

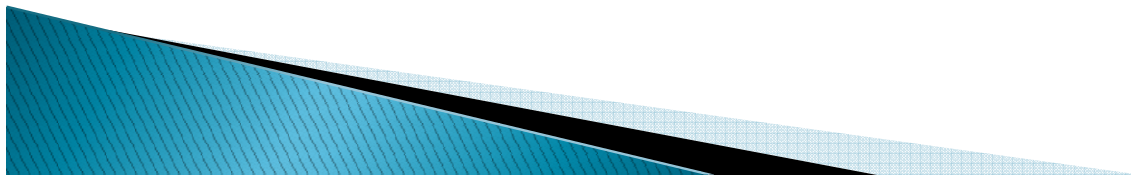
# Risk Assessment in Social Networks Based on Anomalous Behavior Detection

Advisors: Prof. Elena Ferrari, Prof. Barbara Carminati,  
Naeimeh Laleh  
Insubria University, Varese Como, Italy



# Problem statement

- Social networks might represent unsafe place as users interact with never-met person, which could potentially be risky users.
- Users cannot avoid serious consequences of interacting with risky users by just properly setting their privacy settings.
- We propose a model for risk assessment based on anomalous behavior detection in online social network.

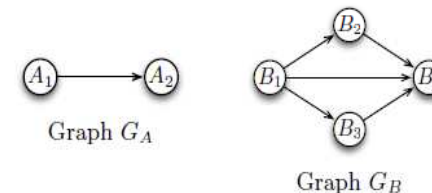


# Related Work

- ▶ Trust evaluation models

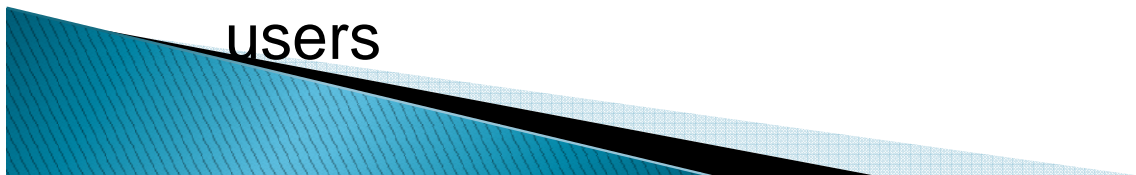
- Network-based trust models

- A trust network is a graph where nodes represent agents and edges represent trust relations



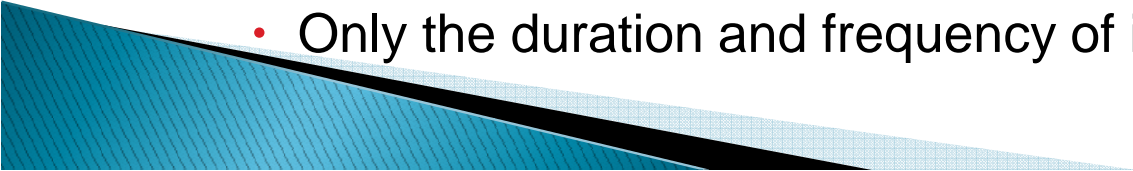
- ▶ Interaction-based trust models

- They consider user actions and interactions
  - Evaluate trust based on communication behavior of users



# Related Work

## ▶ Hybrid trust models

- **Explicit social trust:** based on consciously established social ties. Each time two users interact, they exchange their friend lists and save them as friendship graphs.
  - **Implicit social trust:** based on frequency and duration of contact between two users
    - **Familiarity:** the length of the interactions/ contacts between the two nodes
    - **Similarity of the nodes:** the degree of coincidence of the two nodes' familiar circles
  - Only the duration and frequency of interactions
- 

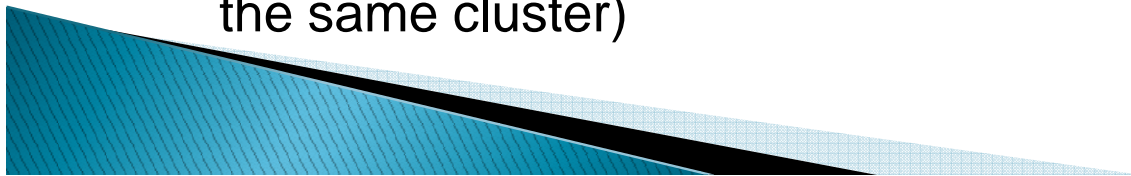
# Overall Approach

- ▶ Our risk model is hybrid based on interactions and network structure
- ▶ Unsupervised anomaly detection techniques
  - Based on the assumption that anomalies are very rare compared to normal users
- ▶ Why not supervised and semi-supervised?
  - We cannot define a normal behavior for the whole universe of users in the social network, but we can group similar users (1st clustering phase) and then study the behavior to detect anomalous one (2st clustering phase). We compute the distribution of behavior of each user across all other users.




