KDD2018 ODD v5.0

中国科学院 CHINESE ACADEMY OF SCIENCES

Carnegie Mellon SCHOOL OF COMPUTER SCIENCE

EagleMine: Vision-Guided Mining in Large Graphs

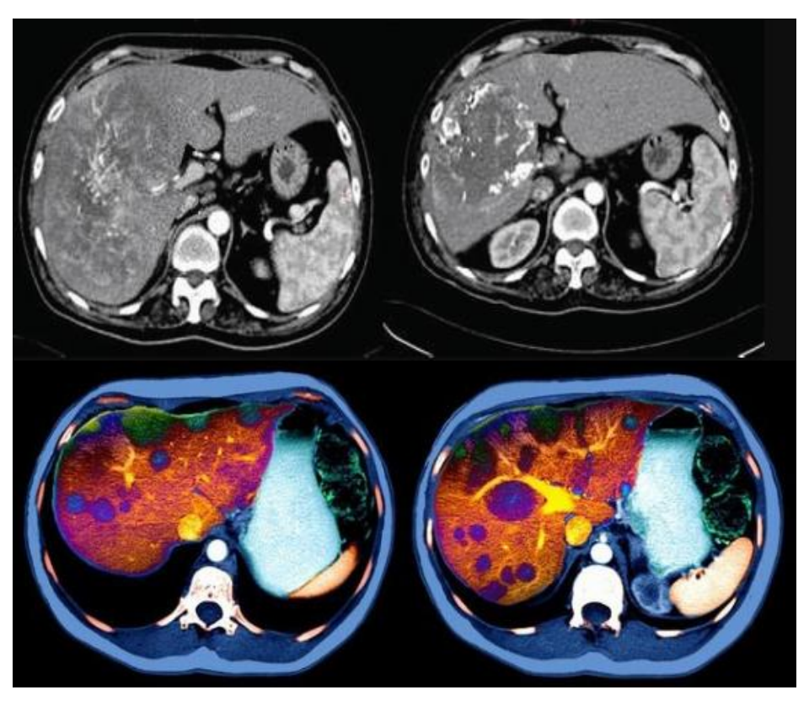
Wenjie Feng[1,2], Shenghua Liu[1], Christos Faloutsos[3], Bryan Hooi[3], Huawei Shen[1], and Xueqi Cheng[1]

[1]Institute of Computing Technology ICT, CAS [2]University of Chinese Academy of Sciences [3]Computer Science Dept., CMU

