Evaluation of the Sample Clustering Process on Graphs

Citation for published version (APA):
https://doi.org/10.1109/TKDE.2019.2904682

Document license:
TAVERNE

DOI:
10.1109/TKDE.2019.2904682

Document status and date:
Published: 01/07/2020

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.
Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 03. Oct. 2022
Evaluation of the Sample Clustering Process on Graphs

Jianpeng Zhang, Yulong Pei, George Fletcher, and Mykola Pechenizkiy

Abstract—An increasing number of networks are becoming large-scale and continuously growing in nature, such that clustering on them in their entirety could be intractable. A feasible way to overcome this problem is to sample a representative subgraph and exploit its clustering structure (namely, sample clustering process). However, there are two issues that we should address in current studies. One underlying question is how to evaluate the clustering quality of the entire sample clustering process. Another non-trivial issue is that multiple ground-truths exist in networks, thus evaluating the clustering results in such scenario is also a challenging task. In this paper, we first utilize the set-matching methodology to quantitatively evaluate how differently the clusters of the sampled counterpart correspond to the ground-truth(s) in the original graph, and propose several new quality metrics to capture the differences of clustering structure in various aspects. Second, we put forward an evaluation framework for the general problems of evaluating the clustering quality on graph samples. Extensive experiments on various synthetic and real-world graphs demonstrate that our new quality metrics are more accurate and insightful for the sample clustering evaluation than conventional metrics (e.g., NMI). Thus the evaluation framework is effective and practical to assess the clustering quality of the sample clustering process on massive graphs.

Index Terms—Graph clustering, graph sampling, multiple ground-truths, evaluation framework, quality metrics

1 INTRODUCTION

Graph (network) clustering [1], [2], [3] is one of the most important research topics in graph mining. Its aim is to group the vertices into clusters with dense intra-cluster link and sparse inter-cluster connectivity. The major challenge of this work is that most networks in practice are large-scale such that analyzing them in their entirety can be intractable and impractical. These real-world graphs might contain an overwhelming amount of vertices and edges. How to process these big graphs effectively is very critical to data scientists. One direction is to design scalable (possibly parallel) clustering algorithms to analyze them. It requires either large working memory (e.g., most in-memory algorithms such as Blondel [4] need to iterate over all the nodes in the memory) or powerful parallel processors (e.g., parallelization implementation). However, such computing resources are not easily accessible in all application scenarios. Another feasible way to circumvent this problem is to take a small sample of the graph and conduct the traditional analysis on the sampled subgraph. Through sampling a representative subgraph, clustering analysis can be performed on the sampled counterpart instead of the original graph, in order to detect the inherent clustering structure.

In this paper, we first illustrate why sampling can be used as a powerful analysis tool for clustering tasks. It is a way to shrink the size of the graphs while still preserving their structural properties (e.g., clustering structure) [5], [6]. There are mainly three reasons that make sampling necessary for graph analysis: (i) the graph is either too massive in size to be handled for inefficient algorithms or difficult to access in the memory of a commodity computer [7]. For instance, Facebook has reported to have more than one billion active users. With 8 bytes for user ID and 100 friends per user, storing the raw edges might take 1 billion*100*8 bytes = 800 GB. In-memory graph algorithms such as Blondel are infeasible for such large dataset unless the extremely large memory is available; (ii) some graphs are represented as streaming graphs [8], [5] (i.e., nodes and edges are added or deleted in an arbitrary order). It needs to re-cluster the whole graph for each update which leads to the expensive computational cost. Thus, sampling from infinite streaming graphs is required (preferably by single-pass under the limited memory); and (iii) when analyzing graphs whose complete structures are hidden from public view (such as friendship networks in social media and connection networks among drug users), it is a severe restriction and most analytical algorithms cannot run on such a restricted setup since they need random access of vertices/edges in the graph [9]. In these scenarios, for reasons of accessibility or efficiency, a sample must be collected, and then topological property of the original graph can be estimated from the sampled counterpart.

Therefore, the focus of this work is to explore the underlying questions of the sampling-based methodology. The first and foremost problem is how to evaluate the clustering quality on samples of graphs. However, little attention has
been paid to how to evaluate the clustering quality on it, which is important not only to measure the effectiveness of the sample clustering process, but also to give insights on the relationships of clustering structure between the original graph and the sampled one. There are many open questions on how to design an appropriate evaluation scheme in order to get credible and comparable results. Therefore, we have a strong motivation to develop a standard evaluation (benchmark) framework to measure the clustering quality on graph samples.

The problem setting is shown in Fig. 1. For a given graph $G$, we sample a representative subgraph $S$ and then generate a set of possible clusterings $\Pi(S)$ of $S$. We aim to evaluate the clustering quality of $\pi(S) \in \Pi(S)$ with respect to the multiple ground-truth clusterings $\Pi(G)$ of $G$ and investigate how well the entire clustering process (namely, sample clustering process) performs. The second issue is that, in practice, there may be more than one valid ground-truth and we may not have access to the authentic ground-truth, e.g., in the context of multiple social network views. A variety of credible clusterings exist in the network. Several reasonable but uncorrelated clusterings might explain the network well from various perspectives. This fact could make the sample clustering process find one ground-truth while the available information in hand gives another ground-truth, thus identifying and evaluating such kind of clusterings is also a challenging task.

Our framework is devised to evaluate the clustering quality of the entire sample clustering process. It is based on supervised metrics [10] (i.e., comparing identified clusters against the ground-truth) to measure the clustering quality, and it is tailored to handle the graphs with multiple ground-truths. To sum up, our main contributions are as follows:

1) We utilize the set-matching methodology to quantitatively evaluate how different the clusters of the sampled graph are with respect to the ground-truth(s) in the original one, and propose several new quality metrics to capture the differences of clustering structure in various aspects.

2) We propose an evaluation framework for the general problems of assessing the clustering quality on graph samples. The framework is based on proposed quality metrics and is capable of evaluating the clustering quality of the sample clustering process. Also, we extend it to handle the cases of multiple ground-truths. Thus we can infer the "optimal" clustering result with respect to multiple ground-truths, and vice versa.

3) We show in extensive experiments on various synthetic and real-world graph that proposed metrics are more accurate and insightful for the sample clustering evaluation, and the evaluation framework is effective and practical to assess the clustering quality of the entire sample clustering process on massive networks.

To the best of our knowledge, this is the first framework to evaluate the clustering quality of the entire sample clustering process which integrates the sampling with the clustering together. However, it is an essential procedure to analyze the massive networks and will guide our understanding and study of improved sampling methods and sample clustering solutions.

The paper is organized as follows. In Section 2 we give the problem definition and then provide a brief review of the background knowledge in Section 3. In Section 4 we address the main issues of evaluating the sample clustering process. We propose new quality metrics and an evaluation framework in Sections 5 and 6, respectively. Experimental evaluation is given in Section 7 and then we discuss the remarkable results in Section 8. The paper concludes with a summary and future work in Section 9.

