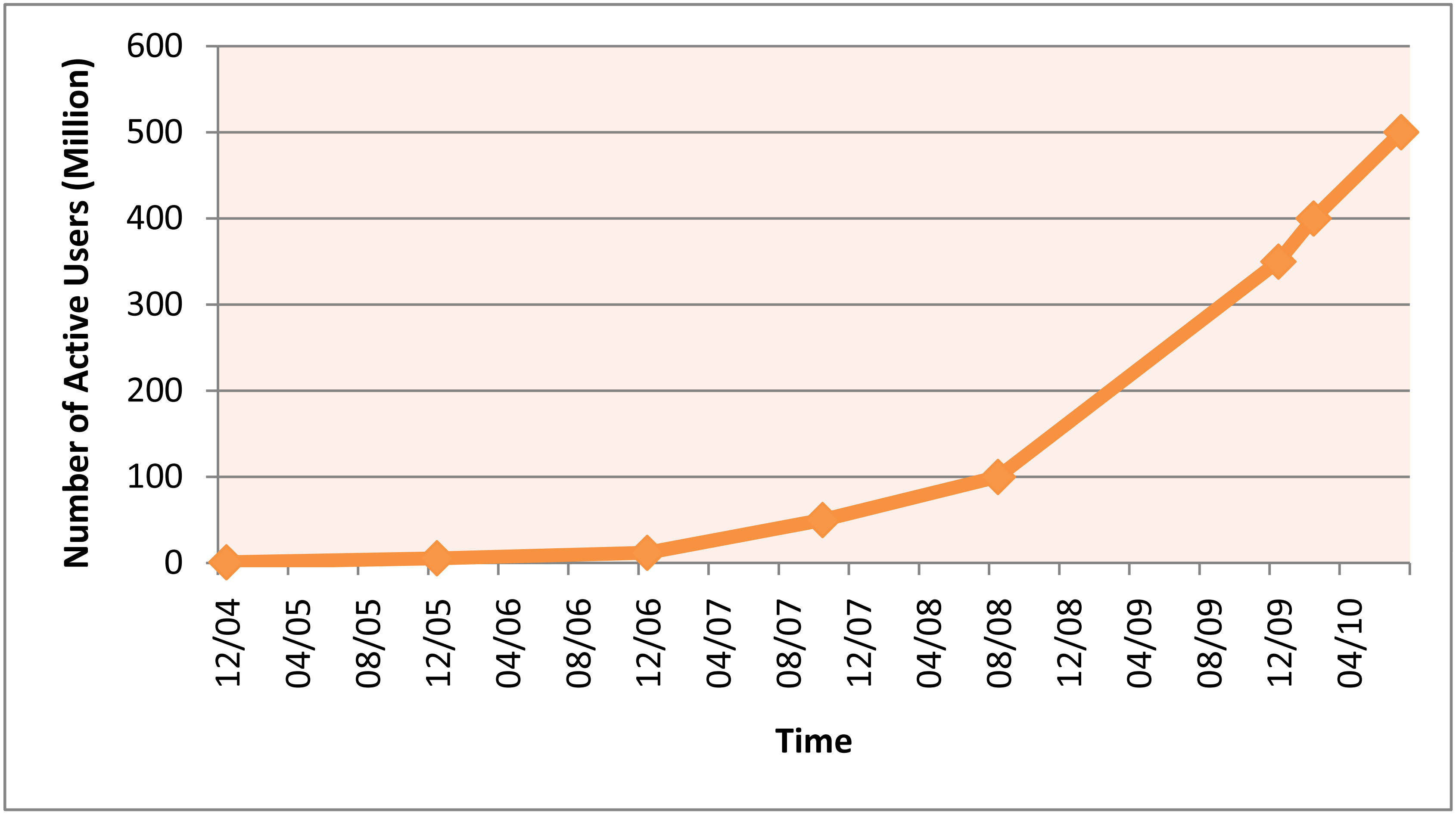
Social Media Mining

Chapter 5

EVOLUTION PATTERNS IN SOCIAL MEDIA

Growth of Facebook Population

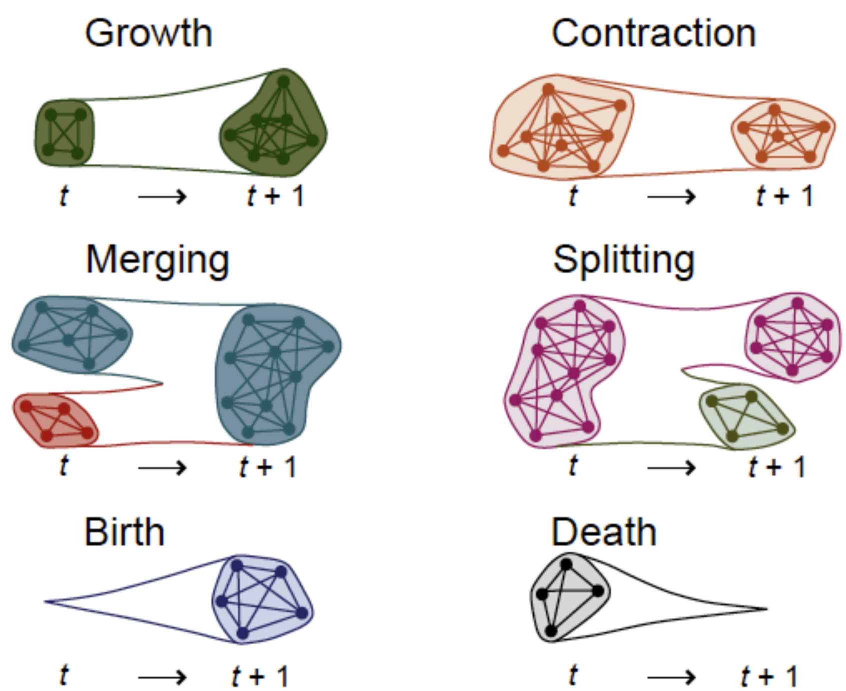


Evolutions in Social Media

- Social media networks are highly dynamic
- Interesting patterns in dynamic networks
  - Decreasing probability of new connections between two nodes with increasing distance
  - Many new connections occur as *triadic closures*
  - Segmentation of dynamic networks into 3 regions
    - Singletons
    - Isolated communities with a star structure
    - A giant component anchored by a well-connected core region
  - Density increases with the network growth
  - Average distance between nodes shrinks

Community Evolution

- Communities also expand, shrink , or dissolve in dynamic networks

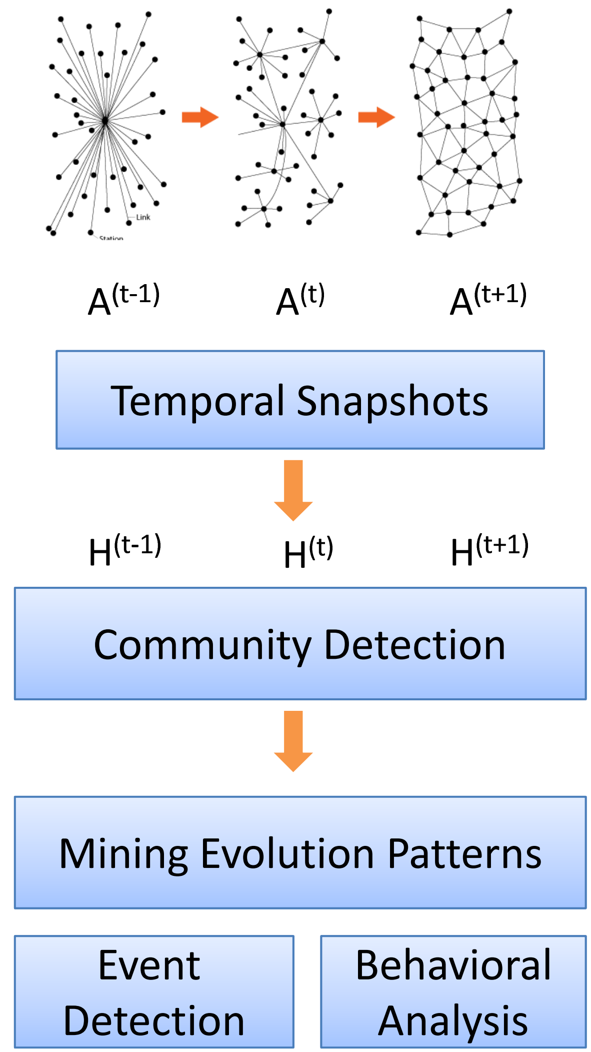


