VANDERBILT UNIVERSITY

TRUST
Team for Research in Ubiquitous Secure Technology

TRUST Autumn 2011 Conference

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)

VANDERBILT UNIVERSITY

Privacy Rights Clearinghouse

Empowering Consumers. Protecting Privacy.

Home | Why Privacy | About Us | Fact Sheets | Latest Issues | Speeches & Testimony

Browse Privacy Topics

- Privacy Basics
- Background Checks & Workplace
- Banking & Finance
- Credit & Credit Reports
- Debt Collection
- Education
- Harassment & Stalking
- Identity Theft & Data Breaches
- Insurance
- Junk Mail/Faxes/Email
- Medical Privacy
- Online Privacy & Technology
- Privacy When You Shop
- Public Records & Info Brokers

Fact Sheet 8a: HIPAA Basics: Medical Privacy in the Electronic Age

Send to Printer

Posted April 2003
Revised June 2011

[Also see our FAQ on medical privacy.]

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.

VANDERBILT UNIVERSITY

MEDLAW.COM

EMTALA and Healthlaw Resources For
Healthcare Professionals, Hospitals, and Their Attorneys

Search

Browse...

| About Us | Contact Us | Privacy Policy | Terms of use |

- EMTALA Resources
- EMS and Air Medical
- Emergency Preparedness
- Drugs / Pharmacy
- Fraud & Abuse
- Hospital Issues
- Medical Malpractice
- Medical Records Privacy & Security
- CMS / OSHA / CDC
- Product Liability
- Professional Rights
- How We Can Help
- News and Updates
- MedLaw Links (A-M)
- MedLaw Links -- (N-S)
- Publisher's Opinions
- Contact Us
- About Us
- Privacy Policy
- Terms Of Use

[In HIPAA / State Privacy Regulations] Jul 15, 2010
New HIPAA/HITECH Rules Announced:

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.

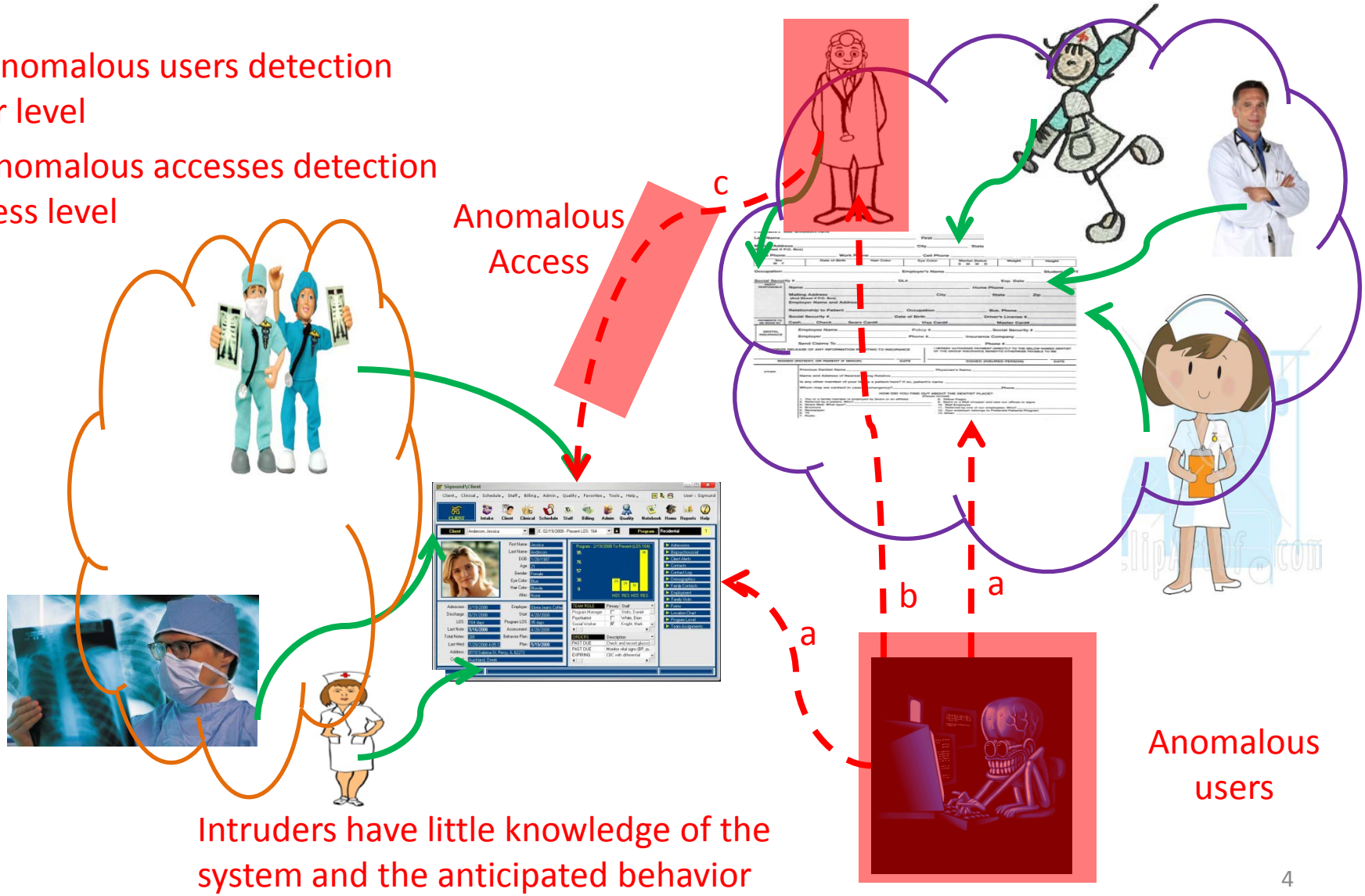
Subscribe now: RSS news feeds provide instant updates without the email

Two Typical Attacks

Intruders have complete knowledge of the system and its policies

(1) Anomalous users detection –user level

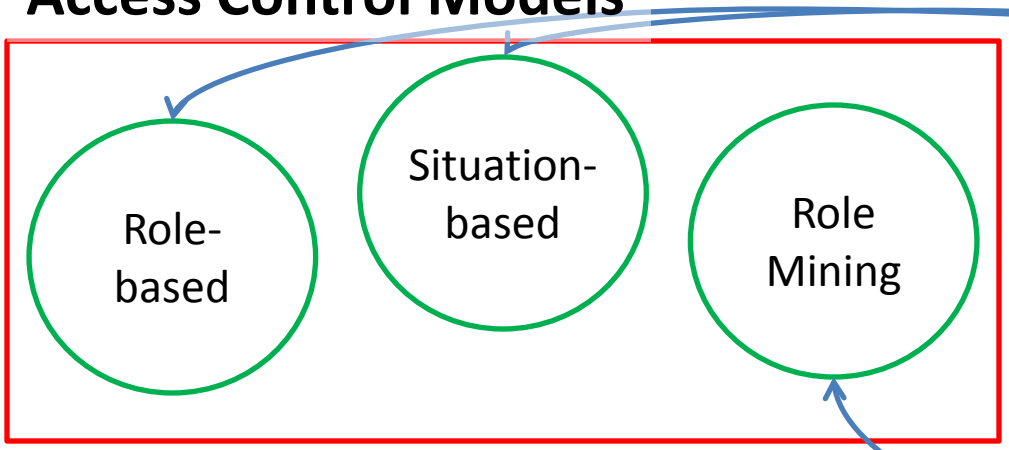
(2) Anomalous accesses detection –access level



