# On Graph Representation for Attributed Hypergraph Clustering (with Appendix)


## Abstract

Attributed Hypergraph Clustering (AHC) aims at partitioning a hypergraph into clusters such that nodes in the same cluster are close to each other with both high connectedness and homogeneous attributes. Existing AHC methods are all based on matrix factorization which may incur a substantial computation cost; more importantly, they inherently require a prior knowledge of the number of clusters as an input which, if inaccurately estimated, shall lead to a significant deterioration in the clustering quality. In this paper, we propose Attributed Hypergraph Representation for Clustering (AHRC), a cluster-number-free hypergraph clustering consisting of an effective integration of the hypergraph topology and node attributes for hypergraph representation, a multi-hop modularity function for optimization, and a hypergraph sparsification for scalable computation. AHRC achieves cutting-edge clustering quality and efficiency: compared to the state-of-the-art (SOTA) AHC method on 10 real hypergraphs, AHRC obtains an average of 13% higher F-measure, 16% higher ARI, 17% higher Jaccard Similarity, 11% higher Purity, and runs 5.5× faster. As a byproduct, the intermediate result of graph representation dramatically boosts the clustering quality of SOTA contrastive-learning-based hypergraph clustering methods, showing the generality of our graph representation.



## CCS Concepts

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

## Keywords

Hypergraph, Clustering, Representation, Sparsification, Modularity, Contrastive Learning

**ACM Reference Format:**

. 2018. On Graph Representation for Attributed Hypergraph Clustering (with Appendix). In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX


## 1 Introduction

Attributed Graph Clustering (AGC) [20] partitions an attributed graph into a collection of disjoint node sets where each node set is called a cluster. In addition to the topological requirement imposed by traditional graph clustering, i.e., nodes in one cluster should be more closely connected to each other than to the nodes in the other clusters, AGC also expects nodes in the same cluster to have similar attributes [7]. Traditional graph clustering can be formulated as optimizations with objective functions such as normalized cut [55], conductance [33], and modularity [5], or addressed by firstly embedding the graph nodes into a vector space using eigenvalue decomposition [62] or graph neural networks [4, 61, 68] (GNNs), and then applying K-Means for clustering. AGC can be reduced to traditional graph clustering by edge re-weighting [49] based on attribute similarity, or by treating each attribute as a node in an augmented graph [77]. Both, as commented in [67], ignore the similarities between nodes that are not directly connected. Alternatively, AGC can be addressed by computing similarities between all pairs of nodes [16], integrating both topological and attributed similarity; such integration can also be achieved in a random walk model [67]. The state-of-the-art quality of AGC is achieved [38, 42] by graph contrastive learning [69], e.g., TriCL [42], which learns unsupervised representations of the graph nodes based on both graph topology and node attributes.

With graph applications engaging more with high-order connections [3], e.g., groups in social networks or author teams in citation networks [64], recent years have witnessed growing research on hypergraphs [3]. Unlike traditional graphs where each edge connects two nodes (thus called dyadic graphs), hypergraphs allow each edge to connect an arbitrary number of nodes (called hyperedges). This paper studies *Attributed Hypergraph Clustering* (AHC), aiming to partition a hypergraph into clusters such that nodes in the same cluster are close to each other with both high connectedness and homogeneous attributes. AHC has wide applications in social community detection [45], metabolic reactions analysis [37], image segmentation [35], and biological analysis [65]; however, existing AHC methods face the following two challenges.

Firstly, existing AHC methods [10, 18, 31, 45] are all matrix factorization based. They require prior knowledge of the number of clusters to produce quality clustering and may incur substantial computation costs. Specifically, the cluster number of a desirable AHC is usually dataset-dependent and unknown in advance, which, if inaccurately estimated, shall lead to a significant deterioration in the clustering quality [63]. Besides, matrix factorization facilitated with Non-negative Matrix Factorization (NMF), Singular Value Decomposition (SVD), or eigendecomposition, leads to substantial computational and memory costs, thereby limiting scalability [31]. The state-of-the-art AHC method AHCKA [45] adopts a greedy iterative method to approximate the eigendecomposition, achieving outstanding performance; however, its algorithm design and performance are still highly sensitive to the cluster number, and the complexity of approximate eigendecomposition may hinder a further improvement on the scalability of AHCKA.




*Conference acronym 'XX, June 03–05, 2018, Woodstock, NY*





ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

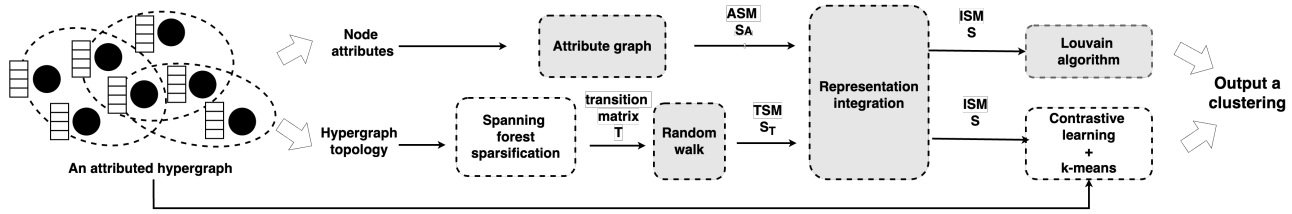



**Figure 1: An overview of our pipeline** AHRC

Secondly, a more effective integration of the hypergraph topology and node attributes is desirable for quality and scalable clustering. Existing works [21, 57, 75] focusing on only one of the two lack a clear pathway for integration. The integration that involves matrix operations such as NMF and SVD incurs high computation costs [10, 18, 31] and is thus not scalable. AHCKA [45] performs a multi-hop random walk where each step has a fixed probability to walk along the attribute graph – the graph where each node $v$ is connected to the nodes whose attribute sets have the largest cosine similarity to that of $v$ – instead of the hypergraph. However, it is unclear why the attribute similarity should be propagated through multi-hop random walks especially when the attribute graph has already considered attribute similarities among all node pairs.

Given the above challenges, this paper considers three questions. Q1) How to integrate the topological and attributed information more effectively to enhance clustering quality? Q2) How to conduct clustering without prior knowledge of cluster numbers while achieving high scalability over large attributed hypergraphs? Q3) Could one generate a graph representation for AHC that is generally applicable, e.g., can improve the clustering quality of existing learning-based methods? We provide positive answers to the above questions with Attributed Hypergraph Representation for Clustering (AHRC), a hypergraph clustering pipeline that achieves cutting-edge efficiency and effectiveness. AHRC comprehensively outperforms the state-of-the-art attributed hypergraph clustering method AHCKA [45]: averaged over 10 real hypergraphs, AHRC obtained 13% higher F-measure, 16% higher ARI, 17% higher Jaccard Similarity, 11% higher Purity, and is 5.5× faster in running time.

Figure 1 overviews the pipeline of our AHRC. Given an attributed hypergraph, AHRC computes an Attribute Similarity Matrix (ASM) $S_A$ and a Topology Similarity Matrix (TSM) $S_T$ to capture the node-wise relationships in terms of attributes and hypergraph topology, respectively, and integrate them into an Integrated Similarity Matrix (ISM) $S$. $S_A$ is derived from the attribute graph while $S_T$ is obtained by firstly sparsifying the hypergraph and then performing a random walk to capture multi-hop relations. AHRC formulates AHC as an optimization on the objective function of multi-hop modularity and then engages the cluster-number-free Louvain for clustering; the intermediated ISM $S$ can alternatively be fed into other clustering methods, e.g., contrastive learning-based clustering method, for general useage. In the design of the pipeline, we find that excluding the attribute similarity from the random walk and our unique presentation integration of the attribute and hypergraph topology are highly effective in enhancing the clustering quality. To make the method scalable, we introduce a sparsification module, which dramatically improves the efficiency without deteriorating the clustering quality. Our contributions are summarized below.

(1) We propose a cluster-number-free AHC method AHRC that represents an attributed hypergraph for clustering by effectively integrating both hypergraph topology and node attributes. The graph representation allows a formulation of a multi-hop modularity as the objective function for optimization and can be of independent and general use in other clustering frameworks.
(2) AHRC adopts spanning forest sparsification to further scale up the pipeline while preserving essential features for clustering.
(3) Extensive experiments justify the outperformance of AHRC over the state-of-the-art (SOTA) AHC methods in both scalability and effectiveness. AHRC is efficient: averaged over all datasets, our AHRC speeds up the SOTA method AHCKA [45] by an average of 5.4× and up to 23×. AHRC is effective: it obtained 13% higher F-measure than AHCKA, 16% higher ARI, 17% higher Jaccard Similarity, 11% higher Purity.
(4) AHRC intermediate graph representation **S** can be of general use. Notable improvements in clustering quality are observed by feeding **S** into the convolutional encoder of two cutting-edge attributed hypergraph contrastive learning models: averaged over 10 real hypergraphs, by using **S**, the best-in-class model TRICL achieved a 21% gain in F-measure, a 107% gain in ARI, a 27% gain in Jaccard Similarity, and an 11% gain in Purity.

The rest of this paper is organized as follows. Section 2 introduces the building blocks of our AHRC: attribute graph construction, hypergraph random walk, and modularity-based clustering. Section 3 presents our proposed attributed hypergraph representation approach. Section 4 describes the sparsification module for scalable computation and contrastive learning for the general use of the graph representation. Section 5 discusses the related work. Section 6 shows the empirical results. Section 7 concludes the paper.

## 2 Preliminary

Let $A$ be a set of attributes. An *attributed hypergraph* $\mathcal{H}(V, E, \text{att})$ has a node set $V$, an edge set $E$ where each edge $e \subseteq V$ is a subset of $V$, and a function $\text{att} : V \mapsto 2^A$ that maps each node $v$ in $V$ to a subset $\text{att}(v) \subseteq A$ of attributes. For each node $v \in V$, define the degree $d_v(\mathcal{H})$ of $v$ as the number of hyperedges in $\mathcal{H}$ that contain node $v$, i.e., $d_v(\mathcal{H}) = |\{e \in E | v \in e\}|$. For a set $C \subseteq V$ of nodes, denote by $\text{vol}_{\mathcal{H}}(C) = \sum_{v \in C} d_v(\mathcal{H})$ the volume of $C$. Denote by $\text{vol}(\mathcal{H}) = \text{vol}_{\mathcal{H}}(V)$ the volume of hypergraph $\mathcal{H}$. Denote by $n = |V|$ number of nodes in $\mathcal{H}$, $m = |E|$ the number of edges, $d = |A|$ the number of attributes. When $\mathcal{H}$ is clear in the context, we denote by $d_v$ the degree of a node $v$ and by $\text{vol}(C)$ the volume of a node set $C$. Denote by $\mathbf{H} \in R^{m \times n}$ the incident matrix of $\mathcal{H}$: for each edge $e_i$, $i \in [m]$ and each node $v_j$, $j \in [n]$, entry $\mathbf{H}[i, j] = [v_j \in e_i]$, i.e., $\mathbf{H}[i, j] = 1$ if $v_j$ is incident to $e_i$ and $\mathbf{H}[i, j] = 0$ if otherwise.

An attributed dyadic graph $G(V, E, \text{att})$ is a special attributed hypergraph where each edge $e \in E$ has exactly two nodes. Represent the graph $G$ as an adjacency matrix $\mathbf{W}$: for two nodes $v_i$ and $v_j$, $\forall i, j \in [n]$, $\mathbf{W}[i, j] = 1$ if there is an edge $(v_i, v_j) \in E$; otherwise $\mathbf{W}[i, j] = 0$. A weighted (dyadic) graph assigns a weight $w(e)$ to each edge $e \in E$, its adjacency matrix has $\mathbf{W}[i, j] = w(e)$ if $e(v_i, v_j) \in E$ and $\mathbf{W}[i, j] = 0$ if no edge in $E$ connects $v_i$ and $v_j$.

**Clique Reduction [39].** Given an attributed hypergraph $\mathcal{H}(V, E, \text{att})$, clique reduction is a standard process that transforms $\mathcal{H}$ to an attributed dyadic graph $G_2(V, E_2, \text{att})$. Specifically, it converts each hyperedge $e \in E$ to a clique of nodes in $e$ and unions the cliques to a dyadic graph $G_2$ with edge set $E_2 = \{(u, v) | \exists e \in E, s.t., u, v \in e\}$. We call $\text{vol}(G_2)$ the dyadic volume of $\mathcal{H}$ and denote it as $\text{vol}_2(\mathcal{H})$. The drawback of clique reduction is the loss of high-order information.

Property 1 (Attributed Hypergraph Clustering [45, 67]). *Given an attributed hypergraph $\mathcal{H}(V, E, \text{att})$, a clustering $\mathscr{C}$ of $\mathcal{H}$, a disjoint partitioning of $V$, is desirable if it satisfies two constraints: 1) nodes in the same cluster are closely connected to each other in terms of structure, while nodes between clusters are structurally separated, and 2) nodes in the same cluster have homogeneous attribute values, while nodes in different clusters may have diverse attribute values.*

**Remarks.** Property 1 shows the high-level objectives of existing AHC methods [45, 67]; however, their solutions assume that the number of clusters $|\mathscr{C}|$ in a desirable clustering is known in advance, which is not valid in reality. This paper focuses on the problem of finding a desirable AHC without a predefined cluster number.