- How to uncover latent community change behind dynamic network interactions?

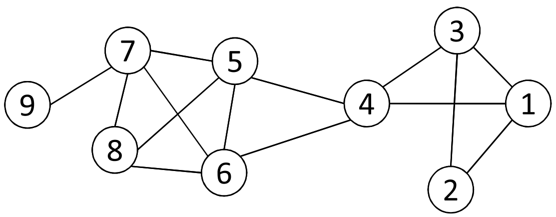
Naïve Approach to Studying Community Evolution

- Take snapshots of a network
- find communities at each snapshot
- Clustering independently at each snapshot

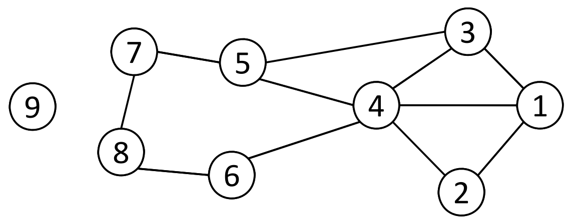
- Cons:
  - Most community detection methods produce local optimal solutions
  - Hard to determine if the evolution is due to the evolution or algorithm randomness



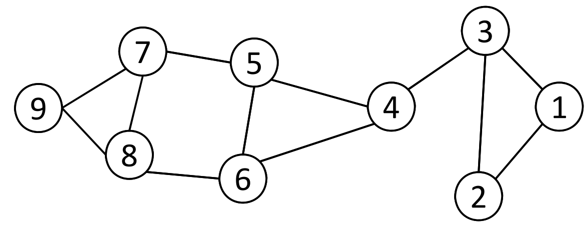
Naïve Approach Example



$A^{(1)}$ at T_1



$A^{(2)}$ at T_2



$A^{(3)}$ at T_3

$$H^{(1)} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad H^{(2)} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad H^{(3)} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

- There is a sharp change at T_2
- This approach may report spurious structural changes