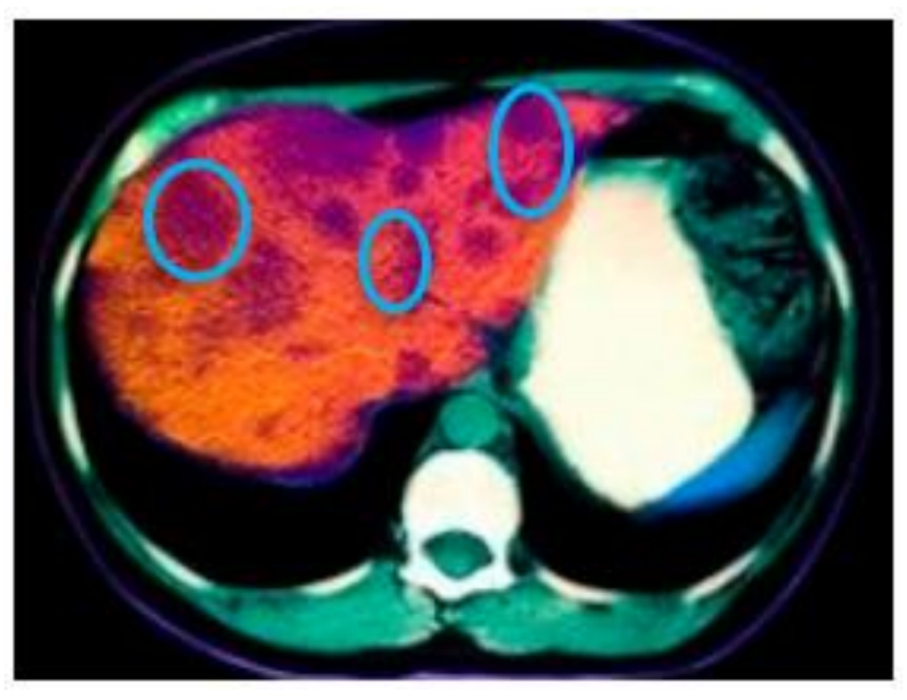
Motivation

Human Healthcare

Chest CT scans

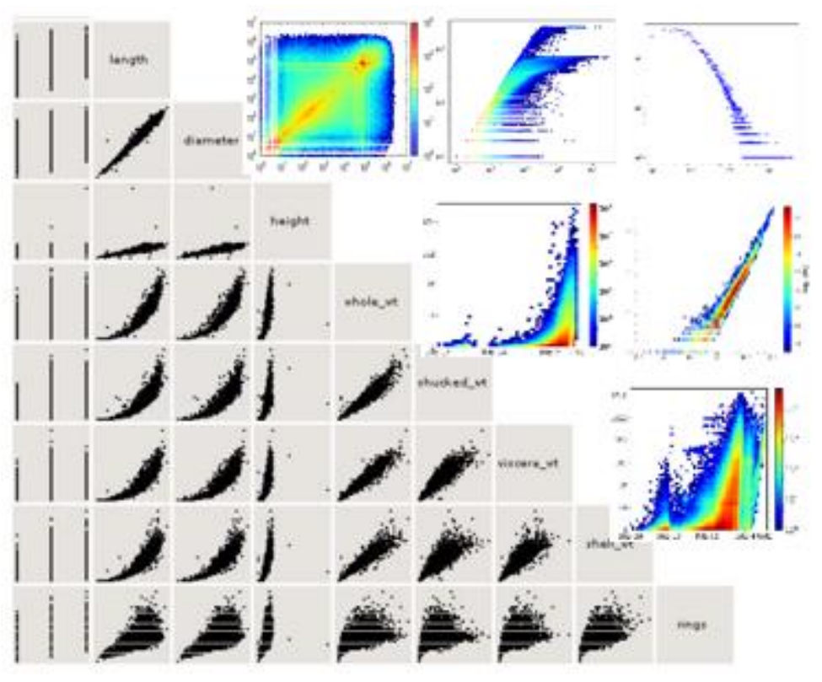


Cancer & Tumor

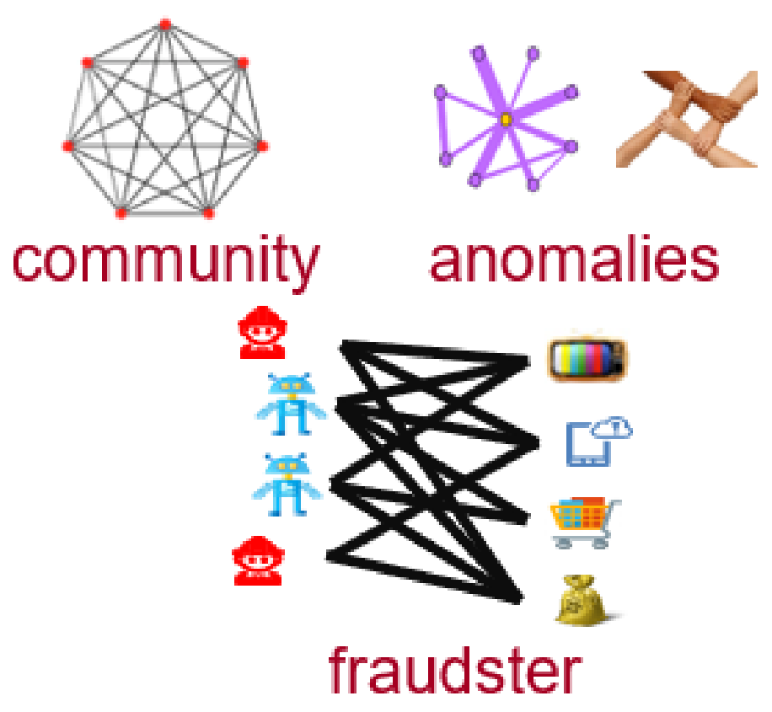


Large graph

View spaces



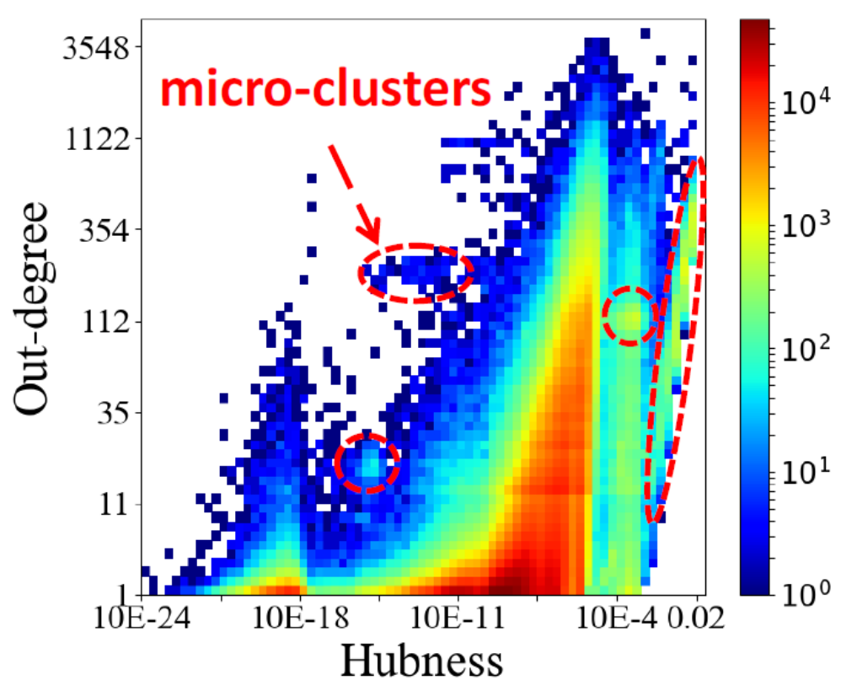
Patterns

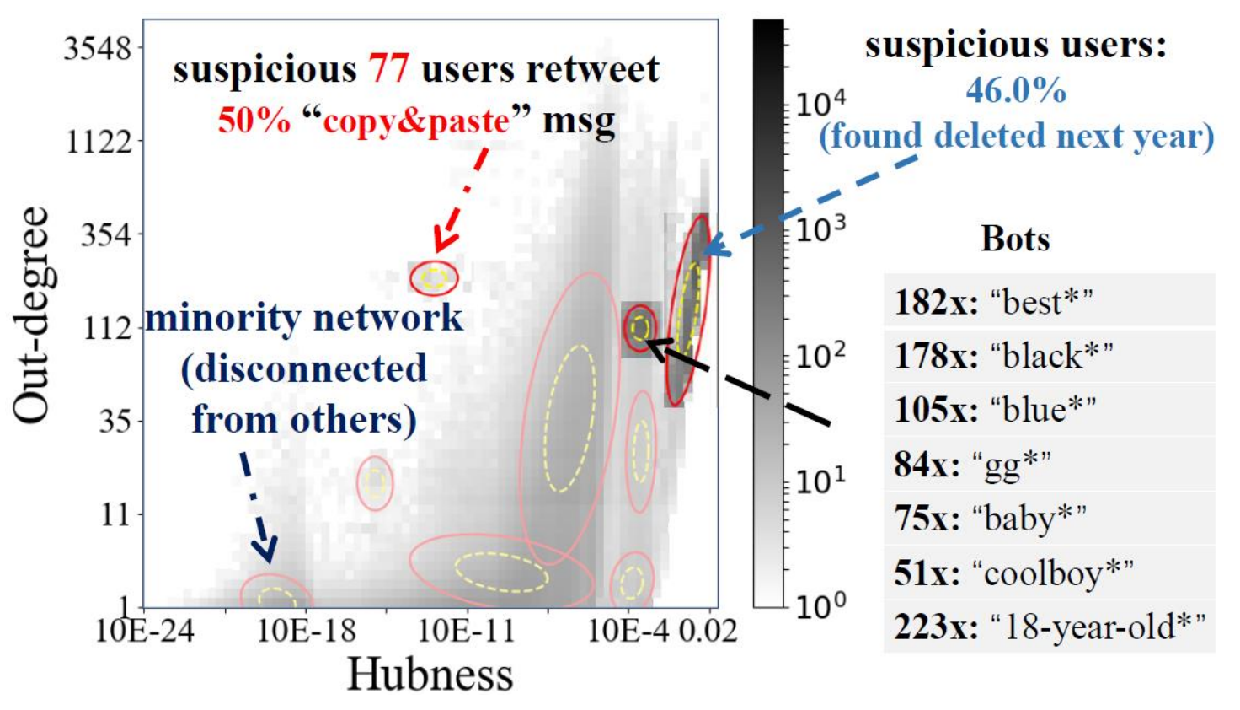


- **How to diagnose very large graph as the healthcare? How to use the vision knowledge in view spaces for patterns mining?**
- **Goal**: For a heat-map of some correlated feature space of graph nodes
  - 1. ***recognize*** and ***monitor*** node groups as human vision does; ▪ 2. ***summarize*** node groups and ***identify*** suspicious micro-cluster.

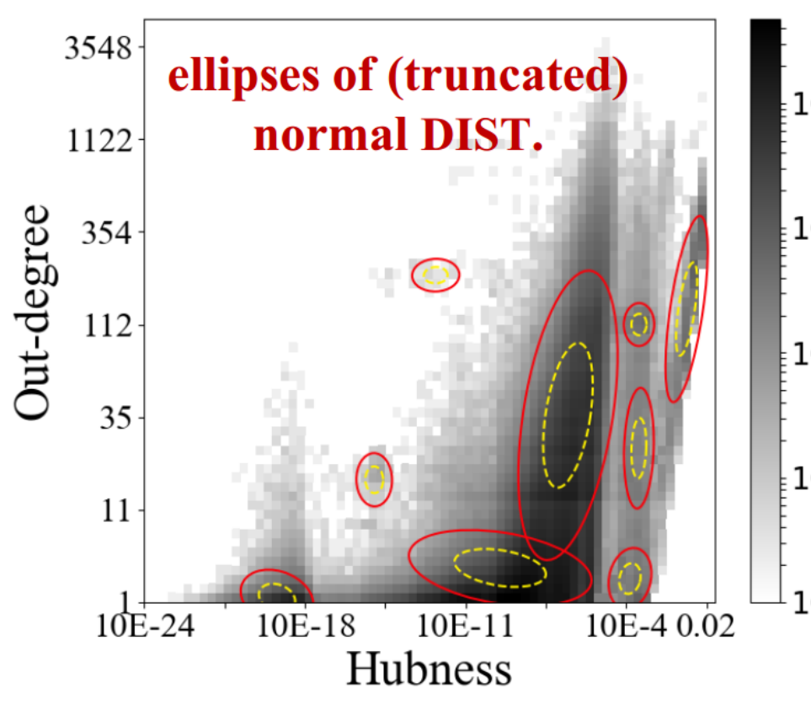
Proposed Model

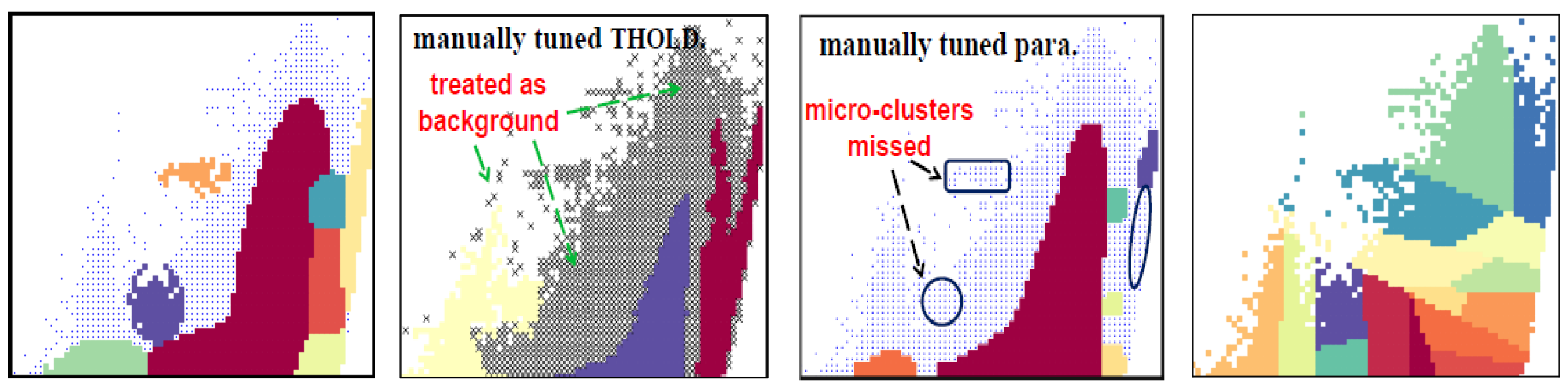
- **1. Graph** $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ (homogeneous / bipartite);
- **2. Correlated features** of nodes. (Degree, PageRank, Spectral, #Triangle, etc.)
- **Goal**: Optimize the **GOF** of node distribution & **consistency** with vision recognition.
- **Histogram** $\mathcal{H}$ of digitalized features, multi-dimensional tensors:
  non-negative value $h_{i_1,\cdots,i_F}$ for the $(i_1,\cdots,i_F)$-th bin.
- **Summarization model** for histogram
  Vocabulary-based summarization model for C node groups
  - **Configurable vocabulary**: distributions $\mathcal{Y}$;
  - **Model parameters**: $\Theta = \{\theta_1,\cdots,\theta_C\}$;
  - **Assignment**: $\mathcal{S} = \{s_1,\cdots,s_C\}$ for each node group;
  - **Outliers**: unassigned bins $\mathcal{O}$.





