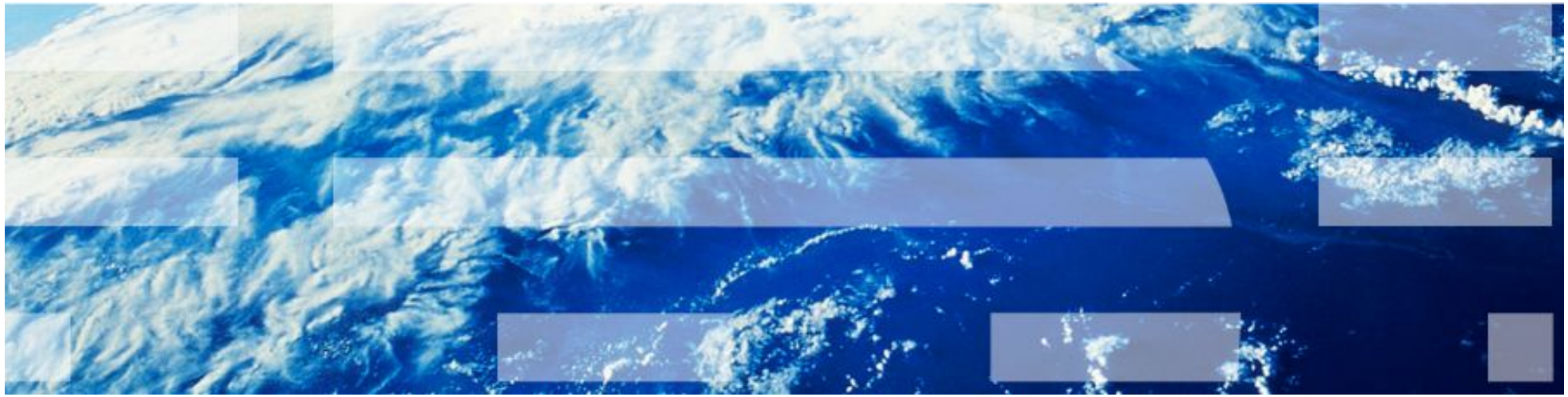
E6893 Big Data Analytics:

# *Community Detection in Social Networks*

**Team Members (with UNI):**
**Siyu Liu(sl4262), Tianyao Hua(th2706), Anke Xu(ax2127)**

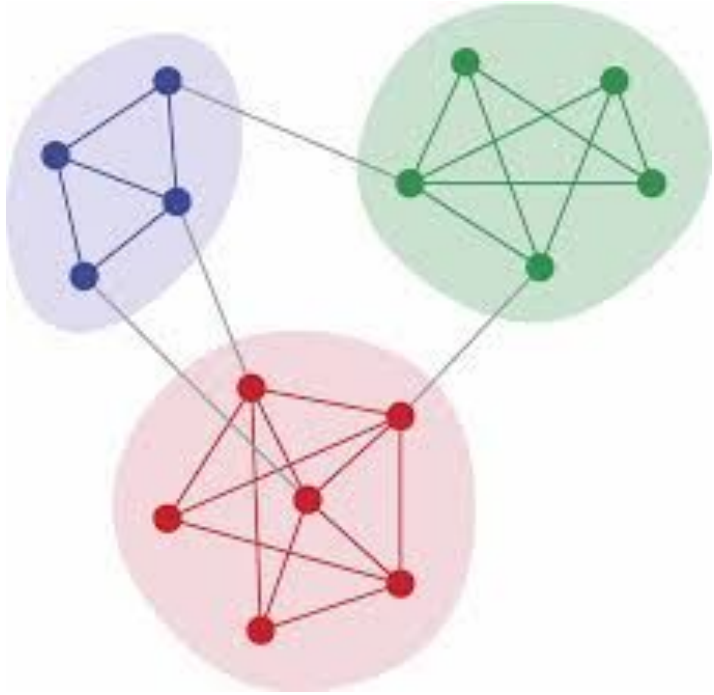


# Motivation

- **What is Community Detection?**
  Reveal the organizational structure of social networks