Evolutionary Clustering in Smoothly Evolving Networks

- Evolutionary Clustering: find a *smooth* sequence of communities given a series of network snapshots
- Objective function: snapshot cost (CS) + temporal cost (CT)

$$Cost = \alpha \cdot CS + (1-\alpha) \cdot CT$$

- Take spectral clustering as an example
  - Snapshot cost : $CS_t = Tr(S_t^T L_t S_t), \quad s.t. \quad S_t^T S_t = I_k$
  - Temporal cost: $CT_t = \|\cancel{S_t - S_{t-1}}\|^2$

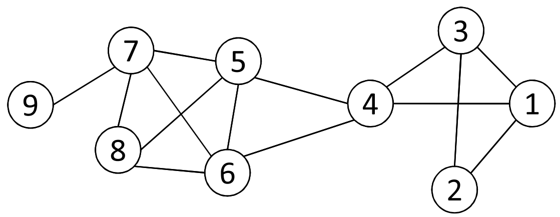
S_t is still a valid solution after an orthogonal transformation

$$CT_t = \frac{1}{2}\|S_t S_t^T - S_{t-1} S_{t-1}^T\|^2$$

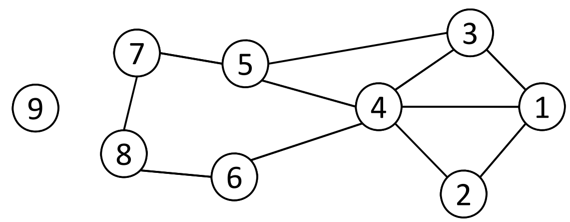
- Community Evolution: $Cost_t = Tr\left[S_t^T \widetilde{L}_t S_t\right]$

  where $\widetilde{L}_t = I - \alpha \cdot D_t^{-1/2} A^{(t)} D_t^{-1/2} - (1-\alpha) \cdot S_{t-1} S_{t-1}^T$

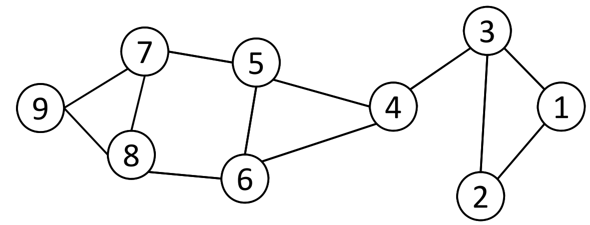
Evolutionary Clustering Example



$A^{(1)}$ at T_1



$A^{(2)}$ at T_2



$A^{(3)}$ at T_3

For T_1

$$S_1 = \begin{bmatrix} 0.33 & -0.44 \\ 0.27 & -0.43 \\ 0.33 & -0.44 \\ 0.38 & -0.16 \\ 0.38 & 0.24 \\ 0.38 & 0.24 \\ 0.38 & 0.38 \\ 0.33 & 0.30 \\ 0.19 & 0.23 \end{bmatrix}$$

For T_2

$$\widetilde{L}_2 = \begin{bmatrix} 0.91 & -0.42 & -0.33 & -0.21 & -0.01 & -0.01 & 0.01 & 0.01 & 0.01 \\ -0.42 & 0.92 & -0.08 & -0.27 & 0.00 & 0.00 & 0.02 & 0.01 & 0.01 \\ -0.33 & -0.08 & 0.91 & -0.22 & -0.25 & -0.01 & 0.01 & 0.01 & 0.01 \\ -0.21 & -0.27 & -0.22 & 0.95 & -0.18 & -0.24 & -0.02 & -0.02 & -0.01 \\ -0.01 & 0.00 & -0.25 & -0.18 & 0.94 & -0.06 & -0.37 & -0.06 & -0.04 \\ -0.01 & 0.00 & -0.01 & -0.24 & -0.06 & 0.94 & -0.07 & -0.45 & -0.04 \\ 0.01 & 0.02 & 0.01 & -0.02 & -0.37 & -0.07 & 0.91 & -0.44 & -0.05 \\ 0.01 & 0.01 & 0.01 & -0.02 & -0.06 & -0.45 & -0.44 & 0.94 & -0.04 \\ 0.01 & 0.01 & 0.01 & -0.01 & -0.04 & -0.04 & -0.05 & -0.04 & 0.97 \end{bmatrix}$$

We obtain two communities based on spectral clustering with this modified graph Laplacian: {1, 2, 3, 4} and {5, 6, 7, 8, 9}