Bots	
182x:	"best*"
178x:	"black*"
105x:	"blue*"
84x:	"gg*"
75x:	"baby*"
51x:	"coolboy*"
223x:	"18-year-old*"





EagleMine Watershed DBSCAN Gmeans

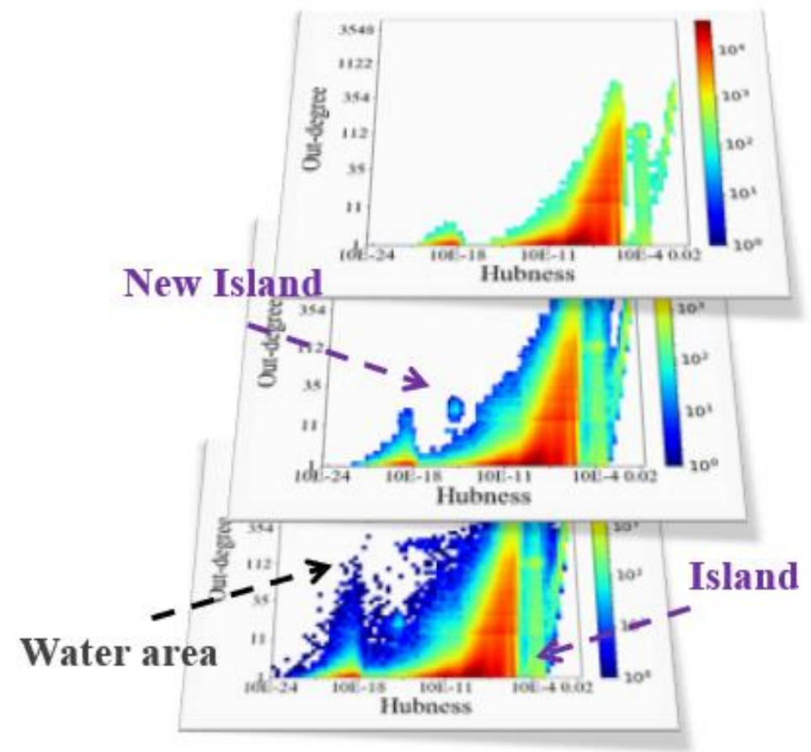
Proposed Method

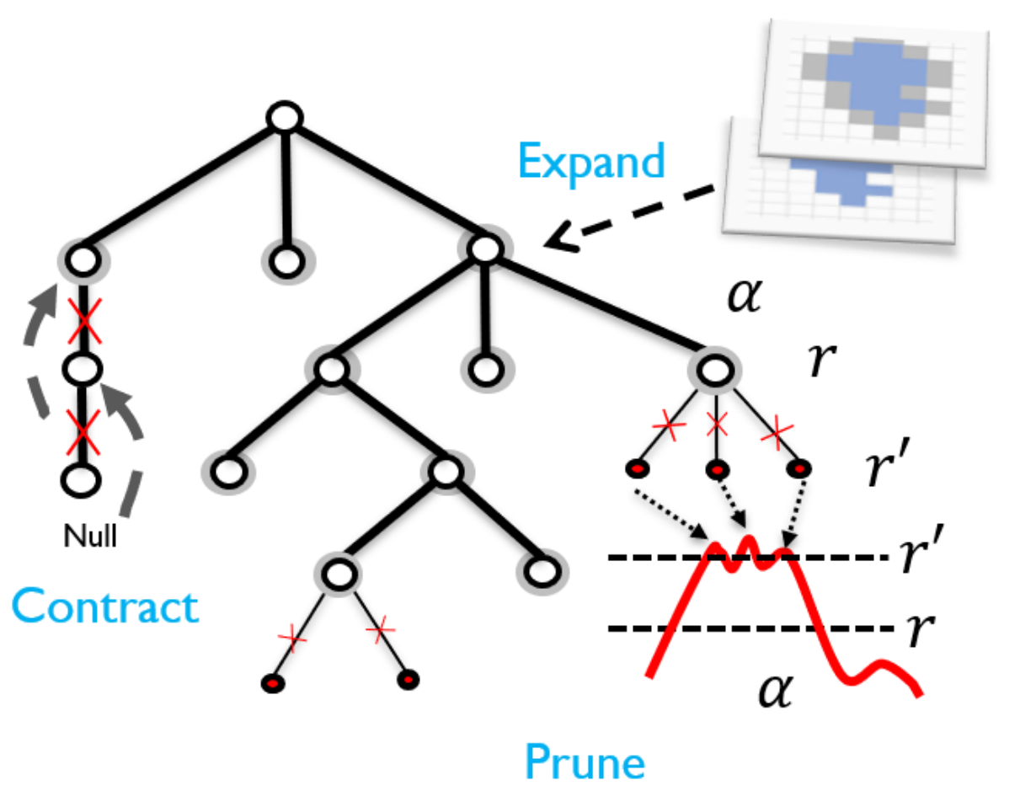
- **Human vision and cognitive system traits:**
  1. **Connected components** can be rapidly detected by eyes;
  2. **Top-to-bottom** recognition and **hierarchical segmentation**;
- **EagleMine ALG.**

