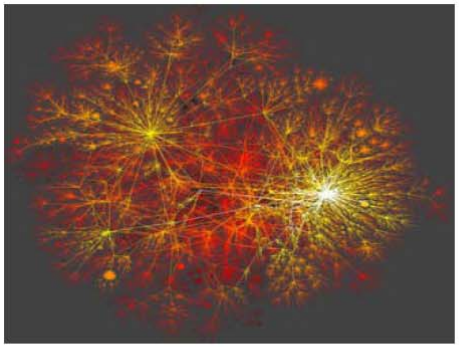
On Community Outliers and their Efficient Detection in Information Networks

Jing Gao[1], Feng Liang[1], Wei Fan[2],
Chi Wang[1], Yizhou Sun[1], Jiawei Han[1]

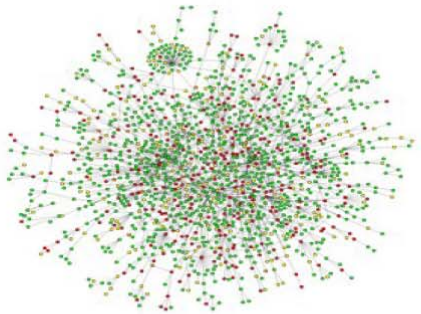
1 University of Illinois
2 IBM TJ Watson

Information Networks

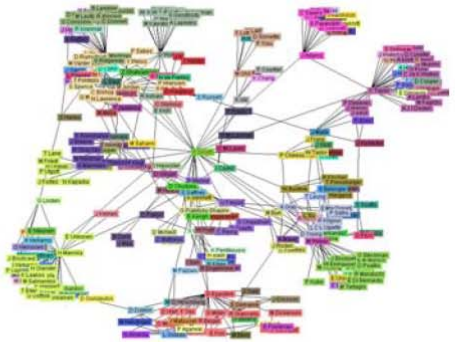
- Node represents an entity
  - Each node has feature values
  - e.g., users in social networks, webpages on internet
- Link represents relationship between entities
  - e.g., two users are linked if they are friends; two webpages are linked through hyper-links
- Information networks are ubiquitous