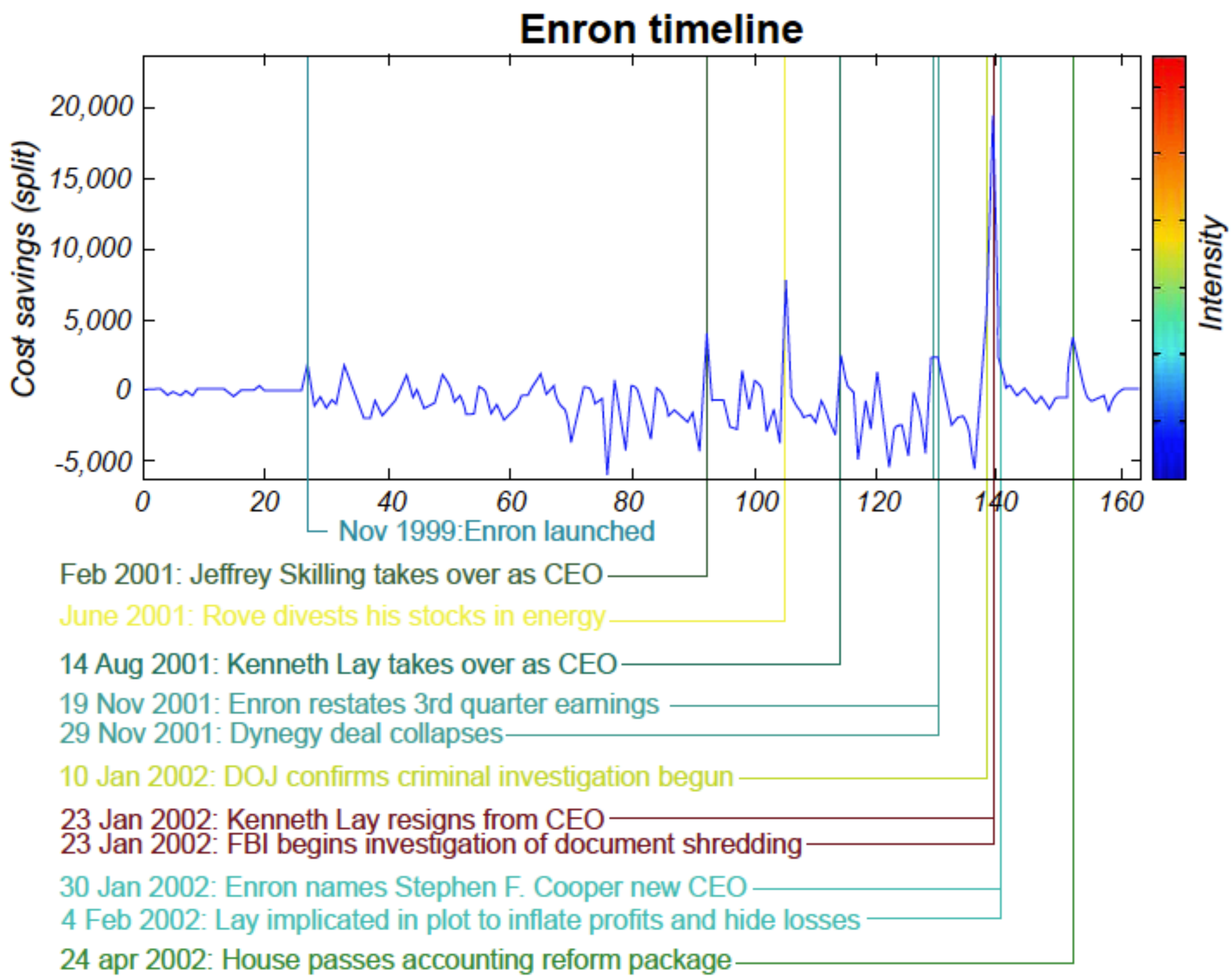
Segment-based Clustering with Evolving Networks

- Independent clustering at each snapshot
  - do not consider temporal information
  - Likely to output specious evaluation patterns
- Evolutionary clustering enforces smoothness
  - may fail to capture drastic change
- How to strike balance between *gradual changes* under normal circumstances and *drastic changes* caused by major events?
- Segment-based clustering:
  - Community structure remains unchanged in a segment of time
  - A change between consecutive segments
- Fundamental question: how to detect the change points?

Segment-based Clustering

- Segment-based Clustering assumes community structure remains unchanged in a segment of time
- GraphScope is one segment-based clustering method
  - If network connections do not change much over time, consecutive network snapshots should be grouped into one segment
  - If a new network snapshot does not fit into an existing segment (when current community structure induces a high cost on a new network snapshot), then introduce a change point and start a new segment