# Anomaly Detection Based on Clustering

- ▶ Our model is based on clustering
  - ***Key assumption.*** normal users belong to large and dense clusters, while anomalies do not belong to any of the clusters or form very small clusters
- ▶ Advantages:
  - No need to be supervised
  - Easily adaptable to on-line mode suitable for anomaly detection
- ▶ Detected anomalies detected are:
  - Users that do not fit into any cluster
  - Small clusters
  - Low density clusters or local anomalies (far from other users within the same cluster)

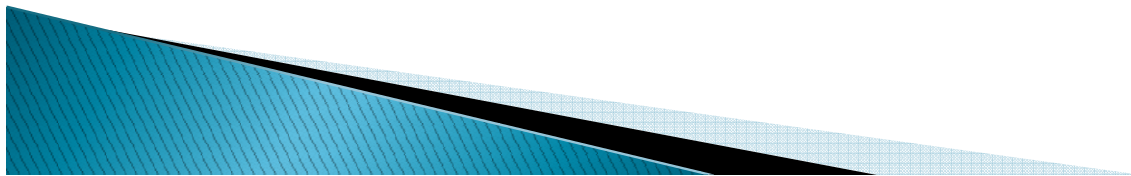


# Our Hybrid Model (Interactions of user)

- ▶ Sent Information (Out-degree Activity)
    - Average of four features
      - Number of likes
      - Number of comments
      - Number of Share
      - Number of post
  - ▶ Received information (In-degree Activity)
    - Average number of likes received on comments
    - Average\_No\_likes\_on\_PostItems
    - Average\_No\_Comments\_on\_PostItems
  - ▶ The popularity of sent and received information
- 

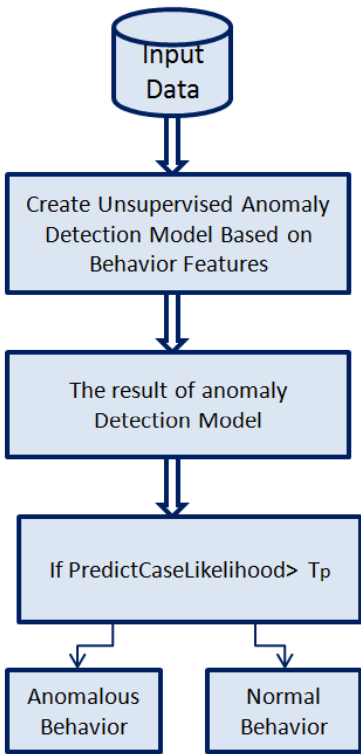
## Our Hybrid Model (Network structure)

- ▶ Consider the number of neighbors (the position of user in the network)
- ▶ Average number of mutual friends with neighbors
- ▶ If the number of neighbors is high and average number of mutual friends is low, the risk of user increased