An Internet Web



Yeast protein interaction network

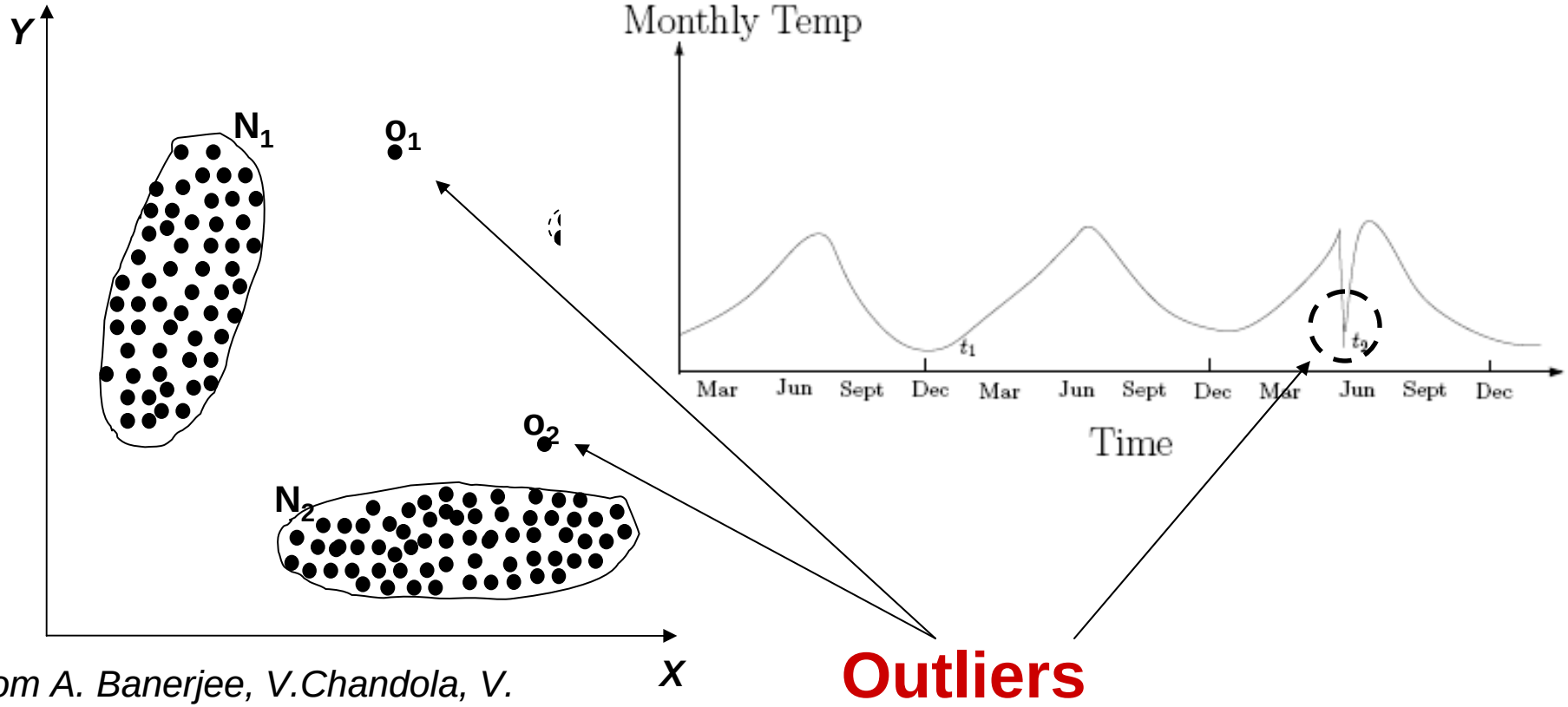


Co-author network

Social network sites

Outlier (Anomaly, Novelty) Detection

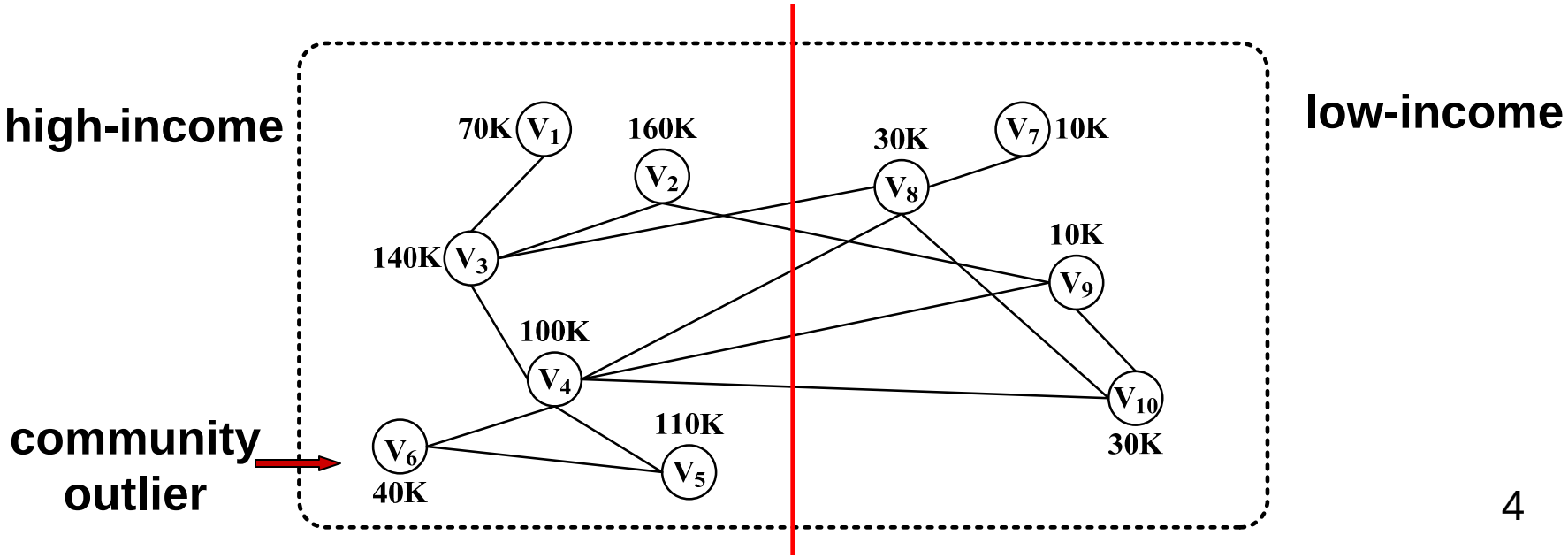
- Goal
  - Identify points that deviate significantly from the majority of the data



from A. Banerjee, V.Chandola, V. Kumar, J.Srivastava. Anomaly Detection: A Tutorial. SDM'08