Intruders have little knowledge of the system and the anticipated behavior

Related Research

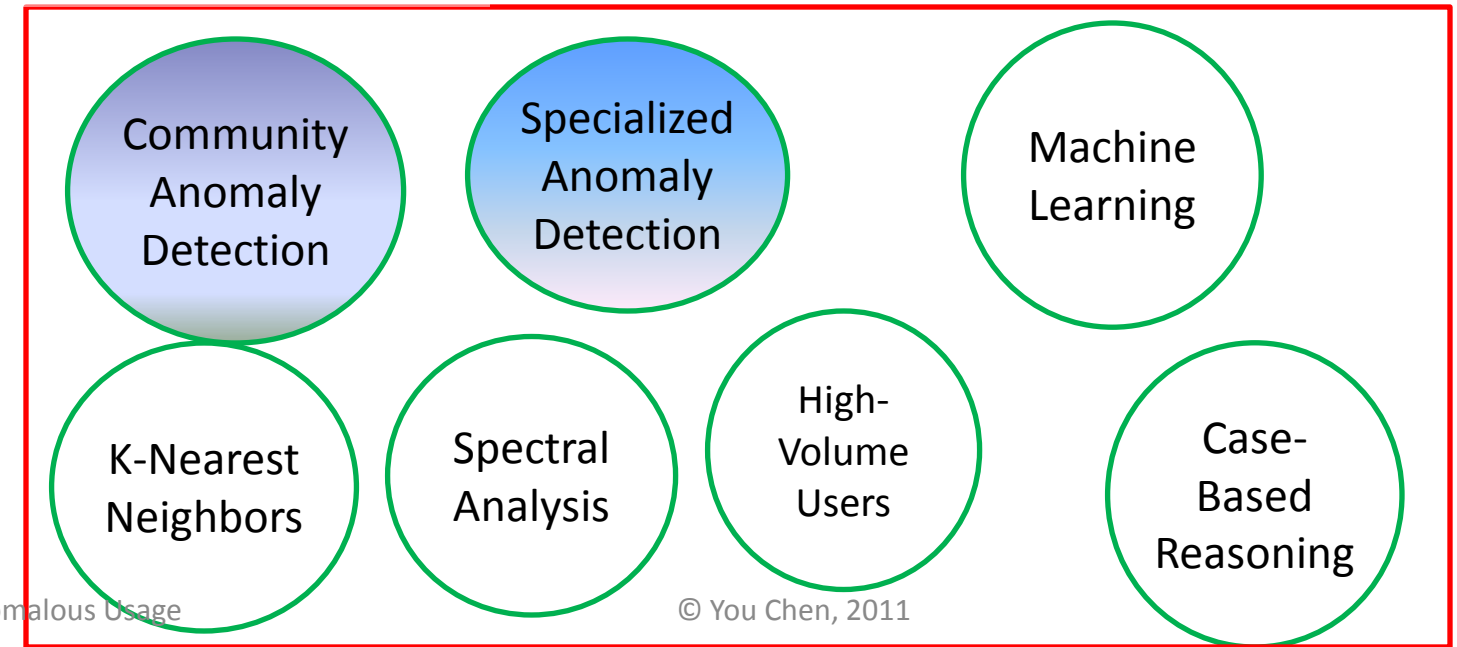
Access Control Models



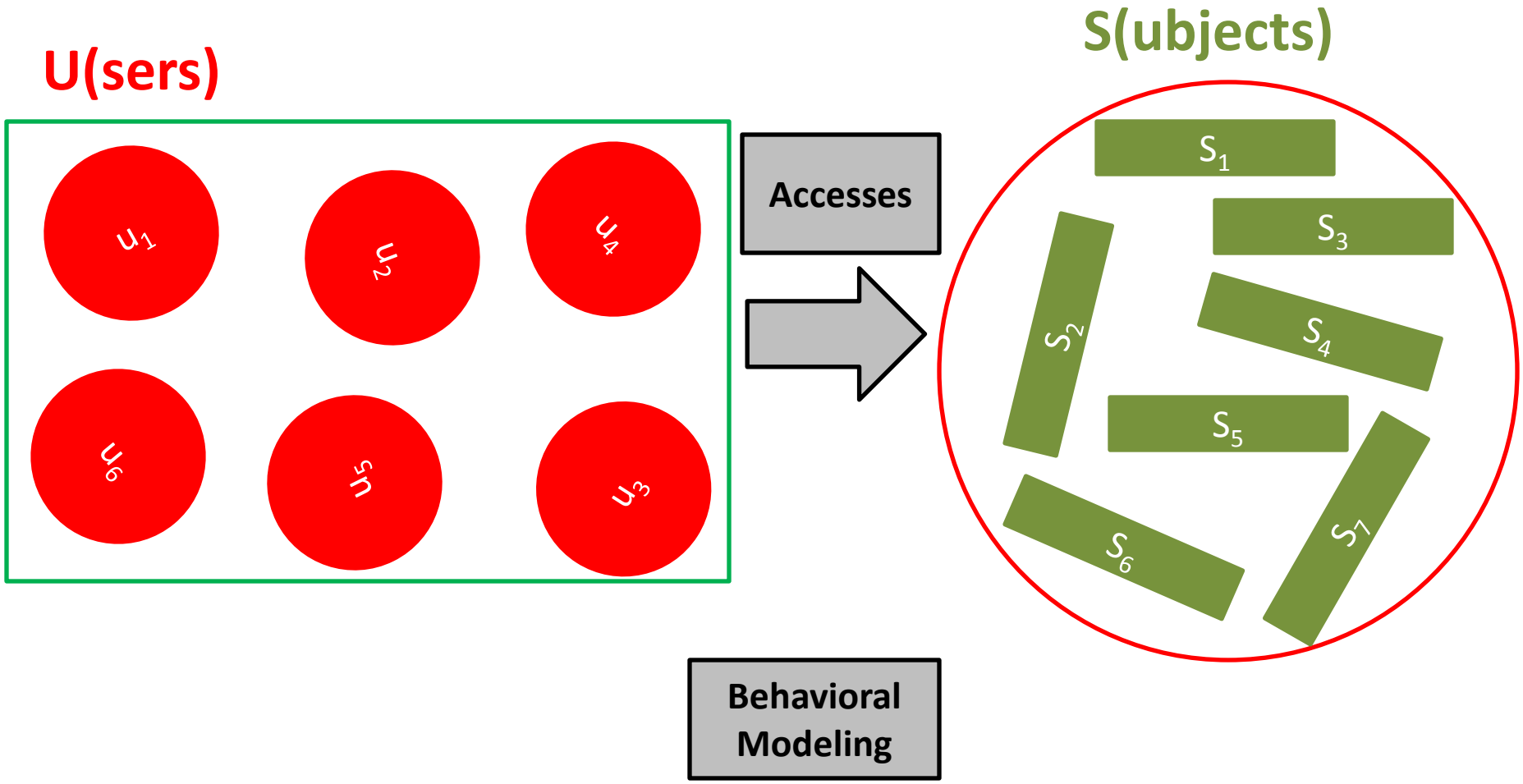
Do not capture the dynamic relationships among users in collaborative information systems

Does not offer stability of access control model over time

Auditing Models



Two general objects of health information system



Where are We Going?

User Level Anomaly Detection

Community Anomaly Detection System (CADS)

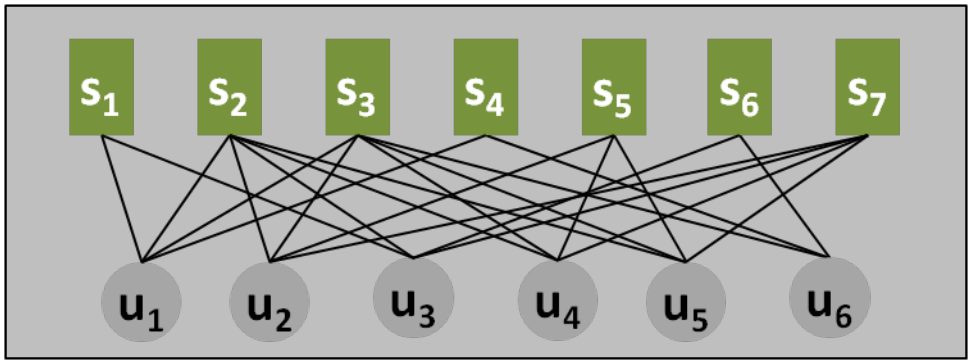
(ACM CODASPY'11)

Access Level Anomaly Detection

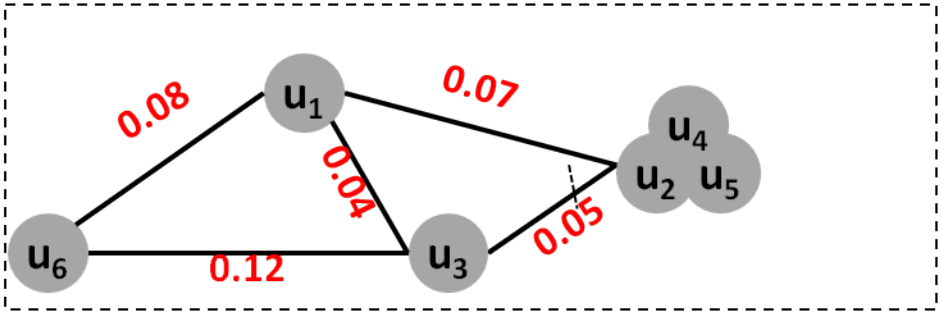
Specialized Network Anomaly Detection (SNAD)

(IEEE ISI'11)

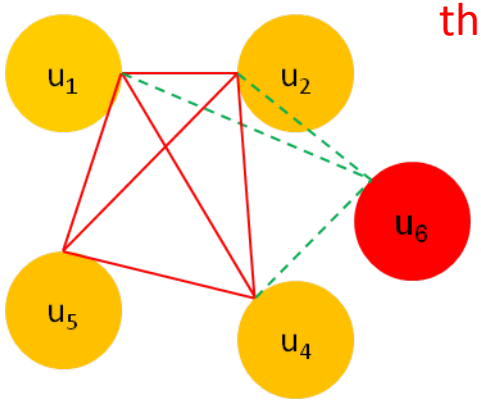
Social Networks are a Novel Approach to Discovery of Electronic Medical Record Misuse



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



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



Example Environments

Electronic Health Records (EHR)

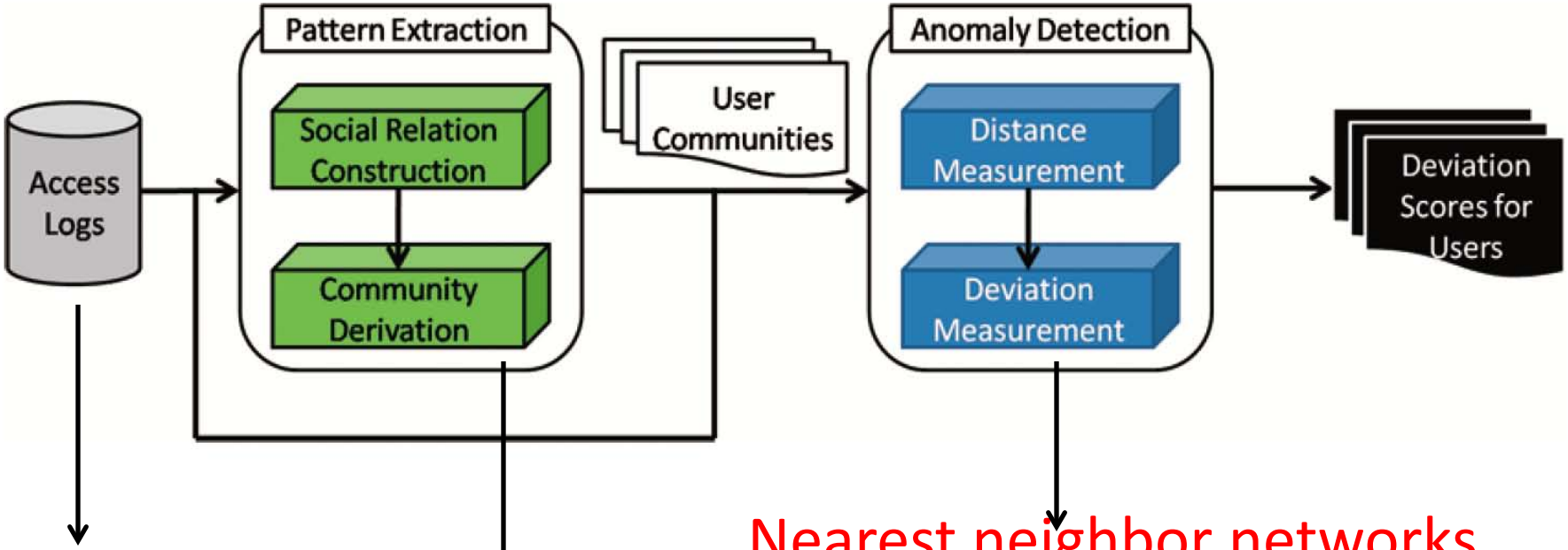
- Vanderbilt University Medical Center "StarPanel" Logs
- 6 months in 2006
- Arbitrary Week
  - ≈ 2,300 users
  - ≈ 35,000 patient records
  - ≈ 66,000 accesses

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)



<user, subject, time>

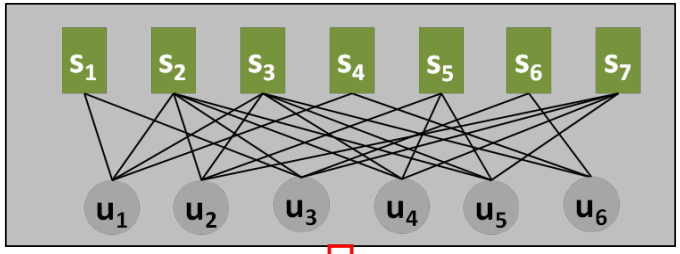
Communities of users

Nearest neighbor networks

Where are We Going?