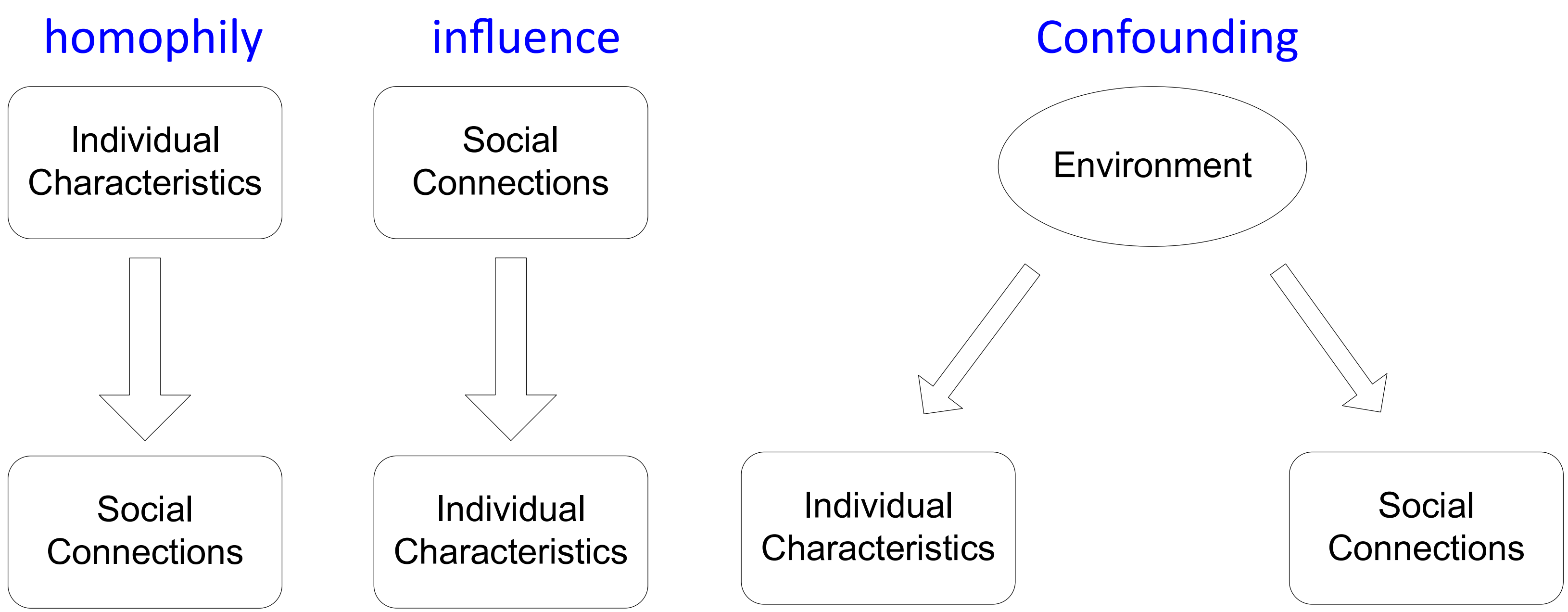
GraphScope on Enron Data



CLASSIFICATION WITH NETWORK DATA

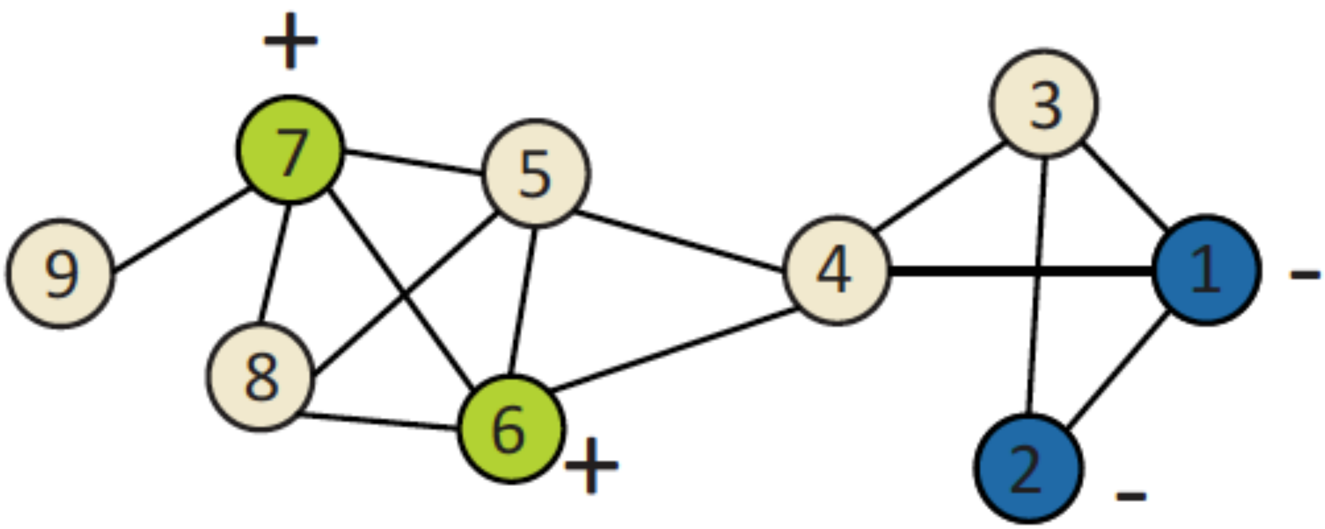
Correlations in Network

- Individual behaviors are correlated in a network environment



Classification with Network Data

- How to leverage this correlation observed in networks to help predict user attributes or interests?



Predict the labels for nodes in yellow

Collective Classification

- Labels of nodes are interdependent with each other
- The label of one node cannot be determined independently; Need Collective Classification
- Markov Assumption: *the label of one node depends on the label of his neighbors*

$$p(y_i|A) = p(y_i|N_i)$$

- Collective classification involves 3 components:

Local Classifier	Relational Classifier	Collective Inference
• Assign initial label	• Capture correlations between nodes	• Propagate correlations through network

Collective Classification

- Local Classifier: used for initial label assignment
  - Predicts label based on node attributes
  - Classical classification learning
  - Does not employ network information
- Relational Classifier: capture correlations based on label info
  - Learn a classifier from the labels or/and attributes of its neighbors to the label of one node
  - Network information is used
- Collective Classification: propagate the correlation
  - Apply relational classifier to each node iteratively
  - Iterate until the inconsistency between neighboring labels is minimized
  - Network structure substantially affects the final prediction

