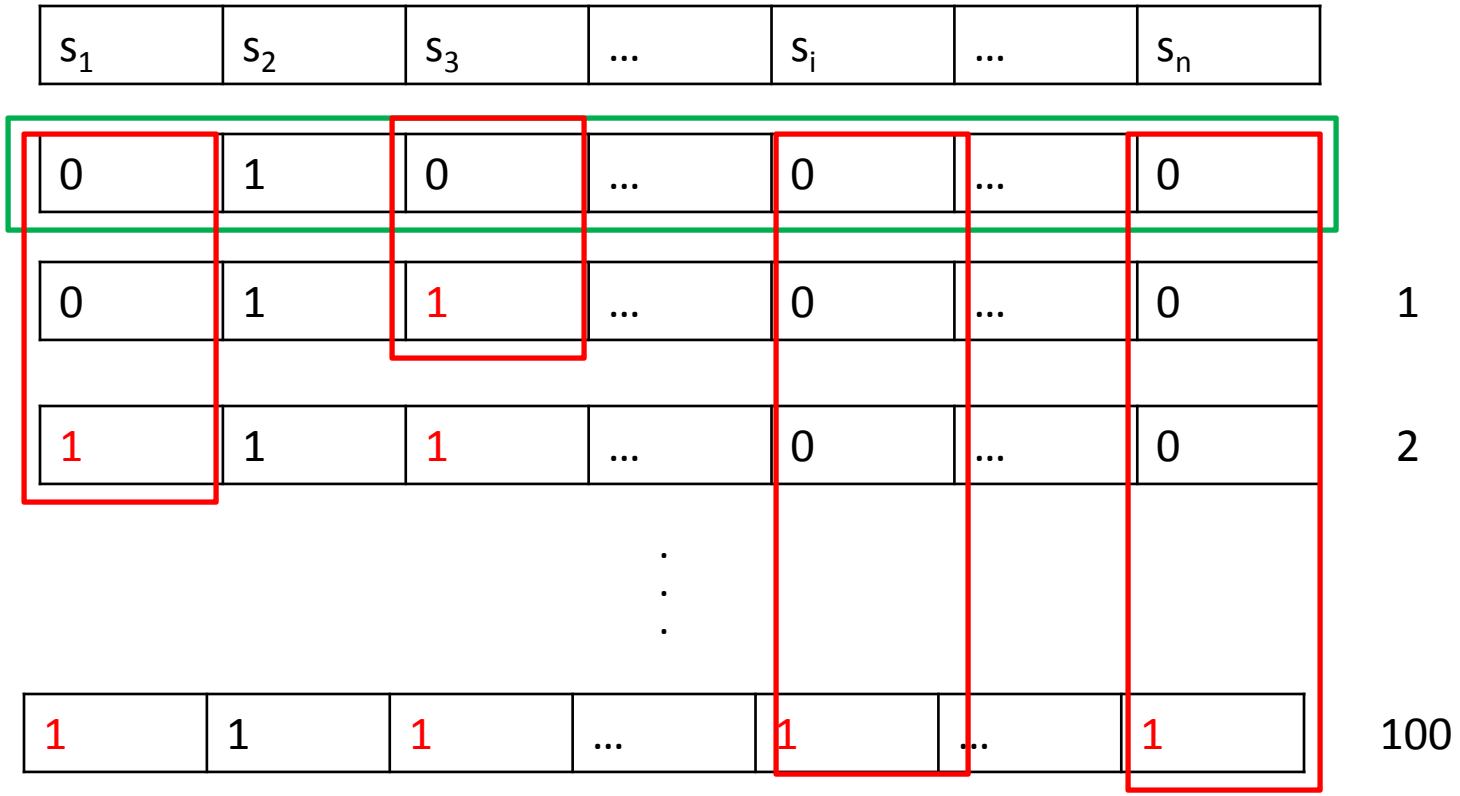
For a random user, verifying how number of simulated access injected into this user influence the performances of SNAD

