Network Outlier Detection in Many Small Graphs

-Zesheng Xing

Common Approaches

Patterning Mining and Sim Rank/Edit Distance

- search for common subcomponents (patterns)
- the graph that has none or few frequent patterns is considered as outliers
- NP problem

Graph Kernels with Clustering

- apply graph kernels to obtain similarity matrix
- apply clustering methods
- number of clusters is a hyperparameter
- good to detect community
- hard to detect outliers

Extended Research on Graph Kernels

Our approaches

- convert similarity matrix to classical statistics
- combine and/or compare different graph kernels

Sim Matrix Approach

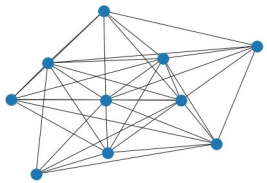
- normalize similarity matrix obtained from graph kernel
- apply dimensional reduction (PCA, t-SNE) to sim matrix
- analyze the result matrix by classical methods (plots, distributions) in target space
- find outliers

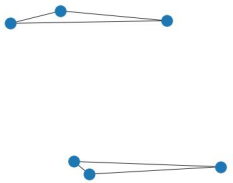
Experiment: Sim Matrix

simulate dataset given different network settings

example:

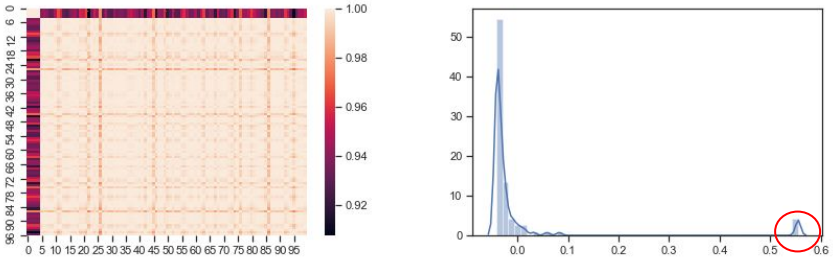
gaussian random partition graph	connected caveman graph
95%	5%



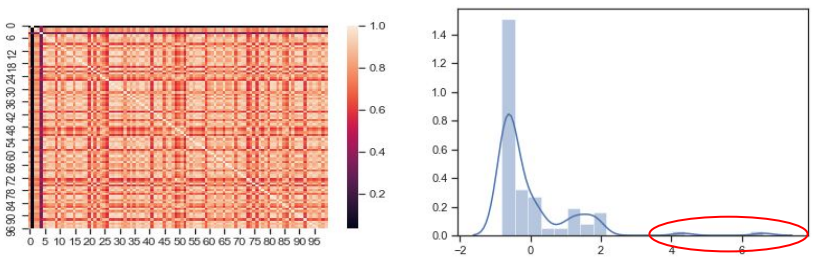


Kernel Performance Example

Shortest Path Graph Kernel [1]



Pyramid Search Graph Kernel [2]



Combine Graph Kernels

Different kernels has their own limits, focus, and advantages

- normalize, weight, and add up similarity matrices obtained from different kernels
- find outliers

Difficulty

- hard to measure the effectiveness of kernels

Attempt to Measure Effectiveness of Kernels

Outliers are hard to define in network settings

- hard to measure if a detected outlier is a desired outlier in reality

Measure the effectiveness of similarity matrix rather than the accuracy of finding outliers

Experiment: Measure Effectiveness

Take out cells randomly from similarity matrix

Recover matrix by completion methods

Calculate the difference(error) between original matrix and recovered matrix

No pattern found between the performance and the error

Conclusion

Sim Matrix method gives a quantitative measure of outliers

- convert to classical distribution problems

Kernel Comparison/Combination fails

- cannot relate graph kernel performance with certain scores
- hard to measure the effectiveness of kernels