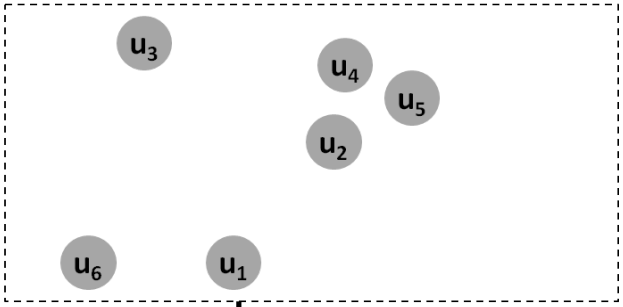
- User Level: Community Anomaly Detection System (CADS) (ACM CODASPY'11)
  - Framework of CADS
  - **An Example of CADS**
  - Experimental Evaluation
  - Limitation

- Access Level: Specialized Network Anomaly Detection (SNAD) (IEEE ISI'11)



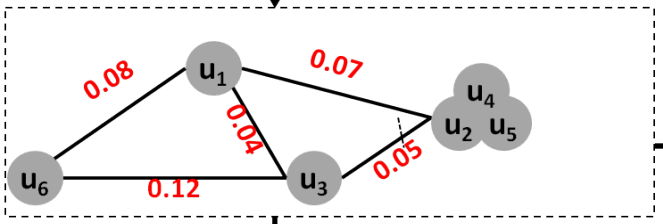
Bipartite graph

Communities Derivation



Communities

Distance Measurement



Nearest Neighbor Network

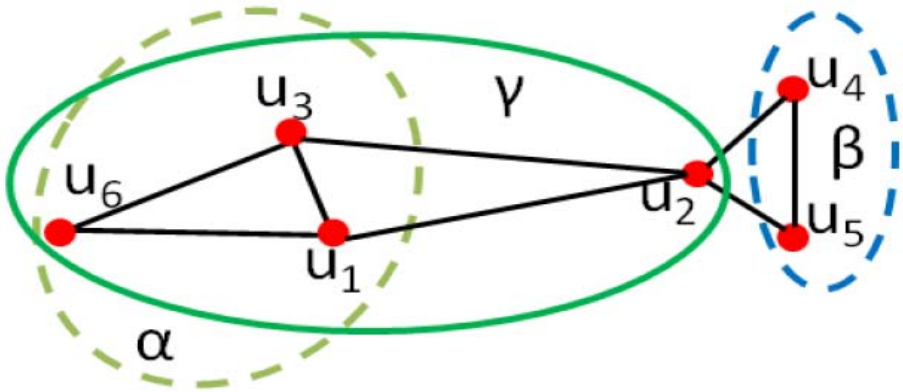
Deviation Measurement

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

How Do We Set "k"-NN?

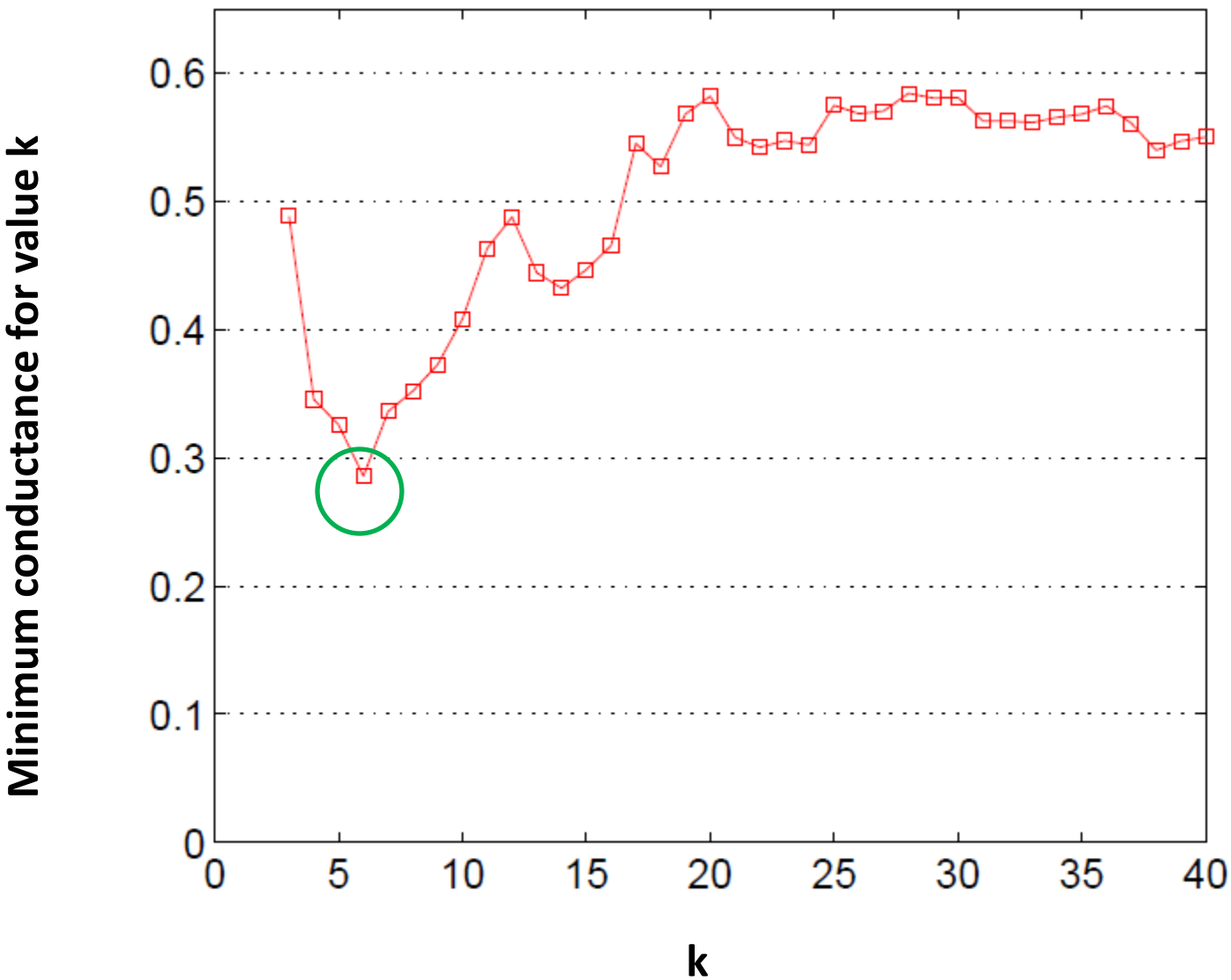
- Conductance- a measure of community quality (Kannan et al)



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

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

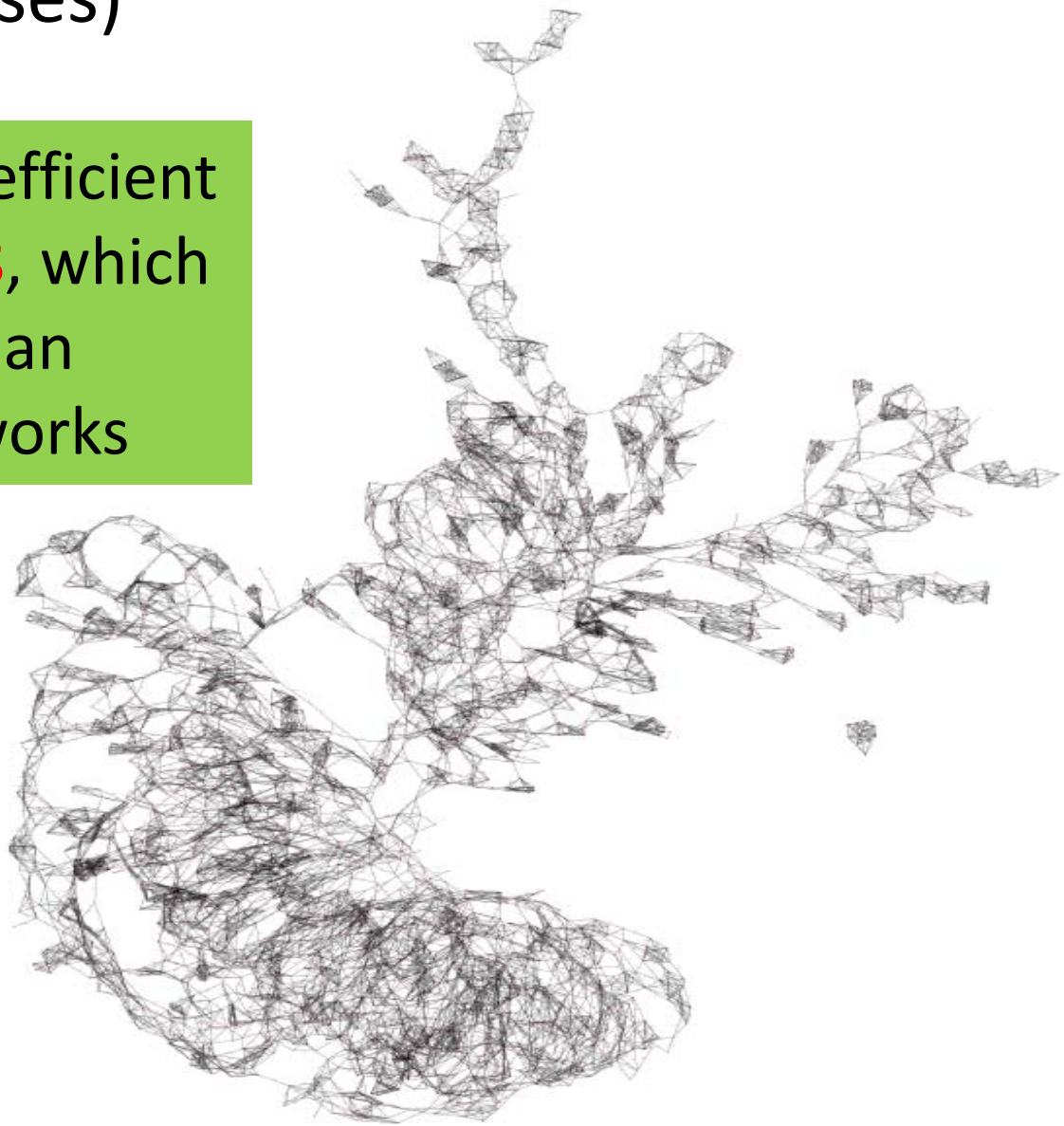
Minimum conductance at k=6



Example 6-Nearest Neighbor Network (1 day of accesses)

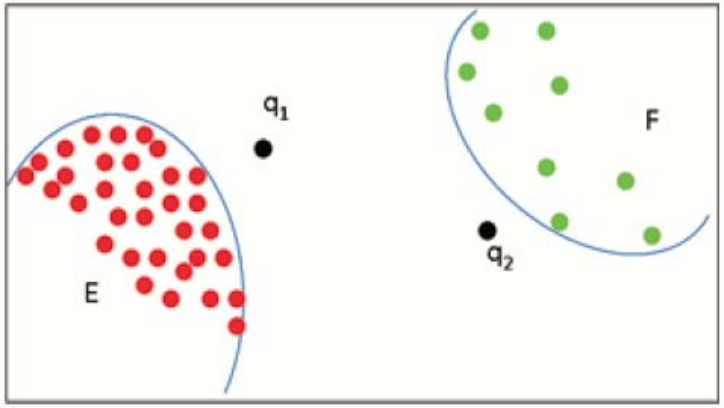
The average cluster coefficient for this network is 0.48, which is significantly larger than 0.001 for random networks

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 → higher density in user's network



Where are We Going?