2 BASIC NOTATION & MOTIVATION

2.1 Basic Notation

Formally, we consider an input network as a graph $G = (V, E)$ with vertex set $V = \{v_1, v_2, \ldots, v_N\}$ and edge set $E = \{e_1, e_2, \ldots, e_M\}$, such that $N = |V|$ is the number of vertices, and $M = |E|$ is the number of edges. We give the description of the entire notations in the Table 1 and some key definitions are described in detail as follows:

**Definition 1 (Graph).** A graph $G$ is defined as $G = (V, E)$ where $V$ is a finite set of vertices and $E \subseteq V \times V$ is a set of edges.

**Definition 2 (Sample).** A sample $S = (V_s, E_s)$ is a sampled subgraph of an original graph $G = (V, E)$, where $E_s \subseteq E$ and $V_s \subseteq V$ such that each vertex mentioned in $E_s$ is in $V_s$.

<table>
<thead>
<tr>
<th>Table 1: Notation of the Graph Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G$</td>
</tr>
<tr>
<td>$V$</td>
</tr>
<tr>
<td>$E$</td>
</tr>
<tr>
<td>$N$</td>
</tr>
<tr>
<td>$M$</td>
</tr>
<tr>
<td>$S$</td>
</tr>
<tr>
<td>$A$</td>
</tr>
<tr>
<td>$p$</td>
</tr>
<tr>
<td>$\Pi(G)$</td>
</tr>
<tr>
<td>$\Pi(S)$</td>
</tr>
<tr>
<td>$\pi(G)$</td>
</tr>
<tr>
<td>$\pi(S)$</td>
</tr>
<tr>
<td>$b_i(G)$</td>
</tr>
</tbody>
</table>
Definition 3 (Clustering). For a graph $G$, a clustering $\pi(G)$ is a finite non-empty set $\pi(G) = \{b_1(G), \ldots, b_k(G)\}$, where each $b_i(G)$ $(i \in [1, k])$ is a subset of $V$, called a cluster of $G$, and it holds that $\bigcup_{b \in \pi(G)} b = V$. Note that the clusters can be either disjoint or overlapping with each other.

Definition 4 (Multiple clusterings). Multiple clusterings for a graph $G$ are a set $\Pi(G)$ of clusterings of $G$, i.e., $\Pi(G) = \{\pi_1(G), \pi_2(G), \pi_3(G), \ldots, \pi_s(G)\}$ where $s$ is the number of multiple clusterings that exist in a graph $G$.

Please note that for the original graph $G$, we can obtain multiple clusterings $\Pi(G)$ from the reliable meta-data of vertices. We treat them as multiple ground-truth clusterings of the original graph $G$. Meanwhile, for the sample $S$, we generate multiple clusterings $\Pi(S)$ which are produced by the clustering process $\mathcal{P}$ on the sample $S$. Nevertheless, their essence is consistent (i.e., they both represent a set of clusterings) and more specific descriptions are given in Table 1.

2.2 Motivation

Our goal is to find a better way to evaluate the entire sample clustering process which integrates the sampling with the clustering process. There are two assumptions: 1) key characteristics (e.g., betweenness centrality, closeness centrality, clustering coefficient, etc.) of the original graph are too costly to calculate for most algorithms within a reasonable time; 2) vertices of a given graph have the meta-data corresponding to the ground-truths to some extent. We aim to evaluate the quality of the clustering result $\pi(S) \in \Pi(S)$ with respect to multiple ground-truth clusterings $\Pi(G)$. More precisely, given

- a graph $G = (V, E)$ and valid multiple ground-truths $\Pi(G)$ of $G$;
- a sampling strategy $\Lambda$ which generates a sampled counterpart $S$ of $G$; and
- a clustering process $\mathcal{P}$, which generates a possible clustering $\pi(S) \in \Pi(S)$ of the sampled graph $S$,

we measure the clustering quality of the entire process under the condition of multiple ground-truths $\Pi(G)$. Such an evaluation framework will guide our understanding and study of improving sampling strategies and sample clustering solutions.

3 Background Knowledge

In this section, we review the background knowledge, including graph sampling, clustering analysis and existing evaluation schemes, which are closely related to the problems we concern about the sample clustering evaluation.

3.1 Graph Sampling

With the advent of large-scale networks, there are several reasons that make it difficult to analyze these networks in their entirety. First, the massive sizes of graphs make them difficult to store and process in the memory and computationally expensive for some inefficient algorithms. Moreover, designing more efficient algorithms and/or leveraging computing power are not always easily available for iterative algorithms [8], [11]. Please note that it also depends on the scenario and computing environment. Several network analytical algorithms were proposed to address the scalability problem [4], [12], [13]. However, these methods are in-memory algorithms and need to iterate all the edges in the memory. It is a big challenge to store and process massive graphs in a single machine’s memory even though in-memory algorithms have the ability to be scalable. For instance, in the Internet network the entire set of user relationships is difficult to store in the memory. These in-memory graph algorithms are simply infeasible for such large dataset unless extremely large memory is available. Second, due to the limitations in data collection mechanism, contemporary graphs are not completely accessible and partial visible to the users [9], [14]. In other words, they can only be accessed through crawling (i.e., a variant of sampling). Thus, graph sampling provides a feasible solution in big-graph analytics and it is an essential procedure for inferring and approximating latent properties such as the clustering structure in large graphs.

All of these factors motivate the need for a better understanding of what is the behavior of the graph sampling on the designated task. Several classic sampling algorithms have been proposed in previous studies [15], [16], [17], [11], [8]. In particular, to preserve the clustering structure, Maiya et al. [11] proposed two sampling algorithms based on the notion that the sample with good expansion properties tends to be more representative of the clustering structure. Newly sampled vertices can be chosen either deterministically or probabilistically. However, they do not preserve the clustering structure well when used in sparse graphs. Simmelian backbones [18] was proposed based on the concept of triadic cohesion. It is capable of extracting structure from a large network such that the intrinsic structure is more pronounced. [6] investigated the visual factors that should be retained to make samples representative in terms of visualization. They studied how sampling strategies influence visualizations of graphs and found that three visual factors have significant effect on the representativeness: cluster quality, high degree vertices, and coverage area. Hence, the challenge of sampling is to create small, but representative samples out of the massive graphs with millions or even billions of vertices and edges. A sampled graph is representative if it preserves selected properties of the original graph. In general, the properties of interest are popular topological properties such as degree distribution, clustering coefficient distribution, cluster sizes, etc. The authors in [11], [8], [19] argue that the measure of representativeness should be consistent with the task being performed. Indeed, we also conjecture that the existing sampling techniques and corresponding measures of representativeness are inadequate for our sample clustering evaluation: how well the genuine clustering structure is represented by the clustering results of graph samples. It further confirms that there is a great need to design new quality measures to assess the representativeness of the clustering structure.

3.2 Graph Clustering

Graph clustering [2] is one of the most important issues of graph mining and a fundamental problem in many fields.
including: bioinformatics, statistics, and data analytics. The
discovery of clustering structures in networks is important
as they often correspond to real social groups, affiliations,
functional properties, etc. A plethora of formalizations exists
for what a good clustering is from different viewpoints, but
the most classical and adopted view is the one based on
the concept of homophily. High homophily means relatively
high internal density and sparse external connections.