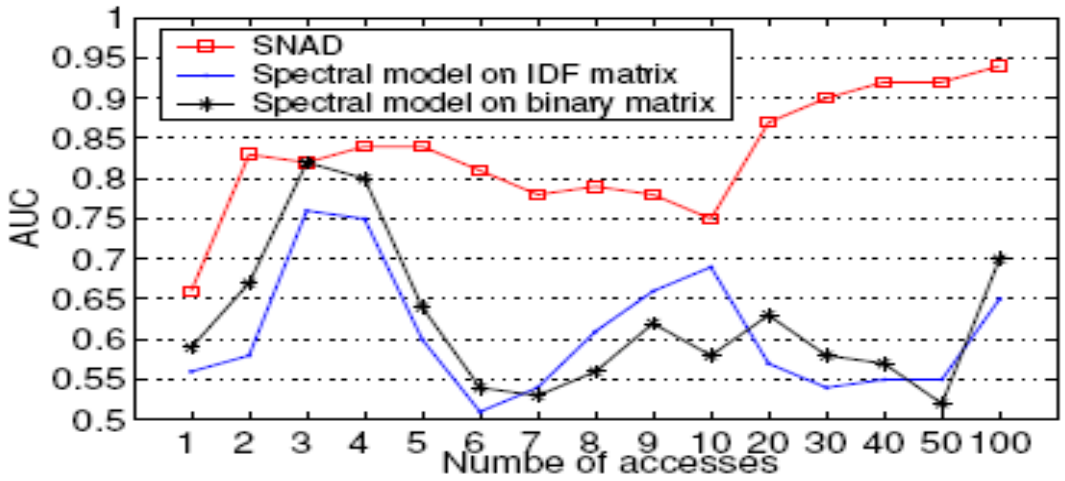
For a fixed number of simulated accesses, verifying how number of intruded users influence the performances of SNAD

The number of simulated accesses and intruded users are both diverse

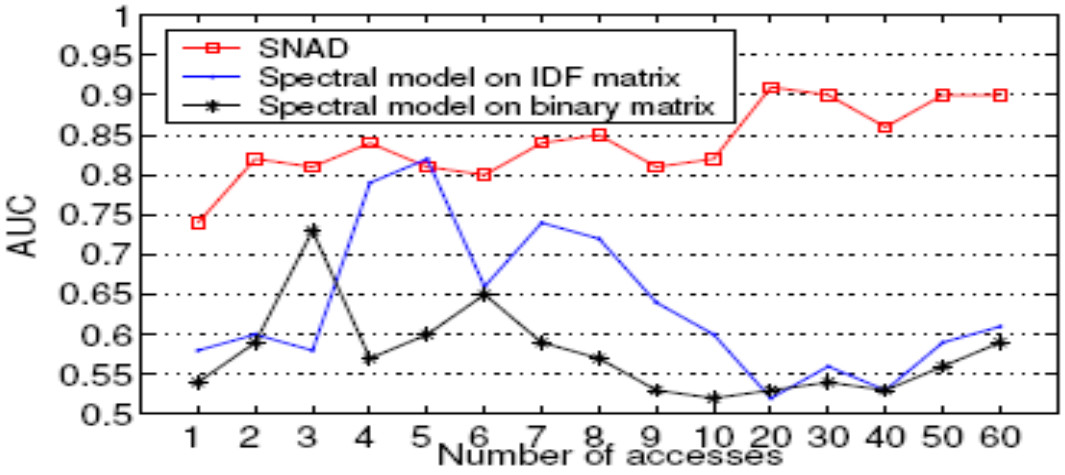
Model Evaluation-setting 1

For a random user, injecting simulated accesses




SNAD
Spectral model on IDF matrix
Spectral model on binary matrix
AUC
Numbe of accesses
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
1 2 3 4 5 6 7 8 9 10 20 30 40 50 100

