Structural Measures of Clustering Quality on Graph Samples

Jianpeng Zhang, Yulong Pei
George Fletcher, Mykola Pechenizkiy

TU/e Technische Universiteit Eindhoven University of Technology

Where innovation starts

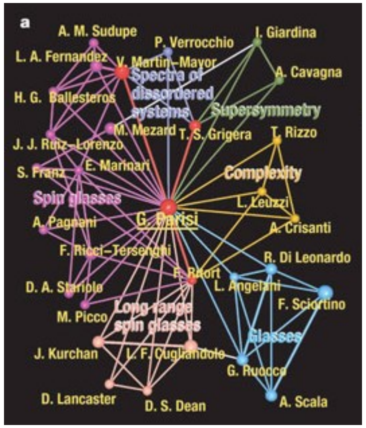
Outline

- Motivation
- Main Issues of Evaluation Metrics
- New Structural Measures
- Experiment and Analysis
- Conclusion & Future Work

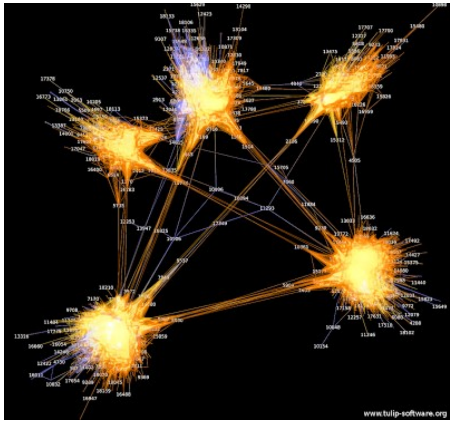
1. Motivation

- **Most graphs in practice are large-scale and/or streaming. They are too large and can not be clustered unless we sample a representative subgraph.**

- **Challenge: how to evaluate the clustering quality in the samples of the graph?**



a) Co-authorship network of the Los Alamos Condensed Matter archive[1]



b) Facebook100 graph of UNC[2]

Technische Universiteit Eindhoven University of Technology

Key Question

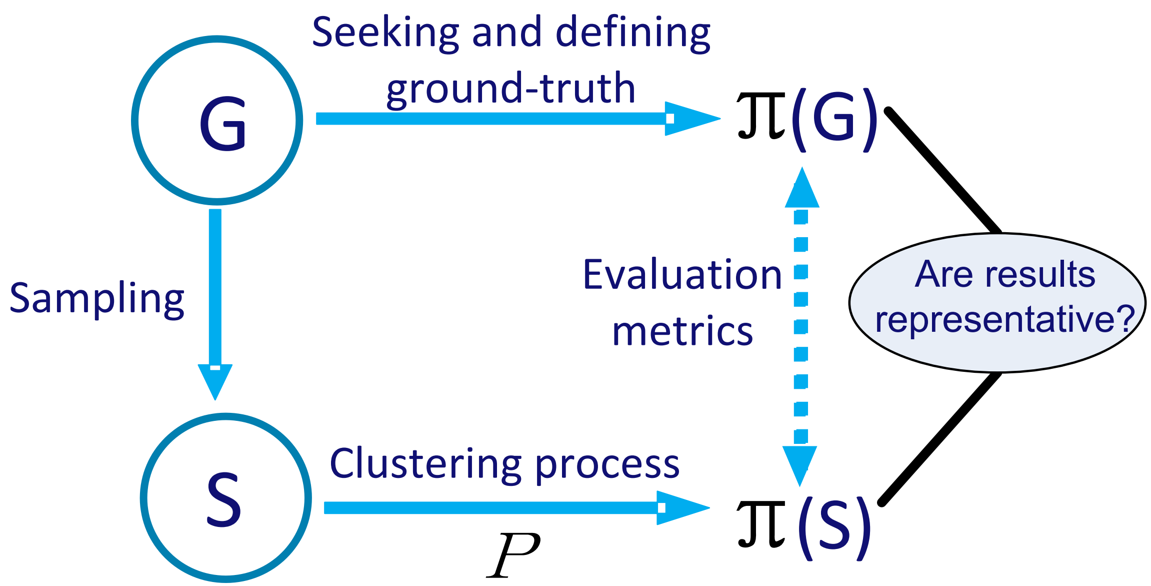


Fig. 1. Problem setting. Let S be a sampled subgraph of a graph G and π(G) be a valid ground-truth clustering. Given clustering π(S) of S induced by process P, what is the quality of π(S) with respect to π(G)?