- **Advantage of Community Detection over Other Clustering**
  - Clustering only considers attributes of nodes
  - CD also considers the structure of edges
    to detect underlying relationship and explainable insights

- **Wide range of application**
  - Targeted advertising
  - Detect suspicious groups in Anti-Money Laundering
  - Healthcare - Infectious Disease Outbreak
  - Business Intelligence



# Dataset, Algorithm, and Tools

- **Dataset**

  Main Source: Stanford Large Network Dataset Collection - Email-Eu-core network
  - All incoming and outgoing email data between members from a large European research institution with 1005 nodes, 25571 edges and 42 labeled communities

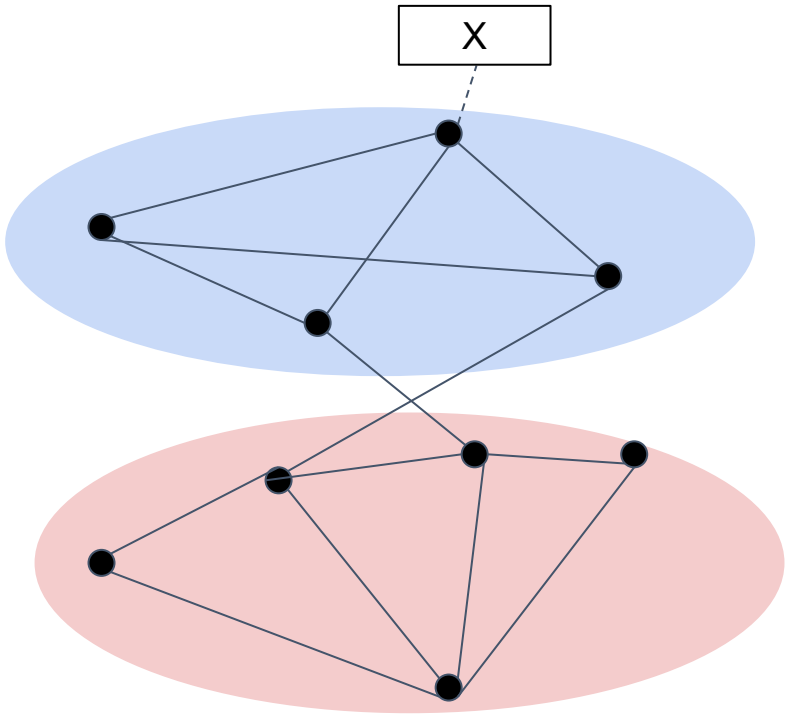
  Other Option: Kaggle Competition - Synthetic Financial Datasets For Fraud Detection
  - A simulated dataset about mobile money transactions based on a sample of real transactions

  Other Option: Facebook, Twitter social network data sets (Still finding)

- **Tools**
  - Python
  - PySpark
  - Hadoop
  - Hive
- **Algorithm**
  - Community from Edge Structure and Node Attributes - CESNA (See Next)

# Dataset, Algorithm, and Tools (cont.)

- **Algorithm-CESNA**




Probabilistic Network Model

Model

- Network denoted by $G(V, E)$, has $C$ communities
- Each node has attributes $X$, a vector
- Each node has affiliation weight $\{F_{uc}\}$
- Network Adjacency Matrix