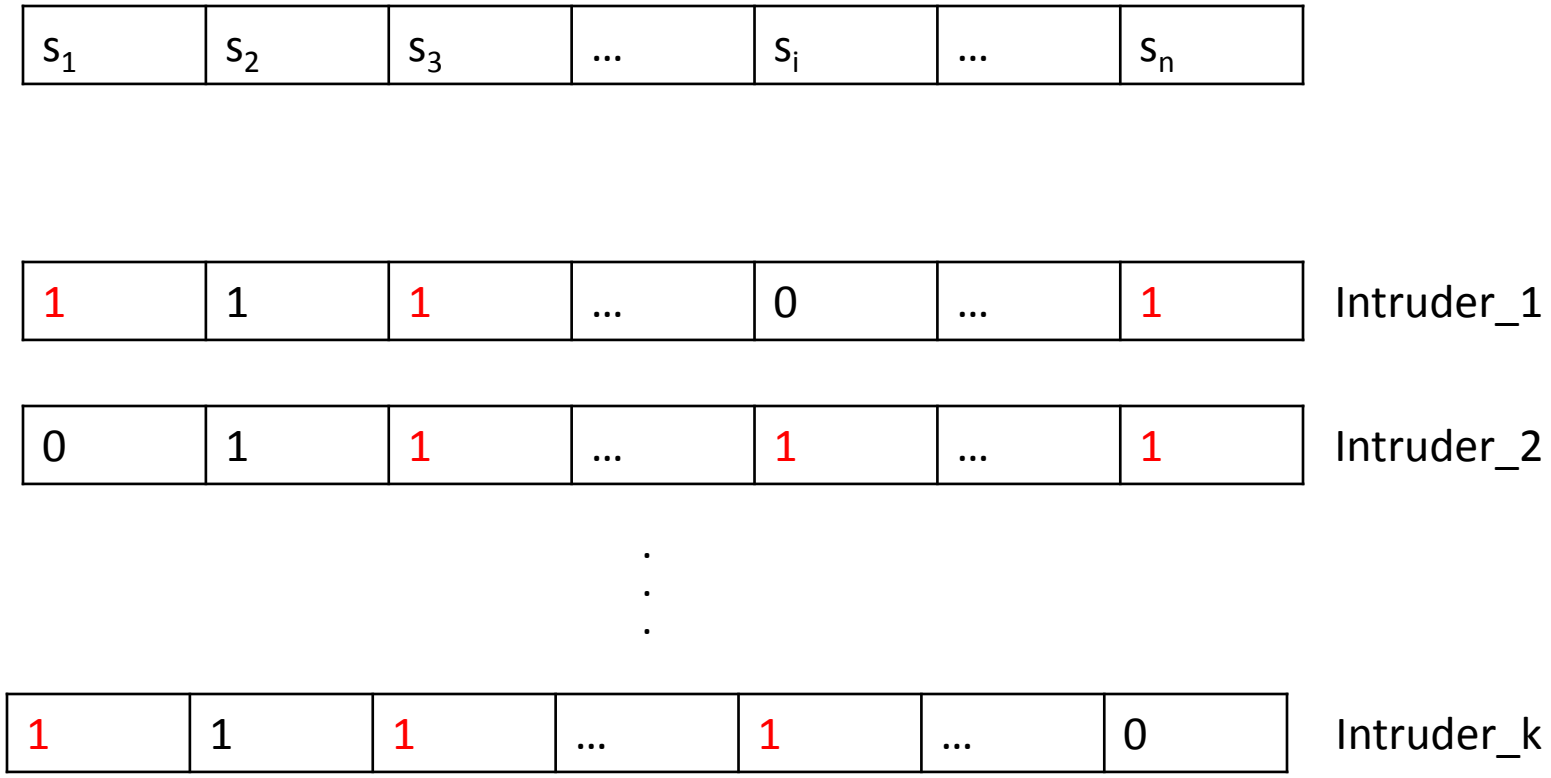
(a) EHR

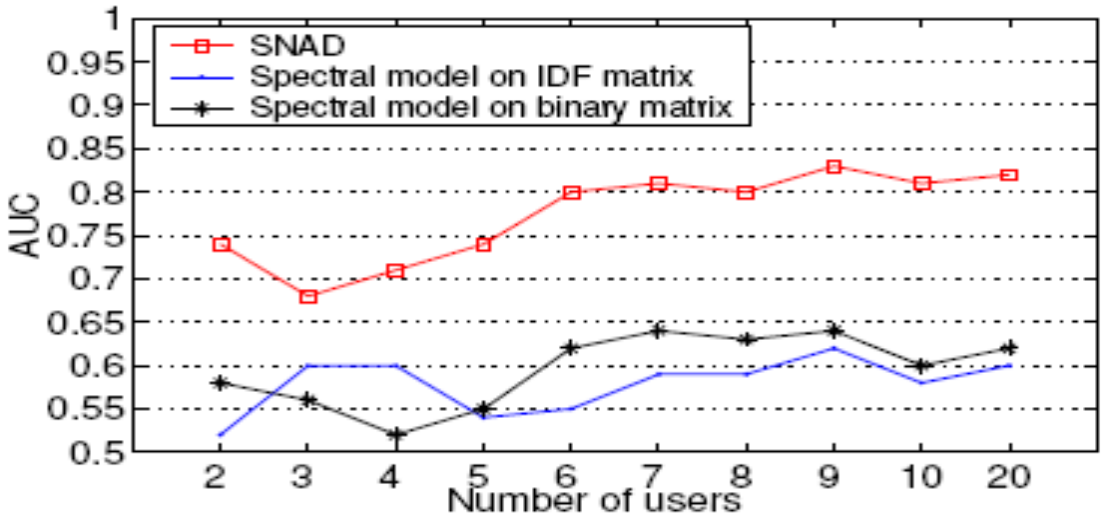

SNAD
Spectral model on IDF matrix
Spectral model on binary matrix
AUC
Number of accesses
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
1 2 3 4 5 6 7 8 9 10 20 30 40 50 60

(b) Wiki