Algorithm **Overview structure**

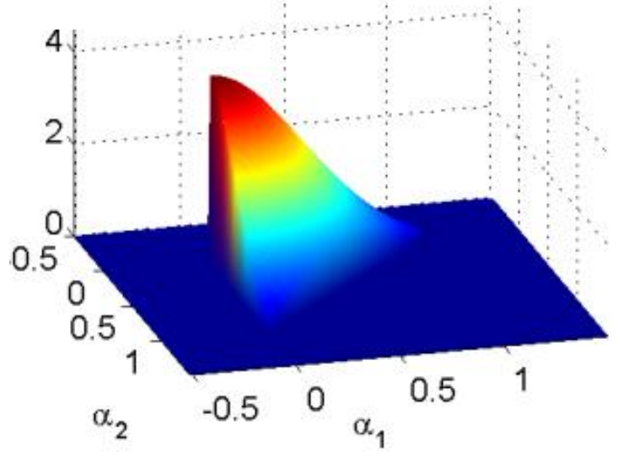
```
Algorithm 1 EagleMine Algorithm
Input: Histogram $\mathcal{H}$ for node features of graph $\mathcal{G}$.
Output: summarization $\{\mathcal{S},\Theta,\mathcal{O}\}$.
  Build a hierarchical tree structure $\mathcal{T}$ for $\mathcal{H}$.
  Describe node of $\mathcal{T}$ with the vocabulary.
  Explore the tree $\mathcal{T}$ and use hypothesis test as metric to determine the best node groups, which are summarized by the model parameters $\Theta$ and the assignment $\mathcal{S}$, as well as the outliers $\mathcal{O}$.
  return summarization $\{\mathcal{S},\Theta,\mathcal{O}\}$.
```

- **Water-level tree** (recognize micro-clusters)
  I. Build waterlevel tree $\mathcal{T}$; II. Refine tree structure;

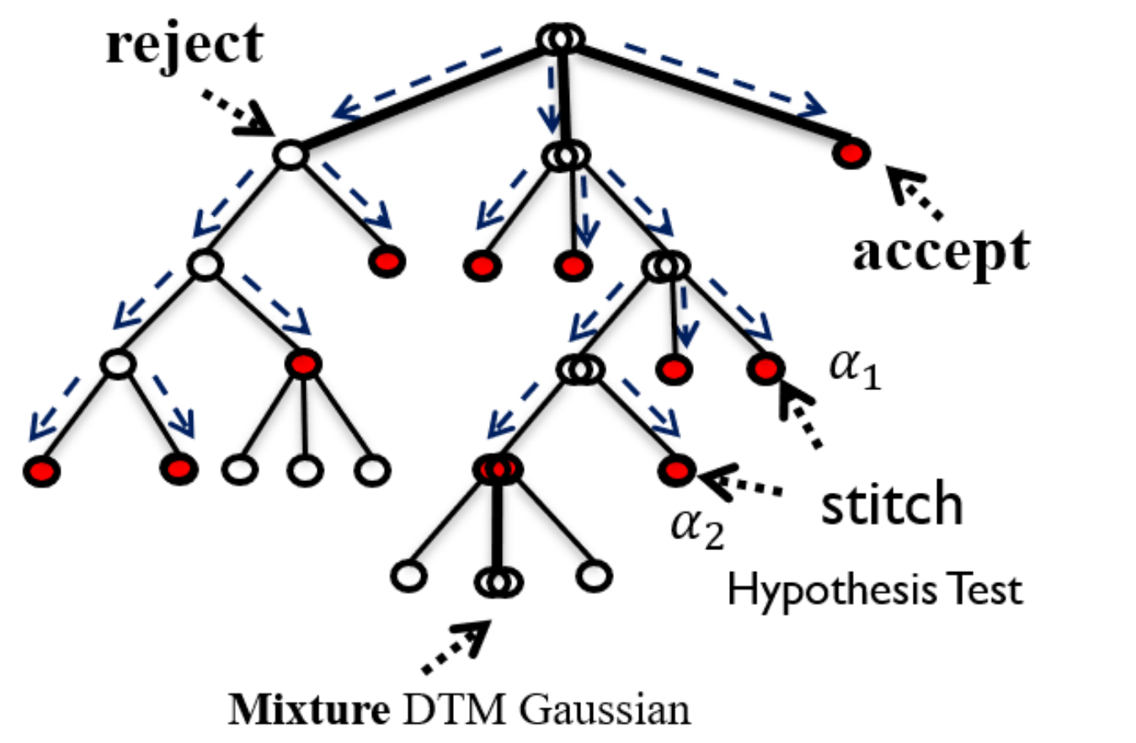




- **Node group summary** (power-specific)
  - *Discretized, Truncated, Multivariate* (**DTM**) gaussian
  - *Mixture* DTM gaussian;



- **Tree exploration**
  1. BFS tree search;
  2. Determine optimal node groups with *Hypothesis Test*;
  3. Islands stitch for enhancing;



- **Micro-cluster Suspicious score**:
  Weighted probability *KL distance* with the majority island.