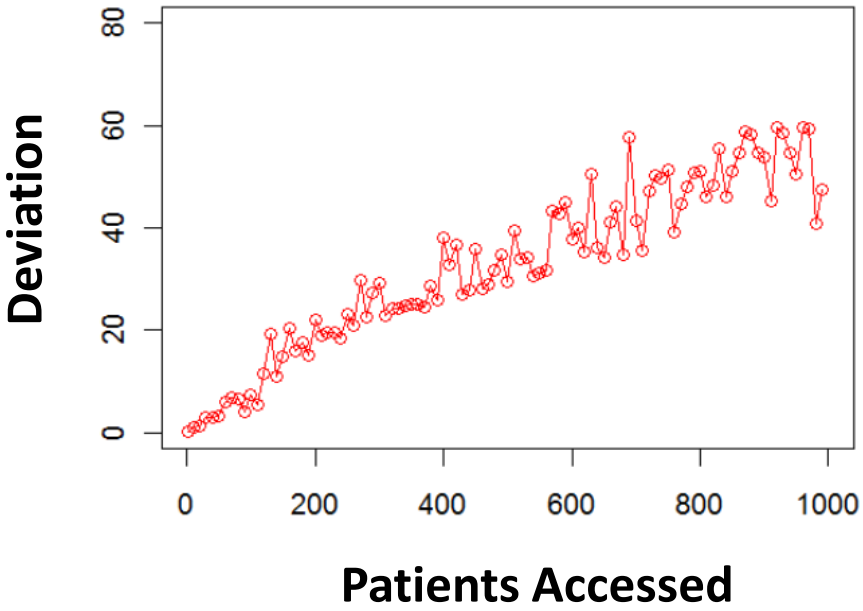
- User Level: Community Anomaly Detection System (CADS) (ACM CODASPY'11)
  - Framework of CADS
  - An Example of CADS
  - **Experimental Evaluation**
  - Limitation

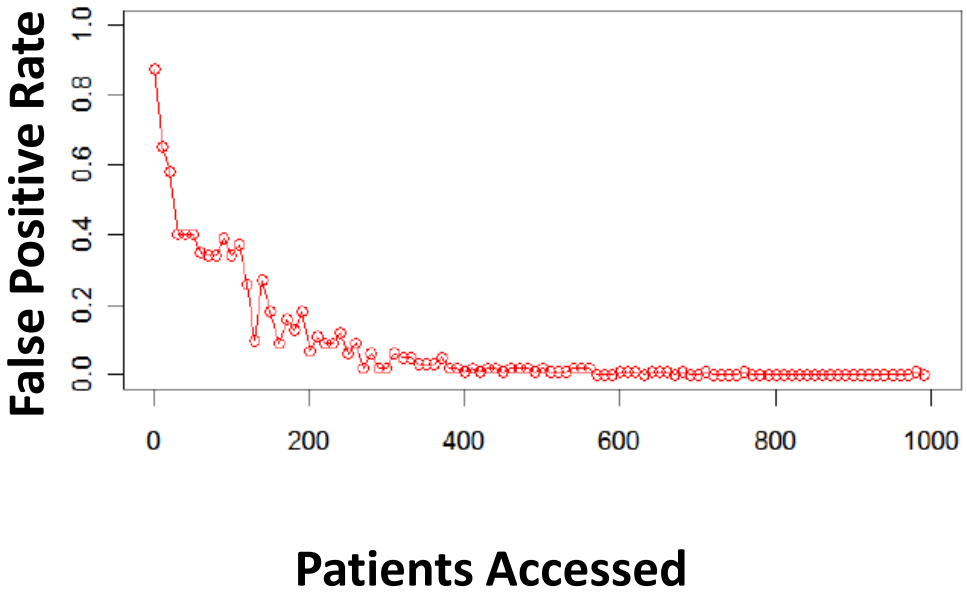
- Access Level: Specialized Network Anomaly Detection (SNAD) (IEEE ISI'11)

Experimental Design

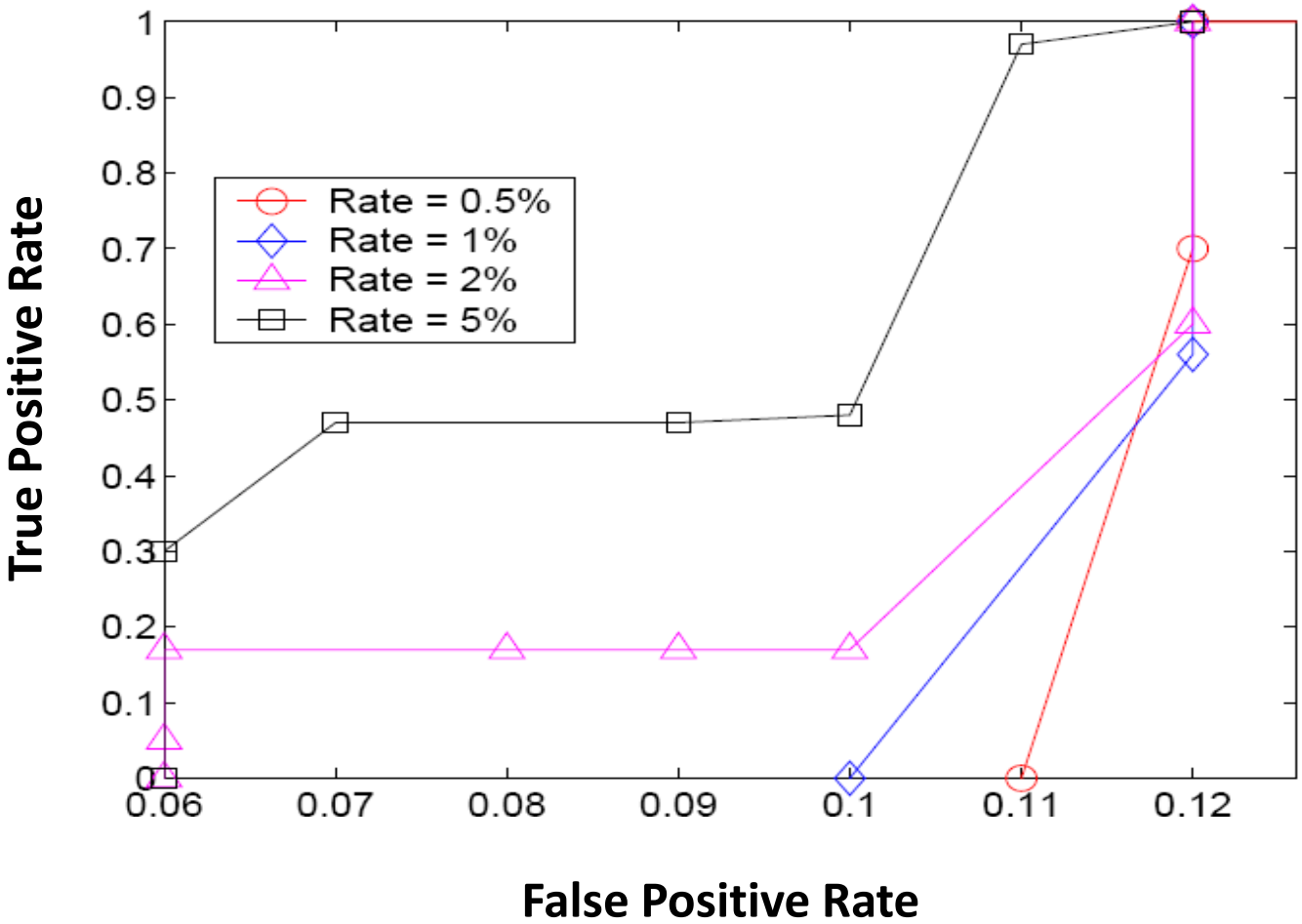
- Datasets are not annotated for illicit behavior
- We simulated users in several settings to test:
  - Sensitivity to number of records accessed
    - Range from 1 to 1,000
  - Sensitivity to number of anomalous users
    - simulated users correspond to 0.5% to 5% of total users
    - Number of records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