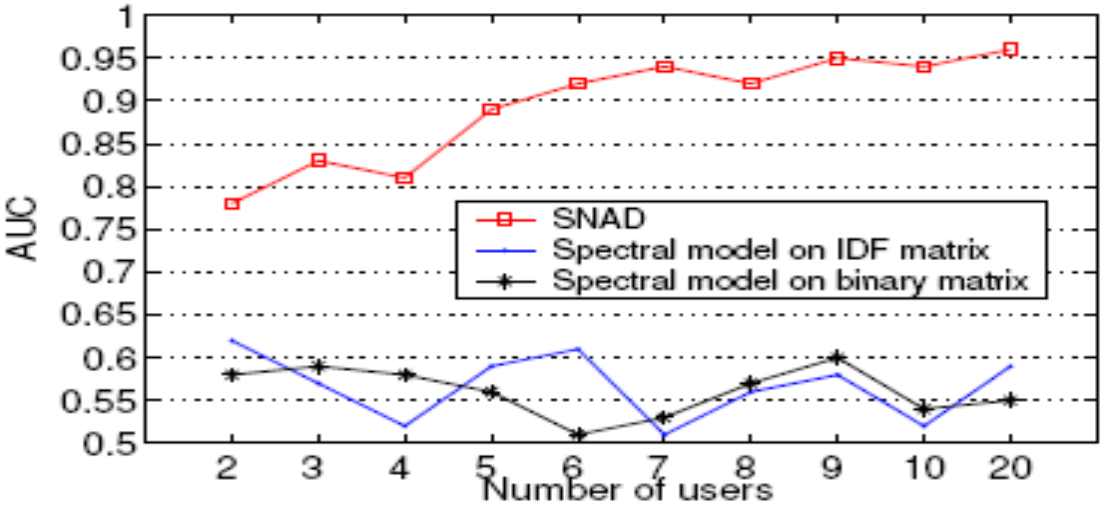
Model Evaluation-setting 2

Fixing number of simulated accesses, number of intruders is random




SNAD
Spectral model on IDF matrix
Spectral model on binary matrix
AUC
Number of users
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
2
3
4
5
6
7
8
9
10
20