2、 Main Issues of Evaluation Metrics

(1) Ground-truth clusters

- How to assess the clustering quality by using ground-truth clusters (i.e., common external properties that the members of given clusters share) is an essential procedure.
- Unsupervised quality metrics (e.g., cut-size) are used as metrics of clustering quality, but we are not sure whether the quality metric gives the expected answers compared with the ground-truth clusters.
- Little attention has been paid to evaluation measures for clustering quality on samples of graphs.

2、 Main Issues of Evaluation Metrics

(2) The validity of quality metrics

- Several classic quality metrics were proposed in the literatures. However, there is no consensus on their quality and how well they perform on different kinds of graphs.
- These metrics try to identify good clusters by quantifying the value of metrics which may not be the most meaningful interpretation for what a good cluster is.

Proposed solution: defining two novel quality metrics called σ-precision and σ-recall based on ground-truth clusters.

3、 New Structural Measures

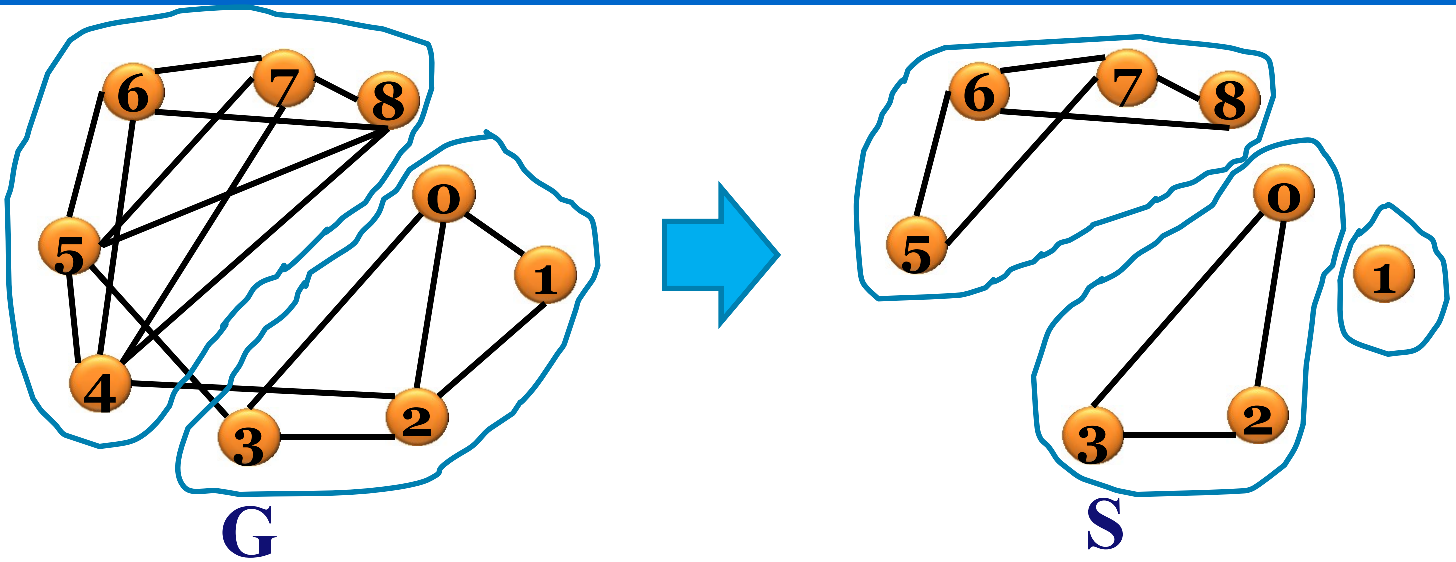
The goal: the clusters in sampling graph S are good representations of the clusters of the original graph G.