Deviation and Detection Rate Increases with Number of Subjects Accessed

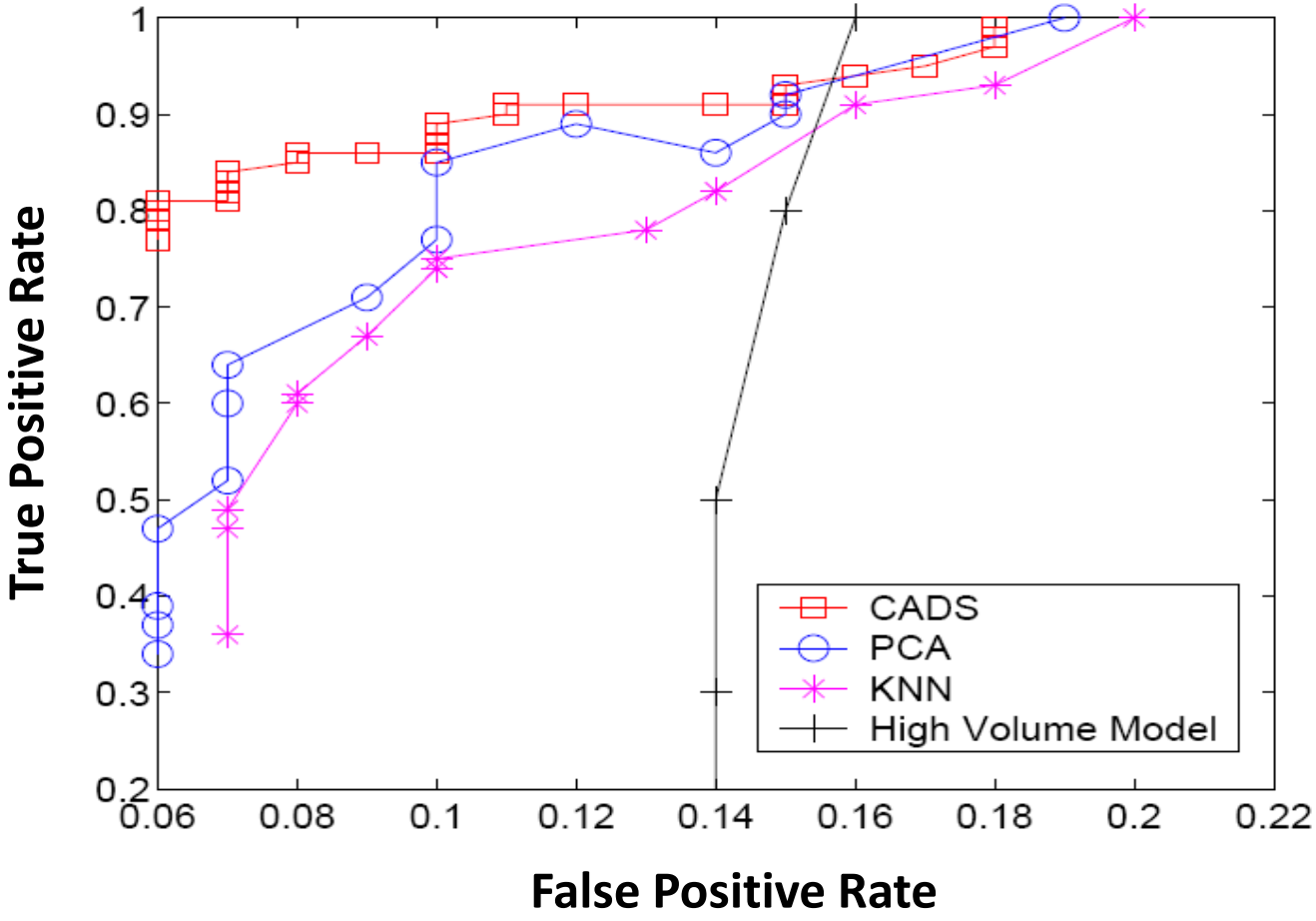




Detection Rate With Various Mix Rates of Real and Simulated Users



CADS Outperforms Competitors (mix rate = 0.5%)



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

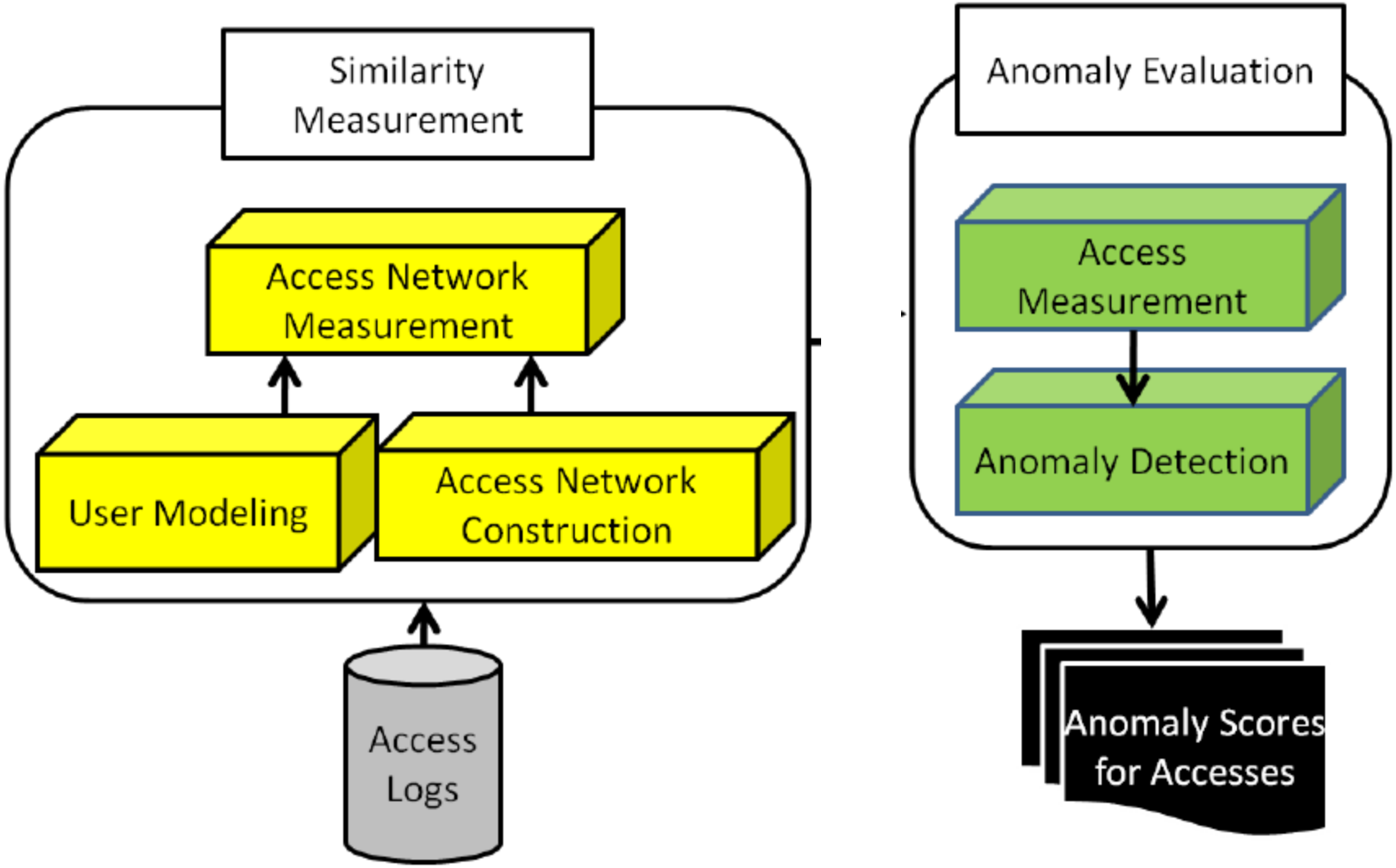
- Simulated users are indicative of misuse of the system…
  …but actual illicit behavior may be more directed.
- "False positives" are not necessarily false!
  (Adjudication by EHR privacy experts under way)
- Need to specialize tool to account for semantics of users and subjects
  - User: {Role, Department, 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

SNAD Framework

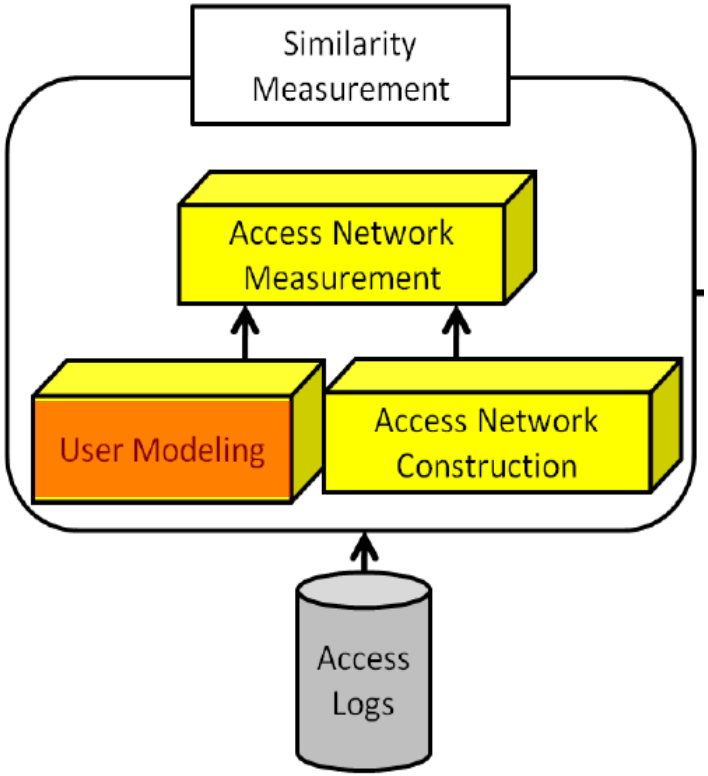


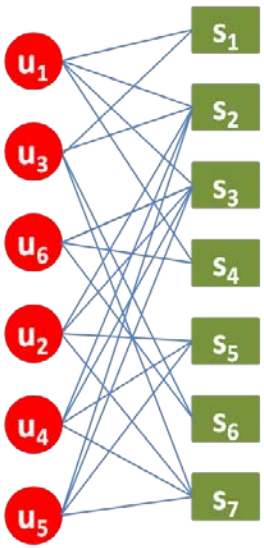
Where are We Going?

- 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



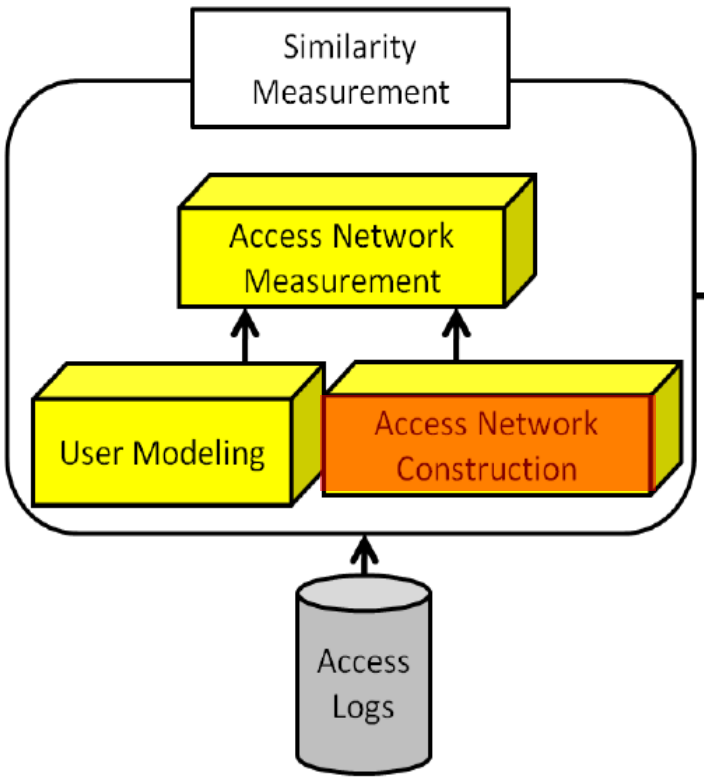


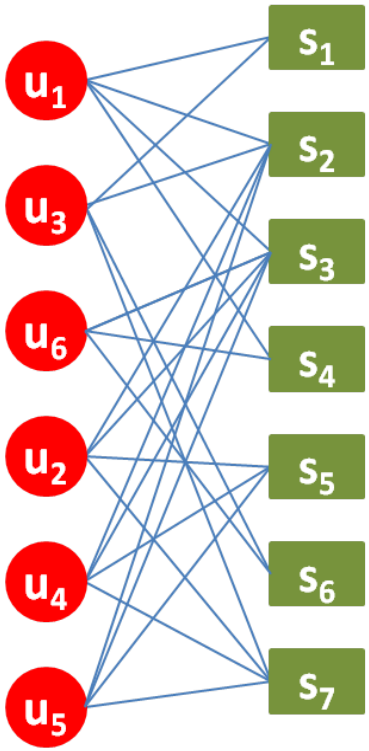
	u_1	u_2	u_3	u_4	u_5	u_6
s_1	1	0	1	0	0	0
s_2	1	1	1	1	1	0
s_3	1	1	0	1	1	1
s_4	1	0	0	0	0	1
s_5	0	1	0	1	1	0
s_6	0	0	1	0	0	1
s_7	0	1	1	1	1	0