Weighted-vote Relational Neighborhood Classifier

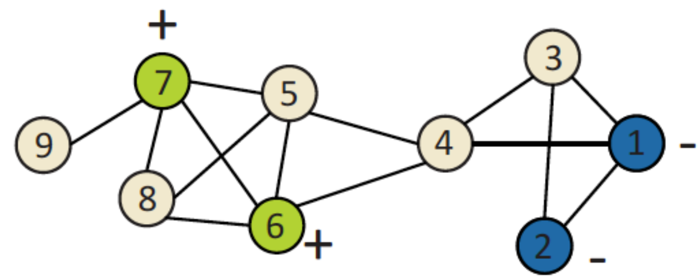
- No local classifier
- Relational classifier
  - prediction of one node is the average of its neighbors

$$
\begin{aligned}
p(y_i = 1|N_i) &= \frac{1}{\sum_{v_j \in N_i} A_{ij}} \sum_{v_j \in N_i} A_{ij} \cdot p(y_j = 1|N_j) \\
&= \frac{1}{|N_i|} \sum_{v_j \in N_i} p(y_j = 1|N_j).
\end{aligned}
$$

- Collective Inference

Example of wvRN

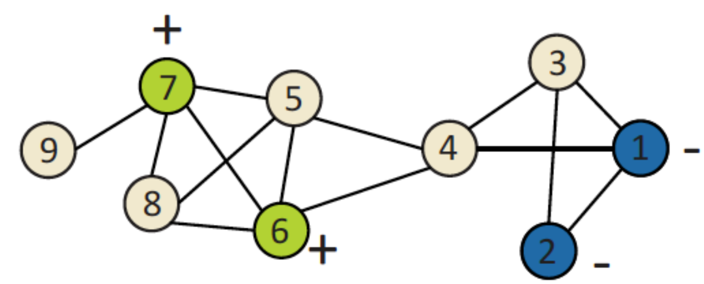
- Initialization for unlabeled nodes
  - $p(y_i=1|N_i)=0.5$
- 1st Iteration:
  - For node 3, $N_3=\{1,2,4\}$
    - $P(y_1=1|N_1) = 0$
    - $P(y_2=1|N_2) = 0$
    - $P(y_4=1 \ |N_4) = 0.5$
    - $P(y_3=1|N_3) = 1/3\ (0 + 0 + 0.5) = 0.17$
  - For node 4, $N_4=\{1,3, 5, 6\}$
    - $P(y_4=1|N_4)= ¼(0+ 0.17+0.5+1) = 0.42$
  - For node 5, $N_5=\{4,6,7,8\}$
    - $P(y_5=1|N_5) = ¼\ (0.42+1+1+0.5) = 0.73$



$$\mathbf{p}^{(1)} = \begin{bmatrix} 0 \\ 0 \\ 0.17 \\ 0.42 \\ 0.73 \\ 1 \\ 1 \\ 0.91 \\ 1.00 \end{bmatrix}$$

Iterative Result

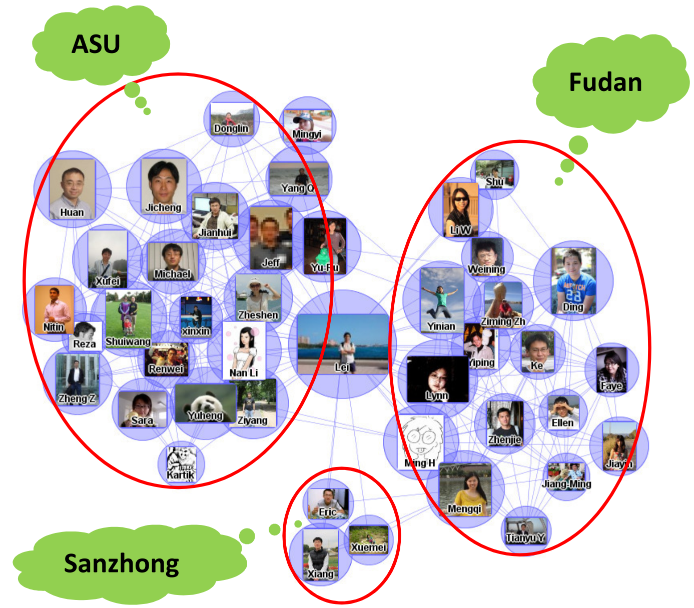
- Stabilizes after 5 iterations
  - Nodes 5, 8, 9 are + ($P_i > 0.5$)
  - Node 3 is – ($P_i < 0.5$)
  - Node 4 is in between ($P_i = 0.5$)



$$\mathbf{p}^{(1)} = \begin{bmatrix} 0 \\ 0 \\ 0.17 \\ 0.42 \\ 0.73 \\ 1 \\ 1 \\ 0.91 \\ 1.00 \end{bmatrix}, \mathbf{p}^{(2)} = \begin{bmatrix} 0 \\ 0 \\ 0.14 \\ 0.47 \\ 0.85 \\ 1 \\ 1 \\ 0.95 \\ 1.00 \end{bmatrix}, \mathbf{p}^{(3)} = \begin{bmatrix} 0 \\ 0 \\ 0.16 \\ 0.50 \\ 0.86 \\ 1 \\ 1 \\ 0.95 \\ 1.00 \end{bmatrix}, \mathbf{p}^{(4)} = \begin{bmatrix} 0 \\ 0 \\ 0.17 \\ 0.51 \\ 0.87 \\ 1 \\ 1 \\ 0.96 \\ 1.00 \end{bmatrix}, \mathbf{p}^{(5)} = \begin{bmatrix} 0 \\ 0 \\ 0.17 \\ 0.51 \\ 0.87 \\ 1 \\ 1 \\ 0.96 \\ 1.00 \end{bmatrix}.$$

Community-Based Learning

- People in social media engage in various relationships
  - Colleagues
  - Relatives
  - Friends
  - Co-travelers



- Different relationships have different correlations with user interests/behavior/profiles

Challenges

- Social media often comes with a network, but no relationship information
- Or relationship information is not complete or refined enough

- Challenges:
  - How to differentiate these heterogeneous relationship in a network?
  - How to determine whether the relationship is useful for prediction?