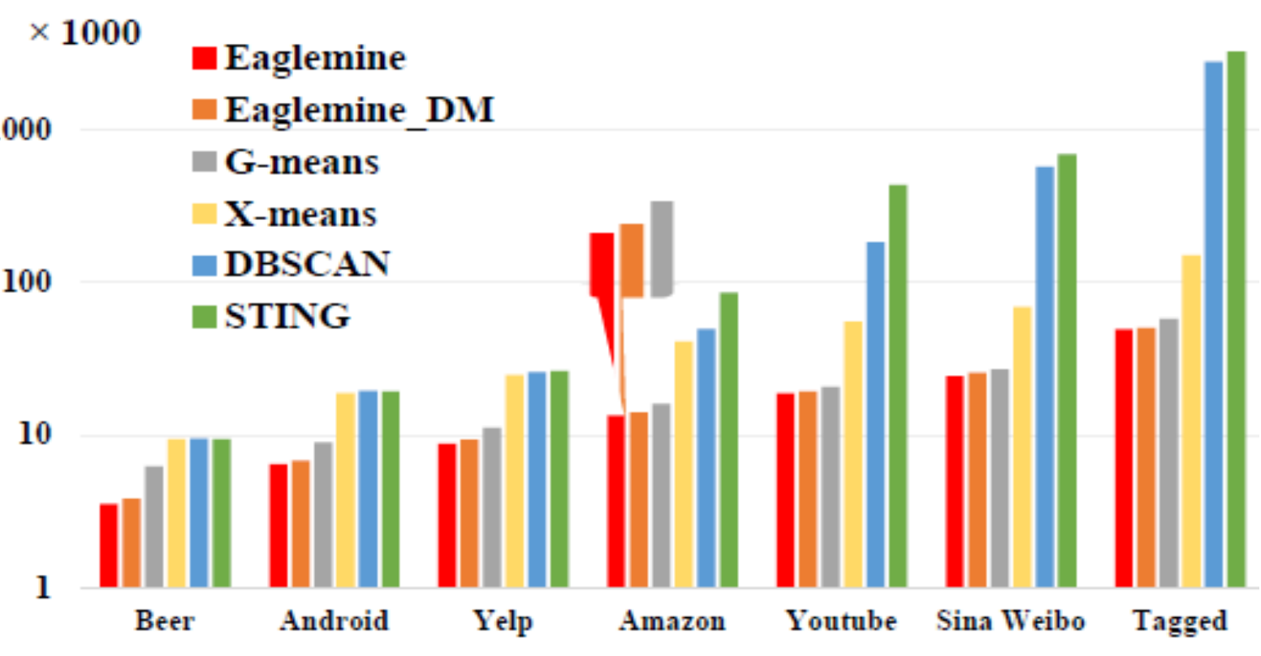
$$\kappa(\theta_i) = \log \bar{d}_i \cdot \sum_{\boldsymbol{b}} N_i \cdot KL(P_{\theta_i}(\boldsymbol{b}) \,||\, P_{\theta_m}(\boldsymbol{b}))$$

- **Time complexity** $O(\frac{\log h_{max}}{\rho} \cdot M + C \cdot T \cdot M)$

M: # of nnz-bin in **H**; T: # of iteration for fitting; ρ: level rising step;

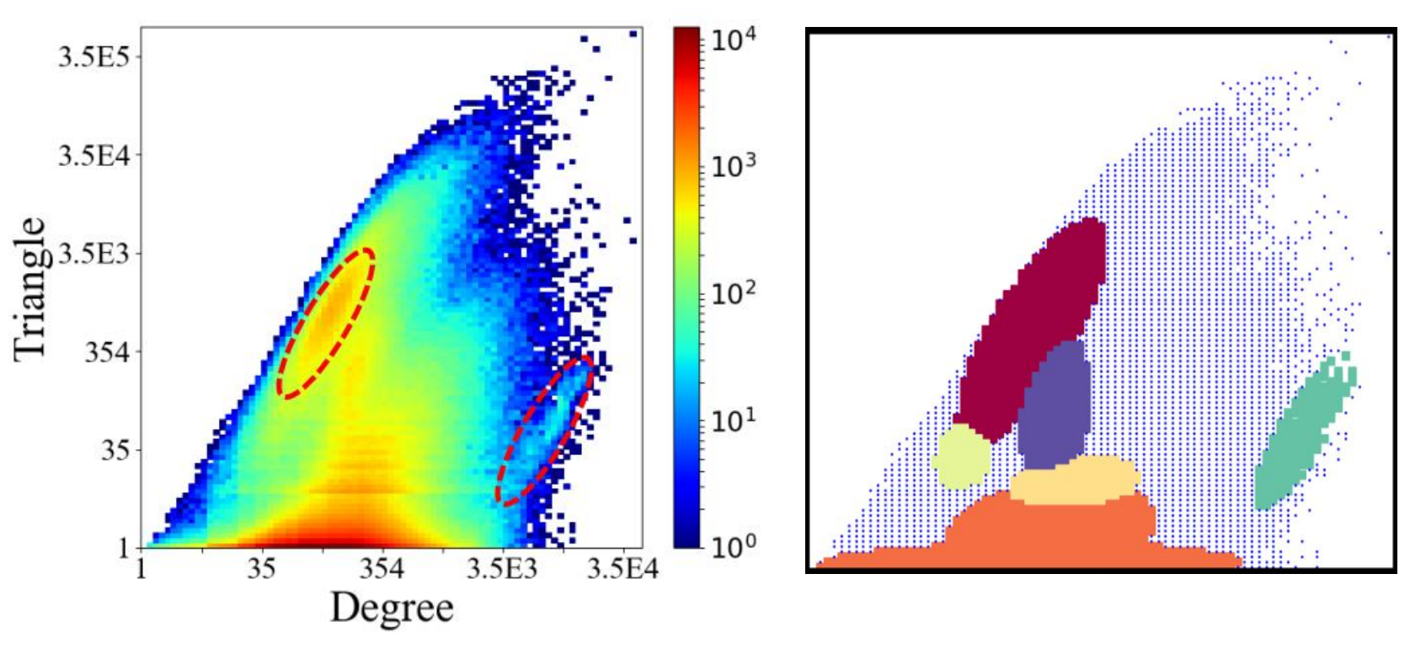
Experimental Results

- **Q1. Quantitative Evaluation**
  EagleMine concisely summarizes the graph nodes distribution in the feature spaces.