1) Firstly, we proposed a basic metric based on the set-matching, i.e., δ-coverage.

$$\delta\text{-}coverage(\pi(G), \pi(S)) = \left| \left\{ b \in \pi(G) \;\middle|\; \exists\, b_S \in \pi(S) \text{ such that } \frac{|b_S \cap b|}{|b_S|} \geq \delta \right\} \right|$$

- **b is a cluster-set of nodes in π(G)**
- **b_s is a cluster-set of nodes in π(S)**
- **δ is a predefined match threshold, e.g., 90%.**

An illustration example



{1}, {0,2,3} ∈ {0,1,2,3}

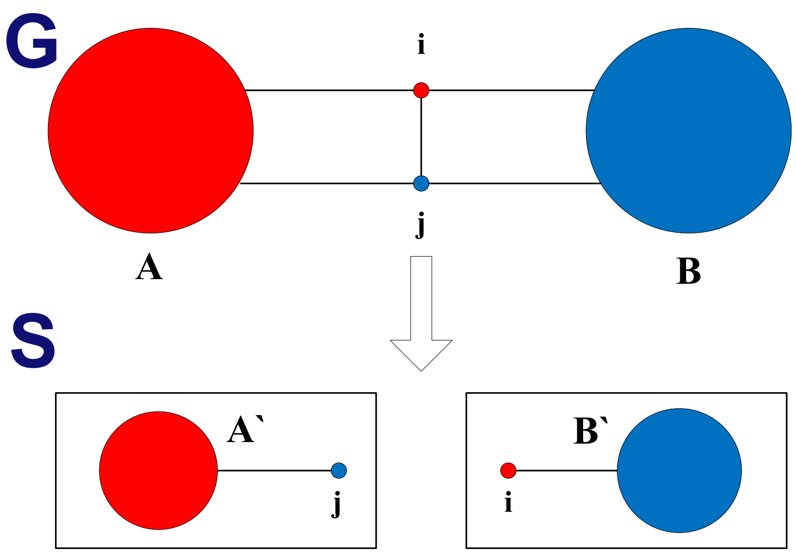
{5,6,7,8} ∈ {4,5,6,7,8}

(δ=1) ⇒ δ-coverage =2

- The coverage of the clustering π(G) is given as the number of clusters in π(G) which are represented by clusters in π(S).
- Higher values of coverage mean the clusters in π(S) are more consistent with and reflective of the ground-truth clusters in π(G).

3、 New Structural Measures

A good example why we need parameter δ?



- Assume A and i belong to one partition block (in red color) while B and j belong to another block (in blue color).
- After sampling and clustering on the sampled graph, we have two clusters: (1) A' and j; (2) B' and i, where A'∈A, B'∈B.
- Intuitively, the cluster blocks in the sampled graph can represent the original graph well

- The main effect of δ-coverage is that the measure is more lenient.
- Based on δ-coverage, we design our new metrics.