## 2.1 Attribute Graph

To capture the attribute similarities among nodes, the techniques of K-Nearest Neighbor (KNN) search have been widely used [29, 45, 46]. Specifically, given an attributed hypergraph $\mathcal{H}(V, E, \text{att})$ and a parameter $K$, for each node $v \in V$, the $K$ nodes $N_K(v_i)$ with the highest attribute similarity with $v$ are computed. For two nodes $v_i, v_j \in V$, measure their attribute similarity with a cosine-similarity function $f(\text{att}(v_i), \text{att}(v_j))$ over their attribute sets. A straightforward similarity matrix can then be derived as follows:

$$\mathbf{M}[i, j] = \begin{cases} f(\text{att}(v_i), \text{att}(v_j)), & \text{if } v_j \in N_K(v_i) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Because $\mathbf{M}$ is not symmetric, AHCKA [45] constructs a symmetric Attribute Similarity Matrix (ASM) $\mathbf{S_A} \in \mathbb{R}^{n \times n}$ by letting $\mathbf{S_A} = \mathbf{M} + \mathbf{M}^\intercal$ and uses $\mathbf{S_A}$ in the random walk for clustering. While computing the exact KNN graph could take quadratic time, fast approximate KNN algorithm [11, 26] has been adopted on large-scale attributed graphs due to its outstanding efficiency and accuracy. The attribute graph is a weighted graph constructed according to $\mathbf{S_A}$.

## 2.2 Hypergraph Random Walk

Random walk captures multi-hop similarities among nodes in a graph [30]. To preserve the high-order information in a hypergraph, the random walk is conducted in two steps [27]. Step 1 walks from a node $v$ to an edge $e$ chosen uniformly at random from all the incident hyperedges of $v$. Step 2 walks from $e$ to a node $u$ chosen uniformly at random from all the nodes in $e$. Formally, given an attributed hypergraph $\mathcal{H}$ with incident matrix $\mathbf{H} \in \mathbb{R}^{m \times n}$, let $\mathbf{T_V} \in \mathbb{R}^{n \times m}$ and $\mathbf{T_E} \in \mathbb{R}^{m \times n}$ be the row-normalized matrices of $\mathbf{H}^\intercal$ and $\mathbf{H}$, respectively. The transition matrix of Step 1 is $\mathbf{T_V}$ and that of Step 2 is $\mathbf{T_E}$. We call $\mathbf{T} = \mathbf{T_V} \times \mathbf{T_E}$ the **hypergraph transition matrix**.

Based on the hypergraph transition matrix defined above, the random walk with restart process on dyadic graph [60] can be generalized to hypergraph, to capture the multi-hop topology in the hypergraph. Formally, from a node $u \in V$, an $\alpha, \gamma$-**Hypergraph Random Walk** moves in $\gamma$ steps where in each step:

- With probability $\alpha$, terminates at the current node and then jumps back to the source node $u$;
- With probability $1 - \alpha$, transits from the current node $v_i$ to a node $v_j$ based on the hypergraph transition matrix $\mathbf{T}$.

## 2.3 Modularity-based Clustering

On dyadic graphs, a widely adopted line of clustering optimizes the modularity function proposed by Newman-Girvan [50]. Given an unweighted dyadic graph $G(V, E)$ and a random graph model [2] that preserves the degree distribution of $G$, the NG modularity $\text{NG}(C)$ of a subset $C$ of nodes in $G$ is defined as follows.

$$\text{NG}(C) = \frac{|E(C)| - Exp[|E(C)|]}{m} = \frac{2|E(C)|}{\text{vol}(G)} - \left(\frac{\text{vol}(C)}{\text{vol}(G)}\right)^2 \quad (2)$$

where $E(C) = \{(u, v) \in E | u, v \in C\}$ is the set of edges with both ends in $C$. For a clustering $\mathscr{C}$, the modularity for $\mathscr{C}$ is the sum of modularity for each cluster $C \in \mathscr{C}$, i.e., $\text{NG}(\mathscr{C}) = \sum_{C \in \mathscr{C}} \text{NG}(C)$. The NG modularity measures the difference between the actual number of innercluster edges of $G$ and the expected number of innercluster edges of a random graph. A higher modularity score indicates a more pronounced clustering structure: nodes within the same cluster of $\mathscr{C}$ are more closely connected in $G$ than that would be anticipated in a random graph.

Modularity-based clustering is highly popular [5, 13, 21, 32, 51] especially in large-scale graph applications because it requires no prior knowledge of the cluster number, i.e., it decides the cluster number automatically, and moreover, its algorithm, e.g., Louvain [5], achieves both high scalability and clustering quality [74].

Our proposed clustering method is established based on the above building blocks, which will be introduced in Section 3.

# 3 Clustering Attributed Hypergraph

In this section, we introduce the backbone (modules shaded in Figure 1) of the hypergraph clustering pipeline of Attributed Hypergraph Representation for Clustering (AHRC) in two parts. Section 3.1 elaborates attributed hypergraph representation (AHR) which integrates hypergraph topology and attribute information into an Integrated Similarity Matrix (ISM) $\mathbf{S}$. Section 3.2 formulates, based on $\mathbf{S}$, an integrated multi-hop modularity, as the objective function for modularity-based clustering.

## 3.1 Attributed Hypergraph Representation

The topological similarity between nodes in a graph is computed based on $\alpha, \gamma$-Hypergraph Random Walk introduced in Section 2,

which derives the Topology Similarity Matrix (TSM) $\mathbf{S_T} \in \mathbb{R}^{n \times n}$

$$\mathbf{S_T} = \alpha \sum_{l=0}^{\gamma} (1-\alpha)^l \mathbf{T}^l, \tag{3}$$

where entry $\mathbf{S_T}[i, j]$ is the probability that an $\alpha$,$\gamma$-hypergraph random walk from $v_i$ terminates at $v_j$ under hypergraph transition matrix $\mathbf{T}$ defined in Section 2.2. $\mathbf{S_T}$ captures multi-hop topological similarity by considering random walks up-to-$\gamma$ lengths. Specifically, $\alpha$ controls the probability of restarting the random walk from the initial node at each step, balancing local and global topological information. $\mathbf{T}^l$ represents the probability of transitioning from one node to another in exactly $l$ steps. The summation $\sum_{l=0}^{\gamma}(1-\alpha)^l\mathbf{T}^l$ captures the contribution of walks of different lengths (from 0 to $\gamma$ hops) to the overall topological similarity.

The parameter $\gamma$ can be infinite, but it is practically set to a constant for an efficient approximation [45]. Our empirical studies suggest that $\gamma = 2$ strikes a balance between the computation cost and effectiveness and thus is set as a default value. Lemma 1 shows the computational time and space complexities of $\mathbf{S_T}$ when $\gamma = 2$. The proof of Lemma 1 indicates that the main cost in computing $\mathbf{S_T}$ arises from the large dyadic volume $\text{vol}_2(\mathcal{H})$. To mitigate this issue, Section 4.1 will show a sparsification process to reduce $\text{vol}_2(\mathcal{H})$.

Lemma 1. *Given a hypergraph $\mathcal{H}$ and let $\gamma = 2$, the computation of $\mathbf{S_T}$ takes $O(\frac{\text{vol}_2(\mathcal{H})^2}{n})$ time and $O(\frac{\text{vol}_2(\mathcal{H})^2}{n})$ memory space in the average case.*

Proof. Given a hypergraph $\mathcal{H}$, the number of non-zero entries in the transition matrix $\mathbf{T}$ is $O(\text{vol}_2(\mathcal{H}))$ because any two nodes have non-zero transition probability if they have at least one common incident hyperedge. Since $\mathbf{T}$ is a sparse matrix, the time complexity of computing matrix power $\mathbf{T}^2$ is $O(\frac{\text{vol}_2(\mathcal{H})^2}{n})$ [70]. Since the sparse matrix power takes the main computational cost, the overall time complexity of computing $\mathbf{S_T}$ is thus to be $O(\frac{\text{vol}_2(\mathcal{H})^2}{n})$. The space overhead is also determined by the densest matrix $\mathbf{T}^2$. For nodes $v_i, v_j \in V$, we call $v_j$ a 1-hop neighbor of $v_i$ if entry $\mathbf{T}[i, j] > 0$. As there are $O(\text{vol}_2(\mathcal{H}))$ non-zero entries in $\mathbf{T}$, the average number of 1-hop neighbors of a node is $O(\frac{\text{vol}_2(\mathcal{H})}{n})$. Similarly, we call node $v_j$ to be the 2-hop neighbor of $v_i$ if $\mathbf{T}^2[i, j] > 0$. The average number of 2-hop neighbors of a node is thus expected to be $O((\frac{\text{vol}_2(\mathcal{H})}{n})^2)$. Thus, summing up over $n$ nodes, the overall space complexity is expected to be $O(\frac{\text{vol}_2(\mathcal{H})^2}{n})$ in the average case. □

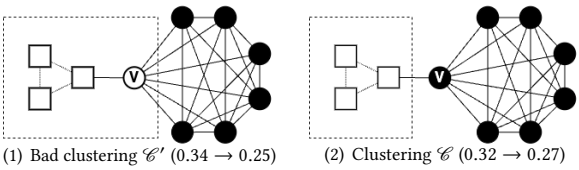
On the other hand, we employ a fast approximate KNN search to construct the attribute graph and the corresponding Attribute Similarity Matrix (ASM) $\mathbf{S_A}$ based on Section 2.1.

**Integrated Similarity.** With Topology Similarity Matrix (TSM) and Attribute Similarity Matrix (ASM), we now compute the integrated similarity between nodes. We first show the steps of the integration and then elaborate on the rationales behind the integration.

Algorithm 1 shows the pseudo code for computing the integrated similarity. Given the topological similarity matrix $\mathbf{S_T}$ and attribute similarity matrix $\mathbf{S_A}$, Line 1 performs row normalizations on both $\mathbf{S_T}$ and $\mathbf{S_A}$ to convert each row into probability distributions. Consequently, entry $\mathbf{S_T}[i, j]$ (resp. $\mathbf{S_A}[i, j]$) denotes the topological (resp. attributed) similarity of $v_j$ from the perspective

**Algorithm 1:** Integrator

**Input:** Topological similarity matrix $\mathbf{S_T}$, attributed similarity matrix $\mathbf{S_A}$
**Output:** Integrated similarity matrix $\mathbf{S}$
1 Compute row normalization matrices $\mathbf{S_T} \leftarrow \text{norm}(\mathbf{S_T})$ and $\mathbf{S_A} \leftarrow \text{norm}(\mathbf{S_A})$;
2 Compute $\mathbf{S'} \leftarrow \mathbf{S_T} \times \mathbf{S_A}$;
3 **for** each entry $\mathbf{S'}[i, j]$ **do** Let $\mathbf{S}[i, j] \leftarrow \sqrt{\mathbf{S'}[i, j]}$;
4 **return** $\mathbf{S}$;



(1) Bad clustering $\mathscr{C}'$ (0.34 → 0.25) (2) Clustering $\mathscr{C}$ (0.32 → 0.27)

**Figure 2: Example on how transformation affects modularity**

of $v_i$. Line 2 defines $\mathbf{S'} \in \mathbb{R}^{n \times n}$ as $\mathbf{S'} = \mathbf{S_T} \times \mathbf{S_A}$ where entry $\mathbf{S'}[i, j] = \sum_{v_r \in V} \mathbf{S_T}[i, r] \cdot \mathbf{S_A}[r, j]$ is the weighted sum of the product of $\mathbf{S_T}[i, r]$ and $\mathbf{S_A}[r, j]$ over all intermediate nodes $v_r$. Line 3 transforms $\mathbf{S'}$ to Integrated Similarity Matrix (ISM) $\mathbf{S}$ by applying a square root transformation [52], i.e., $\mathbf{S}[i, j] = \rho(\mathbf{S'}[i, j]) = \sqrt{\mathbf{S'}[i, j]}$, for each pair $i, j \in [n]$.

**Interpretation.** $\mathbf{S'}$ propagates the multi-hop topological similarity across the attribute graph. Specifically, for two nodes $v_i$, $v_j$, and an intermediate node $v_r$, the topological similarity between $v_i$ and $v_r$ is passed on to $v_j$ if $v_j$ has a similar set of attributes with $v_r$. In other words, if $v_j$ and $v_r$ are similar by nature (attribute-wise), they exchange the information of their topological neighbors in the computation of $\mathbf{S'}$. $\mathbf{S'}[i, j]$ reflects a similarity between $v_i$ and $v_j$ in terms of both topology and attributes. Integrated Similarity Matrix (ISM) $\mathbf{S}$ is eventually computed by applying a square root transformation on $\mathbf{S'}$ for a better similarity distribution. Specifically, in the presence of unbalanced graph structures [73], existing optimization methods adopting objectives (e.g., cut ratio [41], conductance [48] and modularity [22]) empirically does not perform well. In other words, they tend to favor graphs with balanced ground truth clusterings (i.e., each cluster has similar volume). By applying the concave square root function, large values become less influential, leading to a more even distribution of similarities and consequently, a better clustering quality [23]. The choice of the smooth function is not exclusive; we conducted experiments (Exp 6 in Section 6) which suggests square root function is an ideal candidate.

Example 1. *To illustrate the impact of the square root transformation on modularity-based clustering, consider Figure 2. The graph consists of a 3-clique and a 7-clique, connected by a single edge. Naturally, each clique would be its own cluster. Without the transformation, node $v$ might be incorrectly assigned to the 3-clique cluster because the clustering $\mathscr{C}'$ in Figure 2 (1) has a higher modularity ($0.34 > 0.32$). However, after applying the square root transformation, the modularity of $\mathscr{C}'$ decreases significantly to 0.25, making it lower than that of the true clustering of $\mathscr{C} = \{3\text{-}clique, 7\text{-}clique\}$ in Figure 2 (2).*

Lemma 2 shows the time and space complexities of Algorithm 1.