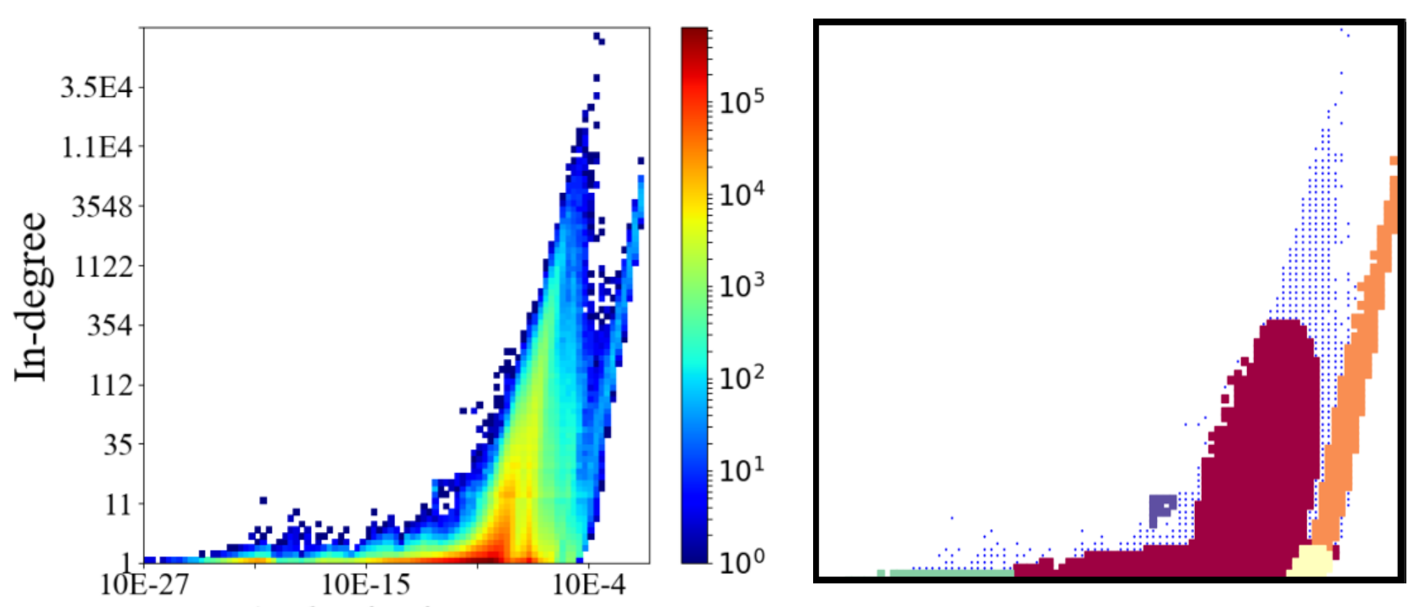


MDL Quantitative Evaluation

- **Q2. Qualitative Evaluation**
  EagleMine accurately identify micro-clusters that agree with human vision judgement.