Social Dimension

- Social Dimension:
  - Latent dimensions defined by social connections
- Each dimension represents one type of relationship

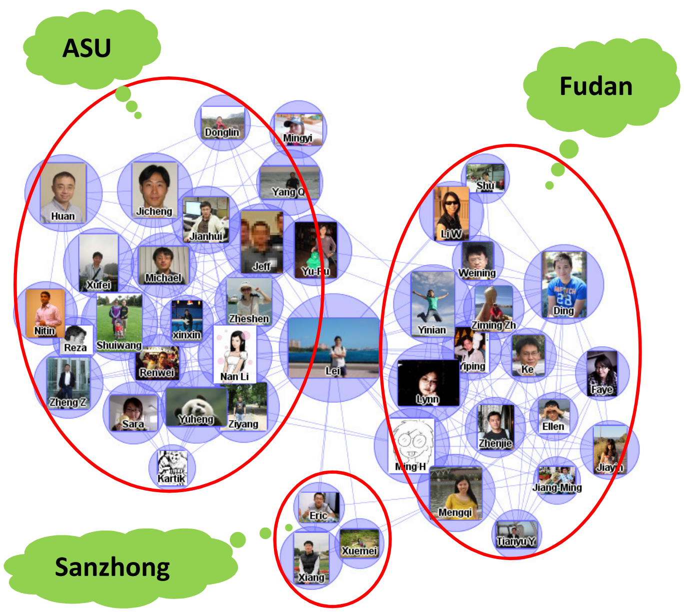


Table 5.1: Social Dimensions

Actors	ASU	Fudan	Sanzhong
Lei	1	1	1
$Actor_1$	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$

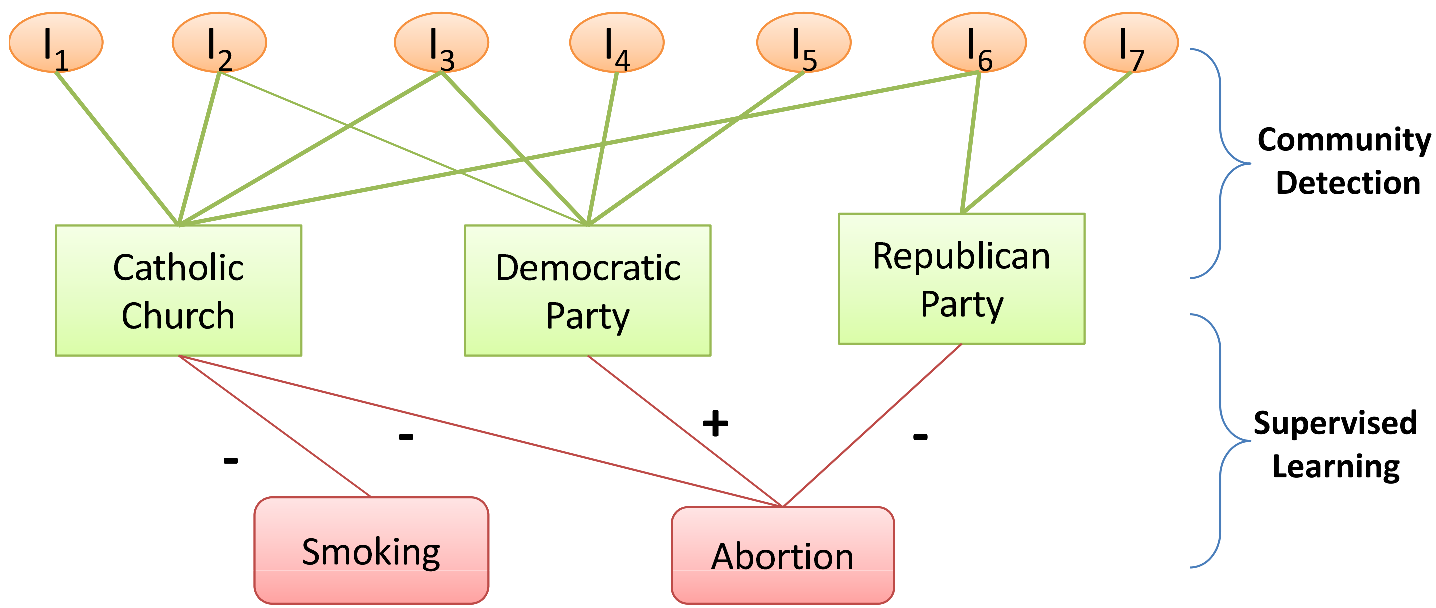
A Learning Framework based on Social Dimensions

- People involved in the same social dimension are likely to connect to each other, thus forming a community

- Training:
  1. Extract meaningful social dimensions based on network connectivity via community detection
  2. Determine relevant social dimensions (plus node attributes) through supervised learning