Lemma 2. *Given a topological similarity matrix $\boldsymbol{S_T}$ and an attribute similarity matrix $\boldsymbol{S_A}$, Algorithm 1 takes $O(\frac{K\cdot \mathrm{vol}_2(\mathcal{H})^2}{n})$ time and $O(\frac{K\cdot \mathrm{vol}_2(\mathcal{H})^2}{n})$ memory space.*

Proof. We first prove that the space complexity of $\mathbf{S}$ is $O(K \cdot \mathrm{vol}_2(\mathcal{H})^2)$, where $K$ is the parameter for the KNN algorithm. Since the average number of non-zero entries per row in sparse matrices $\mathbf{S_T}$ and $\mathbf{S_A}$ is $\frac{\mathrm{vol}_2(\mathcal{H})^2}{n^2}$ and $K$, respectively, a node can access $O(\frac{K\cdot \mathrm{vol}_2(\mathcal{H})^2}{n^2})$ number of nodes on $\mathbf{S_T}\times\mathbf{S_A}$. The number of non-zero entries in $\mathbf{S}$ (the memory cost) is thus $O(\frac{K\cdot \mathrm{vol}_2(\mathcal{H})^2}{n})$.

Then, we prove the time complexity. Line 1 normalizes matrices $\mathbf{S_T}$ and $\mathbf{S_A}$, taking $O(\frac{\mathrm{vol}_2(\mathcal{H})^2}{n}+K\cdot n)$ time. Since both $\mathbf{S_T}$ and $\mathbf{S_A}$ are sparse matrices with $O(\frac{\mathrm{vol}_2(\mathcal{H})^2}{n})$ and $O(Kn)$ numbers of non-zero entries, respectively, the complexity of Line 2 is $O(\frac{K\cdot \mathrm{vol}_2(\mathcal{H})^2}{n})$ [70]. Line 3 performs a transformation on $\mathbf{S}$, taking $O(\frac{K\cdot \mathrm{vol}_2(\mathcal{H})^2}{n})$ time. Overall, Algorithm 1 takes $O(\frac{K\cdot \mathrm{vol}_2(\mathcal{H})^2}{n})$ time. □

**Remarks.** We represent the attributed hypergraph with $\mathbf{S}$ which combines both graph topology and attribute information. Specifically, consider two nodes $v_i$ and $v_j$. If $v_i$ and $v_j$ are closely connected topologically and share homogeneous attributes, the probability of a random walk connecting them should be high, leading to a large value of $\mathbf{S}[i,j]$. Conversely, if $v_i$ and $v_j$ are distant with dissimilar attributes, and there is no node that is similar to $v_j$ (in terms of attributes) and topplogically close to $v_i$, the value of $\mathbf{S}[i,j]$ should be small. Section 3.2 defines an integrated multi-hop modularity for clustering.

## 3.2 Integrated Multi-hop Modularity

The definition of NG modularity fails to capture the constraints of AHC (as described in Property 1), as it considers neither multi-hop topology nor attribute information. To address this issue, we propose an objective function, called Integrated Multi-Hop Modularity (IMM). Specifically, given an attributed hypergraph $\mathcal{H}$, Section 3.1 computes a similarity matrix $\mathbf{S}$ using our proposed attributed hypergraph representation (AHR). Regard $\mathbf{S}$ as the adjacency matrix of a weighted dyadic graph where entry $\mathbf{S}[i,j]=0$ indicates there is no edge between nodes $v_i$ and $v_j$. We call this weighted dyadic graph the *representative graph* of $\mathcal{H}$, on which we define the integrated multi-hop modularity, denoted as IMM, as follows.

Definition 1 (Integrated Multi-Hop Modularity). *Given an attributed hypergraph $\mathcal{H}$, a clustering $\mathscr{C}$, and a similarity matrix $\boldsymbol{S}$ under the* AHR *model, the integrated multi-hop modularity of the clustering $\mathscr{C}$ is defined as:*

$$\mathrm{IMM}(\mathscr{C}) = \sum_{C\in\mathscr{C}} \frac{\sum_{v_i,v_j\in C}\boldsymbol{S}[i,j]}{\sum_{v_i,v_j\in V}\boldsymbol{S}[i,j]} - \left(\frac{\sum_{v_i\in C,v_j\in V}\boldsymbol{S}[i,j]}{\sum_{v_i,v_j\in V}\boldsymbol{S}[i,j]}\right)^2. \quad (4)$$

Note that $\sum_{v_i,v_j\in C}\mathbf{S}[i,j]$ is the sum of similarities between all pairs of nodes within cluster $C$, and $\sum_{v_i\in C,v_j\in V}\mathbf{S}[i,j]$ is the sum of similarities of all nodes in $C$ with its neighbors. Recall that the IMM does not require a predefined cluster number $k$. The subsequent process is to partition all nodes within $G$, aiming to find the clustering $\mathscr{C}=\{C_1,C_2,\cdots,C_{|\mathscr{C}|}\}$ such that their IMM score is maximized. Exact modularity optimization is NP-hard [8], leading to approximation approaches such as Louvain [5].

**Algorithm 2:** AHRC

**Input:** Attributed hypergraph $\mathcal{H}(V,E,\mathrm{att})$, attribute similarity matrix $\mathbf{S_A}$, decay factor $\alpha$, number of iteration $\gamma$, sparsification parameter $\tau$, and boolean *spax*: switch of the sparsification
**Output:** Clustering $\mathscr{C}$

1 $\mathbf{H}\leftarrow$ incident matrix of $\mathcal{H}$;
2 Compute row normalization matrices $\mathbf{T_V}\leftarrow \mathrm{norm}(\mathbf{H}^\intercal)$ and $\mathbf{T_E}\leftarrow\mathrm{norm}(\mathbf{H})$;
3 Compute transition matrix $\mathbf{T}\leftarrow\mathbf{T_V}\times\mathbf{T_E}$;
4 **if** *spax* **then**
5 Perform matrix sparsification $\mathbf{T}\leftarrow\mathrm{Sparsifier}(\mathbf{T},\tau)$;
6 Compute topological similarity matrix $\mathbf{S_T}\leftarrow\alpha\sum_{l=0}^{\gamma}(1-\alpha)^l\mathbf{T}^l$;
7 Compute integrated similarity matrix $\mathbf{S}\leftarrow\mathrm{Integrator}(\mathbf{S_T},\mathbf{S_A})$;
8 $\mathscr{C}\leftarrow\mathrm{Louvain}(\mathbf{S})$;
9 **return** $\mathscr{C}$;

**Remarks.** Different from the classic NG modularity function, which merely relies on edges (1-hop relation), the IMM function (Definition 1) captures multi-hop relations under our AHR model, which encodes high-order information in $\mathbf{H}$. Specifically, the random walk takes into account the paths that start from and end at nodes within the same cluster: given a cluster $C$, IMM computes the difference between the actual possibility that paths on the data graph stay within $C$ and, the expected possibility that the paths on the random graph stay within $C$. A higher modularity score indicates that the actual paths within clusters are (probabilistically) more than what would be expected in a random graph, suggesting a good clustering structure. Paths can capture higher-order relationships between nodes, providing a richer representation of connectivity in the data graphs. Implicitly generalizing modularity from edges to paths 1) allows the extraction of more informative features on the data graphs, 2) is an alternative approach to overcome the resolution limit of modularity (struggling to identify small clusters) [17], and thus 3) often brings higher clustering quality [24].

## 3.3 The AHRC Algorithm

The process of AHRC starts with an attributed hypergraph as input, computes the integrated similarity matrix as representation to capture both graph topological and attribute information using the AHR model, and performs clustering based on this representation.

Algorithm 2 shows the pseudo code of the AHRC. It takes as input an attributed hypergraph $\mathcal{H}(V,E,\mathrm{att})$, the attribute similarity matrix $\mathbf{S_A}$ of $\mathcal{H}$, a decay factor $\alpha$, and the number of iteration $\gamma$ for hypergraph random walk. Two additional inputs, a sparsification parameter $\tau$ and a boolean indicator *spax* will also be taken when AHRC applies the proposed spanning forest sparsification process (will be described in Section 4.1). In Line 1, AHRC first extracts the incident matrix $\mathbf{H}$ of $\mathcal{H}$ followed by computing the row normalization matrices $\mathbf{T_V}$ and $\mathbf{T_E}$ in Line 2. Then, the transition matrix $\mathbf{T}$ of $\alpha,\gamma$-Hypergraph Random Walk is computed (Line 3). After that, in Line 6, AHRC computes the TSM $\mathbf{S_T}$ according to Equation 3. Algorithm 1 is then called in Line 7 to compute the ISM $\mathbf{S}$. Based on the obtained $\mathbf{S}$, the Louvain method is then applied to

do the clustering (Line 8). Line 9 returns the resulting clustering $\mathscr{C}$. Lemma 3 analyzes the time complexity of the AHRC algorithm without spanning forest sparsification.

Lemma 3. *When $\gamma = 2$, the time complexity of Algorithm 2 without spanning forest sparsification is $O(\frac{K \cdot \text{vol}_2(\mathcal{H})^2}{n})$.*

Proof. Since the number of non-zero entries in sparse matrix $\mathbf{H}$ is $\text{vol}(\mathcal{H})$, Line 2 takes $O(\text{vol}(\mathcal{H}))$ time. Both sparse matrices $\mathbf{T_V} \in \mathbb{R}^{n \times m}$ and $\mathbf{T_E} \in \mathbb{R}^{m \times n}$ have $\text{vol}(\mathcal{H})$ non-zero entries, their multiplication takes $O(\frac{\text{vol}(\mathcal{H})^2}{n})$ time [70] in Line 3. According to Lemma 1, Line 6 takes $O(\frac{\text{vol}_2(\mathcal{H})^2}{n})$ time. Line 7 then calls Algorithm 1 to multiply matrices $\mathbf{S_T}$ and $\mathbf{S_A}$, taking $O(\frac{K \cdot \text{vol}_2(\mathcal{H})^2}{n})$ time according to Lemma 2. Since the matrix $\mathbf{S}$ has $O(\frac{K \cdot \text{vol}_2(\mathcal{H})^2}{n})$ non-zero entries, the Louvain method called in Line 8 is thus to be $O(\frac{K \cdot \text{vol}_2(\mathcal{H})^2}{n})$. Therefore, the total time complexity of the AHRC algorithm is $O(\frac{K \cdot \text{vol}_2(\mathcal{H})^2}{n})$. □

## 4 Scalability and Generalizability

This section introduces the module of Spanning forest sparsification (Figure 1) which addresses the main scalability bottleneck of AHRC and the module of contrastive learning as a general application of our graph representation $\mathbf{S}$.

### 4.1 Spanning Forest Sparsification

Lemma 3 indicates that the main scalability bottleneck of AHRC is the large dyadic volume $\text{vol}_2(\mathcal{H})$ resulting from the dense hypergraph transition matrix $\mathbf{T}$. It's a natural idea to consider how to reduce $\text{vol}_2(\mathcal{H})$ through sparsification. In this section, we propose a linear-time graph sparsification method called Spanning Forest Sparsification. Recall the definition of $\mathbf{T}$ and hypergraph random walk in Section 2.2, we have a Lemma as follows.

Lemma 4. *Given a hypergraph transition matrix $\boldsymbol{T}$, define a binary matrix of $\boldsymbol{B_T}$ such that $\boldsymbol{B_T}[i,j] = 1$ if $\boldsymbol{T}[i,j] > 0$ and $\boldsymbol{B_T}[i,j] = 0$ otherwise. Then, $\boldsymbol{B_T}$ is a symmetric matrix.*

Proof. To prove that $\mathbf{B_T}$ is a symmetric matrix, we show that $\mathbf{B_T}[i,j] = \mathbf{B_T}[j,i]$ for every $i, j$. For any pair of nodes $v_i$ and $v_j$, assume $\mathbf{T}[i,j] > 0$. It implies that $v_i$ can transit to $v_j$ through a one-step random walk. According to the definition of hypergraph random walk, $v_i$ can transit to $v_j$ if and only if they share at least one incident hyperedge. Thus, $v_i$ and $v_j$ share at least one incident hyperedge, and through this shared hyperedge, $v_j$ can also transit to $v_i$, leading to $\mathbf{T}[j,i] > 0$. Therefore, we have $\mathbf{T}[j,i] > 0$ if $\mathbf{T}[i,j] > 0$. We can prove that $\mathbf{T}[i,j] > 0$ if $\mathbf{T}[j,i] > 0$ similarly by reversing the roles of $v_i$ and $v_j$. Thus, for every $i, j$, we have $\mathbf{T}[j,i] > 0$ if and only if $\mathbf{T}[i,j] > 0$, and therefore $\mathbf{B_T}[j,i] = \mathbf{B_T}[i,j]$. □

In this section, we represent $\mathbf{T}$, and hence the random walk, as a weighted directed graph [27] $G$ with edge set $\{e(i,j) | \mathbf{T}[i,j] > 0\}$, where the edge weight of $e(i,j)$ is the transition probability from node $v_i$ to $v_j$. According to Lemma 4, $G$ has an undirected structure: for any node pairs $(v_i, v_j)$, there is an edge from $v_i$ to $v_j$ if and only if there is another edge from $v_j$ to $v_i$. $G$ has asymmetric weights: due to the asymmetry of the hypergraph random walk, the weights of edge $e(v_i, v_j)$ may not equal to that of $e(v_j, v_i)$.