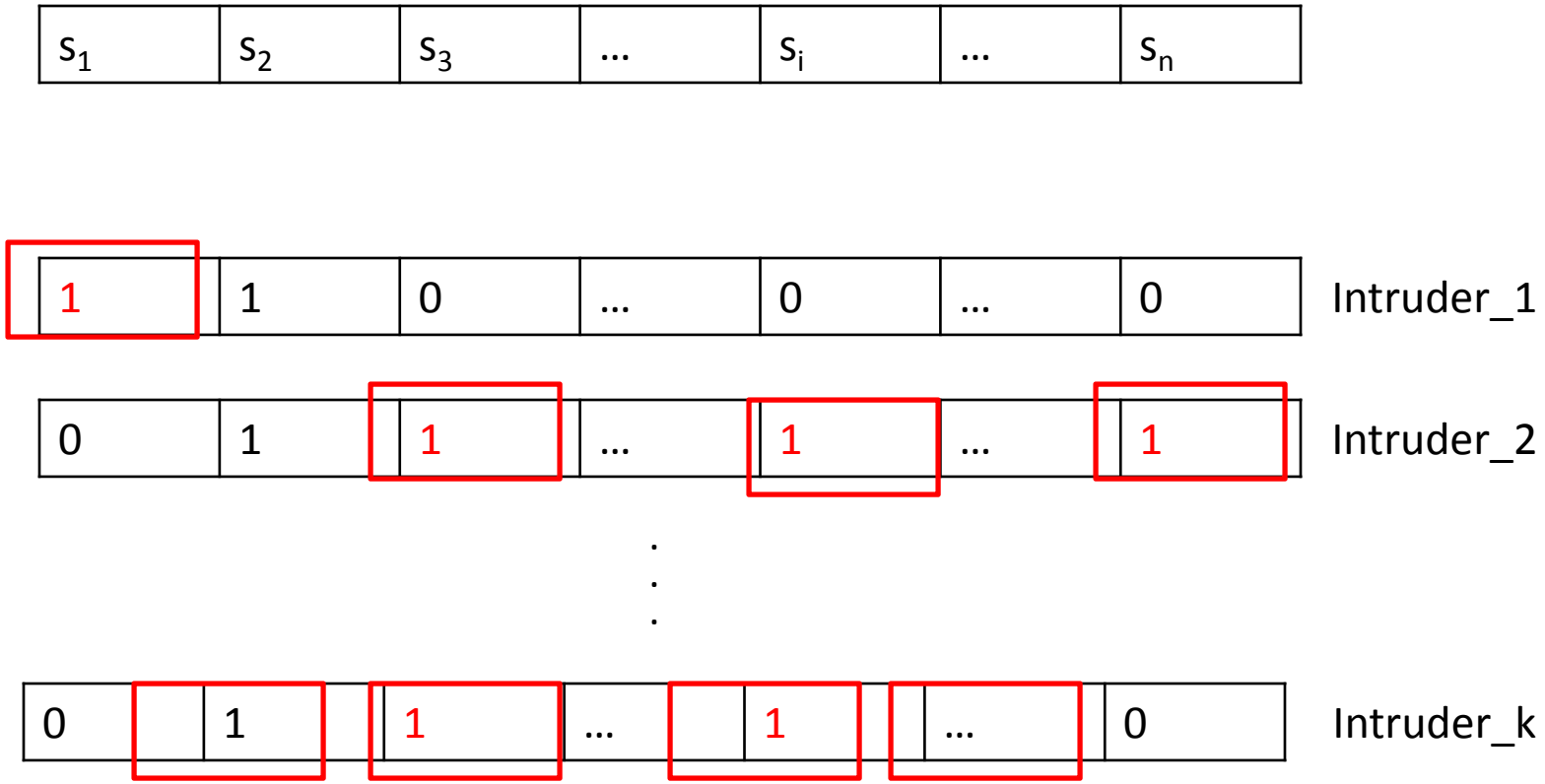
(a) EHR

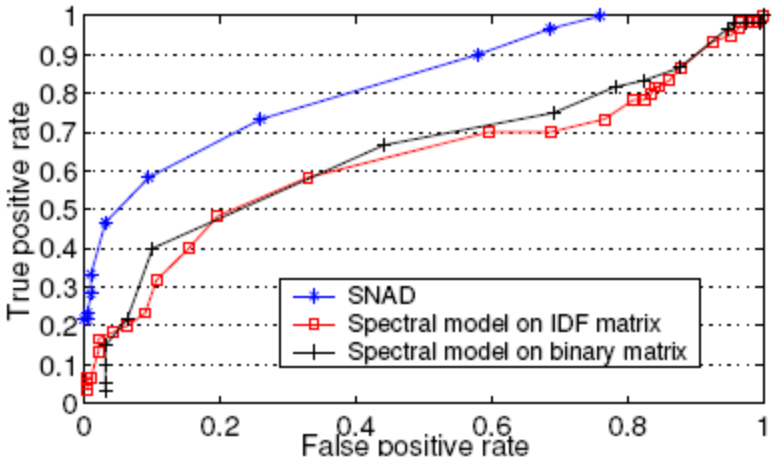

SNAD
Spectral model on IDF matrix
Spectral model on binary matrix
AUC
Number of users
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
2
3
4
5
6
7
8
9
10
20

