EagleMine: Vision-Guided Mining in Large Graphs

Wenjie Feng1,2, Shenghua Liu1, Christos Faloutsos3, Bryan Hooi3, Huawei Shen2, and Xueqi Cheng1

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

Motivation

- 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 $g=(V, E)$ (homogeneous / bipartite);
2. Correlated features of nodes. (Degree, PageRank, Spectral, #Triangle, etc.)

Goal: Optimize the GOF of node distribution & consistency with vision recognition.

Histogram $H$ of digitalized features, multi-dimensional tensors: non-negative value $h_{i_1, \cdots, i_p}$ for the $(i_1, \cdots, i_p)$-th bin.

Summarization model for histogram
Vocabulary-based summarization model for C node groups
- Configurable vocabulary: distributions $\gamma_i$;
- Model parameters: $\Theta = \{\theta_1, \cdots, \theta_C\}$;
- Assignment $S = \{s_1, \cdots, s_C\}$ for each node group;
- Outliers: unassigned bins $O$.

Proposed Method

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

Algorithm 1 EagleMine Algorithm

1. Build a hierarchical tree structure $T$ for $G$.
2. Describe node of $T$ with the vocabulary;
3. Explore the tree $T$ and use hypothesis test as metric to determine the best node groups, which are summarized by the model parameters $\Theta$ and the assignment $S$, as well as the outliers $O$.
4. Return summarization $(S, \Theta, O)$.

Experimental Results

1. Quantitative Evaluation
   EagleMine concisely summarizes the graph nodes distribution in the feature spaces.
2. Qualitative Evaluation
   EagleMine accurately identify micro-clusters that agree with human vision judgment.

Conclusions

1. Automated summarization for histogram of node feature with distribution vocabularies, and find the graph node groups and outliers.
   Effectsiveness: 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