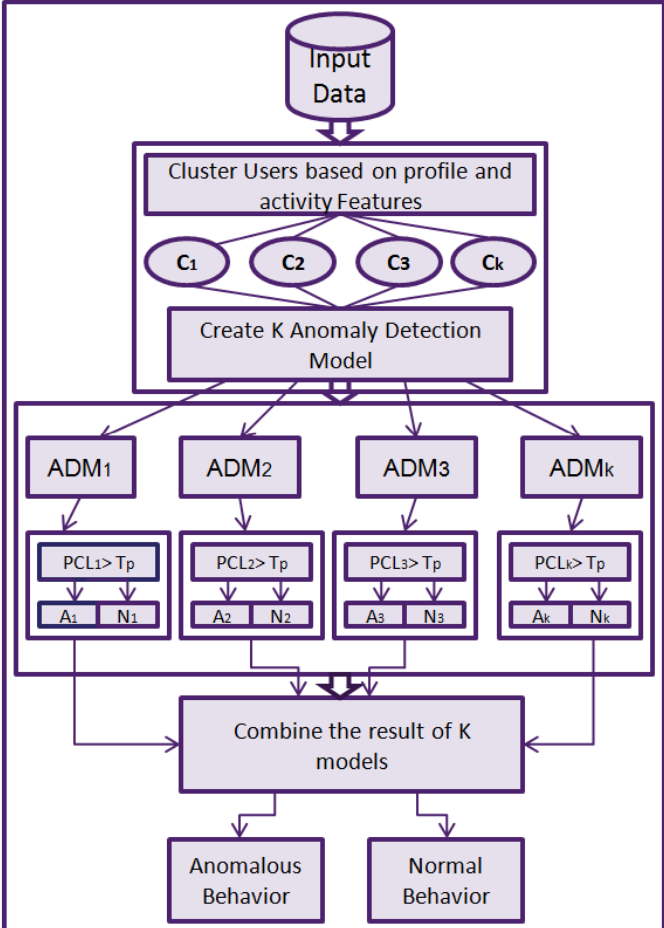




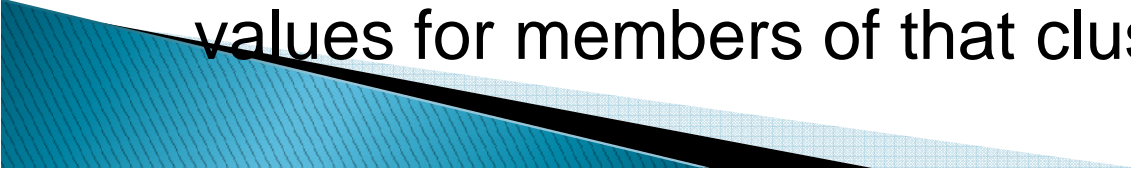
### Simple unsupervised Anomaly Detection Model



### Two phase Unsupervised Anomaly Detection Model

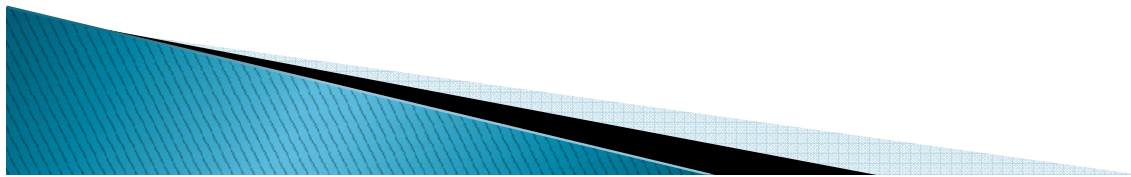


# Probability based clustering

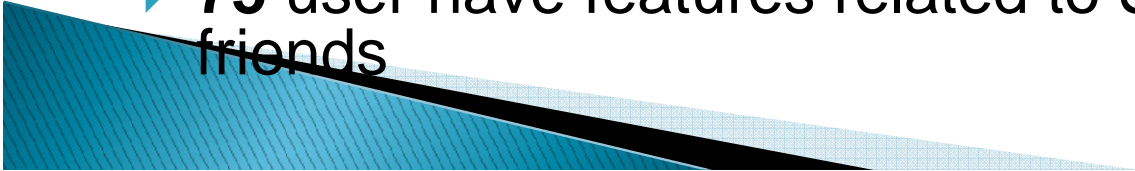
- ▶ In our anomaly detection model in both phases, we use probability based clustering.
  - ▶ Probabilistic cluster is a distribution over the data that each cluster has a different distribution and each user would have a certain set of feature values if it were known to be a member of that cluster.
  - ▶ There is a set of  $k$  probability distributions, representing  $k$  clusters, that govern the feature values for members of that cluster.
- 

# Use Multivariate Gaussian Distribution

- ▶ Applying Multivariate Gaussian Distribution to our Anomaly Detection model
- ▶ The normal Gaussian model is a special case of multivariate Gaussian distribution
- ▶ We use EM algorithm to construct a maximum likelihood to estimate parameters such Mean and standard deviation.