However, discovering those edge-dense clusters in graphs
is a very challenging task since graph clustering has been
proven to be a \(NP\)-hard problem [2]. Thus it is very difficult
to apply exact algorithms, which only could be applicable to
small graphs. Thus it is common to apply approximation
rather than exact algorithms to this \(NP\)-hard problem, with a
lower time complexity. Many heuristics have been proposed
to find clusters which maximize both intra-cluster density
and inter-cluster sparsity. Many researchers have studied
several different types of clustering problems, e.g., over-
lapping [1], non-overlapping [4], hierarchical [20], multi-
level [21] clustering, and proposed corresponding clustering
algorithms to handle these problems. However, since those
algorithms are inexact, there is no guarantee that the cluster-
ing results obtained by them are the best ones. There is a vast
amount of research in the field of graph clustering and com-
munity detection, a discussion of them is beyond the scope of
this paper but is detailed surveyed in [2].

3.3 Clustering Evaluation Scheme
The field of clustering on networks suffers from a lack of fair
evaluation that can tell us which algorithms work on a given
type of networks and which don’t. The lack of fair evaluation
is partly caused by the fact that the clustering problem is not
precisely defined [2]. Thus there is no consensus on how they
compare to each other and how well they perform on
different kinds of graphs. To evaluate and compare the clus-
tering quality, there are two evaluation schemes to assess the
clustering results of graphs.

Supervised Evaluation Scheme. Its aim is to evaluate the
clustering with the help of ground-truth information. When
we have the ground-truth available, the goal is to compare
how well a specific clustering process detects the authentic
clusters by using supervised metrics. It is non-trivial to quan-
tify the correctness between the clustering results and the
“gold-standard” ground-truth. Basically, three types of
supervised metrics exist for evaluating the clustering in
graphs [22] [2]. The first is based on Information Theory [23],
measuring the mutual information between two clusterings.
The second is based on Set Matching [24], [22], i.e., mapping
each cluster from one clustering to the most relevant cluster
in the other clustering and computing some variants of preci-
sion-recall metrics [24] or other relevant measures [25]. The
third is based on Pair Counting [22], counting the number of
pairs of vertices which are classified in the same (different)
clusters in two clusterings; and more consensus of pairs indi-
cates that the clusterings are more consistent.

Unsupervised Evaluation Scheme. The scheme is based on
the use of unsupervised quality metrics to evaluate the good-
ness of a given clustering [26], [27], [3]. It simplifies the evalu-
ation task as it turns the inherently vague definition of
clustering into defining a set of desirable qualities for a clus-
ter. This unsupervised scheme has a long tradition in graph
clustering and has been widely used for validating the clus-
tering structure without requiring any additional informa-
tion. Note that for the supervised scheme, clustering
algorithms are judged on their ability to recover the ground-
truth. In such cases, there is a clear definition of rank: the
best clustering algorithm is the one that recovers the most
information. However, judging clustering algorithms based
solely on their ability to recover a “gold standard” clustering
is incomplete. It ignores other properties of clustering, such
as modularity [27], conductance [3] and coverage [2]. Current
studies have given much attention to them in order to deter-
mine the best clustering for a specific application. Usually
they claim to obtain the “best” clustering results according to
the considered unsupervised metrics. However, researchers
found significant differences among the results of different
quality metrics and they do not share a common consensus
about what a good clustering should look like [28], [22]. In
the following section we will give a further discussion on the
difficulties on the existing clustering evaluation.

4 MAIN ISSUES WHEN EVALUATING THE
CLUSTERING QUALITY ON GRAPH SAMPLES
4.1 Multiple Ground-Truths in Networks
The definition of the ground-truth is given by those clusters
which correspond to authentic clusters in the graph, and the
members of the same authentic clusters should share the
common/latent property (e.g., interest, role and affilia-
tion) [29]. Commonly, the clustering algorithm should iden-
tify clusters based on the topological structure and the
extracted clusters should be evaluated based on their com-
mon property, i.e., ground-truth.

In general, the authentic clustering of a given graph
should be divided into three categories: no ground-truth
clustering, single ground-truth and multiple ground-truths.
As Fig. 2 illustrates, if a network is generated randomly, it
does not contain any clustering structure and no ground-
truth clustering is embedded in it.\(^3\) Otherwise, the graph
should have single ground-truth or multiple ground-truths.

\(^3\) Random graphs should not have a meaningful clustering struc-
ture. They can be treated as null graphs and be used to examine
whether algorithms are capable to recognise the absence of clustering
structure. Example graphs can be Erdős Rényi random graphs, the con-
figuration model, etc [2].
Please note that based on the accessibility of the ground-truth(s) in graphs, we subdivide them into accessible ground-truth(s) and inaccessible ones. Concretely, if we can access to the authentic ground-truth(s) of the graph, this is the ideal scenario and we can use them to evaluate the clustering quantitatively. However, it is very difficult (or even impossible) to obtain the authentic ground-truth(s) in real-world graphs, even in small ones. Recent work [30], [29] addresses this non-trivial issue by treating observed meta-data as a proxy of the ground-truth clustering. The meta-data could consist of node’s attributes, e.g., the interest groups of users, the departments of students, the research fields of scholars, etc. Some of meta-data are closely correlated with the authentic clustering structure and we can evaluate the clustering quality with quality guaranteed by using them. Such effort is a step in the right direction, but if the meta-data does not contain any useful information about the clustering structure, the clustering quality is not guaranteed by treating the meta-data as the ground-truth. There are several studies [2], [31], [32] that have investigated and quantified the relationship between the meta-data and clustering structure, and a careful exploration of their relationship with network structure can yield insights of genuine worth.

Based on the statement above, although some mathematical ways to quantify the difference between clustering results and the ground-truth(s) have been proposed, discovering the “optimal” clustering result is still not solved appropriately. Many reasonable but non-correlated clustering results may explain the network well from various perspectives. Meanwhile, the clustering result could reveal one hidden ground-truth but the meta-data represents the alternative ground-truth. Thus, how to evaluate such kind of clusterings in a meaningful manner remains an open research question. Accordingly we propose the formal definition based on the set-matching methodology to evaluate how differently the sample clustering results correspond to the multiple ground-truths quantitatively.

### 4.2 Validity of Clustering Evaluation on Graph Samples

Because of the potential high complexity of graphs and the raw definition of clustering structure according to the intra-cluster density versus inter-cluster sparsity paradigm, the comparison of different clustering approaches is inherently difficult. A large amount of various clustering results might be produced by different clustering approaches or even a single approach with varying parameters. We expect to find the most reasonable clustering result on the sampled graph if the ground-truth information is known.

A large number of quality metrics have been proposed in the literature based on various theoretical foundations [22] [2]. However, they do not share a common consensus for the same clustering. Each of them has its applicable scope and it could fail beyond the scope of the metric itself. So it is uncertain (even impossible to say) which quality metric presents the best choice. Furthermore, we would be interested in reasoning why one sample clustering solution performs better or worse on a given type of networks. How to ensure the validity of the clustering evaluation becomes the key issue of clustering applications. Much research has focused on the comparison of clustering algorithms, but less on the comparison of quality metrics. However, comparing the clustering algorithms under the premise of not confirming the validity of quality metrics is not unarguable [31], [30]. Last but not least, evaluation the clustering quality on samples of a graph is still lacking. It will guide us to better understand the impact of sampling and clustering on big graphs and give insights on the difference of clustering structure between the original graph and the sampled counterpart.