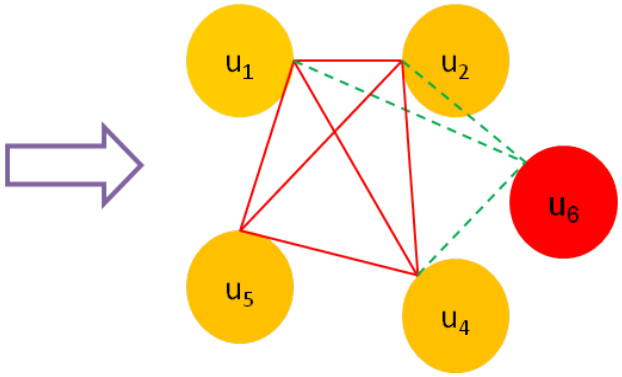
	u_1	u_2	u_3	u_4	u_5	u_6
s_1	0.15	0	0.15	0	0	0
s_2	0.15	0.15	0.15	0.15	0.15	0
s_3	0.15	0.15	0.00	0.15	0.15	0.24
s_4	0.15	0	0	0	0	0.24
s_5	0	0.15	0	0.15	0.15	0
s_6	0	0	0.15	0	0	0.24
s_7	0	0.15	0.15	0.15	0.15	0

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

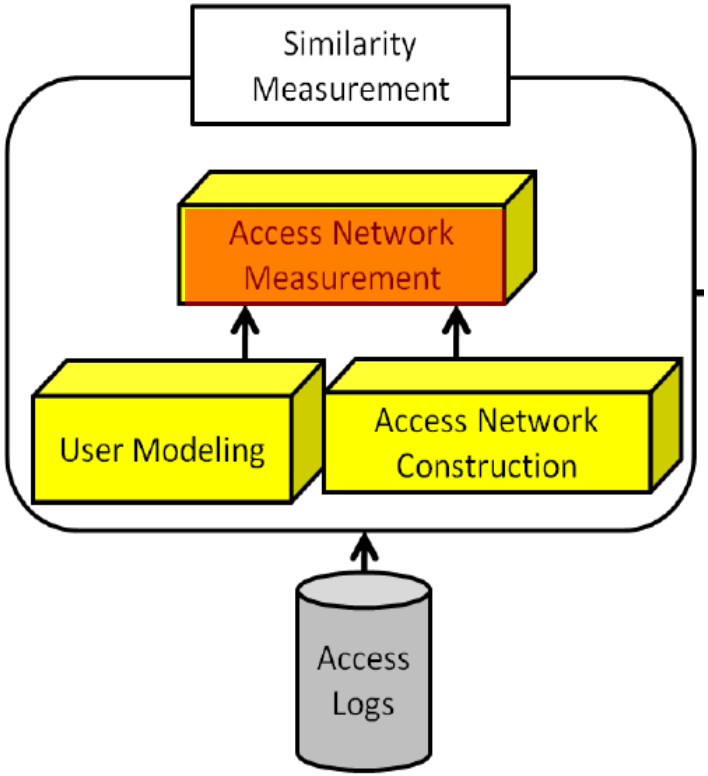
Access Network Construction







Access Network Measurement

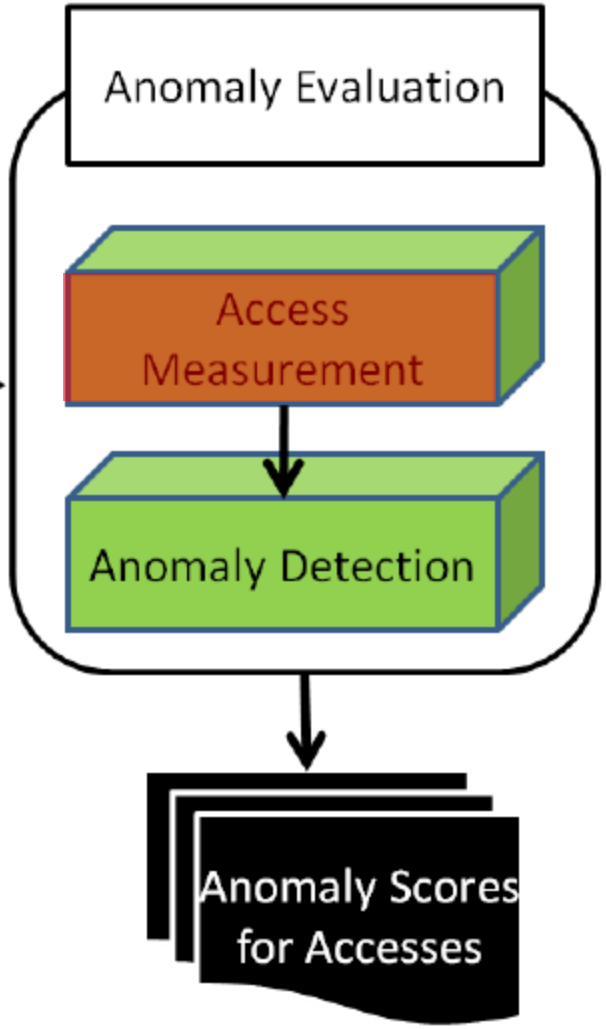


	u_1	u_2	u_3	u_4	u_5	u_6
s_1	0.15	0	0.15	0	0	0
s_2	0.15	0.15	0.15	0.15	0.15	0
s_3	0.15	0.15	0.00	0.15	0.15	0.24
s_4	0.15	0	0	0	0	0.24
s_5	0	0.15	0	0.15	0.15	0
s_6	0	0	0.15	0	0	0.24
s_7	0	0.15	0.15	0.15	0.15	0

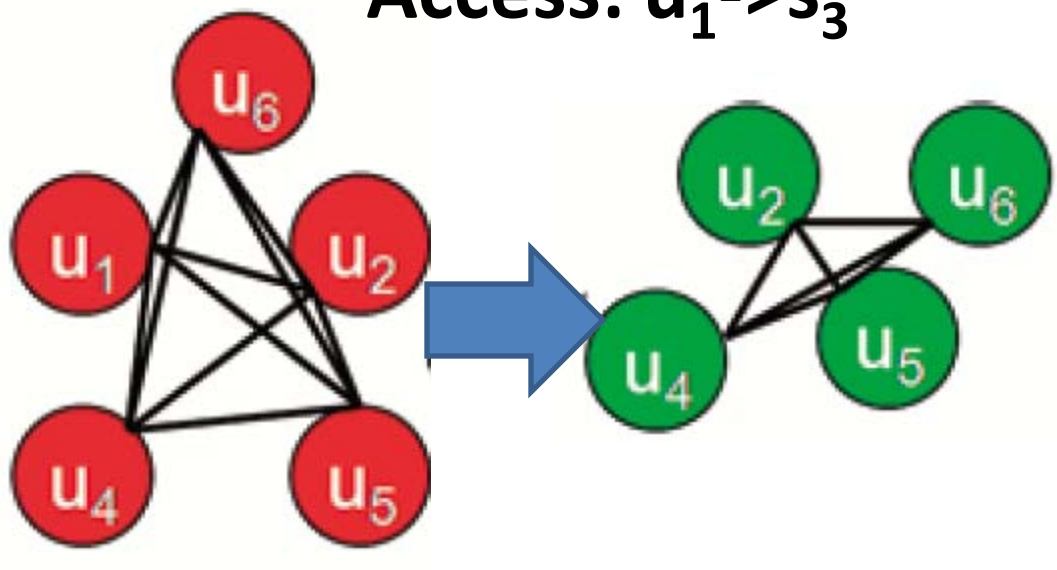
	u_1	u_2	u_4	u_5	u_6
u_1	1.00				
u_2	0.50	1.00			
u_4	0.50	1.00	1.00		
u_5	0.50	1.00	1.00	1.00	
u_6	0.58	0.29	0.29	0.29	1.00

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

Measuring Accesses for Changes in 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

Where are We Going?