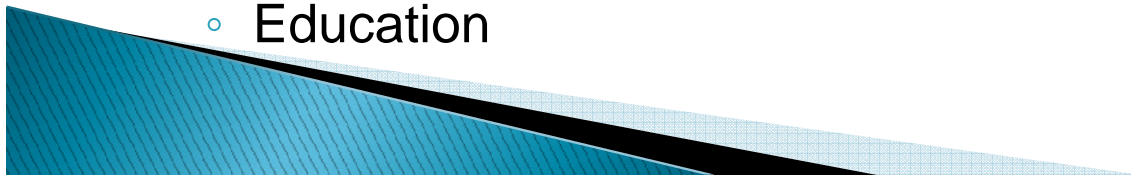


# Input Data, Facebook Users

- ▶ Around **569,829** user have gender
    - **17572** have more than 20 % profile information
    - Around **7000** have more than 75 % profile information
      - Gender,
      - Education,
      - Hometown city and country,
      - Current Location city and country
  - ▶ **19,000,000** user have features related to likes and comments
  - ▶ **13000** user have features related to friendship connection
  - ▶ **75** user have features related to Share, Post, number of friends
- 

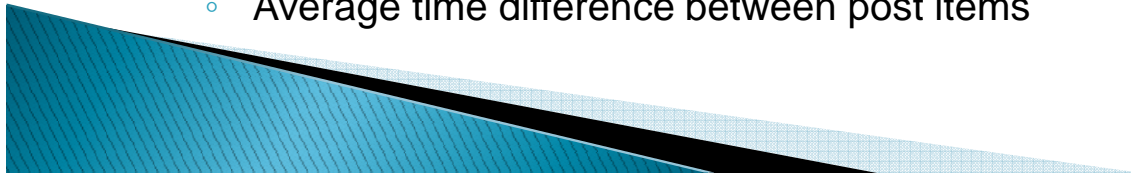
# Users Interactions and Profile Information

- ▶ Grouping features:
  - Number of friends
  - Received information (In-come links)
    - Average number of likes received on comments
    - Average\_No\_likes\_on\_PostItems
    - Average\_No\_Comments\_on\_PostItems
  - Sent Information (Out-come links)
    - Average of four features
      - Number of likes
      - Number of comments
      - Number of Share
      - Number of post
  - Gender
  - Education



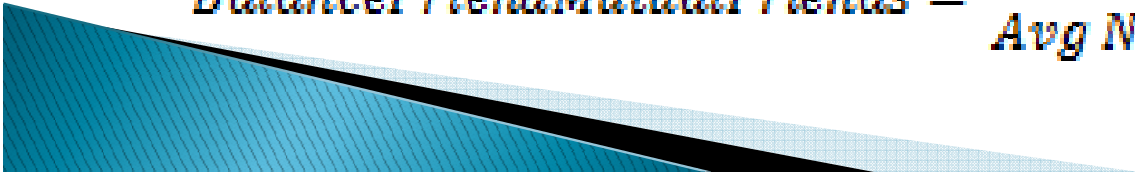
# Behavior Features

- Number of neighbors
- Average number of mutual friends with all his friends
- How many percent of profile is public
- Number of likes
- Average popularity of likes
- Number of comments
- Number of likes on comments
- Number of shared Items
- Number of posts
- Average number of likes on Post Items
- Average number of comments on Post Items
- Average popularity of Shared items
- Average popularity of post Items
- Average time difference between comments
- Average time difference between shared items
- Average time difference between post items



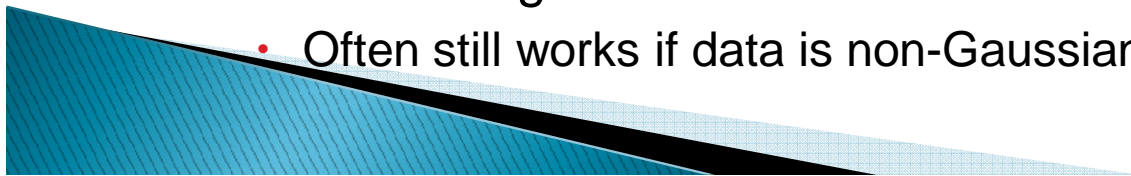
## Features Related to Anomaly Model

- ▶ To detect special type of anomaly, we need to extract new features to help you anomaly detection capture this anomalous behavior.
- ▶ Assume there is an anomalous behavior that we want to capture it (User has a lot of friends, but the average number of his mutual friends is low).
  - We extract a new feature and this would be a feature that would help us in anomaly detection model to capture this sort of anomaly.