Since current approaches and the quality metrics could not handle these issues appropriately, there is a great need to find other, more accurate ways to evaluate the clustering quality on graph samples. Intuitively, clustering on the sample of a graph can suffer from failures mainly in two aspects. For one thing, it only detects a subset of the ground-truth clusters and leads to low “recall” which means how well the ground-truth clusters are detected. For another thing, it detects most of the ground-truth clusters but several spurious clusters also exist, and it causes low “precision” which means how well the detected clusters correspond to the ground-truth clusters. An ideal evaluation framework would be able to detect each of these deficiencies and give corresponding reasons. For instance, if an algorithm detects many spurious clusters instead of finding the desired ones, this deficiency can be measured and tracked. The main purpose of this paper is to develop an evaluation framework that is able to measure the “precision” and “recall” which need to be precisely defined with the help of ground-truth information. In order to better understand our evaluation framework, first we give the basic definition of new quality metrics.

### 5 Quality Metrics based on the Set Matching Methodology

In this section we will define our new metrics based on set-matching methodology and we will also discuss whether those metrics behave consistently with what is expected of a good clustering based on ground-truth, 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 definition of new metrics.

#### 5.1 Basic Definition

For a better understanding of the new metrics, we first give the basic cover definition, and then we further refine it because it is unable to tolerate any noise in the clusters of \( \pi(S) \). The cover of the clustering \( \pi(G) \) of \( G \), namely, graph-cover is given as the number of clusters in \( \pi(G) \) which are represented by clusters in \( \pi(S) \). In other words, the cover of \( \pi(G) \) by \( \pi(S) \) is

\[
\text{graph-cover}(\pi(G), \pi(S)) = \{ |b(G) \in \pi(G) | \exists b(S) \in \pi(S) \text{ such that } b(S) \subseteq b(G) \}.
\]

(1)

where \( |A| \) denotes the cardinality of set \( A \). The values of graph-cover range from 0 to \( |\pi(G)| \), and higher values mean that the clusters in \( \pi(G) \) are more successfully represented by the clusters in \( \pi(S) \).

In fact, it is easy to find that there exists a hard constraint in the definition of graph-cover for \( \pi(G) \). In extreme cases, there is a bad case shown in Fig. 3 for the graph-cover where
A and B are two sets of vertices, and i and j are two individual vertices. According to the ground-truth of the original graph, A and i belong to one cluster (in red color) while B and j belong to another cluster (in blue color). After sampling and clustering on the original graph, suppose we have two clusters: (1) A' and j; (2) B' and i, where A' ⊆ A, B' ⊆ B. Now there is no exact containment between clusters from the original graph and the sampled one. However, intuitively, the clusters in the sampled graph can represent the original graph well (we assume |A| ≥ |A'| ≥ 1, |B| ≥ |B'| ≥ 1). Thus, we need to loosen the restriction and define a lenient “soft” cover graph.

Therefore, as a first step, we need to define a match function, namely δ-match(·) to calculate the match degree of clusters in π(G) and π(S). Assume that given two clusters b(G) ∈ π(G) and b(S) ∈ π(S), and a predefined threshold δ, the match function can measure the correctness of the relation between cluster b(G) and b(S), and the two clusters are considered a match if δ-match(b(G), b(S)) ≥ δ, i.e., the degree of match is not less than a given threshold. Note that the match function is independent of any specific measure and there are many possible ways to define it. We assume that the match function should hold the properties as follows:

1) self-maximization: δ-match(b(G), b(G)) should have the maximum match for cluster b(G). Likewise, the δ-match(b(S), b(S)) is analogous for cluster b(S).
2) monotonicity: δ-match(b(G), b(S)) monotonically increases with the increase of |b(G) ∩ b(S)|, when |b(G)| and |b(S)| is fixed. Analogously, δ-match(b(G), b(S)) monotonically increases with the decrease of |b(G)| and |b(S)| when |b(G) ∩ b(S)| is fixed.
3) non-symmetry: δ-match(b(G), b(S)) is not necessary to be equal to δ-match(b(S), b(G)), and it is an important feature for the asymmetric function (e.g., in a directed graph the shortest path from vertex a to vertex b may be different from b to a).

Specifically, we can define the match function according to the need for different requirements that hold the properties. Here we give one typical match function to reveal the match degree of the clusters between π(G) and π(S).

5.1.1 Match Function Definition
Let ρ = \frac{|b(G) \cap b(S)|}{|b(G)||b(S)|}, where b(G) is the cluster of π(G) and b(S) is the cluster of π(S). We denote the match function as follows:

δ-match(b(G), b(S)) = \begin{cases} \frac{|b(G) \cap b(S)|}{|b(G)||b(S)|} & \text{if } \rho \geq \delta, \\ 0 & \text{otherwise}. \end{cases}

Note that in this case δ-match(b(S), b(G)) is not equal to δ-match(b(G), b(S)) and it is defined analogously as:

δ-match(b(S), b(G)) = \begin{cases} \frac{|b(G) \cap b(S)|}{|b(S)||b(G)|} & \text{if } \rho \geq \delta, \\ 0 & \text{otherwise}. \end{cases}

where |A| denotes the cardinality of cluster A and δ is a purity threshold. From the constraint item ρ ≥ δ, we can observe that the cluster b(G) and b(S) should satisfy one of two conditions: 1) more than δ percent of the vertices in b(S) come from the same cluster b(G); 2) more than δ percent of the vertices in b(G) are located in b(S). It guarantees that cluster b(G) is a good match with the cluster b(S) under the condition of purity constraint δ.

5.1.2 δ-Graph-Cover and δ-Sample-Cover
Now we can define the “soft” graph-cover, namely δ-graph-cover of π(G), by using the defined match function. Formally, the δ-graph-cover of π(G) by π(S) can be defined as follows:

δ-graph-cover(π(G), π(S)) = \sum_{b(G) \in π(G)} \max_{b(S) \in π(S)} δ-match(b(G), b(S)).

Higher values indicate that clusters in π(G) are more accurately represented and covered by clusters in π(S). Ideally, the maximum value of δ-graph-cover of π(G) should be equal to \( p \ast |π(G)| \) where p is the sample rate of node size, and it means that all the sampled (excluding unsampled) vertices of clusters in the original graph are perfectly grouped and matched by the clustering result π(S) of the sampled graph.

Similarly, the δ-sample-cover of the clustering π(S) is analogous, replacing b(S) with b(G). Formally, the δ-sample-cover of π(S) by π(G) is defined as follows:

δ-sample-cover(π(G), π(S)) = \sum_{b(S) \in π(S)} \max_{b(G) \in π(G)} δ-match(b(S), b(G)).

Higher values indicate that the clusters in π(S) are more consistent with and reflective of the ground-truth clusters in π(G). Note that the δ-sample-cover values of π(S) range from 0 to |π(S)| regardless of the value of the sample rate p.

5.2 δ-Precision and δ-Recall
In order to quantify the cluster structure in π(S) with respect to the ground-truth π(G) in G, we define the δ-precision of π(S) with respect to π(G) as

δ-precision(π(G), π(S)) = \frac{δ-sample-cover(π(G), π(S))}{|π(S)|}.