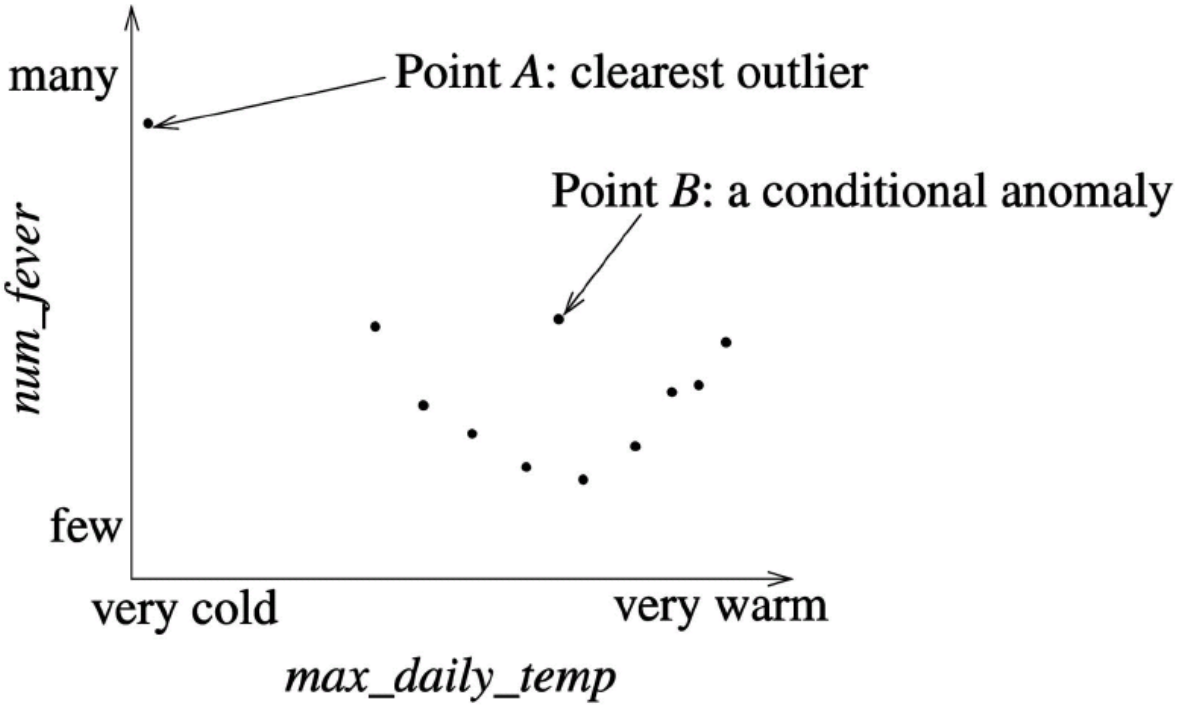
Community Outliers

- Definition
  - Two information sources: links, node features
  - There exist communities based on links and node features
  - Objects that have feature values deviating from those of other members in the same community are defined as community outliers



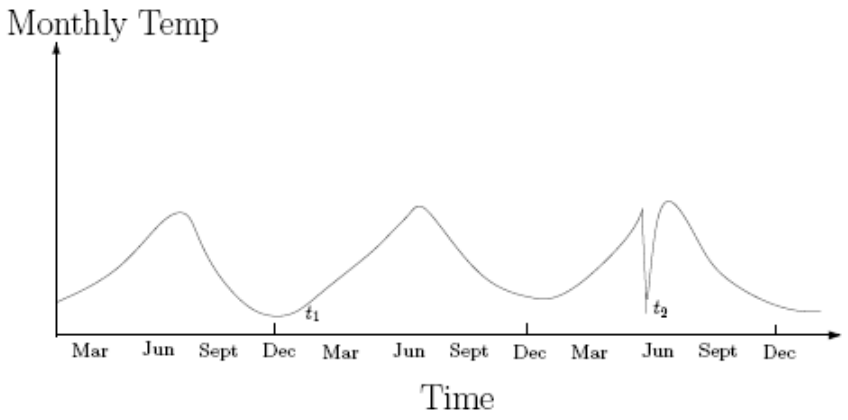
Contextual (Conditional) Outliers

- Global vs. Contextual
  - Global: identify outliers among all the data
  - Contextual: identify outliers within a subset of data defined by contextual features

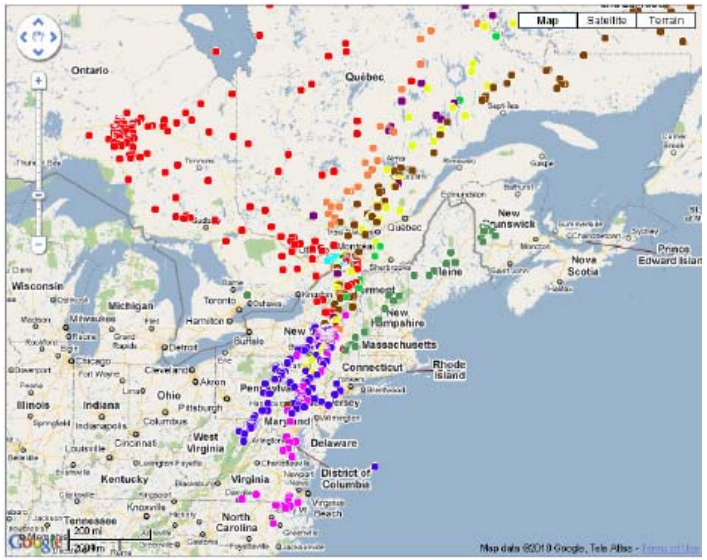


Examples

- Contexts
  - a subset of features, temporal, spatial, or communities in networks (in this paper)