$$\text{BalanceFriendMutualFriends} = \frac{\text{NO. Friends}}{\text{Avg No Mutual friends}}$$


# Choose Features

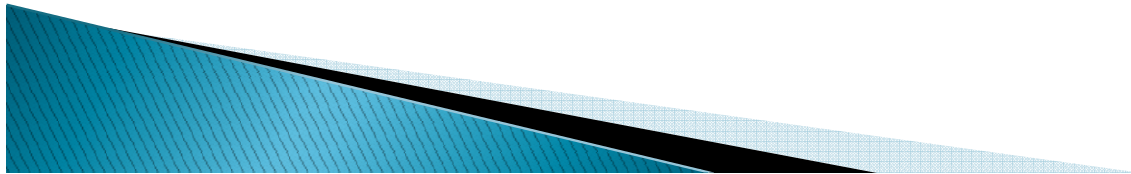
- ▶ We have 5 mix features (key features for anomaly detection):
  - Balance Number Of Friend Mutual Friend
  - Balance In Out
  - Balance Comment Like
  - Balance Share Like Comment
  - Balance Like Popularity
- ▶ Non-Gaussian features
  - Plot a histogram of data to check it has a Gaussian description
    - Often still works if data is non-Gaussian





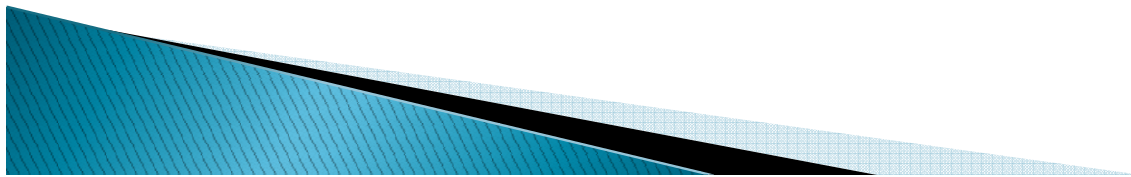
# Two Phase Model in Two Cases

- ▶ **1- Four Grouping Features**
  - Gender
  - Education
  - Number of friends
  - How many percent of profile is public
- ▶ **2- Six Grouping Features**
  - Gender
  - Education
  - Number of friends
  - How many percent of profile is public
  - Average Income Activity
  - Average Outgo Activity