Fig. 3. A bad case for the basic cover definitions.
\(\delta\)-precision is the average value of \(\delta\)-sample-cover for all the clusters in \(S\), and it ranges from 0 to 1 which implies that in principle the performances of different sample clustering processes can be compared with each other. 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

\[
delta\text{-recall}(\pi(G), \pi(S)) = \frac{\delta\text{-graph-cover}(\pi(G), \pi(S))}{|\pi(G)|}.
\] (7)

\(\delta\)-recall is the average value of \(\delta\)-graph-cover for all the clusters in \(G\). Higher values indicate that clusters in \(S\) more successfully cover clusters in \(G\). Note that the maximum value of \(\delta\)-recall is equal to the sample rate \(p\), which means that the \(\delta\)-recall value is relevant to the sample rate. The larger the sample rate is, the larger the upper bound of \(\delta\)-recall is. Specifically, when we consider the entire graph, i.e., do not sample on the graph, the maximum value of \(\delta\)-recall is equal to 1.

### 5.3 \(\delta\)-Precision-Ground-Truth and \(\delta\)-Recall-Ground-Truth

As we have demonstrated, multiple ground-truths may exist in real-world graphs. Now suppose that the graph \(G\) may have multiple ground-truths, and \(\Pi(G)\) is a set of ground-truth clusterings for \(G\). We already define the \(\delta\)-precision and \(\delta\)-recall to measure which clustering result \(\pi(S) \in \Pi(S)\) is more accurate and reliable with respect to the ground-truth \(\pi(G) \in \Pi(G)\). However, researchers might be concerned that given the clustering result \(\pi(S) \in \Pi(S)\), which ground-truth \(\pi(G)\) will it dig out from the multiple clusterings \(\Pi(G)\) of \(G\)?

In order to quantify the “optimal” ground-truth \(\pi(G)\) in \(\Pi(G)\) with respect to a possible clustering result \(\pi(S) \in \Pi(S)\) of \(S\), we define the ground-truth with maximum \(\delta\)-precision score from the given clusterings \(\Pi(G)\) as \(\delta\)-precision-ground-truth (\(\delta\)-PGT). Likewise, we denote \(\delta\)-recall-ground-truth (\(\delta\)-RGT) in the same manner. Formally we define them as follows:

\[
\delta\text{-PGT}(\pi(S), \Pi(G)) = \arg \max_{\pi(G) \in \Pi(G)} \delta\text{-precision}(\pi(G), \pi(S)),
\] (8)

\[
\delta\text{-RGT}(\pi(S), \Pi(G)) = \arg \max_{\pi(G) \in \Pi(G)} \delta\text{-recall}(\pi(G), \pi(S)).
\] (9)

We are interested in the specific ground-truth that produces the maximum \(\delta\)-PGT or \(\delta\)-RGT rather than the value of itself (hence, we employ \(\arg\max\) instead of \(\max\) above). We can utilize \(\delta\)-PGT (\(\delta\)-RGT) to infer the “optimal” ground-truth \(\pi(G) \in \Pi(G)\) with maximum \(\delta\)-precision (\(\delta\)-recall) value from the given multiple ground-truths \(\Pi(G)\). Note that \(\delta\)-PGT is not necessarily the same with \(\delta\)-RGT for a specific clustering result \(\pi(S)\). Also, since the \(\delta\)-PGT and \(\delta\)-RGT are not immutably with different purity threshold \(\delta\), a feasible solution is to consider the average score of \(\delta\)-precision (\(\delta\)-recall) with varying \(\delta\) and infer the optimal-true-ground based on the overall performance, i.e.,

\[
P\text{GT}(\pi(S), \Pi(G)) = \arg \max_{\pi(G) \in \Pi(G)} \frac{\sum_{\delta \in \Delta} \delta\text{-precision}(\pi(G), \pi(S))}{|\Delta|},
\] (10)

\[
\text{RGT}(\pi(S), \Pi(G)) = \arg \max_{\pi(G) \in \Pi(G)} \frac{\sum_{\delta \in \Delta} \delta\text{-recall}(\pi(G), \pi(S))}{|\Delta|},
\] (11)

where \(\Delta\) is the set of values of \(\delta\) \([0, 1]\).

### 5.4 Additional Statistical Metrics

To further analyze the clustering result on the samples, we also propose two complementary metrics (i.e., ANC and NLS) to give more insightful information to infer how the sample clustering process performs. ANC is in charge of the accuracy for the number of clusters between the original graph and the sampled counterpart while NLS is responsible for how good the sample clustering process is in terms of preserving the clustering structure.

#### 5.4.1 Accuracy for the Number of Clusters (ANC)

The accuracy for the number of clusters is the relative accuracy for the number of clusters between \(\pi(G)\) and \(\pi(S)\), and we can define it as:

\[
\text{ANC}(\pi(G), \pi(S)) = 1 - \frac{\text{abs}(|\pi(G)| - |\pi(S)|)}{\max(|\pi(G)|, |\pi(S)|)}.
\] (12)

where \(|A|\) denotes the cardinality of set \(A\), and \(\text{abs}\) denotes the absolute value of a real number. The ANC value ranges from 0 to 1. Higher values indicate that the number of clusters in \(\pi(S)\) is closer to the number of clusters in \(\pi(G)\).

#### 5.4.2 Normalized L1 Similarity (NLS)

For each cluster \(b(G) \in \pi(G)\), we calculate the ratio of how many vertices in \(b(G)\) are matched by the cluster in \(\pi(S)\), so we define the \(\text{cluster-match-ratio}\) as follows:

\[
\text{cluster-match-ratio}(b(G), \pi(S)) = \frac{\max_{b(S) \in \pi(S)} |b(S) \cap b(G)|}{|b(G)|},
\] (13)

where \(|A|\) denotes the cardinality of set \(A\). The \(\text{cluster-match-ratio}\) value ranges from 0 to 1. A higher value indicates that the matched cluster in \(\pi(S)\) is closer to the cluster \(b(G)\) in \(\pi(G)\).

We visualize the scores by means of \(\text{cluster-match-rank-plot}\), which gives an overview of the match quality of clusters between the sampled counterpart and the original graph. We compute the \(\text{cluster-match-ratio}\) for each cluster \(b(G)\) in \(\pi(G)\) and sort the clusters in order of descending \(\text{cluster-match-ratio}\) scores. We plot \(\text{cluster-match-ratio}\) scores versus the cluster ranks, sorted by \(\text{cluster-match-ratio}\) scores such that the horizontal axis is the relative cluster rank (i.e., the ratio of the cluster’s rank to the number of clusters). Ideally for a sample clustering process, each cluster \(b(G)\) in the original graph \(G\) should be sampled and grouped to the desired size which is equal to \(|b(G)| \times p\) (\(p\) is the sample rate). The distribution should be evenly distributed and the mean value is equal to \(p\).