**Algorithm 3:** Sparsifier

**Input:** Transition matrix $\mathbf{T}$ and sparsification parameter $\tau$
**Output:** Sparsified transition matrix $\mathbf{T'}$

1 Initialize cumulative matrix $\mathbf{M^+} \leftarrow \mathbf{0}$;
2 Initialize residual matrix $\mathbf{M^-} \leftarrow \mathbf{T} + \mathbf{T^\intercal}$;
3 **for** $i \leftarrow 1$ *to* $\tau$ **do**
4  Find maximum spanning forest $\mathbf{F_i} \leftarrow \text{Kruskal}(\mathbf{M^-})$;
5  Update $\mathbf{M^-} \leftarrow \mathbf{M^-} - \mathbf{F_i}$ and $\mathbf{M^+} \leftarrow \mathbf{M^+} + \mathbf{F_i}$;
6 $\mathbf{T'} \leftarrow \mathbf{0}$;
7 **for** each non-zero entry $\mathbf{M^+}[i,j]$ **then** $\mathbf{T'}[i,j] \leftarrow \mathbf{T}[i,j]$;
8 **return** $\mathbf{T'}$;

Given a graph $G$ with aforementioned properties, our method constructs the union of a set of edge-disjoint maximum spanning forests, resulting in a sparsified graph that 1) is a structurally connected subgraph of $G$, 2) maintains the asymmetric edges weights of $G$, and 3) preserve significant relationships carrying large edge weights.

Given a transition matrix $\mathbf{T}$ and a parameter $\tau$, we first compute the symmetric matrix $\mathbf{T} + \mathbf{T^\intercal}$, combining transition probabilities in both directions between nodes, yielding an adjacency matrix of an undirected graph, denoted as $G_{sym}$. This symmetrization allows us to apply Kruskal's algorithm [36], which is designed to find the maximum spanning tree (forest) on undirected graphs. Next, on $G_{sym}$, we generate the union of a set of edge-disjoint maximum spanning forests $F_1, F_2, \cdots, F_\tau$. Each $F_i$ is a maximum spanning forest on $G_{sym}$ after removing those edges in $F_1, F_2, \cdots, F_{i-1}$. The union of these forests is denoted as $F = \cup_{i \in [\tau]} F_i$. $F$ can be computed iteratively. Specifically, we maintain two graphs: a residual graph $G^-$ and a cumulative graph $G^+$. Initially, $G^-$ is the same as $G_{sym}$ and $G^+$ has the same node set as $G_{sym}$ but starts with an empty edge set. In each iteration $i$, we generate a maximum spanning forest $F_i$ for $G^-$ by Kruskal's algorithm. We then update $G^-$ by subtracting the set of edges in $F_i$ from $G^-$, and update $G^+$ by taking the union of the edge sets of $F_i$ and $G^+$. The process repeats $\tau$ times. After termination, the edge set of $G^+$ is the edge set of $F$. Finally, for each undirected edge $e(i,j)$ in $F$, we retain a pair of directed edges $e(i,j)$ and $e(j,i)$ in the sparsified transition matrix $\mathbf{T'}$.

Algorithm 3 presents the pseudo code of the sparsification method, where the residual graph $G^-$, cumulative graph $G^+$, and maximum spanning forest $F_i$ is represented by the adjacency matrix $\mathbf{M^-}$, $\mathbf{M^+}$, and $\mathbf{F_i}$, respectively. Algorithm 3 takes the transition matrix $\mathbf{T}$ and a parameter $\tau$ as inputs. In Line 1, the matrix $\mathbf{M^+}$ is initialized as a zero matrix with the same shape as $\mathbf{T}$. In Line 2, the matrix $\mathbf{M^-}$ is initialized to be $\mathbf{T} + \mathbf{T^\intercal}$. Lines 3-5 iteratively find the maximum spanning forest $\mathbf{F_i}$ by calling the Kruskal's algorithm and update the matrices $\mathbf{M^-}$ and $\mathbf{M^+}$. Lemma 5 analyzes the time complexity of the AHRC algorithm with sparsification.

Lemma 5. *When $\gamma = 2$, the time complexity of Algorithm 2 using the spanning forest sparsification method is $O(\text{vol}_2(\mathcal{H}) \log n)$.*

Proof. Since the transition matrix $\mathbf{T}$ has $O(\text{vol}_2(\mathcal{H}))$ number of non-zero entries according to the proof of Lemma 1, Algorithm 3 takes $O(\tau \cdot \text{vol}_2(\mathcal{H}) \log n)$ time [36] for performing the sparsification. Therefore, in Algorithm 2, Line 4-5 takes $O(\tau \cdot \text{vol}_2(\mathcal{H}) \log n)$ time by calling Algorithm 3. Given that $\tau$ forests have $O(\tau n)$ edges

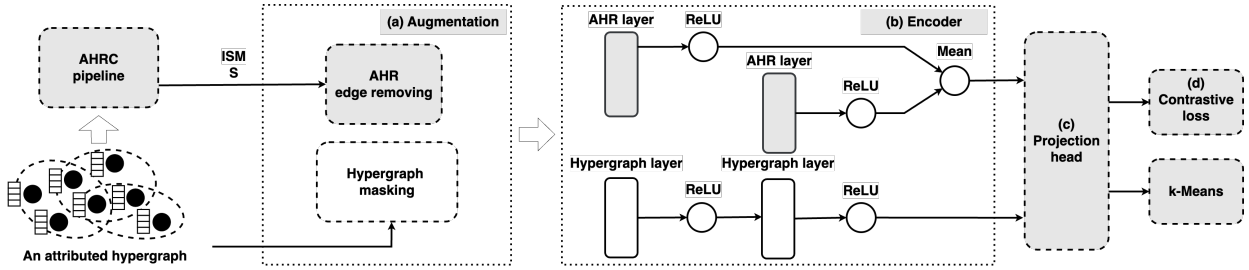



**Figure 3:** TCL+: AHRC **for Contrastive Learning based Clustering**

in total, the sparsified transition matrix has $O(\tau n)$ number of non-zero entries. According to Lemma 1, Line 6 of Algorithm 2 then takes $O(\tau^2 n)$ time to compute matrix $\mathbf{S_T}$ with $O(\tau n)$ non-zero entries. Line 7 multiplies matrices $\mathbf{S_T}$ and $\mathbf{S_A}$, taking $O(\tau K n)$ time according to Lemma 2. The resulting matrix $\mathbf{S}$ has $O(\tau K n)$ non-zero entries. The Louvain algorithm called in Line 8 thus takes $O(\tau K n)$ time. Therefore, the overall time complexity of Algorithm 2 is $O(\tau \cdot \text{vol}_2(\mathcal{H}) \log n + (\tau + K)\tau n)$. Assuming $\tau$ and $K$ are small constants, the overall time complexity of AHRC with sparsification is $O(\text{vol}_2(\mathcal{H}) \log n)$. □

## 4.2 Enhancing Contrastive Learning with AHRC

In the context of GNN-based attributed graph clustering, Graph Contrastive Learning (GCL) [69] has emerged as a popular framework. It learns an encoding function that takes node attributes and graph topology as input, and produces node embeddings as output. These embeddings can then be used for clustering by applying k-Means algorithm. In this section, we show how the attributed hypergraph representation generated by our AHR model enhances existing GCL methods.

We proposed two models TCL+ and GRC+, as enhanced variants of the state-of-the-art GCL methods TRICL [42] and GRACE [78], respectively, based on our AHRC. Figure 3 elaborates the module of contrastive learning + k-means in Figure 1 when it comes to TCL+. It consists of four major components: graph augmentation, encoder, projection head, and contrastive loss. Typically, the encoder is composed of one or more graph neural network layers built on the underlying graph structure. Under our AHRC, we propose an AHR layer that enriches embeddings with the comprehensive multi-hop topological and attribute information captured by the AHRC.

AHR **Layer.** The AHR layer is built on the Integrated Similarity Matrix (ISM) $\mathbf{S}$ under our AHRC pipeline. Given an attributed hypergraph, we first compute ISM $\mathbf{S}$ using AHRC. Regarding $\mathbf{S}$ as a weighted directed graph $G(V, E)$, graph convolution is then applied to the underlying unweighted structure of $G$ to generate node embeddings for the contrastive learning process. Specifically, the AHR Layer iteratively propagates embeddings through the unweighted structure of $G$, updating the embedding of each node by aggregating the embedding of its adjacent nodes. Let $\mathbf{z}_v^{(i)}$ be the embedding of node $v \in V$ at the $i$-th AHR layer, defined as:

$$\mathbf{z}_v^{(i)} = f(\mathbf{z}_v^{(i-1)}, \{\mathbf{z}_u^{(i-1)} : e(u, v) \in E\}) \tag{5}$$

where $f$ is the aggregation rule. Then, our proposed models, TCL+ and GRC+, incorporate the AHR layers to the encoders of TRICL and GRACE, respectively.

TCL+ **Model Architecture.** Figure 3 overviews the architecture of our proposed model TCL+. We briefly introduce it in the following.

(1) **Graph augmentation.** Given an attributed hypergraph $\mathcal{H}$, we first compute an ISM $\mathbf{S}$ using the AHRC pipeline. Next, we augment $\mathbf{S}$ by performing random *edge removing* [78], which we refer to as AHR edge removing, to generate two alternate views on $\mathbf{S}$. In these views, a portion (controlled by a hyper-parameter $p_d$) of non-zero entries in $\mathbf{S}$ are randomly set to be zero. Additionally, we perform hypergraph masking to augment the hypergraph topology and attributes by performing random *membership masking* [42] and *node feature masking* [78] to generate two alternate views of the hypergraph.

(2) **Encoder.** The encoder produces embeddings for the views generated in (1). As Figure 3(b) shows, TCL+ employs a two-level encoder, each of which consists of one AHR layer and one hypergraph layer. All layers use the element-wise mean pooling aggregation rule as a special instance of Equation 5. Specifically, the AHR layer can be represented in the matrix form:

$$\mathbf{Z}^{(i)} = \sigma(\mathbf{D}^{-1}\mathbf{B}\mathbf{Z}^{(i-1)}\mathbf{W}^{(i)}) \tag{6}$$

where $\mathbf{B}$ is a binary adjacency matrix such that $\mathbf{B}[i, j] = 1$ if $\mathbf{S}[i, j] > 0$ and $\mathbf{B}[i, j] = 0$ otherwise, representing the unweighted structure of $G$. The initial node embeddings $\mathbf{Z}^{(0)}$ are set to the node attributes. $\mathbf{D}$ is the diagonal degree matrix where $\mathbf{D}[i, i] = \sum_j \mathbf{B}[i, j]$, $\mathbf{W}^{(i)}$ is the trainable weight for the $i$-th layer, and $\sigma$ is the activation function $\text{ReLU}(x) = \max(0, x)$. The hypergraph layer applied on the hypergraph topology follows the same structure as that of [42].

(3) **Projection head and contrastive loss.** With the embeddings from encoder, we project them by performing non-linear transformation [12] using the same projection heads as [42]. For node embeddings generated by both hypergraph layers and AHR layers, we adopt the same objective function as [42, 78]. The overall loss is computed as follows:

$$\mathcal{L}_n = \mathcal{L}_H + w_s \cdot \mathcal{L}_A \tag{7}$$

where $\mathcal{L}_H$ and $\mathcal{L}_A$ is the contrastive loss on the node embeddings generated by the hypergraph layers and AHR layers, respectively. $w_s$ is the weight balancing two losses.

GRC+ **Model Architecture.** GRACE is a state-of-the-art GCL method on dyadic graphs, employing a two-layer encoder. On hypergraphs, GRACE can be applied on a dyadic graph $G'$ that

is converted from a given attributed hypergraph through clique reduction. Our proposed model, GRC+, enhances GRACE by replacing the layers built on $G'$ with the AHR layers built on the ISM **S** under our AHC model.

## 5 Related Works

**Attributed Hypergraph Clustering.** Graph clustering [22, 54] has been extensively studied on dyadic graphs. Traditional clustering optimizes objective functions such as modularity [14], conductance [6], normalized cut [56], etc. Exact modularity optimization is computationally hard, leading to approximation approaches [5, 15, 19, 51]. Among these, Louvain [5] has been widely used in industry due to its scalability and clustering quality [74]. A hypergraph can be transformed into a dyadic graph using clique reduction [1, 40], and then dyadic graph clustering methods can be applied. However, this approach loses high-order information in hypergraphs. Other methods represent a hypergraph using a hypergraph random walk transition matrix [27, 75] or a normalized Laplacian [43, 44], transforming it into a weighted dyadic graph. Another line of research [13, 21] models a hypergraph with a random hypergraph model, then clustering the hypergraph by iteratively maximizing the modularity-based objective scores.