temporal contexts

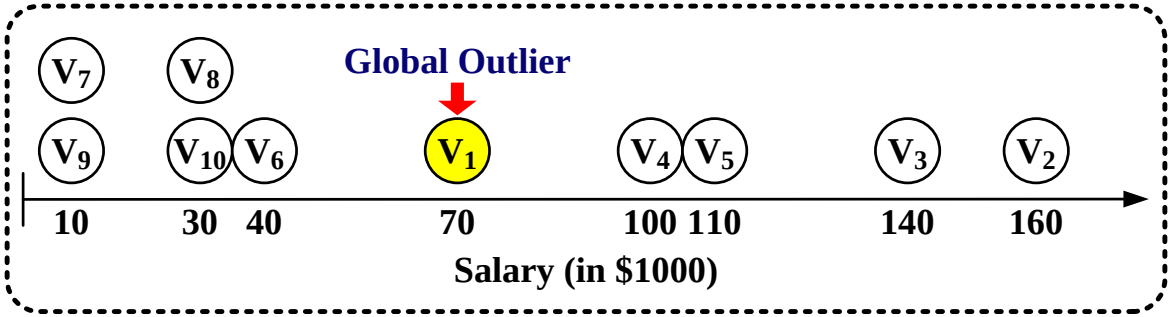


spatial contexts

Outliers in Information Networks

1) Global outlier:

only consider node features

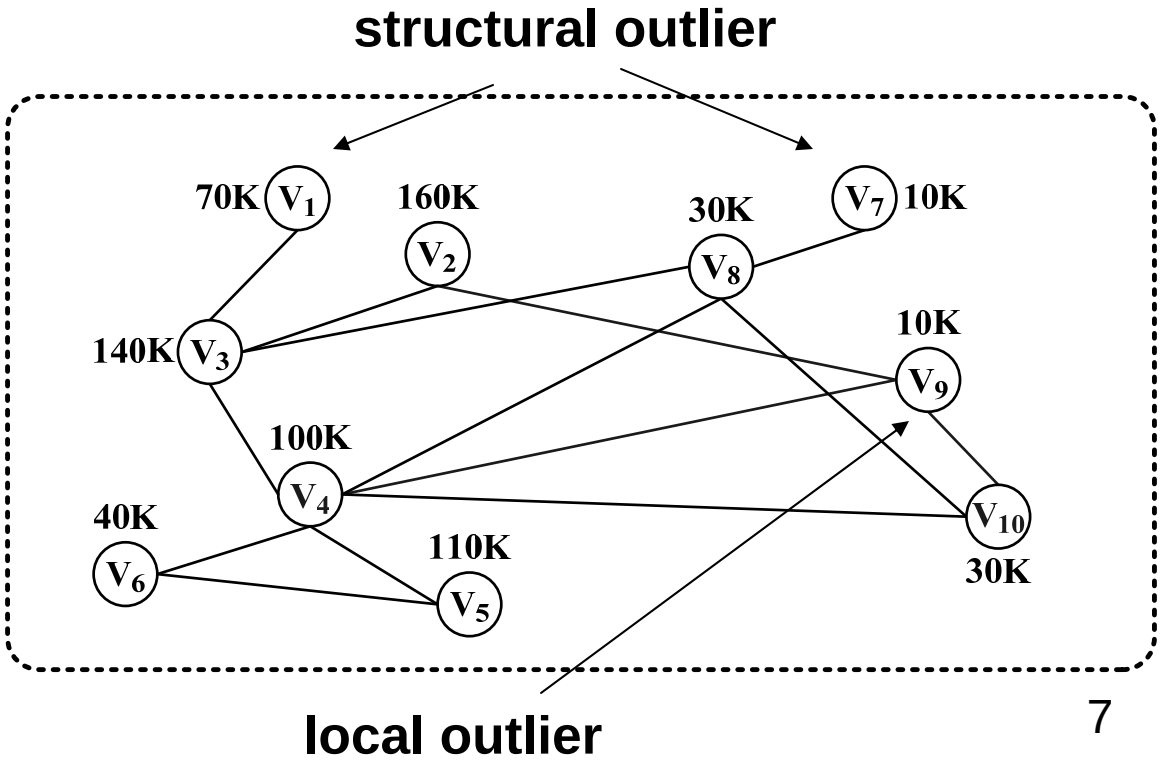


2) Structural outlier:

only consider links

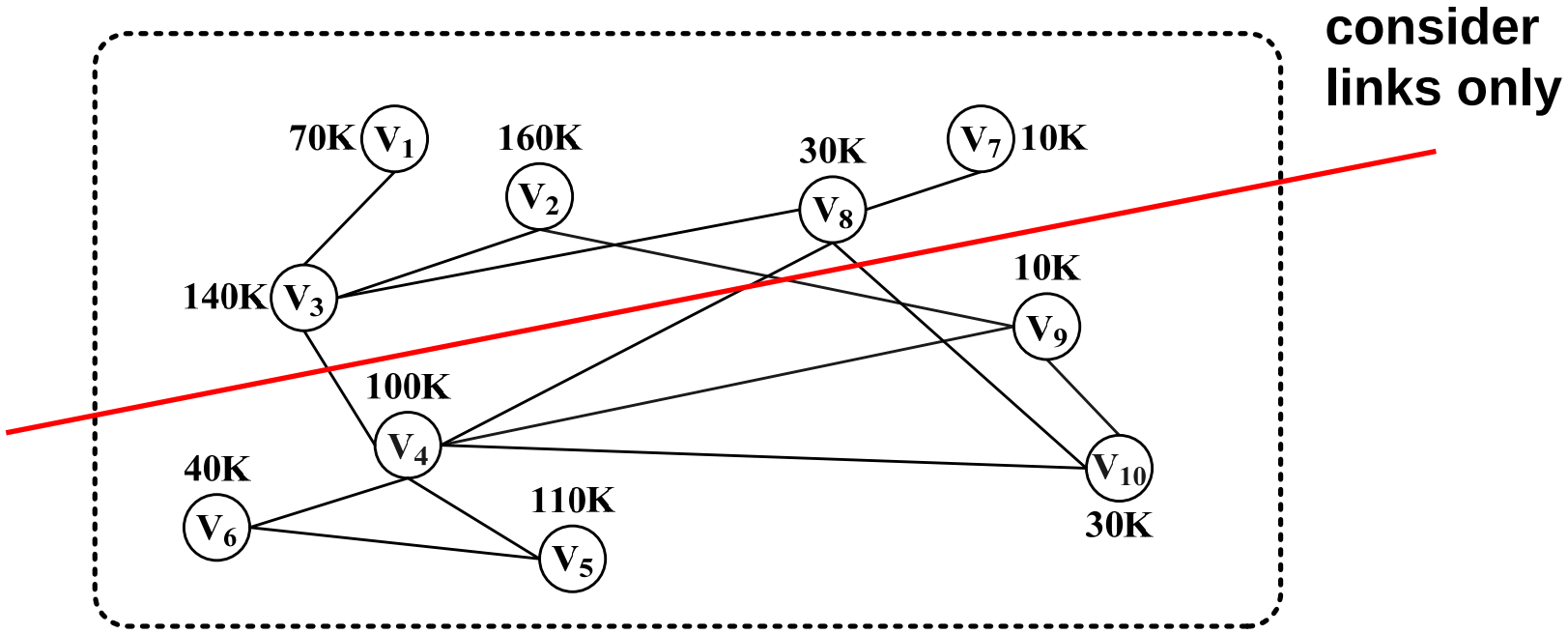
3) Local outlier:

only consider the feature values of direct neighbors

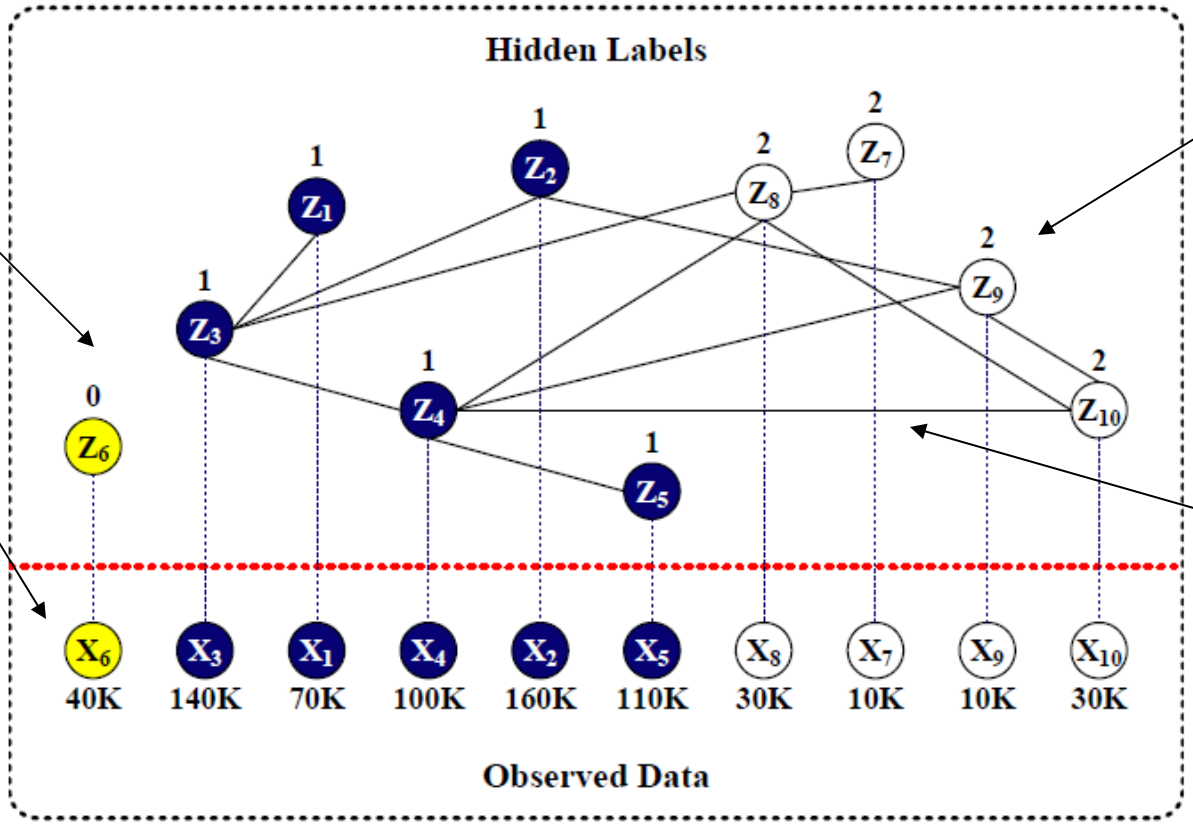


Unified Model is Needed

- Links and node features
  - More meaningful to identify communities based on links and node features together
- Community discovery and outlier detection
  - Outliers affect the discovery of communities



A Unified Probabilistic Model (1)



$\Theta = (\theta_1, \ldots, \theta_K)$

K: number of communities

high-income: mean: 116k std: 35k

low-income: mean: 20k std: 12k

model parameters

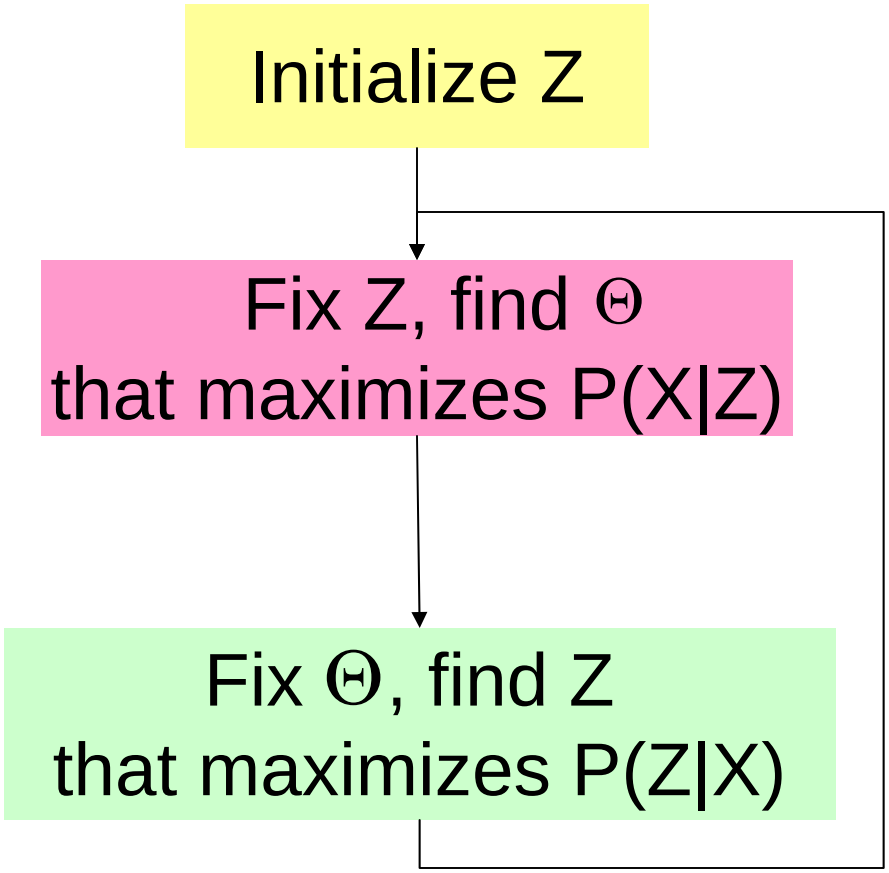
A Unified Probabilistic Model (2)

- Probability
  - Maximize $P(X) \propto P(X|Z)P(Z)$
  - P(X|Z) depends on the community label and model parameters
    - eg., salaries in the high or low-income communities follow Gaussian distributions defined by mean and std
  - P(Z) is higher if neighboring nodes from normal communities share the same community label
    - eg., two linked persons are more likely to be in the same community
    - outliers are isolated—does not depend on the labels of neighbors

Modeling Continuous or Text Data

- Continuous
  - Gaussian distribution
  - Model parameters: mean, standard deviation
- Text
  - Multinomial distribution
  - Model parameters: probability of a word appearing in a community