- Prediction
  - Apply the constructed model in Step 2 to social dimensions of unlabeled nodes
  - No iterative inference

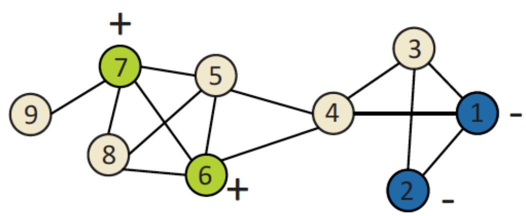
Underlying Assumption



- Assumption:
  - the label of one node is determined by its social dimension
  - $P(y_i|A) = P(y_i|S_i)$
- Community membership serves as latent features

Example of SocioDim Framework

- One is likely to involve in multiple relationships, thus soft clustering is used to extract social dimensions



$$S = \begin{bmatrix} 0.33 & -0.44 \\ 0.27 & -0.43 \\ 0.33 & -0.44 \\ 0.38 & -0.16 \\ 0.38 & 0.24 \\ 0.38 & 0.24 \\ 0.38 & 0.38 \\ 0.33 & 0.30 \\ 0.19 & 0.23 \end{bmatrix}$$

Spectral Clustering

Table 5.2: Communities are Features

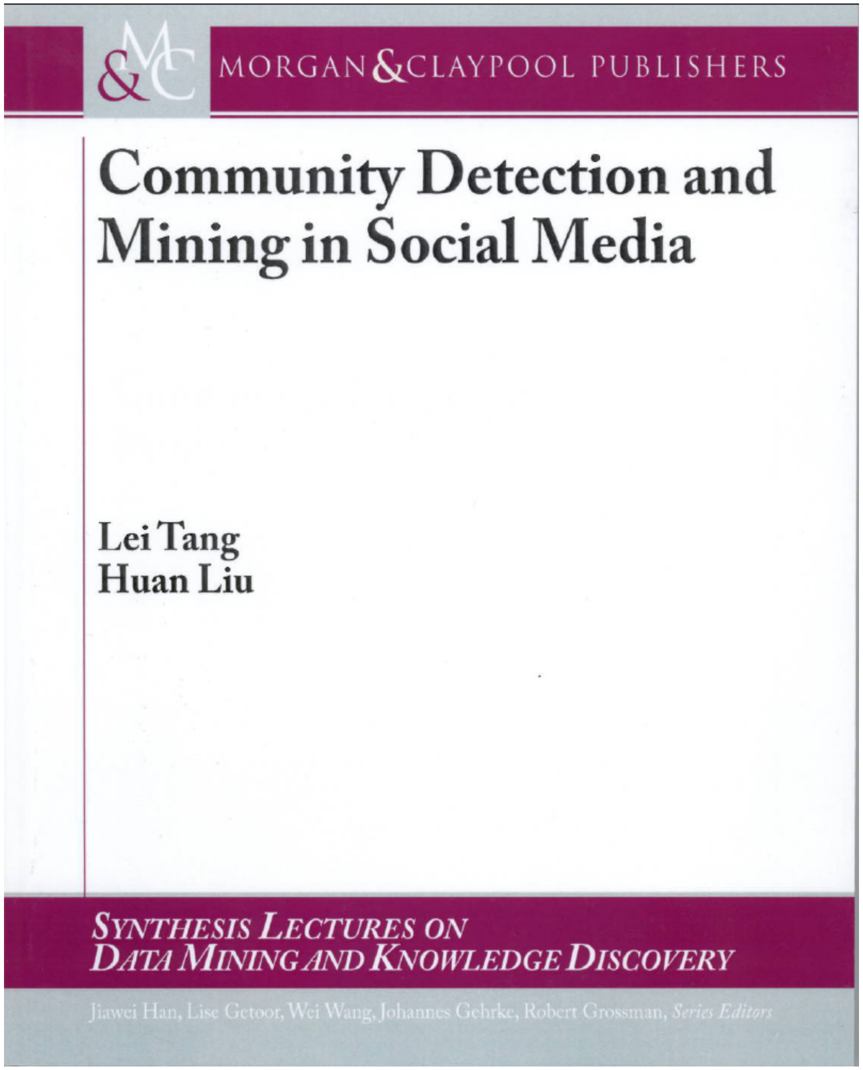
Node	S_1	S_2	Label	Pred. Score	Pred.
1	0.33	-0.44	–		
2	0.27	-0.43	–		
3	0.33	-0.44	?	-0.53	-
4	0.38	-0.16	?	-0.13	-
5	0.38	0.24	?	0.42	+
6	0.38	0.24	+		
7	0.38	0.38	+		
8	0.33	0.30	?	0.50	+
9	0.19	0.23	?	0.38	+

$$y = sign(0.16S_1 + 1.39S_2 + 0.03).$$

Support Vector Machine

Collective Classification vs. Community-Based Learning

	Collective Classification	Community-Based Learning
Computational Cost for Training	low	high
Computational Cost for Prediction	high	low
Handling Heterogeneous Relations		✔
Handling Evolving Networks	✔	
Integrating network information and actor attributes	✔	✔



Book Available at

- Morgan & claypool Publishers
- Amazon

If you have any comments, please feel free to contact:

- **Lei Tang**, Yahoo! Labs, ltang@yahoo-inc.com
- **Huan Liu**, ASU huanliu@asu.edu