Attributed hypergraph clustering (AHC) has also been studied. Existing AHC methods [10, 18, 31, 45] predominantly rely on matrix factorization techniques, leading to high computational and memory costs, thereby limiting scalability [31]. Additionally, they require prior knowledge of the number of clusters to produce quality clustering. However, the number of clusters for a desirable AHC is usually dataset-dependent and unknown beforehand, without knowing which, the performance can drop dramatically [63]. Specifically, JNMF [18] first represents an attributed hypergraph by a hypergraph Laplacian and an attribute matrix. It then integrates both by adopting a Non-negative Matrix Factorization (NMF) objective function that consists of an NMF part for the hypergraph topology and an NMF part for the attributes. [31] transforms an attributed hypergraph to an attributed dyadic graph and then extends GNMF [10], an NMF-based high-dimensional data clustering method, to do the clustering. Specifically, GNMFA uses an adjacency matrix obtained through clique reduction, GNMFC uses an adjacency matrix obtained through clique reduction with self-loops removed, and GNMFL represents the attributed hypergraph by hypergraph normalized Laplacian. However, the clique reduction hinders clustering effectiveness due to the information loss, and the NMF operations incur high computational and memory costs. To address the issue, GRAC [31] represents hypergraph topology by a less costly hypergraph Laplacian and then performs hypergraph convolution on node attributes to obtain a similarity matrix such that they better integrate the topological and attribute information. However, the following Singular Value Decomposition (SVD) operations for clustering are still expensive, limiting its scalability. The state-of-the-art method AHCKA [45] first exploits attributes by using the KNN algorithm to compute an attribute graph. Then, it performs a fixed-length random walk over the original hypergraph and the constructed attribute graph jointly to obtain the node similarities integrated both attribute and topological information. Clusters are computed by approximating the eigenvectors of the obtained similarity matrix in a greedy iterative process. However, the inherent complexity limitations of the eigendecomposition may hinder a further improvement of the scalability. We experimentally prove that our AHRC outperforms AHCKA in efficiency.

**Measure the Similarity Among Nodes.** The measures of node-wise topological [25, 53, 58, 63, 71] and attribute [29, 46, 67, 76] similarity in graphs have been extensively studied. These similarities can be stored in a similarity matrix, serving as an input for clustering algorithms such as Louvain and spectral method. A line of methods uses Gaussian similarity [63], L2 distance [71], or random walk [25, 53, 58] to compute node-wise similarities based on graph topology. However, these methods fail to capture attribute information. Another line of methods [67, 76] augments the data graph by treating the attributes as 'nodes' and establishing a set of node-attribute associations. They then perform random walks on the augmented graph to integrate both topological and attribute information. However, these methods suffer from high computational costs on large graphs with multi-dimensional attributes. Additionally, considering all attributes with potential inconsistencies can diminish clustering effectiveness [45]. To capture sufficient attribute information while reducing computational cost and noise, methods [29, 46] employ the KNN algorithm to measure the attribute similarity. AHCKA [45] integrates the attribute graph computed by the KNN algorithm with hypergraph topology through a joint random walk process. However, it is unclear why the attribute similarity should and could be transmitted through random walk, especially when the attribute graph has considered attribute relations among all node pairs. We experimentally prove that our AHRC outperforms AHCKA in effectiveness.

**Graph Contrastive Learning.** Graph contrastive learning (GCL) methods [42, 66, 72, 78] learn an encoding function that takes node attributes and graph topology as input and produces node embeddings as output. This process involves using GNN layers to propagate and aggregate information based on the underlying graph structure. The embeddings can then be used for clustering by applying the k-Means algorithm. On dyadic graphs, the state-of-the-art GRACE [78] generates two graph views by randomly removing a portion of edges and node attributes, then learns node embeddings by maximizing the agreement that is measured by contrastive loss on node embeddings in these two views. However, GRACE cannot be directly applied to hypergraphs, necessitating the representation of hypergraphs into dyadic graphs through techniques such as clique reduction. There is still a lack of efficient and effective hypergraph representation methods. Among hypergraph contrastive learning methods [42, 66, 72], TRICL [42] achieves the state-of-the-art performances. It aggregates information directly on the underlying hypergraph structure, which can be represented by a hypergraph transition matrix. However, such structures capture the local topology of 1-hop neighbors. Given that existing models are rather shallow – GRACE consists of two layers and TRICL has one layer only – limiting their ability to capture global information. Under our AHR model, the AHR layer we proposed captures multi-hop relationships between nodes in both topological and attribute senses, effectively integrating global information to enhance the performance of contrastive learning on hypergraphs.

| Name | Dataset | $n$ | $m$ | $\text{vol}(\mathcal{H})$ | $\text{vol}_2(\mathcal{H})$ | $d$ |
|---|---|---|---|---|---|---|
| *C13* | C13-C [18] | 693 | 545 | 3,475 | 26,908 | 4,728 |
| *WIK* | Wiki [31] | 1,999 | 2,184 | 16,321 | 91,479 | 4,973 |
| *NTU* | NTU2012 [42] | 2,012 | 2,012 | 10,060 | 20,120 | 100 |
| *COA* | Cora-A [45] | 2,708 | 1,072 | 4,585 | 17,136 | 1,433 |
| *COC* | Cora-C [45] | 2,708 | 1,579 | 4,786 | 5,687 | 1,433 |
| *CIC* | Citeseer-C [45] | 3,312 | 1,079 | 3,453 | 6,007 | 3,703 |
| *NEW* | 20News [45] | 16,242 | 100 | 65,451 | 34,234,847 | 100 |
| *PBC* | Pubmed-C [31] | 19,717 | 7,963 | 34,629 | 186,155 | 500 |
| *DBA* | DBLP-A [45] | 41,302 | 22,363 | 99,561 | 906,564 | 1,425 |
| *MAG* | MAGPM [45] | 2,353,996 | 1,082,711 | 17,279,202 | 517,767,530 | 1,000 |

**Table 1: Datasets**

## 6 Experiments

This section evaluates the performance of our proposed AHRC method on 10 real-world attributed hypergraphs with ground truth clustering. Table 1 shows the statistics of them. $n$ and $m$ denote the number of nodes and hyperedges, respectively. $\text{vol}(\mathcal{H})$, $\text{vol}_2(\mathcal{H})$, and $d$ denote the volume, dyadic volume, and number of attributes of hypergraph $\mathcal{H}$, respectively. All the experiments were conducted on a CPU server (Intel Xeon Gold 6230 CPU 2.10GHz, 376GB RAM, and Ubuntu 5.8.0-38-Generic), and a GPU server with an NVIDIA RTX A6000 48GB GPU. All methods were run 10 times to report the average. The cut-off running time was set to be 12 hours. The implementation is available at anonymous GitHub repository[1].

**Baselines.** We compare our AHRC with 6 state-of-the-art algorithmic attributed hypergraph clustering methods that are introduced in Section 5: GNMFA, GNMFC, GNMFL [10], JNMF [18], GRAC [31], and AHCKA [45]. Additionally, we compare our contrastive learning methods TCL+ and GRC+ with state-of-the-art contrastive learning methods TRICL [42] and GRACE [78]. TRICL is a one-layer model with variants using different objective combinations. For a fair comparison, we use and refer to TRICL as the variant with objective computed on node embeddings.

**Parameters.** For all baselines, we adopt the default parameter values as suggested in their respective papers. Given that the number of clusters $k$ in a desirable clustering is often not available in reality, our AHRC produces a reasonable $k$ based on Property 1. For consistent comparison [21], this value of $k$ will be used for baselines that require a predefined $k$. For our AHRC, unless otherwise specified, we set the default values of parameter $\alpha$ to 0.2 following [45] and $\gamma$ to 2 based on sensitivity analysis. The hyperparameters of our contrastive learning models TCL+ and GRC+ are summarized in Table 7. Due to space limit, more sensitivity analysis on hyperparameters is provided in the detailed report [1].

### 6.1 Effectiveness and Scalability

In this section, we show the experimental results on clustering quality and scalability. Clustering quality is evaluated in the alignment to the ground truth clustering with 6 widely used metrics: F-measure [47], Adjusted Rand Index (ARI) [28], Jaccard Similarity [28], Purity [47], Balanced Accuracy [9], and Normalized Mutual Information (NMI) [34]. For all the above metrics, a larger score indicates better clustering quality.

**Exp 1.** Table 2 shows the clustering performance of our AHRC and 6 algorithmic baselines on datasets with ground truth. Top-2 scores for each dataset are highlighted with bold&underline and bold, respectively. '\' denotes no result due to time-out or out-of-memory reason. Baselines GNMFA, GNMFC, and GNMFL fail on dataset *NEW* as they regard the entire graph as a single cluster. In general, our AHRC achieves the best overall performance. Specifically, on F-measure, AHRC surpasses all 6 baselines (in top-down order as listed in Table 2 unless otherwise specified) by 142%, 119%, 40%, 73%, 39%, and 13%, respectively, averaged over all datasets. In terms of ARI, AHRC outperforms all 6 baselines by 8,086%, 533%, 147%, 6,558%, 83%, and 16%, respectively. On Jaccard Similarity, AHRC is 182%, 153%, 56%, 96%, 52%, and 17% higher than the 6 baselines. For Purity, AHRC outperforms the baselines by 125%, 70%, 25%, 72%, 34%, and 11%, respectively. On Balanced Accuracy, AHRC outperforms baselines by 25%, 20%, 12%, 16%, 9%, and 3%, respectively. On NMI, AHRC is 4,635%, 222%, 61%, 2,356%, and 24% higher than GNMFA, GNMFC, GNMFL, JNMF, and GRAC, respectively. AHRC obtains a similar (by an average of 0% difference) NMI to AHCKA.

**Exp 2.** Table 3 shows the clustering performance of our TCL+, GRC+, and 2 contrastive learning baselines on 9 datasets with ground truth. Due to out-of-memory reason, none of the methods could run on the largest dataset *MAG*. In general, our TCL+ achieves the best overall performance among all 4 methods. Averaged across all datasets, TCL+ obtains 866%, 45%, and 107% higher ARI than GRACE, GRC+, and TRICL, respectively, and outperforms them by 130%, 32%, and 22% in terms of NMI, respectively.

To better demonstrate the effectiveness of our models and AHR layer, we group TCL+ with TRICL, and GRC+ with GRACE, highlighting the best score within each group for each dataset. Specifically, our TCL+ constantly outperforms TRICL on all datasets across all 6 metrics. Specifically, TCL+ achieves 21% higher F-measure, 107% higher ARI, 27% higher Jaccard Similarity, 11% higher Purity, 6% higher Balanced Accuracy, and 22% higher NMI compared to TRICL, averaged over all datasets. Comparing our GRC+ and GRACE, GRC+ constantly surpasses GRACE on all datasets except *PBC* across all 6 metrics: averaged over all datasets, GRC+ achieves 61%, 556%, 78%, 31%, 15%, and 73% higher F-measure, ARI, Jaccard Similarity, Purity, Balanced Accuracy, and NMI, resp., than GRACE.

**Exp 3.** Table 4 shows the time and memory costs of AHRC and the algorithmic baselines on 10 datasets. On the largest dataset *MAG*, the baselines GRAC failed with out-of-time errors, and the other baselines failed with out-of-memory errors. In terms of running time, AHRC is in general the fastest among all 7 methods, averaged across all datasets, showing the effectiveness of our method. Specifically, over all datasets, GNMFA, GNMFC, GNMFL, JNMF, GRAC, and AHCKA is on average 126, 126, 141, 62, 78, and 5 times slower than our AHRC, respectively. In terms of memory cost, averaged over all datasets, GNMFA, GNMFC, GNMFL, JNMF, and GRACE take 27, 26, 25, 26, and 10 times more memory space than our AHRC, respectively. Our AHRC takes similar memory space to AHCKA.

Table 5 shows the training time for a single epoch of our TCL+, GRC+, and 2 contrastive learning baselines. Compared with TRICL, our TCL+ takes 146% longer for training, averaged across all datasets. It is because TCL+ employs a more complex model with a two-level encoder, where each level consists of a hypergraph layer and an AHR layer, while TRICL uses a one-level encoder with a single hypergraph layer only. Compared with GRACE, our GRC+ trains 12% faster, averaged over all datasets. GRACE fails to handle the