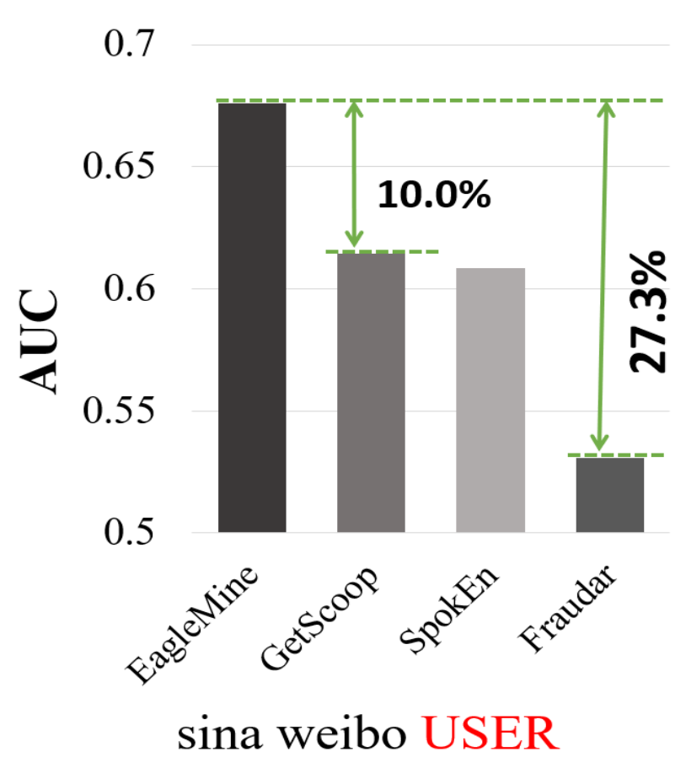


Tagged Friendship

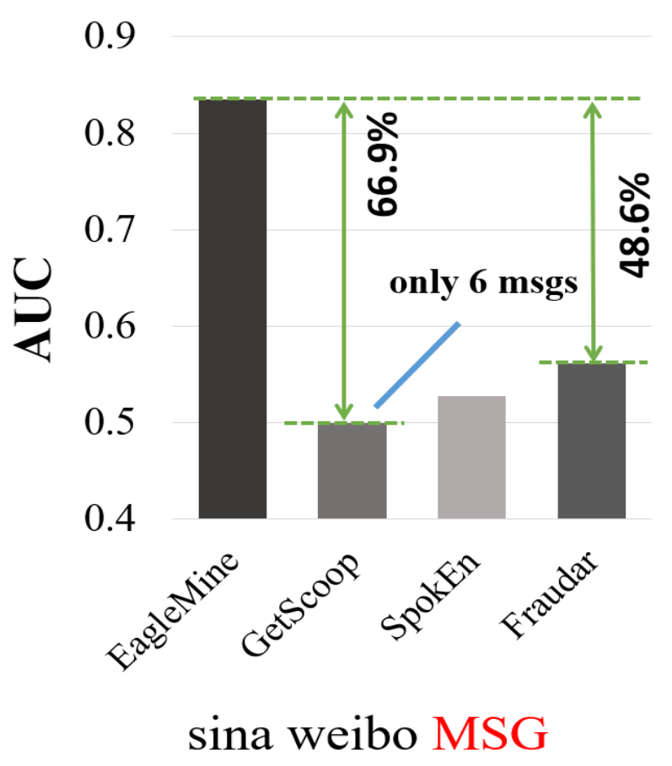


Sina_Weibo msg

- **Q3. Anomaly Detection**
  EagleMine efficiently spot explainable anomaly detection

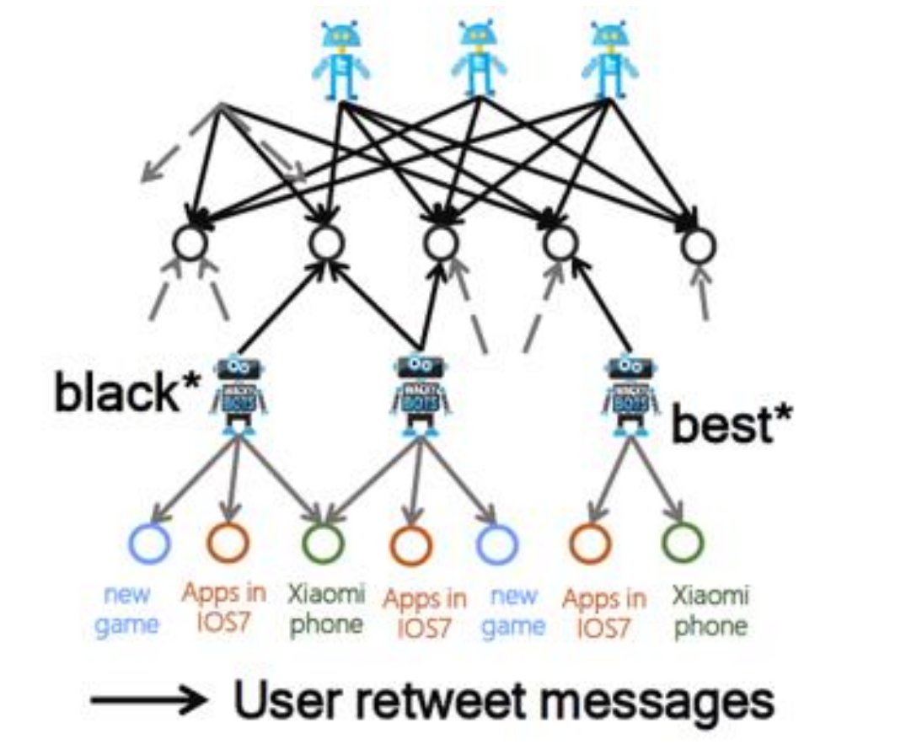


sina weibo USER



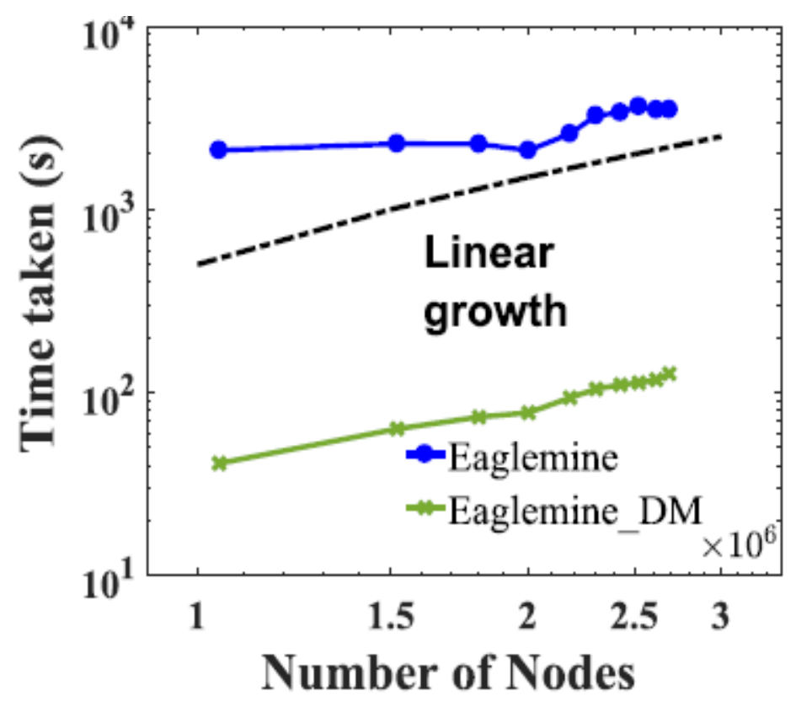
sina weibo MSG

Real-world dataset: sina-weibo
# users: 2.75M, # msgs: 8.08M
# edges: 50.1M



Jellyfish structure anomaly pattern

- **Q4. Scalability**
  EagleMine is scalable with regard to the size of dataset.



Conclusions

- ***Automated summarization*** for histogram of node feature with distribution vocabularies, and find the graph node groups and outliers.
- ***Effectiveness***: achieves better summarization than competitors.
- ***Anomaly detection***: spot explainable anomalies with higher accuracy.
- ***Scalability***: runs linear in # of node, can handle multi-dimensional features.

Code and Data: https://github.com/wenchieh/eaglemine

Main Contact: fengwenjie@software.ict.ac.cn liushenghua@ict.ac.cn christos@cs.cmu.edu