3、 New Structural Measures

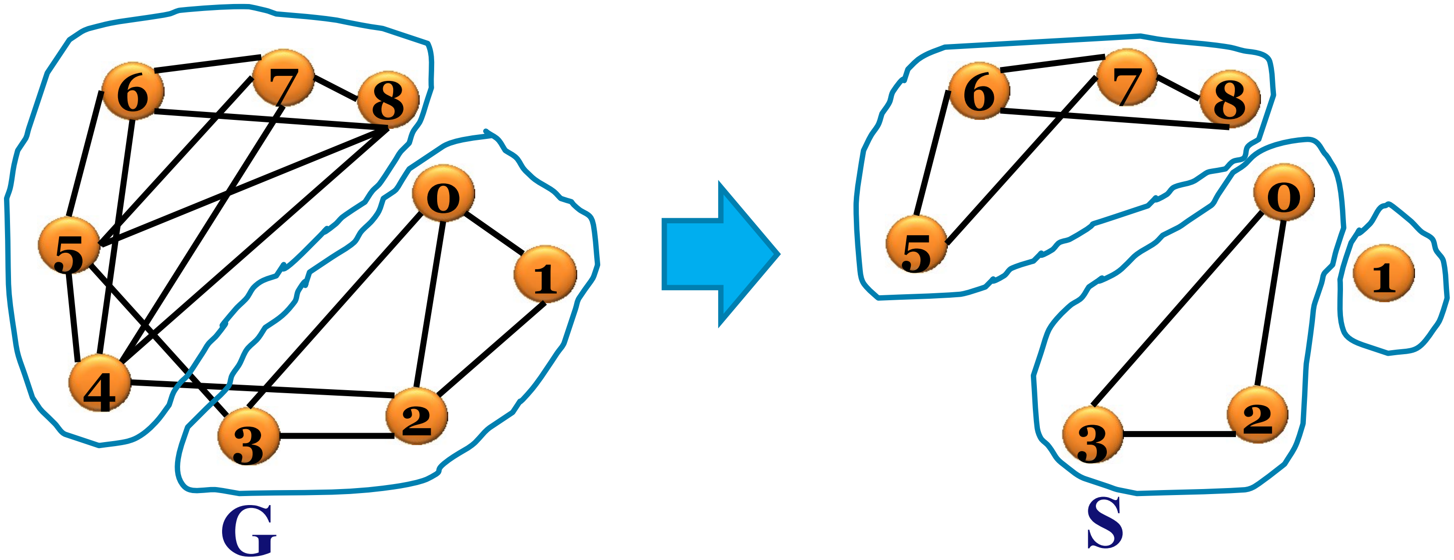
2) δ-precision and **δ-recall**

$$\delta\text{-}precision(\pi(G),\pi(S)) = \frac{|\delta\text{-}coverage(\pi(G),\pi(S))|}{|\pi(S)|}$$

$$\delta\text{-}recall(\pi(G),\pi(S)) = \frac{|\delta\text{-}coverage(\pi(G),\pi(S))|}{|\pi(G)|}$$

- Higher values of **δ-precision** mean that the clusters in S are more precisely representative of the clusters in G.
- Higher values of **δ-recall** indicate that clusters in S more successfully cover clusters in G.

An illustration example



{1}, {0,2,3} ∈ {0,1,2,3}

{5,6,7,8} ∈ {4,5,6,7,8}

δ=1

δ-coverage =2

δ-precision =2/3
δ-recall =2/2

Infer whether results are representative?

Technische Universiteit
Eindhoven
University of Technology

4、Experiments and Analysis

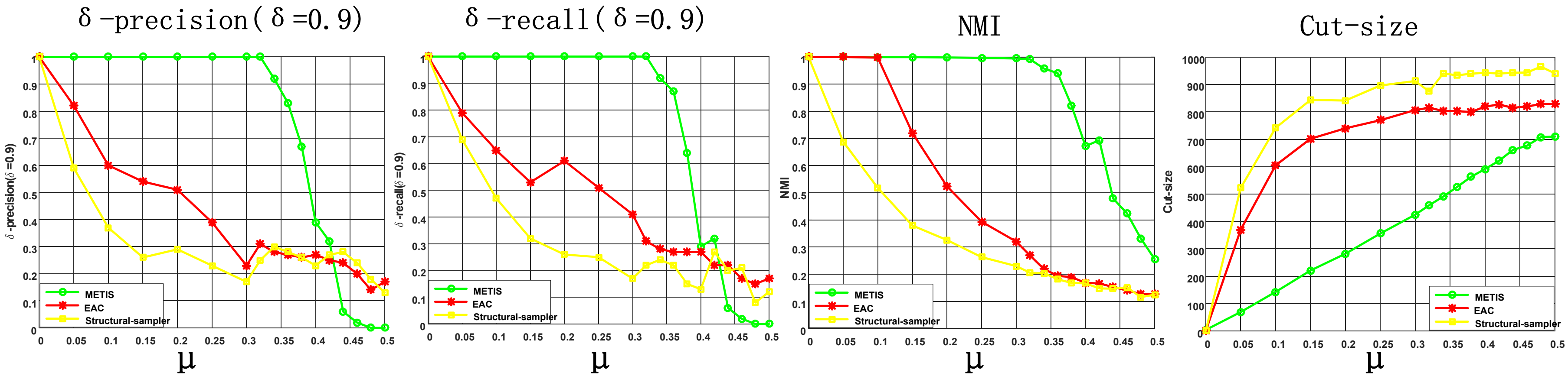
- **Case study: streaming graph(a sequence of consecutively arriving nodes/edges)**
- **Clustering algorithm: Structural-sampler[3], EAC[4] and METIS [5].**

a) **Quality test: we compare our new metrics compared with the classic metrics ,i.e., cut-size and NMI.** We run each algorithm ten times, and then compare the average value of those metrics.