The \(\text{cluster-match-rank-plot}\) is valuable to compare different sample clustering solutions and to detect which one preserves the original graph’s clustering structure well. Nevertheless, it is also essential to have a single value to quantify
the overall performance. A simple solution is the Normalized L1 Similarity (NLS), which can be defined as follows:

\[ NLS(p(G), \pi(S)) = 1 - \frac{1}{|b(G)|} \sum_{b(G) \in \pi(G)} \frac{|\text{cluster-match-ratio}(b(G), \pi(S)) - p|}{\max(\text{cluster-match-ratio}(b(G), \pi(S)), p)} \]  

(14)

where \( p \) is the sample rate. It means if all of the clusters in \( \pi(G) \) are scaled to the desired size and perfectly matched by the clusters in \( \pi(S) \), NLS should be equal to 1. Note that NLS only considers how well clusters in \( \pi(G) \) are matched by the clusters in \( \pi(S) \), and it inevitably shows a bias toward the clustering result which consists of less clusters, but larger cluster-size in \( \pi(S) \). Thus, it makes more sense to compare different clustering results with the same number of clusters when using NLS metric. Furthermore, we illustrate a tiny graph to evaluate the sample clustering process, please refer to Section 4 in the Supplementary Material [33].

6 EVALUATION FRAMEWORK

Based on the proposed metrics, the procedure of the supervised evaluation framework of the entire sample clustering process is shown in Framework 1.

---

Framework 1. Evaluation Framework of the Sample Clustering Process

**Input:**
- Graph: \( G \);
- The sampling strategy: \( \Lambda \);
- The clustering process: \( \mathcal{P} \);
- The set of multiple ground-truths of \( G \): \( \Pi(G) \).

**Output:**
- Quality metrics: \( \delta \)-precision, \( \delta \)-recall, NLS, ANC and cluster-match-rank-plot;
- The “optimal” ground-truth: \( \delta \)-PGT and \( \delta \)-RGT.

1: Preprocess the graph \( G \), including making edges undirected and removing self-loop, duplicate edges.
2: repeat
3:   Handle each ground-truth clustering
4:   Choose an element (clustering) \( \pi(G) \) from multiple ground-truths \( \Pi(G) \);
5:   Run the sampling strategy \( \Lambda \) employed on the graph and generate the sampled graph \( S \);
6:   Run the clustering algorithm \( \mathcal{P} \) on the sampled graph \( S \) and produce a set of possible clustering results \( \Pi(S) \);
7:   Measure the clustering quality using the proposed metrics according to Eqs. (6) and (14);
8: until any clustering \( \pi(G) \) of the multiple ground-truths \( \Pi(G) \) has been processed;
9: Infer the \( \delta \)-precision(recall)-ground-truth for each clustering result \( \pi(S) \) in \( \Pi(S) \) using Eqs. (8) and (9).
10: Return:
- The values of quality metrics;
- The “optimal” ground-truth: \( \delta \)-PGT and \( \delta \)-RGT.

The framework makes it possible to assess the clustering quality on samples of a graph with respect to multiple ground-truths. Specifically, under this framework, we would like to address the following questions:

RQ1: How does the performance of a specific sample clustering process on the LFR synthetic graphs with the sizes scale up?

RQ2: How does a sample clustering process’s performance on supervised metrics (e.g., \( \delta \)-precision) in comparison with unsupervised metrics (e.g., modularity)? How different is the evaluation ability among those metrics and which one has a strong credibility?

RQ3: How does the sample clustering process affect the clustering structure in the original graph? In other words, what’s the best way to preserve and detect the clustering structure when we sample and cluster on the graph?

RQ4: How can we evaluate the clustering results with regard to the multiple ground-truths if they exist in the original graph?

7 EXPERIMENTS AND EVALUATIONS

This section presents extensive experiments to evaluate the quality of the sample clustering process and attempts to give the answers of those questions mentioned above. First of all, we briefly describe the graphs we used and methodology for the experimental evaluation. Then we present the experiments and discuss the obtained results.

7.1 Datasets

We present a series of experiments on well-known Lancichinetti Fortunato Radicchi (LFR) [34] synthetic graphs and real-world graphs. First, the LFR generator is employed as synthetic benchmarks because it takes a given ground-truth clustering as an input and assumes that both node degrees and cluster sizes have power-law distributions. In LFR benchmarks, a set of parameters need to be given as necessary. The mixing parameter \( \mu \) is vital and reflects the average ratio of external degree to total degree for each node. The larger \( \mu \) is, the less distinct the clustering structure of the benchmark is.

Second, we further evaluate the clustering results on real-world graphs in which the meta-data are available, and the meta-data have been proved to be a good proxy for clustering structure in other studies [13], [35]. They can roughly be classified in two groups: small and large networks. The first five datasets including Karate, Football, Dolphin, Polbooks and Polblogs networks are relatively small. The other graphs are relatively large and they are obtained from the Stanford Large Network Dataset Collection. Besides, we also consider the Facebook university networks (FB100) and Weddell-Sea [30], These networks have multiple ground-truths. Due to the space limit, please refer to the Supplementary Material [33] for more details.

7.2 Experimental Setup

The experiments are carried out on a Linux server running CentOS. Each run employs a single core with 2.40 GHz CPU and 4GB memory. The overall experiments are performed on 2.40 GHz CPU

4. They can be obtained in http://www-personal.umich.edu/~mejn/netdata/

5. FB100 consists of 100 separate networks of Facebook users at US universities from 2005.
and at most 32 GB main memory. We have implemented the evaluation framework by mixing C++ with Python.\textsuperscript{6}

7.2.1 Algorithm Selection
For the sampling part, we produce the graph samples by using different sampling strategies including: Induced Random Edge (IRE) \cite{IRE}, Induced Random Vertex (IRV) \cite{IRV}, Random Walk (RW) \cite{RW}, Metropolis Subgraph (MS) \cite{MS} and Metropolized Random Walk (MRW) \cite{MRW, Metropolized RW}. Among these methods, IRE and IRV are the most widely-used strategies to perform sampling with the assumption that the entire network can be fully accessed in the memory while others also support sampling through one-hop restricted access (i.e., crawling the neighbors of the seed vertex).

For the clustering part, we use five well-known clustering algorithms, i.e., Blondel \cite{Blondel}, Infomap \cite{Infomap}, Modularity Optimization via Simulated Annealing (Mod_opt) \cite{Mod_opt}, Oslob \cite{Oslob} and Label Propagation (LP) \cite{LP}, to extract the clusters on the samples of a graph. These methods come from a variety of different theoretical foundations and are widely used in the domain of clustering. Some of them are designed to detect disjoint clusters while others are capable to support for discovering overlapping clusters. Nevertheless, these algorithms do the same task, i.e., clustering of graphs in the context of the respective domains. We choose Blondel and Infomap because they are the state-of-the-art clustering algorithms, and we take Oslob into consideration because it claims to handle the case of overlapping clusters particularly well. Note that some algorithms return the hierarchical dendrogram, the most fine-grained level of clusterings returned by them are considered, which is in line with the study of \cite{Oslob, Infomap}. More further details on selected sample and clustering algorithms, see the Supplementary Material \cite{Supplementary Material}.