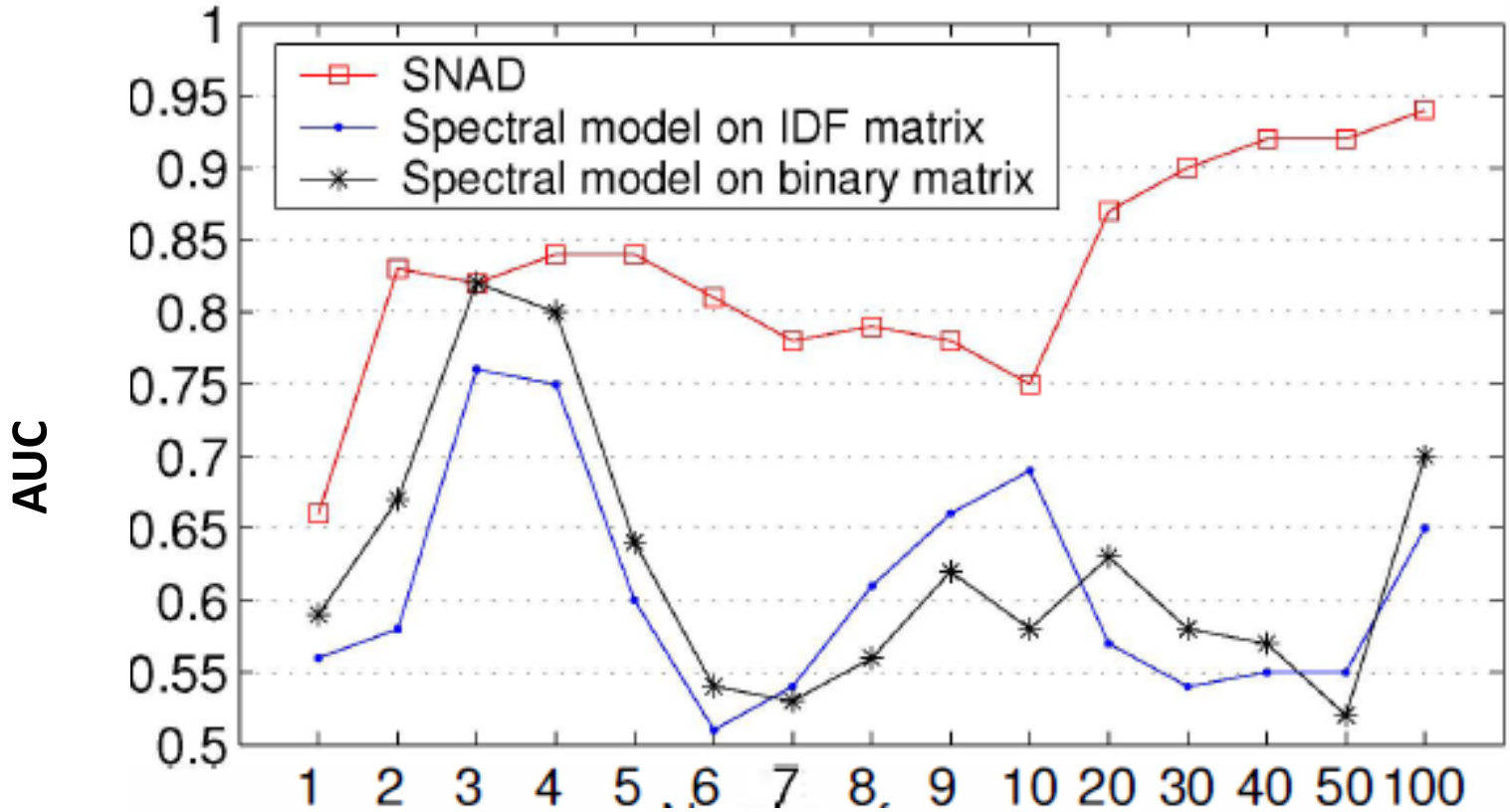
- User Level: Community Anomaly Detection System (CADS) (ACM CODASPY'11)

- Access Level: Specialized Network Anomaly Detection (SNAD) (IEEE ISI'11)
  - Framework of SNAD
  - An Example of SNAD
  - **Experimental Evaluation**
  - Limitation

Experimental Design

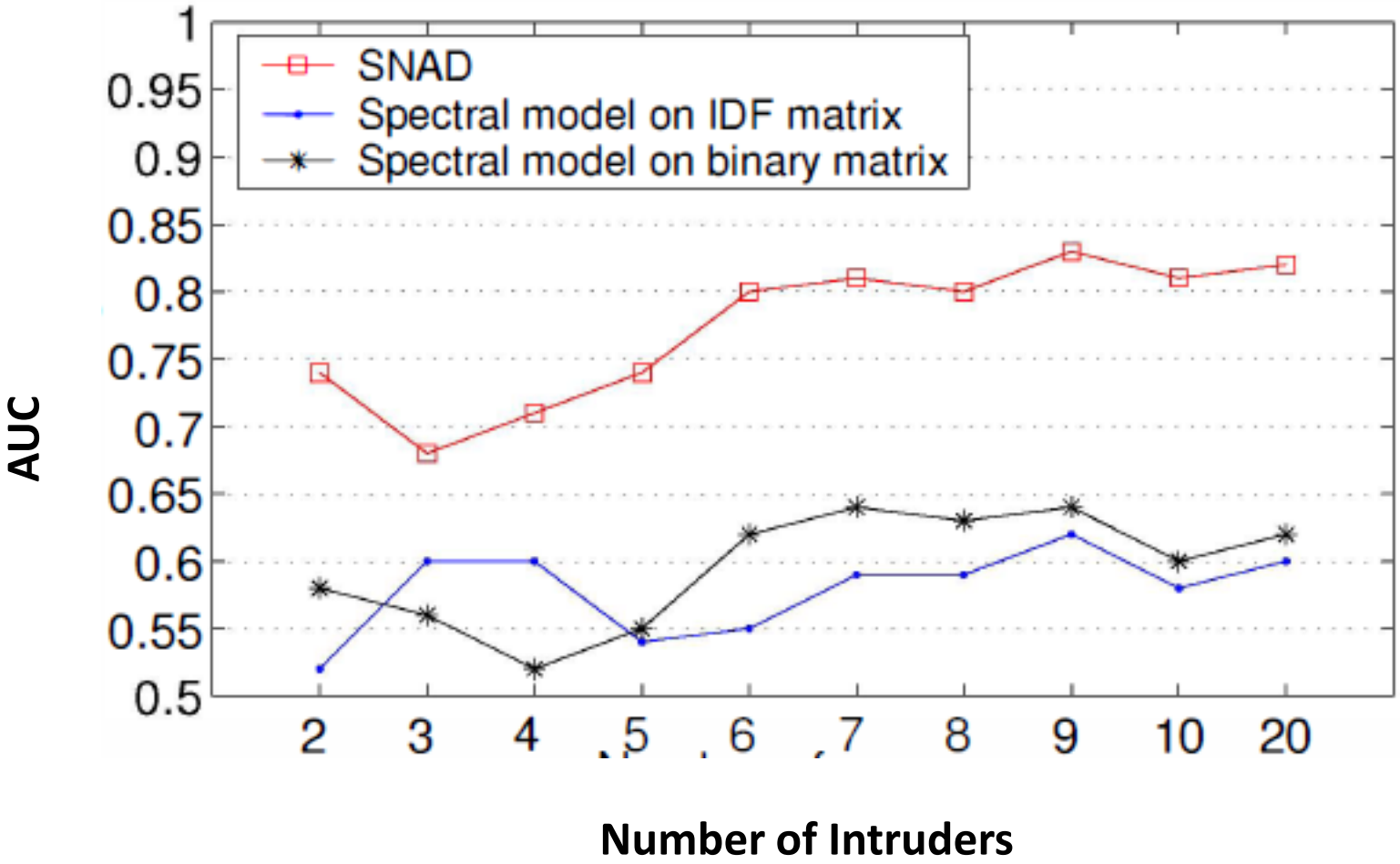
- Datasets are not annotated for illicit behavior
- We simulated users in several settings to test:
  - Sensitivity to number of subjects accessed
    - Range from 1 to 1,00
  - Sensitivity to number of anomalous users
    - Range from 2 to 20
    - Number of subjects accessed fixed to 5
  - Sensitivity to diversity
    - Random number of users and subjects accessed

SNAD: Deviation Rate Increase with Number of Subjects Accessed

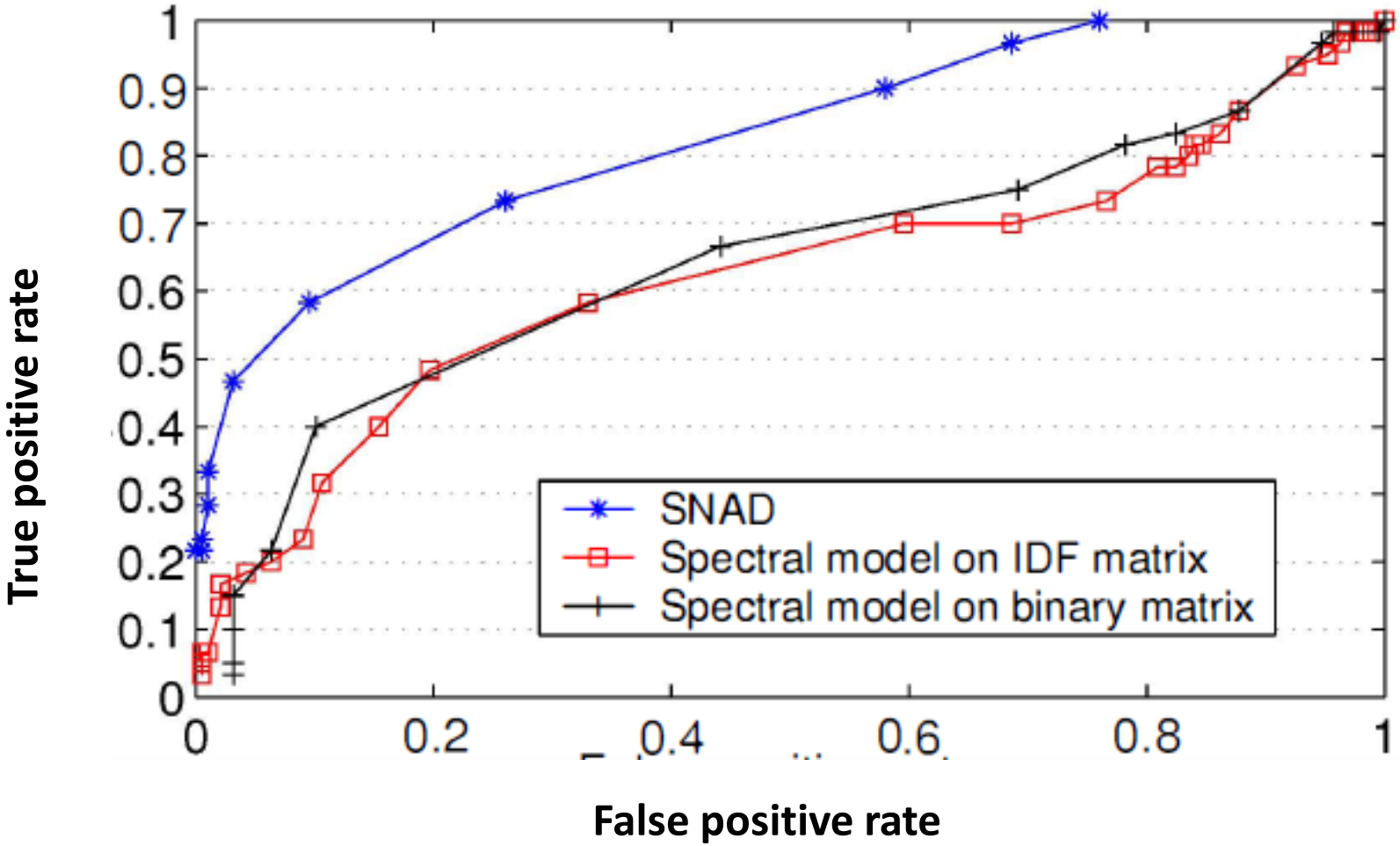


Number of Subjects the Intruder Accesses

SNAD: Deviation Rate Increases with Number of Intruders



SNAD Outperforms Competitors When the Number of Intruders & Accessed Subjects is Random



Where are We Going?