b) **Tuning test**: The task is to **evaluate the impact of the random sampling threshold p on the clustering quality** in sampled graph S. We sample and set all edges with a random value as the same manner as Structural-sampler.

4、Experiments and Analysis

a) Test quality on the benchmark datasets

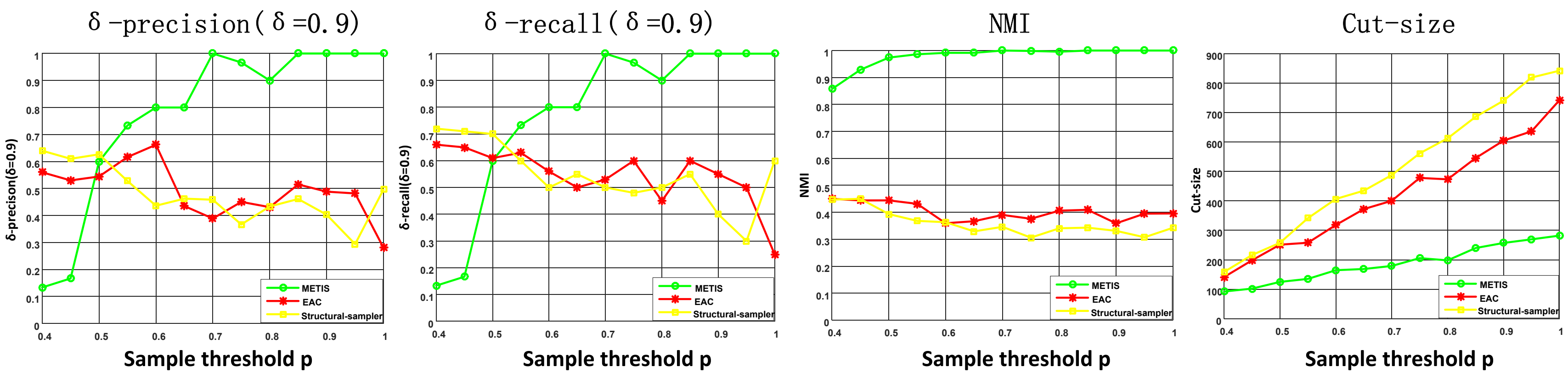


- ✓ Figure shows that when the degree ratio μ changes from 0 to 0.5, the value of the new metrics follows a downward trend.
- ✓ The δ-precision and δ-recall metrics are more insightful expression of clusters structure than the supervised metric NMI.
- ✓ The unsupervised metric cut-size just estimates the edge-cut value between the clusters.

Technische Universiteit Eindhoven University of Technology

4、Experiments and Analysis

b) Tuning test on the benchmark graphs which have relatively distinct cluster structure (μ = 0.2)



- ✓ For the online algorithms, the moderate sampling on the graph with distinct cluster structure makes the cluster structure more clear and obtain higher value of the δ-precision and δ-recall.

- ✓ The cut-size decreases gradually while NMI metric just has a slight increase when the sampling threshold p becomes smaller. They can not capture the structure change appropriately.

Conclusion & Future Work

- We proposed two new structural measures, and they are effectively reflect the match quality of the clusters in the sampled graph with respect to the ground-truth clusters in the original graph.
- The experimental results indicate that classic metrics do not share a common view of what a true clustering should look like.
- Our new metrics have a more insightful results and that could be helpful when used as a complementary standard measures.

Conclusion & Future Work

Our future work

- We want to generalize our evaluation framework and give concrete advise on different sampling strategies and clustering approaches,
- We also plan to extend our study of the fidelity of the evaluation measures.

Thanks

for your attention!!!

Jianpeng Zhang
j.zhang.4@tue.nl