7.2.2 Metrics Selection
We consider several representative metrics widely used for evaluating clusters in the graph and we compare our new metrics with the supervised metrics including: Adjusted Rand Index (ARI) \cite{ARI}, Normalized Mutual Information (NMI) \cite{NMI}, and the variant NMI \cite{variant NMI} proposed by Lancichinetti et al. \cite{variant NMI}. Besides, we also employ three popular unsupervised metrics, i.e., Modularity \cite{Modularity}, Conductance \cite{Conductance}, and Coverage \cite{Coverage} to evaluate the clustering results of the sampled graphs. All of these metrics are normalized which make scores ranging from 0 to 1, where 1 is the optimal score. For the detailed description of each metric, please refer to the Supplementary Material \cite{Supplementary Material}. Note that all of these competing metrics are designed solely to assess the clustering quality on the entire graphs, not particularly on samples. To make them comparable with the proposed metrics, we only utilize them on the subgraph $G(V_s)$ of $G$ (i.e., it has the same vertex-set as the sample).

7.3 Metric Competitiveness Analysis
First, the main objective of this section is to demonstrate that our proposed metrics have the ability to evaluate the pros and cons of the sample clustering process. We consider precision and recall to be the two main aspects of the clustering quality, and each aspect needs to be handled separately without losing the significance of both. Thus, we attempt to evaluate the sample clustering quality for each combination of sampling and clustering methods and show the superiority of the proposed metrics. We sample up to 15 percent of the total number of vertices, and for each sampling strategy, 100 samples are generated by each algorithm on each graph. We compute these quality metrics on each sample, and the experimental results are based on the average values of these metrics.

We found qualitatively similar results on all the LFR benchmarks. Here we only present partial results for LFR benchmarks ($\mu = 0.4$), and results are representative for graphs with different values of $\mu$. Figs. 4 and 5 show the scores of quality metrics for the sampled graphs by using IRV and IRE sampling, respectively. The mixing parameter $\mu$ is set to 0.40 which means the clustering structure of the graphs are not distinct but still can be detected, and the sample rate $p = 0.20$ which is proved to be sufficient to match the graph properties in \cite{LFR}. The rest of plots with different sampling strategies and mixing parameters are revealed in Section 7 of the Supplementary Material \cite{Supplementary Material} and they exhibit similar behaviors. The overall conclusions can be drawn as follows:

- An overall trend is that the clustering qualities on sampled graphs with various sampling strategies keep stable or decrease as the size of the graphs scales up. Also, it shows that sampling strategies have significant influence on the clustering performances. In general, if the sampling strategies such as IRV do not preserve the clustering structure well, then the down-trend becomes more pronounced. As a whole, it verifies that both graph scales and sampling strategies have non-trivial impacts on the performance of the sample clustering process (RQ1, RQ3).

- Our results show that the choice of supervised metrics makes a great difference on the performances of algorithms. There is no consensus on their relative qualities. For instance, in Fig. 4, ARI ranks Infomap the highest while 0.9-precision ranks Blondel the highest. It often occurs when the clustering structure is indistinct or preserved poorly in samples. These results confirm that a more careful study of quality metrics is needed (RQ2).

- For unsupervised metrics, our results show that modularity and coverage are not reliable metrics to evaluate the clustering performance. A sample clustering process’s performance can deteriorate on supervised metrics without impairing them. For example, in Fig. 5 all clustering algorithms return almost similar scores in modularity and coverage. In addition, we find conductance to be an unsupervised metric that has more discrimination ability than others. It further confirms that researchers should carefully claim the best result according to these unsupervised metrics (RQ2).

- Overall, we can observe that the $\delta$-precision and $\delta$-recall metrics are more capable to assess the clustering quality on graph samples compared with other

\textsuperscript{6} The source codes and default parameters of the implementations of all algorithms can be found at https://github.com/feilong0309/EvaluationFramework
supervised metrics. In Fig. 4, the proposed $d$-precision (recall) using different $\delta$ constraints can give different level of insights of the clustering structure. Conversely, the values of some metrics (e.g., coverage) bunch together and we cannot tell the differences of various algorithms regarding small changes. This observation confirms that $d$-precision and $d$-recall have more insights into the clustering structure of the graph (RQ2).

7.4 Evaluation Scenario on Real Graphs with Single Ground-Truth

The previous section shows the competitiveness (effectiveness) of $d$-precision and $d$-recall, and also elaborates that they can give more insightful details about the clustering structure. In this section we need to explore how they perform on real-world networks.

For small real-world graphs, we take a relatively large sample rate (i.e., $p = 0.5$) to show the clustering quality on the samples. However, for large real-world graphs, they are relatively massive in size such that some clustering methods are computationally infeasible in their entirety, so we adopt 15 percent sample rate and analyze clustering results on the obtained samples. For each sampling strategy, we generate 100 samples for small graphs while 10 samples for large graphs, then execute the clustering algorithms and compute quality metrics on each sample. Due to space limitations and the large number of graphs evaluated, we only present partial results for one small graph (Football) and one large graph (DBLP) in Figs. 6 and 7, respectively. They are illustrative for general trends observed in all graphs. We can draw some remarkable conclusions as follows:

- We can observe that small real-world graphs achieve relatively high $d$-precision and $d$-recall scores and the clustering structures of them are detectable by most of the clustering methods on the sampled subgraphs.
- For the large networks, it is much more difficult to obtain satisfactory results. Some of these graphs, e.g., Orkut and Youtube, have only a small percentage of clusters detected fairly well and many detected clusters are not embedded in any of metadata groups, but in some networks such as LiveJournal and DBLP, a good fraction of clusters are well detected. This discovery is in line with the research in the study [29] (RQ2).

We can observe that small real-world graphs achieve relatively high $d$-precision and $d$-recall scores and the clustering structures of them are detectable by most of the clustering methods on the sampled subgraphs. However, for the large networks, it is much more difficult to obtain satisfactory results. Some of these graphs, e.g., Orkut and Youtube, have only a small percentage of clusters detected fairly well and many detected clusters are not embedded in any of metadata groups, but in some networks such as LiveJournal and DBLP, a good fraction of clusters are well detected. This discovery is in line with the research in the study [29] (RQ2).

- The performances of clustering algorithms are comparable in most cases. It is interesting that Blondel and Infomap clustering appears to give better $d$-recall scores in some cases.

Fig. 4. Quality scores of the selected clustering algorithms on the samples of the LFR graphs ($\mu = 0.40$) using induced random vertex sampling (sample rate $p = 0.20$).
scores than other methods in most instances. Meanwhile, Blondel outperforms others in terms of precision nearly for all of our experiments. Oslom and Mod_opt have relatively poor performances in both metrics. However, we found that LP shows wide fluctuations (large standard deviations) and the performance greatly depends on the clustering structure to a great extent. If the clusters are well-defined, LP can perform quite well (see Fig. 7). Otherwise, its performance rapidly deteriorates (RQ3).

Regarding which sample strategies are good for preserving the clustering structure, in general we can observe that IRE and RW sampling have comparably better results than others, and MRW sampling performs the worst almost for each graph when we set the clustering algorithm fixed, and more detailed results of the cluster-match-rank-plot are shown in Supplementary Material [33] (see Figs. S21-S25, available online). This is because IRE and RW sampling are biased towards the vertices of high degree, and it means that the vertices which are located at the center are good representatives of the underlying clustering structure [43]. However, in MRW sampling, each node has the same chance to be sampled into a subgraph regardless of its importance. We believe that this is a striking conclusion. Existence of such cases may motivate researchers to do further analysis to understand the interaction between the sampling design and the clustering structure of the network (RQ3).