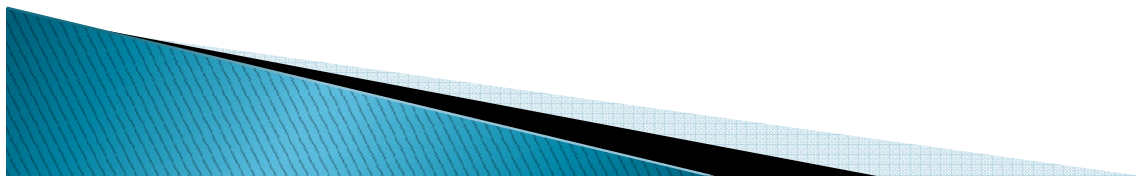
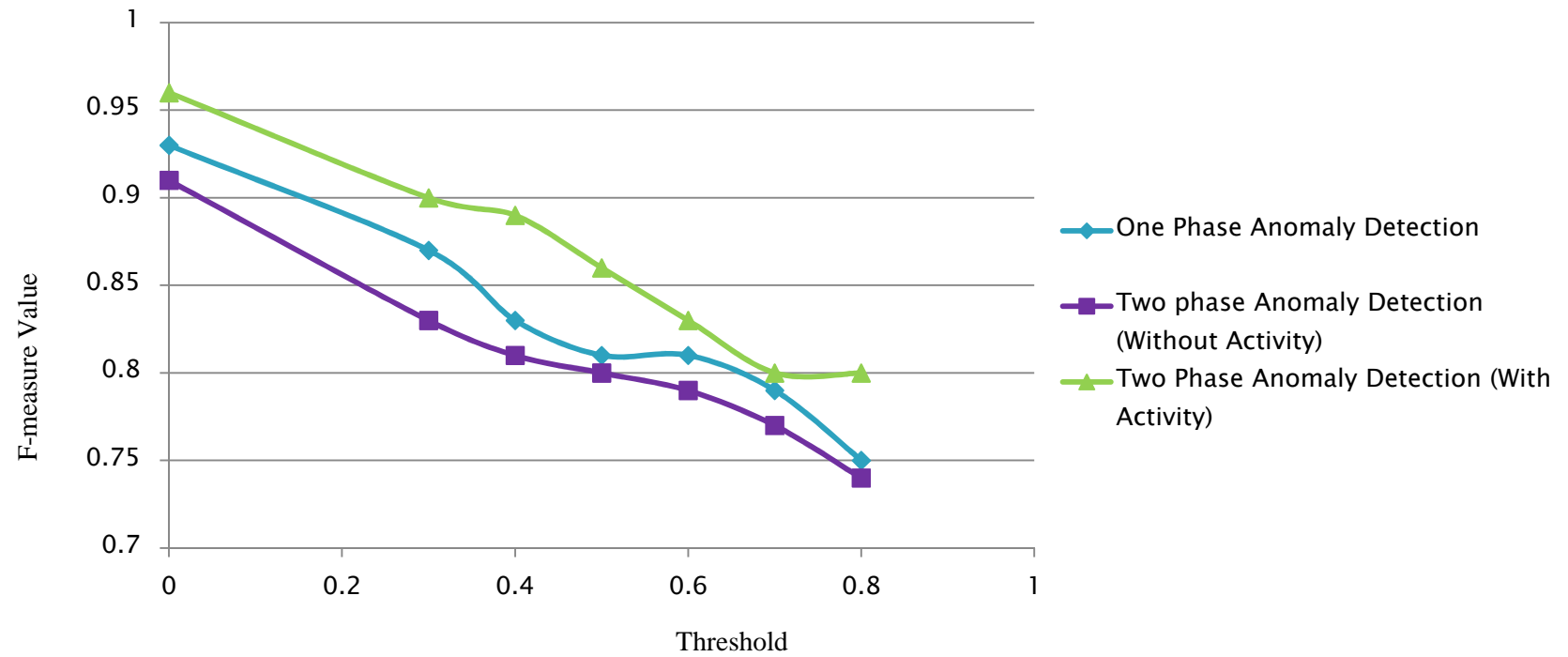