$$P_{uv} = 1 - e^{(-\Sigma_c F_{uc} F_{vc})}$$

$$A_{uv} \sim \text{Bernoulli}(P_{uv})$$

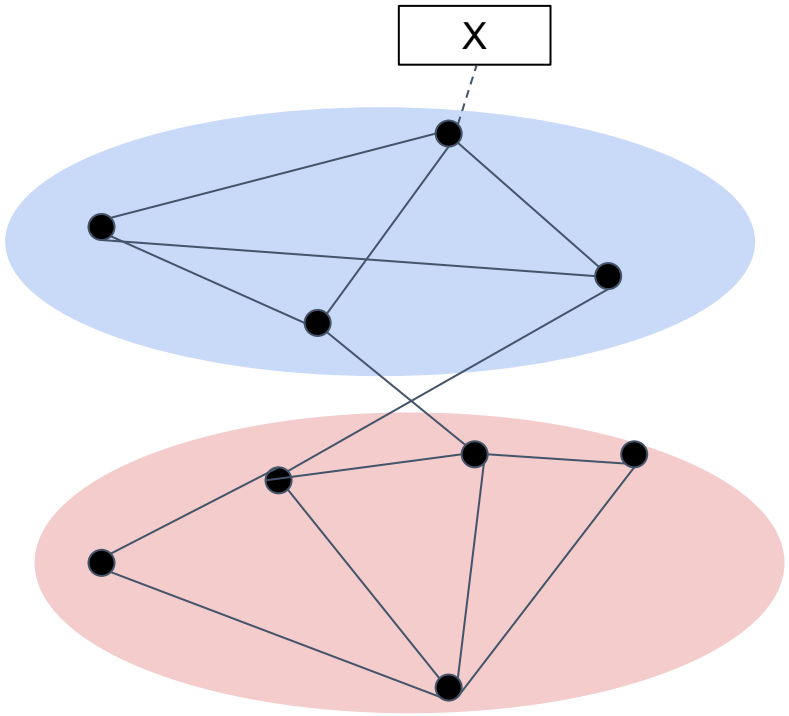
- Attributes

$$Q_{uk} = \frac{1}{1+e^{-\Sigma_c W_{kc} F_{uc}}}$$

$$X_{uk} \sim \text{Bernoulli}(Q_{uk})$$

# Dataset, Algorithm, and Tools (cont.)

- **Algorithm-CESNA**




Probabilistic Network Model

Maximize likelihood

- The optimal estimation of F and W is

$$\hat{F}, \hat{W} = \text{argmax} \log P(G, X|F, W)$$

$$\log P(G, X|F, W) = \mathcal{L}_G + \mathcal{L}_X$$

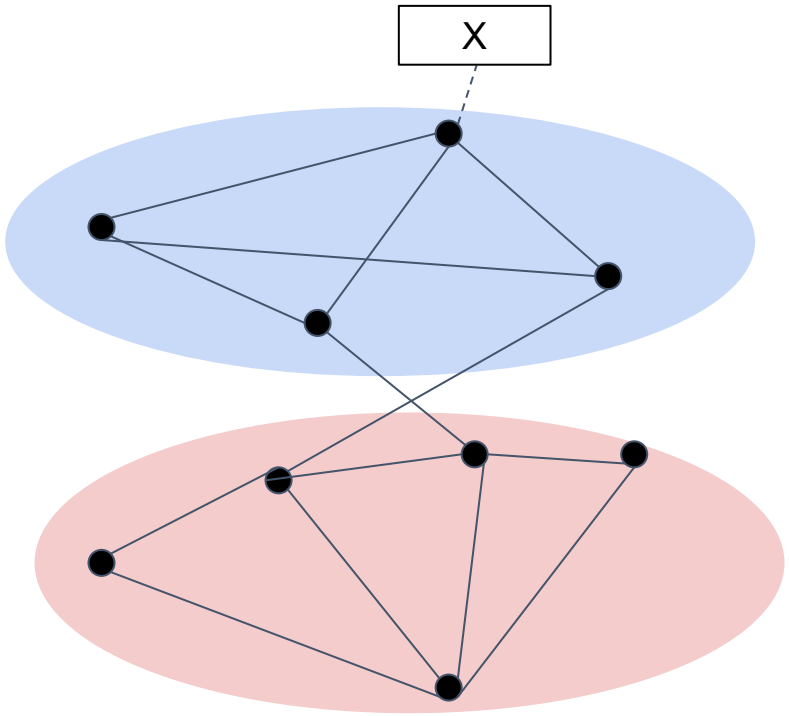
$$\mathcal{L}_G = \Sigma_{(u,v)\in E} \log(1 - e^{-F_u F_v^T}) - \Sigma_{(u,v)\notin E} F_u F_v^T$$

$$\mathcal{L}_X = \Sigma_{u,k}(X_{uk} \log Q_{uk} + (1 - X_{uk}) \log(1 - Q_{uk}))$$