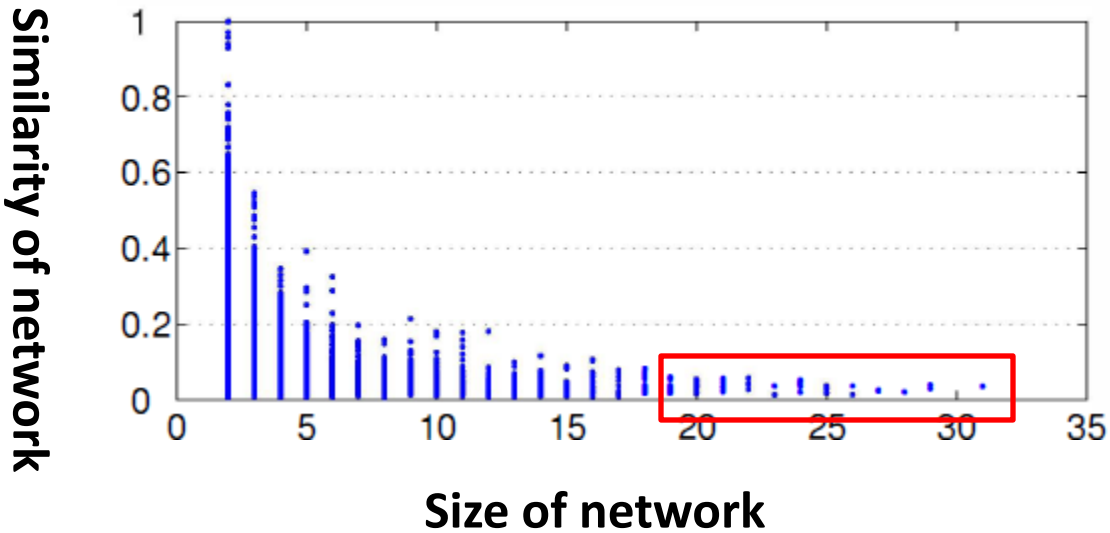
- User Level: Community Anomaly Detection System (CADS) (ACM CODASPY'11)

- 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

- Y. Chen and B. Malin. Detection of anomalous insiders in collaborative environments via relational analysis of access logs. In *Proceedings of the ACM Conference on Data and Application Security Security and Privacy, pages 63–74, 2011. (**CADS**)*
- Y. Chen, S. Nyemba, W. Zhang, and B. Malin. Leveraging social networks to detect anomalous insider actions in collaborative environments. In *Proceedings of the 9th IEEE Intelligence and Security Informatics, pages 119–124, 2011.(**SNAD**)*
- Gallagher R, Sengupta S, Hripcsak G, Barrows R, Clayton P. An audit server for monitoring usage of clinical information systems. Proceedings of the AMIA Symposium. 1998: 1002.
- A. A. Boxwala, J. Kim, J. M. Grillo, and L. O. Machado. Using statistical and machine learning to help institutions detect suspicious access to electronic health records. *Journal of the American Medical Informatics Association, 18:498–505, 2011.*
- Y. Liao and V. R. Vemuri. Use of k-nearest neighbor classifier for intrusion detection. *Computer Security*. 2002; 21(5): 439-448.
- R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM, 51(3):497–515, 2004.*
- Fabbri D, LeFevre K: Explanation-based auditing. In *Proceedings of 38th International Conference on Very*
- *Large Data Bases 2012:to appear.*
- M. Shyu, S. Chen, K. Sarinnapakorn, and L. Chang. A novel anomaly detection scheme based on principal component classifier. In IEEE Foundations and New Directions of Data Mining Workshop. 2003: 172-179.

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

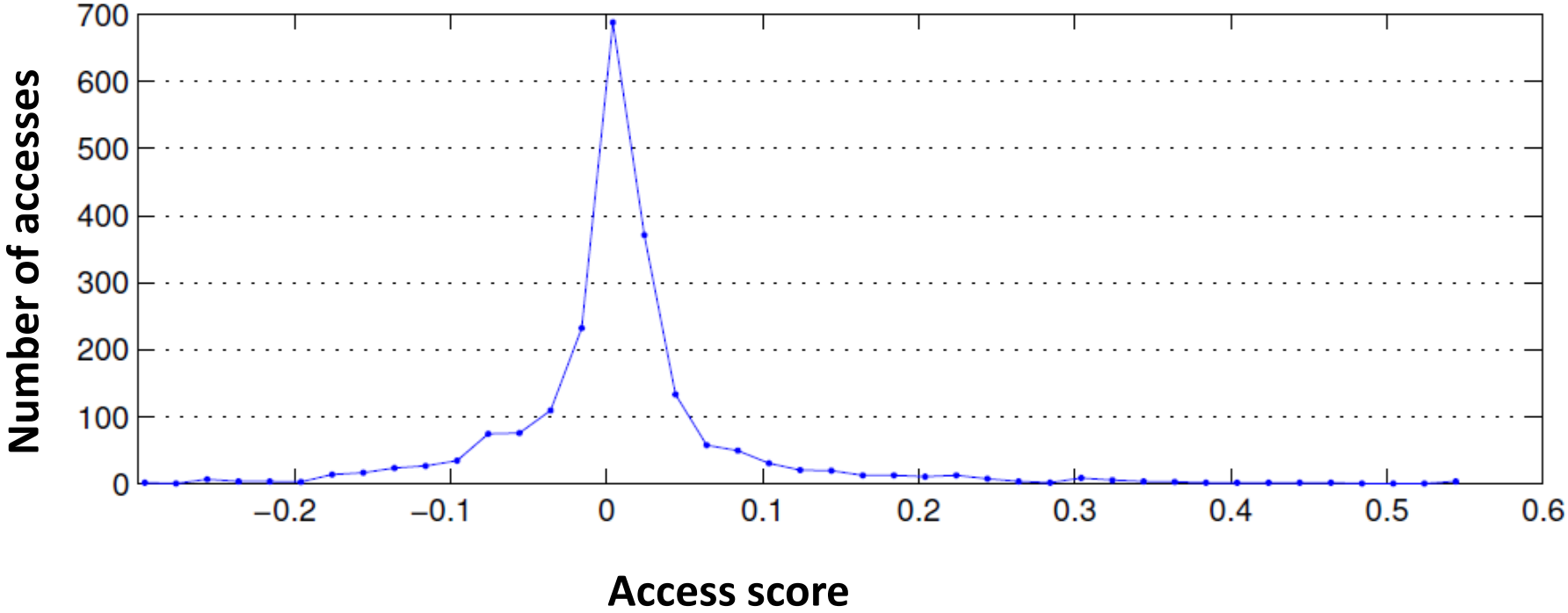
- National Science Foundation
  - CCF-0424422 (TRUST)
  - CNS-0964063
- National Institutes of Health
  - R01LM010207

Questions? Comments?

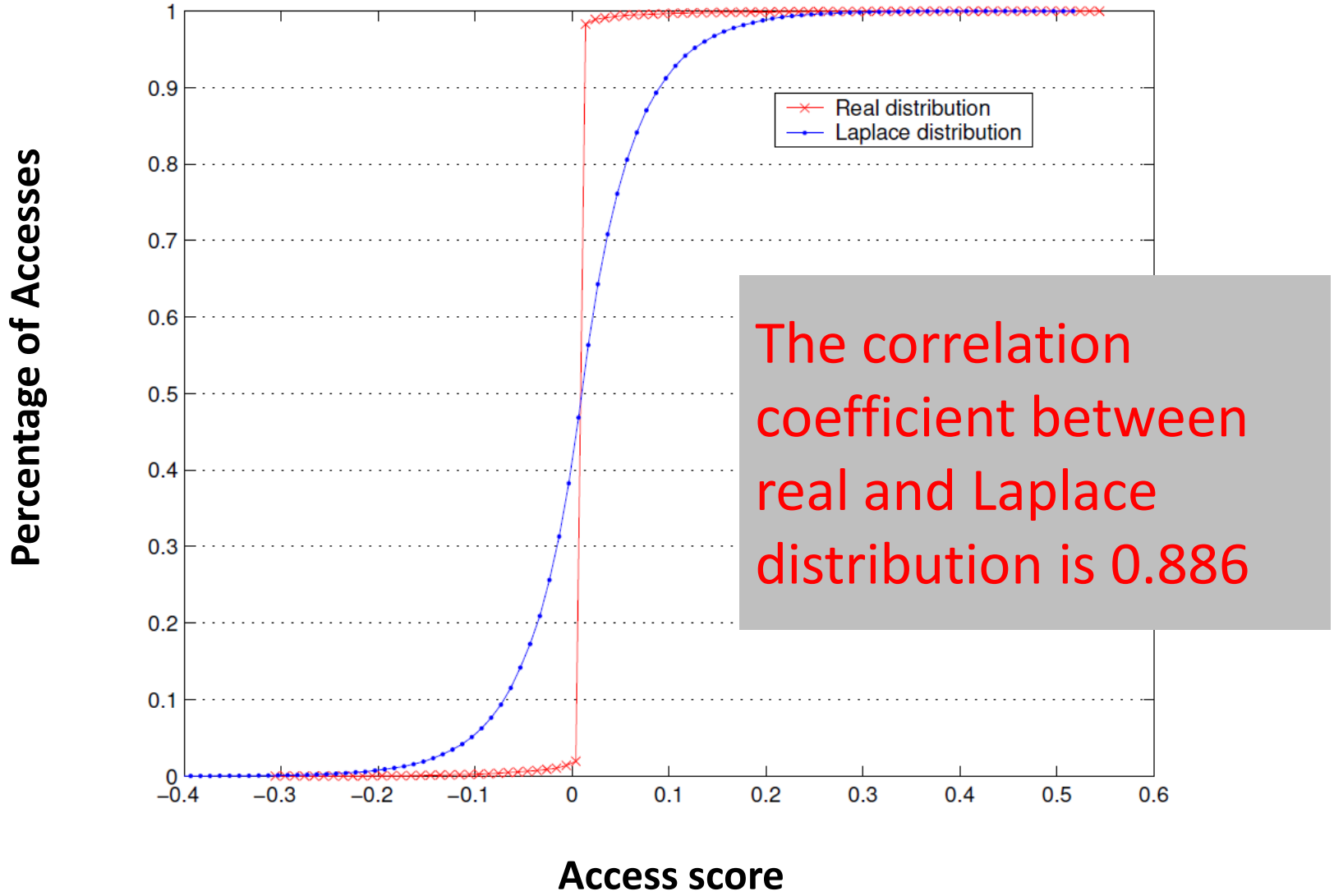
you.chen@vanderbilt.edu

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