# Evaluation of Anomaly Detection model

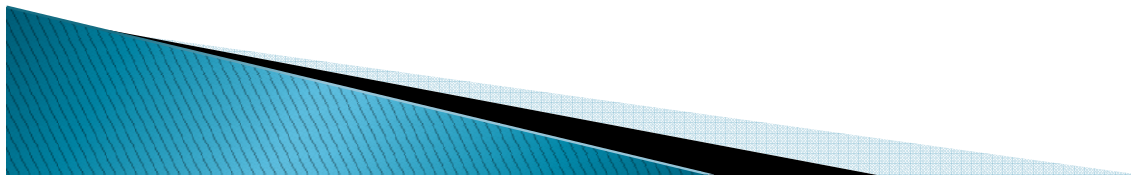
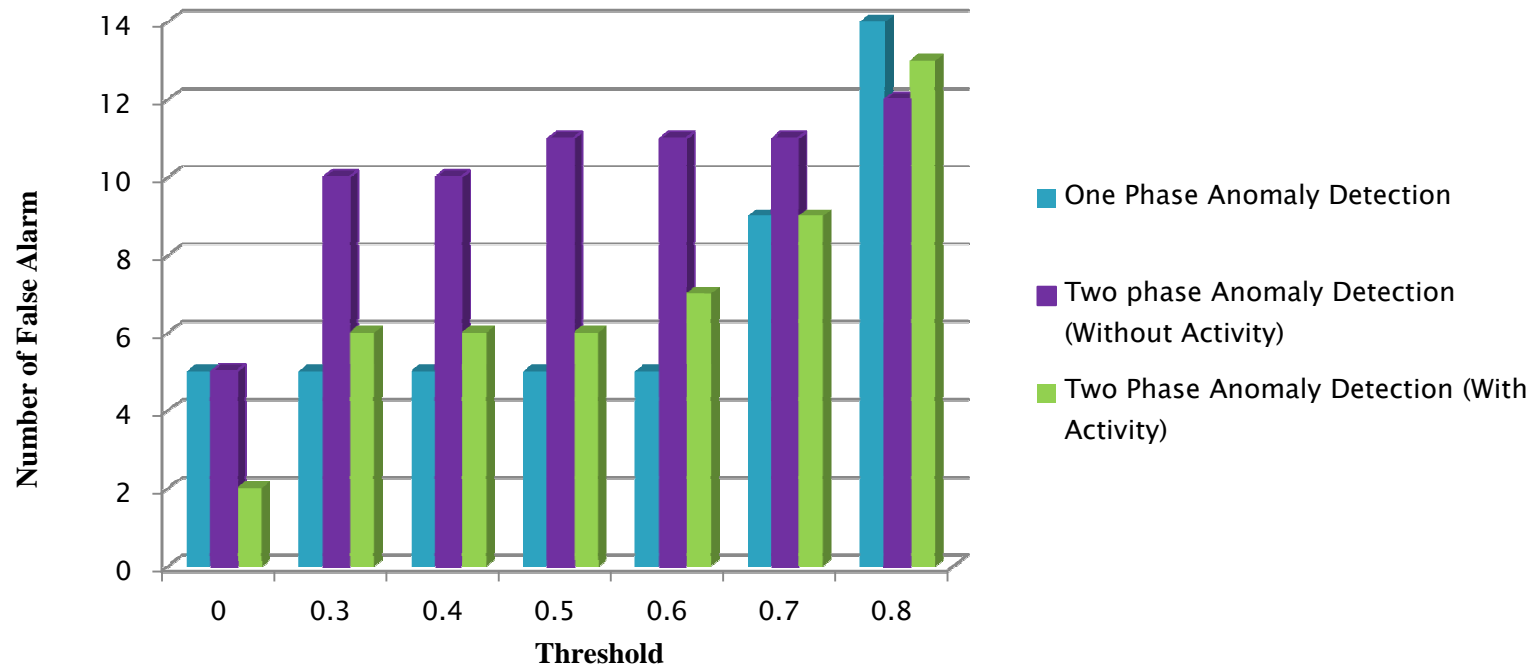
- ▶  $\text{Recall}(R) = \text{TP} / (\text{TP} + \text{FN})$
- ▶  $\text{Precision}(P) = \text{TP} / (\text{TP} + \text{FP})$
- ▶  $\text{F-measure} = 2 * R * P / (R + P)$
- ▶ Standard measures for evaluating anomaly detection problems:
  - Recall (Detection rate) - ratio between the number of correctly detected anomalies and the total number of anomalies
  - False alarm (false positive) rate – ratio between the number of data users from normal class that are misclassified as anomalies




