# Effective Evaluation Measures for Clustering on Graph Samples

**Jianpeng Zhang** J.ZHANG.4@TUE.NL
**Yulong Pei** Y.PEI.1@TUE.NL
**George Fletcher** G.H.L.FLETCHER@TUE.NL
**Mykola Pechenizkiy** M.PECHENIZKIY@TUE.NL
Eindhoven University of Technology, 5600 MB Eindhoven, the Netherlands

**Keywords**: graph clustering, evaluation metrics, streaming setting

## Abstract

Due to the growing presence of large-scale and streaming graphs, graph sampling and clustering play important roles in many real-world applications. One key aspect of graph clustering is the evaluation of cluster quality. However, little attention has been paid to evaluation measures for clustering quality on samples of graphs. As first steps towards appropriate evaluation of clustering methods on sampled graphs, in this work we present two novel evaluation measures for graph clustering called *δ-precision* and *δ-recall*. These new measures effectively reflect the match quality of the clusters in the sampled graph with respect to the ground-truth clusters in the original graph. We show in extensive experiments on various benchmarks that our proposed metrics are practical and effective for graph clustering evaluation.

## 1. Problem Definition

Our goal is to measure the quality of clustering processes on graph samples with respect to the original graph. The problem setting is shown in Fig. 1. For a given graph $G$, we sample a subgraph $S$. Given a clustering $\pi(S)$ of $S$, we want to evaluate the quality of $\pi(S)$ with respect to a valid ground-truth clustering $\pi(G)$ of $G$. More precisely, given

- a graph $G = (V, E)$ and a ground-truth clustering $\pi(G)$ of $G$ (i.e., a partition of $V$);
- a sampling strategy which generates a sampled graph $S$ of $G$; and
- a clustering procedure $\mathcal{P}'$, which generates a clustering $\pi(S)$ of sampled graph $S$

we want to measure the quality of the entire clustering process with respect to ground-truth clustering $\pi(G)$.

Such quality measures will guide our understanding and study of improved sampling methods and sample clustering solutions.

In this paper, we specifically propose to measure

- the clustering quality using our new metrics, *δ-precision* and *δ-recall*, compared with state-of-the-art metrics, including the unsupervised metric *cut-size* (Eldawy et al., 2012) and supervised metric *NMI* (Lancichinetti et al., 2008); and,
- the impact of the sampling process on clustering quality.

## 2. Main Issues of Evaluation Metrics

When the network is a small, cluster quality can be evaluated visually and manually. However, as the size of the graph increases, visual evaluation of the clustering result becomes unfeasible. For those cases, evaluation metrics (e.g., *cut-size*) are used as metric of clustering quality which try to capture the most important characteristics expected from a good cluster. The quality of a clustering algorithm is then estimated in terms of the values of those metrics. However, it is difficult to depict whether a given quality metric gives the expected description for the ground-truth clusters embedded in the graph, because typically those metrics based on the *min-cut* theory do not include explicit ground-truth for comparison. So the evaluation

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

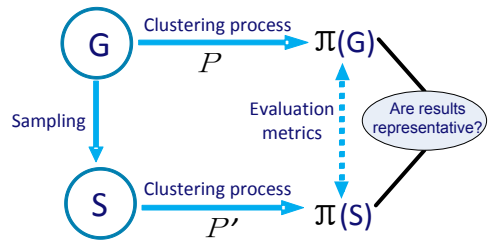



*Figure 1.* Problem setting. Let $S$ be a sampled subgraph of a graph $G$ and $\pi(G)$ be a valid ground-truth clustering of $G$ (i.e, a partitioning of the node set of $G$) induced by process $P$. Given clustering $\pi(S)$ of $S$ induced by process $P'$, what is the quality of $\pi(S)$ with respect to $\pi(G)$?

based on the ground-truth clusters is a significant issue regarding the design of good clustering algorithms. so we design our new metrics based on the ground-truth and the more detail will be given in the next section.

## 3. Novel Quality Metrics: $\delta$-precision and $\delta$-recall

In this section we will define our new metrics and we will also discuss whether those metrics behave consistently with what is expected of good clusterings based on ground-truth clusters, that is, whether or not the clusters in sampling graph $S$ are good representations of the clusters of the original graph $G$. Now we give the definitions of new metrics. Firstly, we define $\delta$-*coverage* as follows:

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

where $\delta$ is a predefined match threshold, e.g., 90%. An intuitive interpretation of $\delta$-*coverage* is the fraction of the clusters in $\pi(S)$ which are valid with respect to the clusters in $\pi(G)$.

Secondly, we define the $\delta$-*precision* of $\pi(S)$ with respect to $\pi(G)$ as

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

The $\delta$-*precision* value ranges from 0 to 1. Higher values of $\delta$-*precision* mean that the clusters in $S$ are more precisely representative of the clusters in $G$.

Similarly, we define the $\delta$-*recall* of $\pi(S)$ with respect to $\pi(G)$ as

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

The $\delta$-*recall* value ranges from 0 to 1. Higher values indicate that clusters in $S$ more successfully cover clusters in $G$.

## 4. Experiments and evaluation

### 4.1. Aim & Methodology

Our proposed evaluation measures are independent of the representation of $G$, but in practice, many large-scale graphs are available in a streaming form, namely, streaming graph. Streaming graph is a sequence of consecutively arriving nodes/edges, but instead of re-clustering the entire graph in each snapshot, the aim of clustering streaming graph is to group the similar vertices in an incremental way which requires to process the graph in one pass under the restriction of limited memory available.

In our experiments, we select several representative streaming clustering approaches as a case study to analyze metric quality. We utilize the standard *LFR* benchmark (Lancichinetti et al., 2008) and it is based on the planted partition model which takes a given ground-truth clustering as input. The benchmark is used to determine how well an algorithm is able to recover the ground-truth clusters. The larger $\mu$ is, the more indistinct the clustering structure of the benchmark graph is. In our study we generate a set of *LFR* benchmark networks with the interval of $\mu$ ranges from 0.1 to 0.5. In our implementation the edges are streamed in the system and the metrics are calculated after all the edges update. Then we applied the new metrics to clusters obtained from three representative clustering algorithms. Three algorithms are used in the experiments. i.e., *METIS* (Karypis & Kumar, 1995), *Structural-sampler* (Eldawy et al., 2012) and *EAC* (Yuan et al., 2013).

### 4.2. Experimental Results

In the experiment, we compare our new metrics with the classic metrics *cut-size* and *NMI*. The experiment results show that the new metrics effectively reflect the match quality of the clusters in the sampled graph with respect to the ground-truth clusters in the original graph. To evaluate our metrics, we compare their results for clusters generated by three different clustering algorithms. The results indicate that classic metrics do not share a common view of what a true clustering should look like, while our new metrics have a more insightful result and provide a better way to evaluate graph clustering algorithms objectively.

## References

Eldawy, A., Khandekar, R., & Wu, K.-L. (2012). Clustering streaming graphs. *IEEE ICDCS* (pp. 466–475).

Karypis, G., & Kumar, V. (1995). Multilevel graph partitioning schemes. *ICPP (3)* (pp. 113–122).

Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical review E*, *78*, 046110.

Yuan, M., Wu, K.-L., Jacques-Silva, G., & Lu, Y. (2013). Efficient processing of streaming graphs for evolution-aware clustering. *ACM CIKM* (pp. 319–328).