7.5 Evaluation Scenario on Real Graphs with Multiple Ground-Truths

We now provide a set of concrete cases to evaluate the clustering quality on networks with multiple ground-truths using our evaluation framework. Our cases are drawn from FB100 networks in which the dorm and graduation year are highly correlated to the clustering structure and the WeddellSea [30] network in which three meta-data: feeding mode, feeding type, zone within the ocean are available. Since different combinations of sampling and clustering algorithms take prohibitively large amounts of time, our evaluation is limited to six networks in prestigious universities from Facebook100 and the WeddellSea network. We execute the procedure outlined in Section 6, and run each sample clustering process 10 trials on each network. Partial results of the evaluation are given in Table 2. We can interpret the results as follows:

8. In our study case, Facebook networks of Harvard, Columbia, Stanford, Yale, Cornell, and Chicago University are employed.

9. Complete results on all the selected Facebook graphs are given in Section 9 of the Supplementary Material [33].
Indeed, we can observe that there exist multiple meaningful clusterings for these networks and the fact that one clustering is uncorrelated with a given meta-data does not rule out the possibility that another could be. For example, in Chicago networks Blondel, Infomap and Oslom algorithms obtain the results which are more consistent with the meta-data of the dorm, whereas the LP algorithm is more in agreement with the meta-data of the graduation year ($\delta = 0.5$). These empirical results emphasize that a lack of correlation between the meta-data and clustering results could possibly exist because of the existence of alternative clustering, which is not correlated with the meta-data we have in hand. Thus, the proposed evaluation framework allows us to analyze the “optimal” clustering result with respect to multiple ground-truths, and vice versa (RQ4).

- By using $\delta$-PGT and $\delta$-RGT, we can infer the ground-truth which is more in line with the clustering result of interest. Generally, empirical results indicate that the meta-data of graduation year is strongly correlated with the clustering structure while the meta-data of
the dorm has a relatively low correlation. These results align with the original analysis of the same datasets by Traud et al [44]. However, it is worth to note that in some networks (e.g., Chicago), the clustering structure is well aligned with both dorm and graduation year metadata, and both of meta-data deem to serve as organizing principles of the network (RQ4).

- It is noteworthy that $\delta$-PGT and $\delta$-RGT are not immutable with the varying purity constraint $\delta$. For one ground-truth, they might obtain relatively low quality scores when $\delta$ is set very close to 1, but they may gain much better results when $\delta$ is more lenient (e.g., $\delta = 0.5$). For another ground-truth, they might keep stable for different values of $\delta$. Thus we can give the overall performance based on Eqs. (10) and (11) to infer the overall "optimal" ground-truth (RQ4).

## 8 Discussion

Our empirical studies indicate that our proposed framework is an essential procedure to analyze large-scale graphs and is practical to evaluate the clusters on graph samples. We can draw some insightful findings as follows:

i) Quality metrics: We examine several unsupervised and supervised quality metrics through a rigorous analysis of the sample clustering process. We found that classic metrics do not share a common view of what a true clustering should look like, and our new metrics give more insightful details and provide a better way to evaluate sample clustering processes objectively. However, a precise understanding of the relationship among these quality metrics is needed.

ii) Clustering algorithms: It is interesting to find that Blondel and Infomap methods perform better than others in most cases. According to our new proposed metrics ($\delta$-precision and $\delta$-recall), they can obtain more than 10 percent better scores on the samples in which the clustering structure is not preserved well by sampling. However, sometimes they lost the ability to approximate an accurate number of clusters (either too much fine-grained clusters or less coarse-grained clusters) using the default setting. It could cause high precision (recall) but low recall (precision) in these scenarios especially when the sample's clustering structure is vague or poorly preserved. Additionally, we observe that the LP algorithm shows the widest variability and the performance greatly depends on the clustering structure to a great extent. It performs poorly when clusters are less clearly defined while it works accurately on graphs with well-defined clusters.

iii) Sampling strategies: We study the impact of varying sampling strategies on the performance of the state-of-the-art clustering algorithms. As is known, there is no sampling algorithm employed explicitly takes the clustering structure into account when performing sampling on the graphs. We conclude that the sampling strategies which are biased towards high-degree vertices such as IRE and RW can preserve the clustering structure better than others while MRW sampling performs the worst. Based on the $\delta$-precision and $\delta$-recall, RW and IRE can obtain approximately 50 percent better clustering results on the sampled subgraphs at the sample rate 15 percent. However, further analysis to understand the interaction between the sampling design and the clustering structure of the network should be addressed. Understanding this interaction may give clues about the feasibility of making inferences about the clustering structure of a given population by using the sample.

## 9 Conclusion

Motivated by the need for better understanding of the clustering process on the samples of a graph, first, we proposed several new quality metrics to quantitatively evaluate the clusters of the sampled graph in various aspects. Second, we outlined a general framework to evaluate the entire sample clustering process. To the best of our knowledge, it is the first framework to evaluate the clustering quality which integrates sampling with clustering simultaneously. The experiments on various synthetic and real-world graphs have shown that $\delta$-precision and $\delta$-recall are good candidates to capture two important aspects (i.e., precision and recall) of the
sample clustering process. Also, the additional metrics, (e.g., β-PGT, β-RGT, ANC and NLS) are insightful for the sample clustering evaluation from different perspectives. Thus the framework is effective and practical to evaluate the clustering quality of the entire sample clustering process on large-scale networks.

In future work, we will explore more possible measurements to quantitatively evaluate the sample clustering process, and give concrete advice on how to choose the sampling strategies and clustering approaches. Also, it is a promising research direction to derive a new sampling approach that take the clustering structure into consideration.

ACKNOWLEDGMENTS

This research of Dr. Zhang is supported by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (Grant No. 61521003) and the National Key Research and Development Program of China (Grant No. 2016YFB0800101). The authors would like to thank Kaijie Zhu, Wilco van Leeuwen, Anil Yaman, and Fuci Chen for a careful reading of the manuscript and many valuable comments.

REFERENCES


Jianpeng Zhang received the PhD degree from the Eindhoven University of Technology in 2018. He is an assistant professor with the National Digital Switching System Engineering & Technological R&D Center (NDSC), China. His research interests include data mining, big data analytics, and social network analysis.

Yulong Pei is working toward the PhD degree in the Department of Mathematics and Computer Science, Eindhoven University of Technology. His current research interests include social network analysis, text mining, and machine learning.

George Fletcher received the PhD degree from Indiana University, Bloomington, in 2007. He is an associate professor with the Department of Mathematics and Computer Science, Eindhoven University of Technology. He studies data-intensive systems, with a focus on the theory and engineering of query languages.

Mykola Pechenizkiy received the PhD degree from the University of Jyvaskyla in 2005. He is a full professor with the Department of Mathematics and Computer Science, Eindhoven University of Technology. He has broad expertise and research interests in data mining and data-driven intelligence, and its application to various (adaptive) information systems serving industry, commerce, medicine, and education.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.