# F-measure Curve



# False Alarm Rate

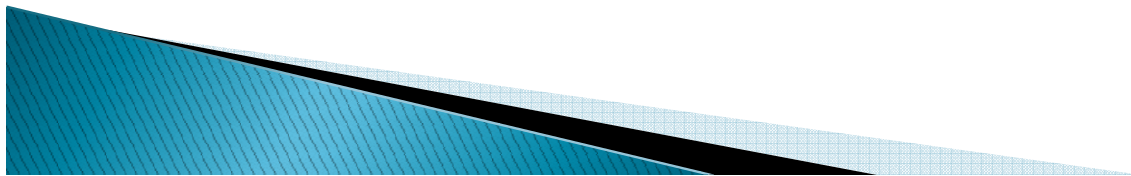


## Future work:

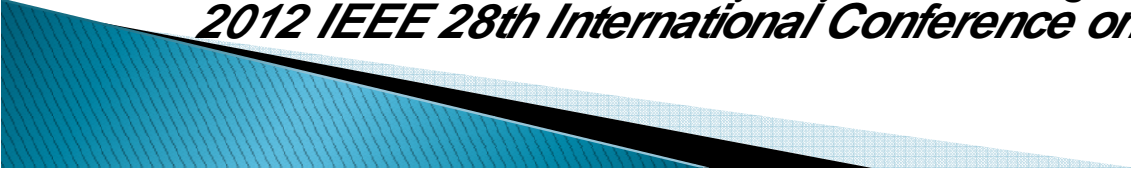
- ▶ Distributed Anomaly Detection
  - ▶ Data may come from many different sources
  - ▶ In a decentralized model, we don't have data in a central site for analysis.
  - ▶ How can we compute a distribution of behavior of each user with respect to others in the network?
  - ▶ Detecting anomalies in such complex systems may require integration of information about detected anomalies from single locations in order to detect anomalies at the global level of a complex system
  - ▶ There is a need for the high performance and distributed algorithms for correlation and integration of anomalies
- 

# References:

- ▶ Sherchan, Wanita, Surya Nepal, and Cecile Paris. "A survey of trust in social networks." *ACM Computing Surveys (CSUR)* 45.4 (2013): 47.
- ▶ Nepal, Surya, Wanita Sherchan, and Cecile Paris. "STrust: a trust model for Social Networks." *Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on*. IEEE, 2011.
- ▶ Adali, Sibel, and Jennifer Golbeck. "Predicting Personality with Social Behavior." *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 2012.
- ▶ Adali, Sibel, Fred Sisenda, and Malik Magdon-Ismael. "Actions speak as loud as words: Predicting relationships from social behavior data." *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012.
- ▶ Li, Ming, and Alessio Bonti. "T-OSN: A Trust Evaluation Model in Online Social Networks." *Embedded and Ubiquitous Computing (EUC), 2011 IFIP 9th International Conference on*. IEEE, 2011.

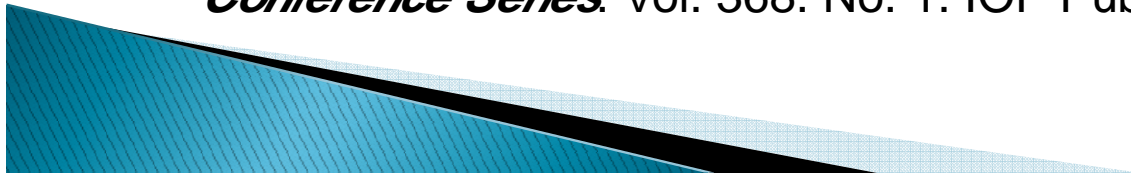


## References:

- ▶ Bouguessa, Mohamed. "Unsupervised Anomaly Detection in Transactional Data." *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*. Vol. 1. IEEE, 2012.
  - ▶ Papadimitriou, Panagiotis, Ali Dasdan, and Hector Garcia-Molina. "Web graph similarity for anomaly detection." *Journal of Internet Services and Applications* 1.1 (2010): 19-30.
  - ▶ DuBois, Thomas, Jennifer Golbeck, and Aravind Srinivasan. "Predicting trust and distrust in social networks." *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. IEEE, 2011.
  - ▶ Akcora, Cuneyt Gurcan, Barbara Carminati, and Elena Ferrari. "User similarities on social networks." *Social Network Analysis and Mining* (2013): 1-21.
  - ▶ Akcora, Cuneyt Gurcan, Barbara Carminati, and Elena Ferrari. "Privacy in social networks: How risky is your social graph?." *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. IEEE, 2012.
- 

# References:

- ▶ Adali, Sibel, et al. "Measuring behavioral trust in social networks." *Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on*. IEEE, 2010.
- ▶ Arnaboldi, Valerio, Andrea Guazzini, and Andrea Passarella. "Egocentric Online Social Networks: Analysis of Key Features and Prediction of Tie Strength in Facebook." *Computer Communications* (2013).
- ▶ Huang, Bert, et al. "A flexible framework for probabilistic models of social trust." *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer Berlin Heidelberg, 2013. 265-273.
- ▶ Smyth, Padhraic. "Probabilistic model-based clustering of multivariate and sequential data." *Proceedings of the Seventh International Workshop on AI and Statistics*. San Francisco, CA: Morgan Kaufman, 1999.
- ▶ Kuusela, Mikael, et al. "Semi-supervised anomaly detection—towards model-independent searches of new physics." *Journal of Physics: Conference Series*. Vol. 368. No. 1. IOP Publishing, 2012.





**Thanks for the consideration**