- Solution-block coordinate ascent
  - Initial guess F and W
  - Fix W, optimize F
  - Fix F, optimize W
  - Repeat until change is small

# Dataset, Algorithm, and Tools (cont.)

- **Algorithm-CESNA**




Probabilistic Network Model

Determine Membership

- $u$ belongs to $c$ if $F_{uc} > \delta$
- Choose threshold?

$$\frac{1}{N} < 1 - e^{-\delta^2}$$

What's the number of $C$?

- Reserve a holdout set

# Dataset, Algorithm, and Tools (cont.)

- **Algorithm-CESNA**

| Method class | Overlapping | Hard membership | Structure+Attributes | nodes per 10 hours |
|---|---|---|---|---|
| Heuristic | No | Yes | No | 100,000 |
| LDA-based | Yes | No | Yes | 85,000 |
| Clique-based | Yes | Yes | No | 100,000 |
| Social circles | Yes | Yes | No | 5,000 |
| **CESNA** | Yes | Yes | Yes | **1,000,000** |

Ref: Jaewon Yang, Julian McAuley, Jure Leskovec. Community Detection in Networks with Node Attributes. 2013, Community Detection in Networks with Node Attributes.

# Expected Contributions and Timeline

- **Expected Contributions**
  We will implement the CESNA algorithm and test it on several social network data sets, and improve it if possible.

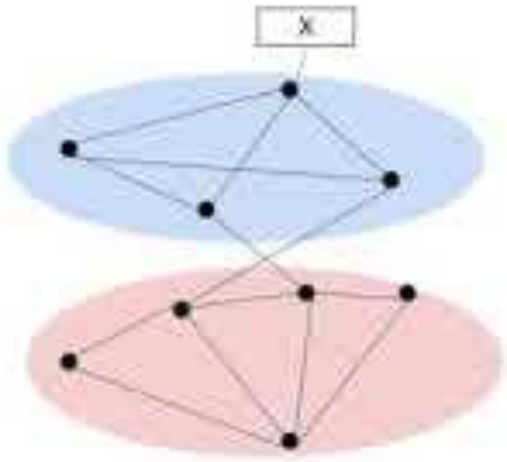
  The final results will be a website for visualizing  the results.

- **Timeline of the Project**
  - Literature review: Look through literatures 11.1-11.8
  - Implementation: Code the algorithm 11.9-11.23.
  - Test & Visualization: Test algo. and Visualize the results of the community detection. 11.23-11.30
  - Report: 12.1-12.6.
  - Further Work: We will extend the application of community detection in Anti-money Laundering dataset. (*If time permits)

# Youtube link: https://youtu.be/v8HZH7oLGQg

## Dataset, Algorithm, and Tools (cont.)

- **Algorithm-CESNA**




Probabilistic Network Model

Model

- Network denoted by $G(V, E)$, has $C$ communities
- Each node has attributes $X$, a vector
- Each node has affiliation weight $\{F_{uc}\}$
- Network Adjacency Matrix

$$P_{uv} = 1 - e^{(-\Sigma_c F_{uc} F_{vc})}$$

$$A_{uv} \sim \text{Bernoulli}(P_{uv})$$

- Attributes

$$Q_{uk} = \frac{1}{1+e^{-\Sigma_c W_{kc} F_{uc}}}$$

$$X_{uk} \sim \text{Bernoulli}(Q_{uk})$$

E6893 Big Data Analytics – Final Project Proposal © 2018 CY Lin, Columbia University