(b) Wiki

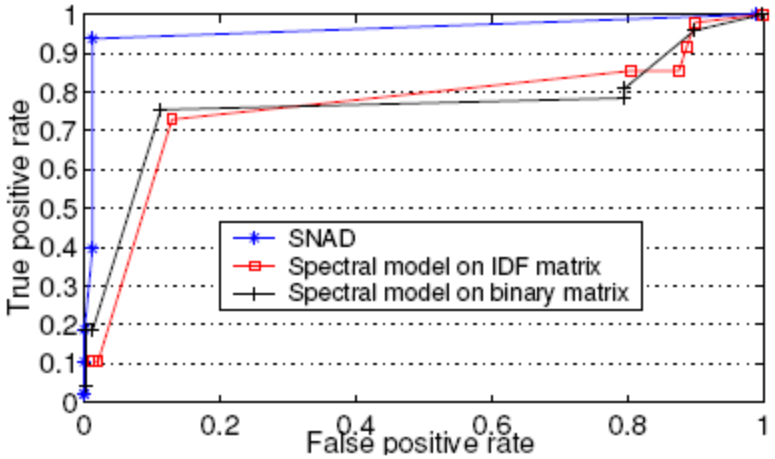
Model Evaluation-setting 3

Fixing number of simulated accesses, number of intruders is random





(a) EHR

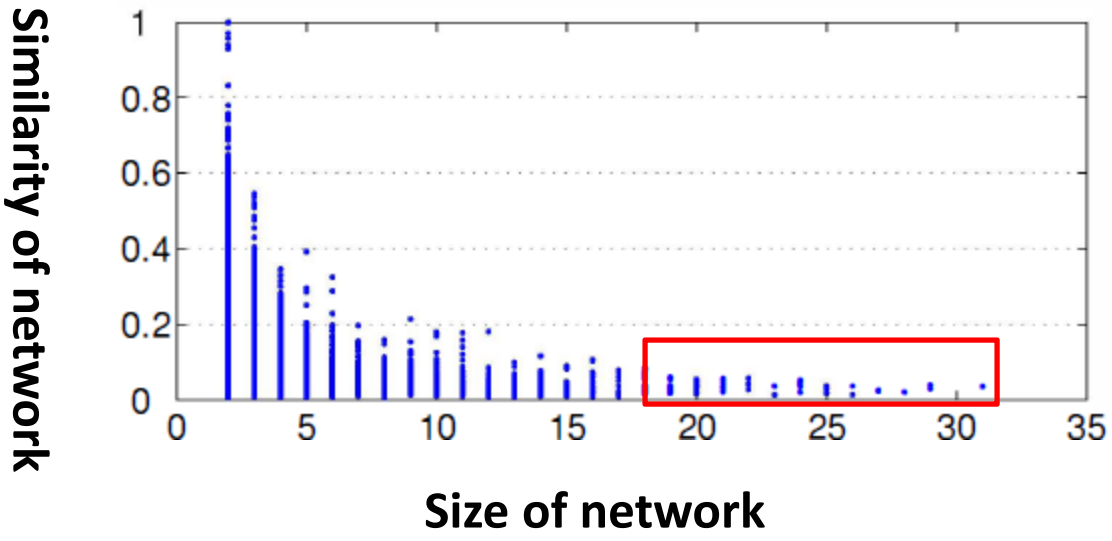


(b) Wiki

Dataset	SNAD	Spectral IDF	Spectral Binary
EHR	0.83±0.03	0.74±0.06	0.69±0.05
Wiki	0.91±0.02	0.76±0.04	0.64±0.04

Limitations

- 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

Thanks!