[1] https://anonymous.4open.science/r/Attributed_Hypergraph_Representation_for_Clustering_Code-4368

| \ | F-measure | | | | | | | | | ARI | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *MAG* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *MAG* |
| GNMFA | 0.35 | 0.06 | 0.23 | 0.22 | 0.19 | \ | 0.13 | 0.42 | \ | 0.24 | 0.02 | 0.01 | 0.00 | 0.02 | \ | 0.00 | 0.20 | \ |
| GNMFC | 0.35 | 0.06 | 0.26 | 0.24 | 0.29 | \ | **0.17** | 0.41 | \ | 0.24 | 0.02 | 0.07 | 0.03 | 0.15 | \ | 0.04 | 0.19 | \ |
| GNMFL | 0.30 | 0.25 | 0.28 | 0.27 | 0.42 | \ | **0.20** | 0.49 | \ | 0.17 | 0.22 | 0.14 | 0.07 | 0.28 | \ | 0.03 | 0.37 | \ |
| JNMF | 0.30 | 0.23 | 0.33 | 0.26 | 0.33 | 0.12 | 0.08 | 0.42 | \ | 0.19 | 0.21 | 0.20 | 0.14 | 0.21 | 0.01 | 0.00 | 0.22 | \ |
| GRAC | 0.32 | **0.38** | 0.38 | 0.33 | 0.28 | 0.14 | 0.14 | 0.56 | \ | 0.20 | **0.36** | 0.27 | 0.22 | 0.11 | 0.07 | 0.05 | 0.45 | \ |
| AHCKA | 0.40 | 0.33 | **0.46** | **0.40** | **0.49** | **0.18** | **0.17** | **0.62** | **0.38** | **0.32** | 0.31 | **0.35** | **0.31** | **0.38** | **0.11** | **0.07** | **0.53** | **0.34** |
| AHRC | **0.41** | **0.35** | **0.55** | **0.51** | **0.46** | **0.28** | **0.17** | **0.62** | **0.43** | **0.34** | **0.32** | **0.46** | **0.41** | **0.36** | **0.20** | **0.06** | **0.53** | **0.37** |
| \ | **Jaccard Similarity** | | | | | | | | | **Purity** | | | | | | | | |
| | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *MAG* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *MAG* |
| GNMFA | 0.21 | 0.03 | 0.13 | 0.12 | 0.11 | \ | 0.07 | 0.26 | \ | 0.54 | 0.11 | 0.29 | 0.26 | 0.23 | \ | 0.12 | 0.47 | \ |
| GNMFC | 0.21 | 0.03 | 0.15 | 0.14 | 0.17 | \ | **0.09** | 0.26 | \ | 0.54 | 0.16 | 0.36 | 0.31 | 0.41 | \ | **0.20** | 0.46 | \ |
| GNMFL | 0.18 | 0.14 | 0.16 | 0.15 | 0.26 | \ | **0.11** | 0.33 | \ | 0.49 | 0.34 | 0.42 | 0.44 | 0.55 | \ | 0.24 | 0.60 | \ |
| JNMF | 0.17 | 0.13 | 0.20 | 0.15 | 0.19 | 0.06 | 0.04 | 0.27 | \ | 0.47 | 0.30 | 0.50 | 0.38 | 0.47 | 0.18 | 0.06 | 0.52 | \ |
| GRAC | 0.19 | **0.23** | 0.24 | 0.19 | 0.16 | 0.08 | 0.07 | 0.38 | \ | 0.56 | **0.42** | 0.56 | 0.45 | 0.40 | 0.19 | 0.14 | 0.69 | \ |
| AHCKA | **0.25** | 0.20 | **0.30** | **0.25** | **0.33** | **0.10** | **0.09** | **0.45** | **0.23** | **0.54** | **0.42** | **0.61** | **0.55** | **0.64** | **0.22** | **0.19** | **0.74** | **0.63** |
| AHRC | **0.26** | **0.21** | **0.38** | **0.34** | **0.30** | **0.17** | **0.09** | **0.45** | **0.28** | **0.57** | **0.42** | **0.71** | **0.67** | 0.62 | **0.37** | **0.19** | **0.74** | **0.59** |
| \ | **Balanced Accuracy** | | | | | | | | | **NMI** | | | | | | | | |
| | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *MAG* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *MAG* |
| GNMFA | **0.71** | 0.66 | 0.51 | 0.50 | 0.51 | \ | 0.50 | 0.66 | \ | 0.46 | 0.24 | 0.05 | 0.02 | 0.02 | \ | 0.00 | 0.30 | \ |
| GNMFC | **0.71** | 0.66 | 0.54 | 0.52 | 0.57 | \ | 0.52 | 0.65 | \ | 0.46 | 0.27 | 0.11 | 0.05 | 0.18 | \ | 0.15 | 0.28 | \ |
| GNMFL | 0.68 | 0.85 | 0.56 | 0.54 | 0.65 | \ | 0.51 | 0.69 | \ | 0.46 | 0.59 | 0.20 | 0.21 | 0.27 | \ | 0.09 | 0.43 | \ |
| JNMF | 0.64 | 0.75 | 0.59 | 0.56 | 0.59 | 0.51 | 0.50 | 0.66 | \ | 0.38 | 0.53 | 0.27 | 0.21 | 0.23 | 0.03 | 0.00 | 0.31 | \ |
| GRAC | 0.68 | 0.90 | 0.62 | 0.59 | 0.56 | 0.53 | 0.52 | 0.73 | \ | 0.46 | 0.69 | 0.36 | 0.34 | 0.22 | 0.19 | **0.16** | 0.53 | \ |
| AHCKA | 0.69 | **0.92** | **0.67** | **0.63** | **0.69** | **0.54** | **0.53** | **0.77** | **0.63** | **0.48** | **0.70** | **0.44** | **0.42** | **0.37** | **0.29** | **0.17** | **0.60** | **0.52** |
| AHRC | 0.69 | **0.92** | **0.73** | **0.69** | **0.67** | **0.58** | **0.53** | **0.77** | **0.67** | **0.52** | **0.71** | **0.49** | **0.46** | **0.36** | **0.27** | 0.14 | **0.61** | **0.49** |

**Table 2: Clustering Quality of Different Algorithmic Methods**

| \ | F-measure | | | | | | | | | ARI | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *C13* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *C13* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* |
| GRACE | 0.17 | 0.21 | 0.10 | 0.28 | 0.24 | 0.28 | \ | **0.21** | \ | 0.05 | 0.05 | 0.06 | 0.06 | 0.04 | 0.01 | \ | **0.07** | \ |
| GRC+ | **0.25** | **0.39** | **0.36** | **0.35** | **0.29** | **0.35** | **0.17** | 0.14 | \ | **0.16** | **0.31** | **0.33** | **0.22** | **0.18** | **0.22** | **0.09** | 0.06 | \ |
| TRICL | 0.16 | 0.32 | 0.31 | 0.44 | 0.39 | 0.45 | **0.19** | 0.16 | 0.57 | 0.03 | 0.21 | 0.28 | 0.33 | 0.28 | 0.35 | 0.12 | 0.07 | 0.46 |
| TCL+ | **0.36** | **0.35** | **0.35** | **0.46** | **0.42** | **0.46** | **0.19** | **0.18** | **0.67** | **0.28** | **0.26** | **0.33** | **0.36** | **0.32** | **0.36** | **0.13** | **0.09** | **0.59** |
| \ | **Jaccard Similarity** | | | | | | | | | **Purity** | | | | | | | | |
| | *C13* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *C13* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* |
| GRACE | 0.09 | 0.12 | 0.05 | 0.16 | 0.14 | 0.16 | \ | **0.12** | \ | 0.28 | 0.34 | 0.26 | 0.44 | 0.34 | 0.30 | \ | **0.26** | \ |
| GRC+ | **0.14** | **0.24** | **0.22** | **0.21** | **0.17** | **0.21** | **0.09** | 0.08 | \ | **0.32** | **0.55** | **0.42** | **0.55** | **0.45** | **0.50** | **0.21** | 0.15 | \ |
| TRICL | 0.09 | 0.19 | 0.18 | 0.28 | 0.24 | 0.29 | 0.10 | 0.09 | 0.40 | 0.26 | 0.49 | 0.40 | 0.61 | 0.56 | 0.59 | 0.21 | 0.18 | 0.70 |
| TCL+ | **0.22** | **0.21** | **0.22** | **0.30** | **0.27** | **0.30** | **0.11** | **0.10** | **0.50** | **0.37** | **0.53** | **0.40** | **0.63** | **0.59** | **0.61** | **0.22** | **0.21** | **0.78** |
| \ | **Balanced Accuracy** | | | | | | | | | **NMI** | | | | | | | | |
| | *C13* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *C13* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* |
| GRACE | 0.61 | 0.58 | 0.65 | 0.54 | 0.52 | 0.51 | \ | **0.53** | \ | 0.22 | 0.27 | 0.44 | 0.22 | 0.12 | 0.10 | \ | **0.17** | \ |
| GRC+ | **0.66** | **0.68** | **0.88** | **0.61** | **0.58** | **0.61** | **0.53** | 0.52 | \ | **0.32** | **0.47** | **0.66** | **0.30** | **0.26** | **0.29** | **0.24** | **0.17** | \ |
| TRICL | 0.58 | 0.66 | 0.88 | 0.65 | 0.62 | 0.66 | 0.54 | **0.53** | 0.74 | 0.17 | 0.45 | 0.67 | 0.43 | 0.41 | 0.37 | **0.32** | 0.19 | 0.59 |
| TCL+ | **0.79** | **0.68** | **0.89** | **0.67** | **0.64** | **0.67** | **0.55** | **0.53** | **0.79** | **0.41** | **0.49** | **0.68** | **0.45** | **0.44** | **0.38** | **0.32** | **0.23** | **0.65** |

**Table 3: Clustering Quality of Different Contrastive Learning Methods**

| \ | *C13* | | *WIK* | | *NTU* | | *COA* | | *COC* | | *CIC* | | *NEW* | | *PBC* | | *DBA* | | *MAG* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Time | Space | Time | Space | Time | Space | Time | Space | Time | Space | Time | Space | Time | Space | Time | Space | Time | Space | Time | Space |
| GNMFA | 23.69 | 0.15 | 112.74 | 0.30 | 4.91 | 0.27 | 31.31 | 0.27 | 27.30 | 0.27 | 164.84 | 0.35 | 295.24 | 2.21 | 414.93 | 4.29 | 1,550.46 | 20.17 | \ | \ |
| GNMFC | 23.90 | 0.15 | 118.46 | 0.27 | 4.87 | 0.19 | 32.54 | 0.23 | 27.30 | 0.23 | 163.70 | 0.31 | 301.69 | 2.14 | 411.21 | 4.27 | 1,501.95 | 20.17 | \ | \ |
| GNMFL | 23.81 | 0.15 | 121.23 | 0.21 | 5.48 | 0.26 | 29.55 | 0.17 | 28.92 | 0.17 | 168.01 | 0.17 | 315.50 | 2.13 | 483.23 | 4.34 | 2,381.92 | 20.11 | \ | \ |
| JNMF | 1.58 | 0.15 | 6.40 | 0.31 | **3.42** | 0.29 | 10.73 | 0.21 | 11.04 | 0.24 | 30.19 | 0.29 | 284.69 | 2.15 | 493.89 | 4.29 | 2,716.49 | 20.11 | \ | \ |
| GRAC | 35.15 | 0.23 | 89.51 | 0.39 | 5.33 | **0.16** | 32.97 | 0.36 | 14.17 | 0.36 | 61.27 | 0.41 | **30.60** | 0.29 | **19.29** | 0.43 | 525.27 | 1.32 | \ | \ |
| AHCKA | **0.74** | **0.03** | **2.93** | **0.14** | 3.62 | 0.19 | **1.98** | **0.02** | **3.47** | **0.13** | **2.67** | **0.04** | **42.48** | **0.13** | 43.86 | **0.19** | **8.74** | **0.21** | **11,064.37** | **21.55** |
| AHRC | **0.16** | **0.03** | **0.51** | **0.04** | **0.15** | **0.02** | **0.48** | **0.02** | **0.72** | **0.02** | **0.46** | **0.02** | 67.61 | **0.16** | **3.91** | **0.07** | **9.33** | **0.19** | **4,235.63** | **32.69** |

**Table 4: Time and Memory Cost of Different Algorithmic Methods (Time in Seconds, RAM in GBs)**

dataset *NEW* due to the dense representation generated by clique reduction. Using the AHR model, our AHR layer captures richer information with less space overhead. It shows the scalability of our approach.

## 6.2 Sensitivity

**Exp 4.** We evaluate the clustering quality of AHRC when parameter $\tau$ varies from 1 to 6 with step size 1. Figure 4 reports the F-measure, ARI, Jaccard Similarity, Purity, Balanced Accuracy, and NMI on 8 datasets due to the limited space. As $\tau$ increases, there are overall upward trends in all 6 metrics across all datasets. Specifically, when $\tau$ increases from 1 to 6, the metric scores on average increase by