Community Outlier Detection Algorithm



Θ: model parameters

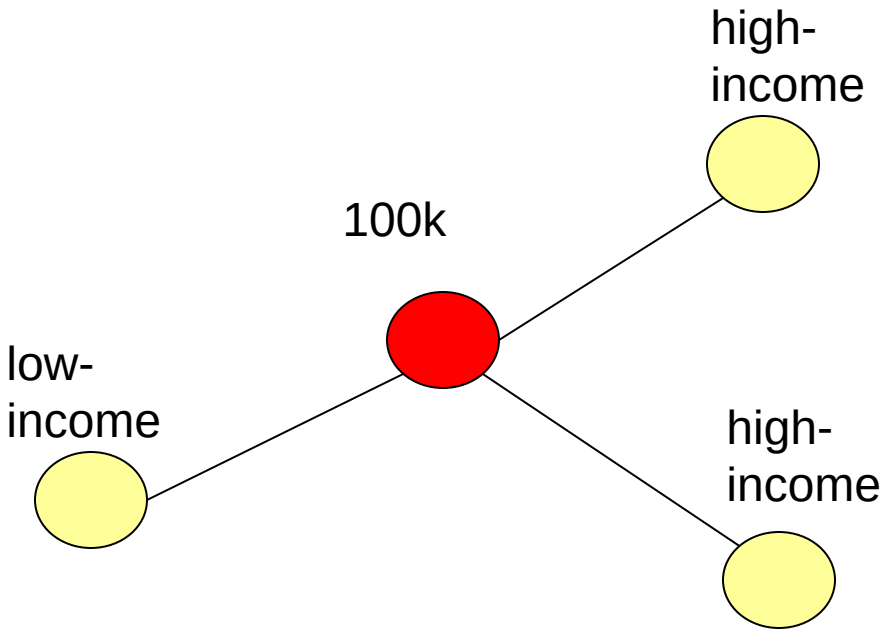
Z: community labels

Parameter estimation

Inference

Inference (1)

- Calculate Z
  - Model parameters are known
  - Iteratively update the community labels of nodes
  - Select the label that maximizes $P(Z|X,Z_N)$



mean:

high-income: 110k

low-income: 30k

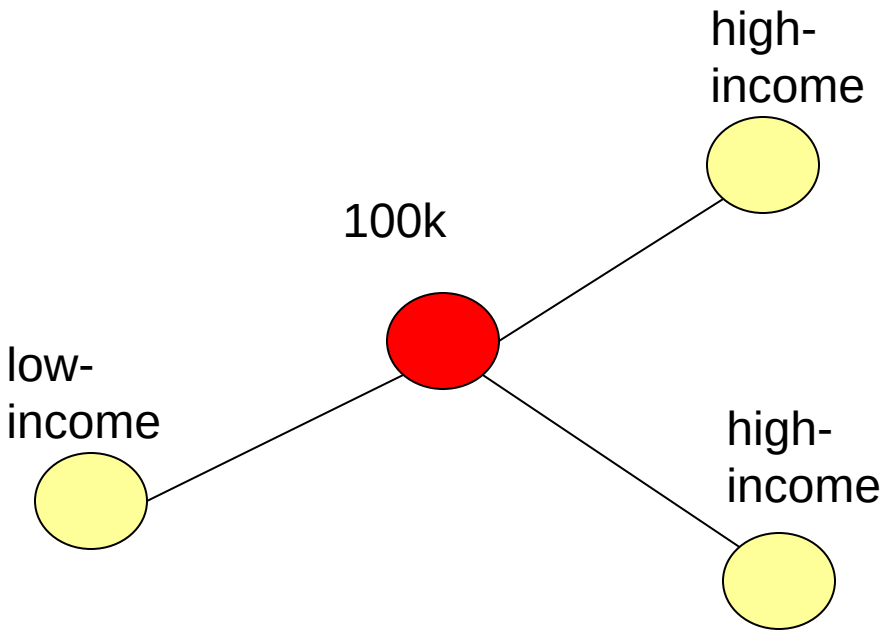
high-income? 80%

low-income? 10%

outlier? 10%

Inference (2)

- Calculate $P(Z|X,Z_N)$
  - Consider both the node features and community labels of neighbors if Z indicates a normal community
  - If the probability of a node belonging to any community is low enough, label it as an outlier



mean:

high-income: 100k

low-income: 30k

high-income:

P(salary=100k|high-income)
P(high-income|neighbors)

low-income:

P(salary=100k|low-income)
P(low-income|neighbors)

outlier:

constant

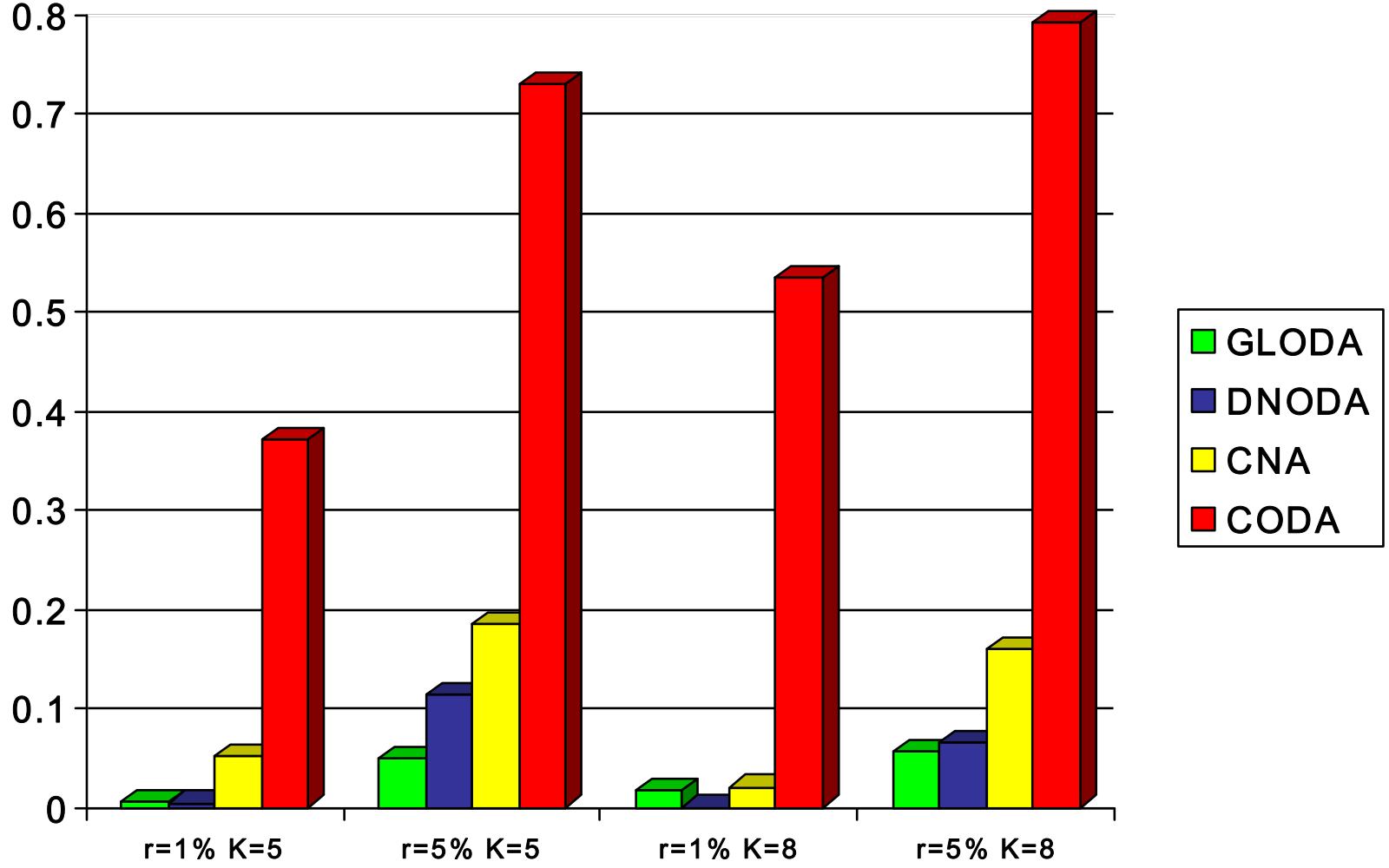
Parameter Estimation

- Calculate model parameters
  - maximum likelihood estimation
- Continuous
  - mean: sample mean of the community
  - standard deviation: square root of the sample variance of the community
- Text
  - probability of a word appearing in the community: empirical probability

Simulated Experiments

- Data
  - Generate continuous data based on Gaussian distributions and generate labels according to the model
  - r: percentage of outliers, K: number of communities
- Baseline models
  - GLODA: global outlier detection (based on node features only)
  - DNODA: local outlier detection (check the feature values of direct neighbors)
  - CNA: partition data into communities based on links and then conduct outlier detection in each community

Precision


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
r=1% K=5
r=5% K=5
r=1% K=8
r=5% K=8
GLODA
DNODA
CNA
CODA

Experiments on DBLP

- Data
  - DBLP: computer science bibliography
  - Areas: data mining, artificial intelligence, database, information analysis
- Case studies
  - Conferences:
    - Links: percentage of common authors among two conferences
    - Node features: publication titles in the conference
  - Authors:
    - Links: co-authorship relationship
    - Node features: titles of publications by an author

Case Studies on Conferences

Communities	Keywords
Data Mining	frequent dimensional spatial association similarity pattern fast sets approximate series
Database	oriented views applications querying design access schema control integration sql
Artificial Intelligence	reasoning planning logic representation recognition solving problem reinforcement programming theory
Information Analysis	relevance feature ranking automatic documents probabilistic extraction user study classifiers

- **Database:** ICDE, VLDB, SIGMOD, PODS, EDBT
- **Artificial Intelligence:** IJCAI, AAAI, ICML, ECML
- **Data Mining:** KDD, PAKDD, ICDM, PKDD, SDM
- **Information Analysis:** SIGIR, WWW, ECIR, WSDM

Community outliers: CVPR CIKM

Conclusions

- Community Outliers
  - Nodes that have different behaviors compared with the others in the community
- Community Outlier Detection
  - A unified probabilistic model
  - Conduct community discovery and outlier detection simultaneously
  - Consider both links and node features

Thanks!

- Any questions?