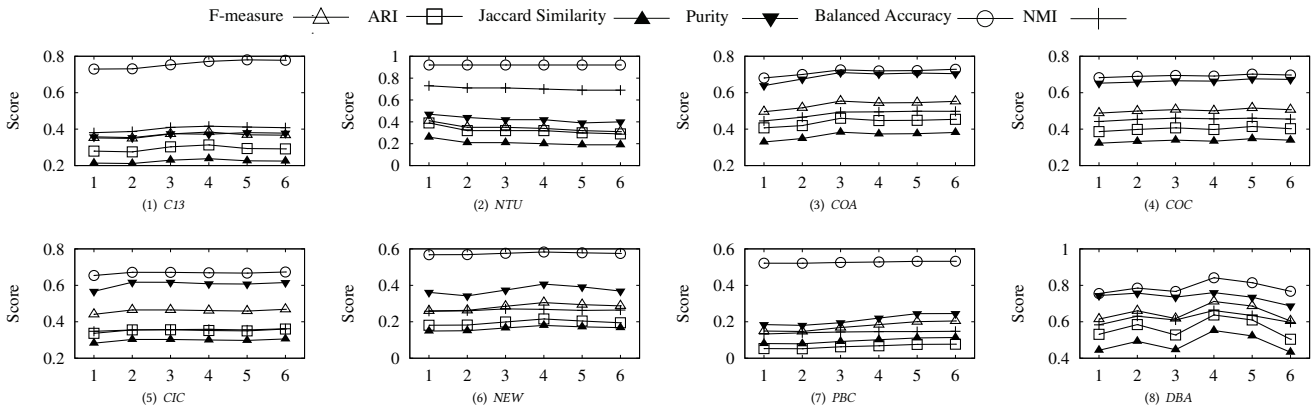



**Figure 4: Sensitivity: Clustering Quality of AHRC on Varying $\tau$**

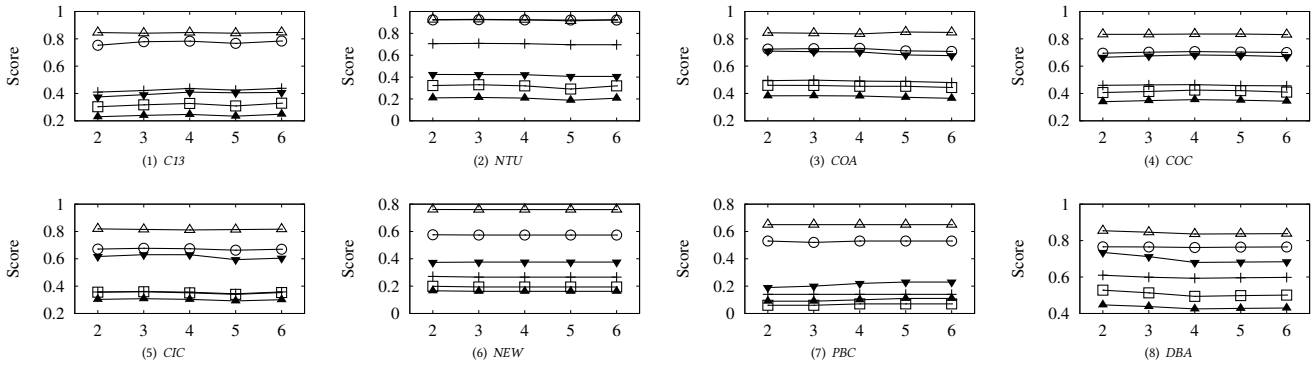



**Figure 5: Sensitivity: Clustering Quality of AHRC on Varying $\gamma$**

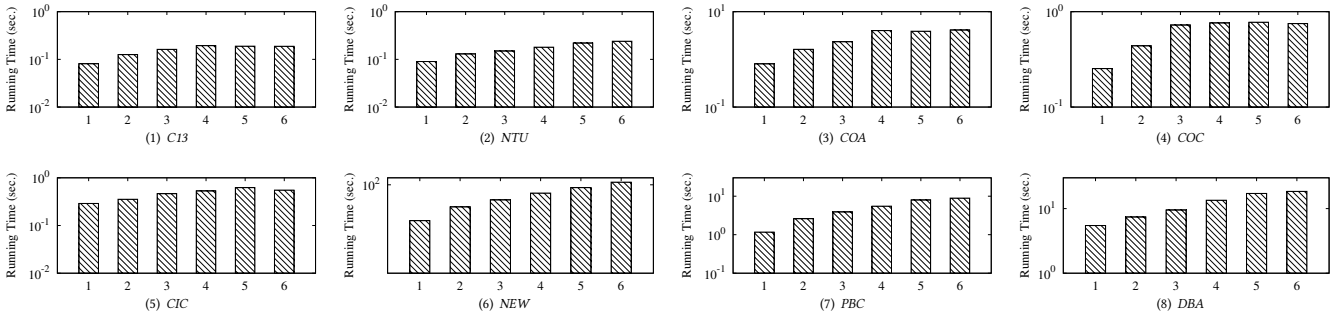



**Figure 6: Sensitivity: Time Cost of AHRC by Varying $\tau$**

| \ | Time (sec.) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *C13* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* |
| GRACE | 3.47 | 6.08 | 7.17 | 7.46 | 8.08 | 12.57 | \ | 161.92 | \ |
| GRC+ | 3.71 | 5.35 | 4.69 | 4.75 | 8.61 | 13.90 | 227.83 | 124.47 | \ |
| TRICL | 5.97 | 7.27 | 3.25 | 11.89 | 6.52 | 10.14 | 62.33 | 188.46 | 342.59 |
| TCL+ | 12.82 | 23.08 | 6.88 | 48.49 | 11.46 | 18.85 | 156.61 | 458.08 | 715.09 |

**Table 5: Time Cost of Different Contrastive Learning Methods**

4%, 1%, 1%, −1%, and 1% at each step, respectively. The scores rise at a faster pace when $\tau$ is small. When $\tau$ becomes larger, scores remain relatively stable at peak. On some datasets such as *NEW* and *DBA*, we observe the decrease in scores when $\tau$ is greater than 4. It is because when $\tau$ is small, the sparsified graph does not capture sufficient node relationships, leading to limited clustering quality. Conversely, when $\tau$ is too large, more noisy or distorted information might be captured, thus affecting clustering effectiveness.

Figure 6 shows the time cost of AHRC when $\tau$ varies from 1 to 6 with step size 1. In general, the running time of AHRC rises as $\tau$ increases. Specifically, when $\tau$ increases from 2 to 6, per step, AHRC takes 53%, 107%, 158%, 205%, and 221% more time, respectively, compared to $\tau = 1$. We observe that when $\tau$ becomes larger, the

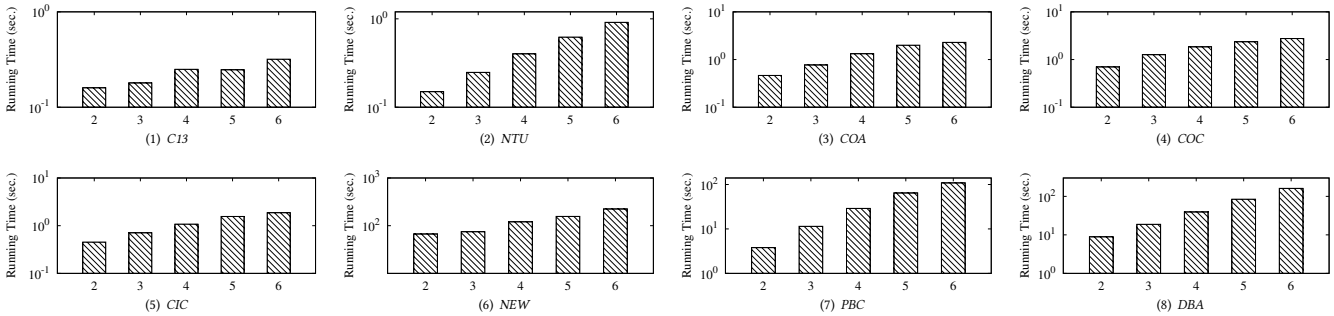



**Figure 7: Sensitivity: Time Cost of AHRC by Varying $\gamma$**

| \ | F-measure | | | | | | | | | ARI | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *MAG* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *MAG* |
| AHRC (LIN) | 0.46 | **0.48** | 0.45 | 0.43 | 0.38 | 0.22 | 0.13 | 0.52 | 0.21 | 0.41 | 0.46 | 0.36 | 0.34 | 0.28 | 0.15 | 0.05 | 0.43 | 0.15 |
| AHRC (LOG) | 0.45 | 0.51 | 0.34 | 0.34 | 0.33 | 0.22 | 0.13 | 0.48 | 0.15 | 0.41 | **0.50** | 0.24 | 0.25 | 0.24 | 0.15 | 0.04 | 0.37 | 0.10 |
| AHRC (EXP) | **0.47** | **0.48** | 0.43 | 0.39 | 0.38 | 0.22 | 0.13 | 0.50 | 0.21 | **0.42** | 0.46 | 0.34 | 0.29 | 0.28 | 0.15 | 0.05 | 0.40 | 0.15 |
| AHRC (SQR) | 0.41 | 0.35 | **0.55** | **0.51** | **0.46** | **0.28** | **0.17** | **0.62** | **0.43** | 0.34 | 0.32 | **0.46** | **0.41** | **0.36** | **0.20** | **0.06** | **0.53** | **0.37** |
| \ | **jaccard similarity** | | | | | | | | | **purity** | | | | | | | | |
| | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *MAG* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *MAG* |
| AHRC (LIN) | **0.30** | 0.31 | 0.36 | 0.27 | 0.23 | 0.12 | 0.07 | 0.35 | 0.12 | **0.68** | 0.52 | 0.58 | 0.57 | 0.50 | 0.32 | 0.17 | 0.61 | 0.47 |
| AHRC (LOG) | 0.29 | **0.35** | 0.24 | 0.20 | 0.20 | 0.13 | 0.07 | 0.32 | 0.08 | 0.66 | **0.56** | 0.47 | 0.47 | 0.43 | 0.32 | 0.16 | 0.64 | 0.40 |
| AHRC (EXP) | **0.30** | 0.31 | 0.34 | 0.24 | 0.23 | 0.12 | 0.07 | 0.33 | 0.12 | **0.68** | 0.52 | 0.57 | 0.52 | 0.50 | 0.30 | 0.17 | 0.61 | 0.47 |
| AHRC (SQR) | 0.26 | 0.21 | **0.38** | **0.34** | **0.30** | **0.17** | **0.09** | **0.45** | **0.28** | 0.57 | 0.42 | **0.71** | **0.67** | **0.62** | **0.37** | **0.19** | **0.74** | **0.59** |
| \ | **balanced accuracy** | | | | | | | | | **NMI** | | | | | | | | |
| | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *MAG* | *WIK* | *NTU* | *COA* | *COC* | *CIC* | *NEW* | *PBC* | *DBA* | *MAG* |
| AHRC (LIN) | 0.68 | **0.92** | 0.65 | 0.65 | 0.62 | 0.56 | 0.52 | 0.69 | 0.56 | 0.57 | 0.76 | 0.41 | 0.41 | 0.31 | 0.24 | 0.13 | 0.55 | 0.27 |
| AHRC (LOG) | 0.66 | **0.92** | 0.60 | 0.60 | 0.60 | 0.56 | 0.52 | 0.68 | 0.54 | 0.57 | **0.78** | 0.34 | 0.34 | 0.29 | 0.24 | 0.13 | 0.50 | 0.24 |
| AHRC (EXP) | 0.68 | **0.92** | 0.64 | 0.62 | 0.62 | 0.55 | 0.52 | 0.68 | 0.56 | **0.58** | 0.76 | 0.40 | 0.38 | 0.31 | 0.25 | 0.14 | 0.54 | 0.27 |
| AHRC (SQR) | **0.69** | **0.92** | **0.73** | **0.69** | **0.67** | **0.58** | **0.53** | **0.77** | **0.67** | 0.52 | 0.71 | **0.49** | **0.46** | **0.36** | **0.27** | **0.14** | **0.61** | **0.49** |

**Table 6: Sensitivity: Clustering Quality of AHRC with Different Transformation Functions**

increasing pace slows down and the time costs remain relatively flat at peak. For example, on datasets *COC* and *CIC*, the time cost of AHRC becomes stable when $\tau$ is greater than 4. It is because the the data graphs are sparse, and nearly all edges are included in the sparsified graph when $\tau$ is large. This result echoes the complexity proved by Lemma 5. To achieve a good balance across different datasets and a trade-off between clustering effectiveness and efficiency, we chose 3 as the default value of $\tau$.

**Exp 5.** We evaluate the clustering quality of AHRC when parameter $\gamma$ varies from 2 to 6 with step size 1. Figure 5 reports the scores for 6 metrics on 8 datasets. As $\gamma$ increases, the trends in all six measures remain generally stable across all datasets. Specifically, when $\gamma$ increases from 2 to 6, the metric scores on average changes by 0%, 1%, −1%, and 1% at each step, respectively. Figure 7 shows the time cost of AHRC when $\gamma$ varies from 2 to 6 with step size 1. In general, the running time of AHRC rises as $\gamma$ increases. Specifically, when $\gamma$ increases from 2 to 6, per step, AHRC takes 76%, 224%, 470%, and 789% more time, respectively, compared to $\gamma = 2$. This result echoes the proof of Lemma 1. Therefore, considering the efficiency, we chose 2 as the default value of $\gamma$.

**Exp 6.** We quantitatively evaluate the clustering performance of AHRC using the square root transformation compared to other transformation methods. We consider 4 candidate transformation functions $\rho : [0,1] \mapsto [0,1]$: (Linear) $\rho(s) = s$, (Linear-over-Logarithm) $\rho(s) = \frac{s}{\log(1/s+1)}$, (Exponential) $\rho(s) = \frac{\exp(s)-1}{\exp(1)-1}$, where $\exp(1) = 2.71828\cdots$, and (Square-Root) $\rho(s) = \sqrt{s}$. Table 6 shows

| | TCL+ | | | GRC+ | | |
|---|---|---|---|---|---|---|
| **Dataset** | $p_d$ | $w_s$ | **learning rate** | $p_e$ | $p_a$ | **learning rate** |
| *C13* | 0.3 | $2^1$ | 5e-04 | 0.5 | 0.0 | 1e-03 |
| *WIK* | 0.3 | $2^{-4}$ | 5e-05 | 0.9 | 0.1 | 1e-03 |
| *NTU* | 0.7 | $2^{-2}$ | 1e-04 | 0.9 | 0.0 | 1e-03 |
| *COA* | 0.5 | $2^0$ | 1e-04 | 0.9 | 0.1 | 1e-03 |
| *COC* | 0.9 | $2^{-1}$ | 5e-04 | 0.5 | 0.0 | 1e-03 |
| *CIC* | 0.9 | $2^1$ | 5e-04 | 0.3 | 0.1 | 1e-03 |
| *NEW* | 0.9 | $2^{-3}$ | 5e-04 | 0.1 | 0.1 | 5e-03 |
| *PBC* | 0.7 | $2^{-2}$ | 5e-03 | 0.9 | 0.0 | 5e-03 |
| *DBA* | 0.9 | $2^1$ | 5e-03 | \ | \ | \ |

**Table 7: Hyperparameter settings on datasets**

the clustering performance of four variants of our method on 9 datasets: AHRC (LIN) denotes our algorithm using the Linear, AHRC (LIN) uses the Linear-over-Logarithm, AHRC (EXP) uses the Exponential, and AHRC (SQR) uses the Square-Root. In general, AHRC (SQR) achieves the best clustering performance. Specifically, on F-measure, AHRC (SQR) outperforms the other 3 variants (in top-down order as listed in Table 6) by 23%, 43%, and 25%, resp, averaged over datasets. On ARI, it outperforms the other 3 variants by 28%, 61%, and 32%, resp. On Jaccard Similarity, AHRC (SQR) is 28%, 53%, and 31% higher. On Purity, it outperforms other variants by an average of 11%, 22%, and 14%, resp. On Balanced Accuracy, it outperforms by 7%, 11%, and 8%, resp. On NMI, it is 16%, 26%, and 16% higher than other variants, resp.

## 7 Conclusions

This paper proposes AHRC, an attributed hypergraph clustering approach with cutting-edge clustering quality and scalability. The performance of AHRC attributes to three new designs of AHRC compared to existing methods: 1) a novel integration of multi-hop hypergraph topology and attributed information, 2) a new formulation multi-hop modularity for clustering, 3) an effective sparsification for improving the scalability, and 4) generalizability to enhance contrastive learning. Our experiments show that AHRC significantly outperforms the state-of-the-art methods on real-world hypergraphs and in particular, it is up to two orders of magnitude faster than the baseline methods.

## References


[1] Sameer Agarwal, Jongwoo Lim, Lihi Zelnik-Manor, Pietro Perona, David J. Kriegman, and Serge J. Belongie. 2005. Beyond Pairwise Clustering. In *CVPR*. 838–845. https://doi.org/10.1109/CVPR.2005.89

[2] William Aiello, Fan Chung, and Linyuan Lu. 2000. A random graph model for massive graphs. In *STOC*. 171–180. https://doi.org/10.1145/335305.335326

[3] Austin R. Benson, Rediet Abebe, Michael T. Schaub, Ali Jadbabaie, and Jon M. Kleinberg. 2018. Simplicial closure and higher-order link prediction. *Proc. Natl. Acad. Sci. USA* 115, 48 (2018), E11221–E11230. https://doi.org/10.1073/PNAS.1800683115

[4] Aritra Bhowmick, Mert Kosan, Zexi Huang, Ambuj K. Singh, and Sourav Medya. 2024. DGCLUSTER: A Neural Framework for Attributed Graph Clustering via Modularity Maximization. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*. AAAI Press, 11069–11077. https://doi.org/10.1609/AAAI.V38I10.28983

[5] Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast Unfolding of Communities in Large Networks. *J. Stat. Mech.* 2008, 10 (04 2008), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

[6] Béla Bollobás and Bela Bollobas. 1998. *Modern graph theory*. Vol. 184. Springer Science & Business Media. https://doi.org/10.1007/978-1-4612-0619-4

[7] Cécile Bothorel, Juan David Cruz, Matteo Magnani, and Barbora Micenková. 2015. Clustering attributed graphs: Models, measures and methods. *Netw. Sci.* 3, 3 (2015), 408–444. https://doi.org/10.1017/NWS.2015.9

[8] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. 2006. Maximizing modularity is hard. *arXiv preprint physics/0608255* (2006).

[9] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The Balanced Accuracy and Its Posterior Distribution. In *ICPR*. 3121–3124. https://doi.org/10.1109/ICPR.2010.764

[10] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S. Huang. 2011. Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 8 (2011), 1548–1560. https://doi.org/10.1109/TPAMI.2010.231

[11] Jie Chen, Haw-ren Fang, and Yousef Saad. 2009. Fast Approximate *k*NN Graph Construction for High Dimensional Data via Recursive Lanczos Bisection. *J. Mach. Learn. Res.* 10 (2009), 1989–2012. https://doi.org/10.5555/1577069.1755852

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1597–1607.

[13] Philip S. Chodrow, Nate Veldt, and Austin R. Benson. 2021. Generative hypergraph clustering: From blockmodels to modularity. *Science Advances* 7, 28 (2021), eabh1303. https://doi.org/10.1126/sciadv.abh1303

[14] Fan Chung and Linyuan Lu. 2002. The average distances in random graphs with given expected degrees. *Natl Acad. Sci.* 99, 25 (2002), 15879–15882. https://doi.org/10.1073/pnas.252631999

[15] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Phys. Rev. E* 70, 6 (2004), 066111. https://doi.org/10.1103/PhysRevE.70.066111

[16] David Combe, Christine Largeron, Elöd Egyed-Zsigmond, and Mathias Géry. 2012. Combining Relations and Text in Scientific Network Clustering. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012, Istanbul, Turkey, 26-29 August 2012*. IEEE Computer Society, 1248–1253. https://doi.org/10.1109/ASONAM.2012.215

[17] Robin Devooght, Amin Mantrach, Ilkka Kivimäki, Hugues Bersini, Alejandro Jaimes, and Marco Saerens. 2014. Random walks based modularity: application to semi-supervised learning. In *23rd International World Wide Web Conference, WWW '14*. 213–224. https://doi.org/10.1145/2566486.2567986

[18] Rundong Du, Barry L. Drake, and Haesun Park. 2019. Hybrid clustering based on content and connection structure using joint nonnegative matrix factorization. *J. Glob. Optim.* 74, 4 (2019), 861–877. https://doi.org/10.1007/s10898-017-0578-x

[19] Jordi Duch and Alex Arenas. 2005. Community detection in complex networks using extremal optimization. *Phys. Rev. E* 72, 2 (2005), 027104. https://doi.org/10.1103/PhysRevE.72.027104

[20] Issam Falih, Nistor Grozavu, Rushed Kanawati, and Younès Bennani. 2018. Community detection in Attributed Network. In *WWW*. ACM, 1299–1306. https://doi.org/10.1145/3184558.3191570

[21] Zijin Feng, Miao Qiao, and Hong Cheng. 2023. Modularity-based Hypergraph Clustering: Random Hypergraph Model, Hyperedge-cluster Relation, and Computation. *Proc. ACM Manag. Data* 1, 3, Article 215 (nov 2023), 25 pages. https://doi.org/10.1145/3617335

[22] Santo Fortunato. 2010. Community detection in graphs. *Phys. Reps.* 486, 3-5 (2010), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

[23] Andrew Gelman and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

[24] Rumi Ghosh and Kristina Lerman. 2008. Community Detection Using a Measure of Global Influence. In *Advances in Social Network Mining and Analysis, Second International Workshop, SNAKDD 2008 (Lecture Notes in Computer Science, Vol. 5498)*. 20–35. https://doi.org/10.1007/978-3-642-14929-0_2

[25] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In *SIGKDD*. 855–864. https://doi.org/10.1145/2939672.2939754

[26] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*. PMLR, 3887–3896.

[27] Koby Hayashi, Sinan G. Aksoy, Cheong Hee Park, and Haesun Park. 2020. Hypergraph Random Walks, Laplacians, and Clustering. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. ACM, 495–504. https://doi.org/10.1145/3340531.3412034

[28] Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *J. Classif.* 2, 1 (1985), 193–218. https://doi.org/10.1007/BF01908075

[29] Caiyan Jia, Yafang Li, Matthew B Carson, Xiaoyang Wang, and Jian Yu. 2017. Node attribute-enhanced community detection in complex networks. *Scientific reports* 7, 1 (2017), 2626.

[30] Jinhong Jung, Namyong Park, Lee Sael, and U Kang. 2017. BePI: Fast and Memory-Efficient Method for Billion-Scale Random Walk with Restart. In *SIGMOD*. ACM, 789–804. https://doi.org/10.1145/3035918.3035950

[31] Barakeel Fanseu Kamhoua, Lin Zhang, Kaili Ma, James Cheng, Bo Li, and Bo Han. 2021. HyperGraph Convolution Based Attributed HyperGraph Clustering. In *CIKM*. ACM, 453–463. https://doi.org/10.1145/3459637.3482437

[32] Bogumił Kamiński, Valérie Poulin, Paweł Prałat, Przemysław Szufel, and François Théberge. 2019. Clustering via hypergraph modularity. *PloS one* 14, 11 (2019), e0224307. https://doi.org/10.1371/journal.pone.0224307

[33] Ravi Kannan, Santosh S. Vempala, and Adrian Vetta. 2004. On clusterings: Good, bad and spectral. *J. ACM* 51, 3 (2004), 497–515. https://doi.org/10.1145/990308.990313

[34] Min-Soo Kim and Jiawei Han. 2009. A particle-and-density based evolutionary clustering method for dynamic networks. *VLDB* 2, 1 (2009), 622–633. https://doi.org/10.14778/1687627.1687698

[35] Sungwoong Kim, Sebastian Nowozin, Pushmeet Kohli, and Chang Yoo. 2011. Higher-order correlation clustering for image segmentation. *NIPS* 24 (2011), 1530–1538.

[36] Jon M. Kleinberg and Éva Tardos. 2006. *Algorithm design*. Addison-Wesley.

[37] Larkshmi Krishnamurthy, Joseph Nadeau, Gultekin Ozsoyoglu, M Ozsoyoglu, Greg Schaeffer, Murat Tasan, and Wanhong Xu. 2003. Pathways database system: an integrated system for biological pathways. *Bioinformatics* 19, 8 (2003), 930–937. https://doi.org/10.1093/bioinformatics/btg113

[38] Gayan K. Kulatilleke, Marius Portmann, and Shekhar S. Chandra. 2022. SCGC : Self-Supervised Contrastive Graph Clustering. *CoRR* abs/2204.12656 (2022). https://doi.org/10.48550/ARXIV.2204.12656 arXiv:2204.12656

[39] Tarun Kumar, Sankaran Vaidyanathan, Harini Ananthapadmanabhan, Srinivasan Parthasarathy, and Balaraman Ravindran. 2020. Hypergraph clustering by iteratively reweighted modularity maximization. *Appl. Netw. Sci.* 5, 1 (2020), 52. https://doi.org/10.1007/S41109-020-00300-3

[40] Tarun Kumar, Sankaran Vaidyanathan, Harini Ananthapadmanabhan, Srinivasan Parthasarathy, and Balaraman Ravindran. 2020. Hypergraph clustering by iteratively reweighted modularity maximization. *Appl. Netw. Sci.* 5, 1 (2020), 52. https://doi.org/10.1007/s41109-020-00300-3

[41] Kevin J. Lang. 2005. Fixing two weaknesses of the Spectral Method. In *NIPS*. 715–722.

[42] Dongjin Lee and Kijung Shin. 2023. I'm Me, We're Us, and I'm Us: Tri-directional Contrastive Learning on Hypergraphs. In *AAAI*. AAAI Press, 8456–8464. https://doi.org/10.1609/AAAI.V37I7.26019

[43] Pan Li and Olgica Milenkovic. 2017. Inhomogeneous Hypergraph Clustering with Applications. In *NIPS*, Vol. 30. 2308–2318.

[44] Pan Li and Olgica Milenkovic. 2018. Submodular hypergraphs: p-laplacians, cheeger inequalities and spectral clustering. In *ICML*. 3014–3023.

[45] Yiran Li, Renchi Yang, and Jieming Shi. 2023. Efficient and Effective Attributed Hypergraph Clustering via K-Nearest Neighbor Augmentation. *Proc. ACM Manag. Data* (2023), 116:1–116:23. https://doi.org/10.1145/3589261

[46] Changshu Liu, Liangjian Wen, Zhao Kang, Guangchun Luo, and Ling Tian. 2021. Self-supervised consensus representation learning for attributed graph. In *Proceedings of the 29th ACM international conference on multimedia*. 2654–2662.

[47] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

[48] Boaz Nadler and Meirav Galun. 2006. Fundamental Limitations of Spectral Clustering. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, Bernhard Schölkopf, John C. Platt, and Thomas Hofmann (Eds.). MIT Press, 1017–1024.

[49] Jennifer Neville, Micah Adler, and David Jensen. 2003. Clustering relational data using attribute and link information. In *Proceedings of the text mining and link analysis workshop, 18th international joint conference on artificial intelligence*. San Francisco, CA: Morgan Kaufmann Publishers, 9–15.

[50] Mark Newman. 2010. *Networks: An Introduction*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199206650.001.0001

[51] Mark EJ Newman. 2006. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* 74, 3 (2006), 036104. https://doi.org/10.1103/PhysRevE.74.036104

[52] RL Ott and Michael Longnecker. 2016. *An introduction to statistical methods and data analysis*. Cengage Learning.

[53] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: online learning of social representations. In *SIGKDD*. 701–710. https://doi.org/10.1145/2623330.2623732

[54] Satu Elisa Schaeffer. 2007. Graph clustering. *Comput. Sci. Rev.* 1, 1 (2007), 27–64. https://doi.org/10.1016/j.cosrev.2007.05.001

[55] Jianbo Shi and J. Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 8 (2000), 888–905. https://doi.org/10.1109/34.868688

[56] Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *TPAMI* 22, 8 (2000), 888–905. https://doi.org/10.1109/34.868688

[57] Yuuki Takai, Atsushi Miyauchi, Masahiro Ikeda, and Yuichi Yoshida. 2020. Hypergraph Clustering Based on PageRank. In *KDD*. 1970–1978. https://doi.org/10.1145/3394486.3403248

[58] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. In *WWW*. 1067–1077. https://doi.org/10.1145/2736277.2741093

[59] Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L. Dyer, Rémi Munos, Petar Velickovic, and Michal Valko. 2022. Large-Scale Representation Learning on Graphs via Bootstrapping. In *ICLR*. OpenReview.net. https://openreview.net/forum?id=0UXT6PpRpW

[60] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. 2006. Fast Random Walk with Restart and Its Applications. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China*. IEEE Computer Society, 613–622. https://doi.org/10.1109/ICDM.2006.70

[61] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2018. Deep graph infomax. *arXiv preprint arXiv:1809.10341* (2018).

[62] Ulrike von Luxburg. 2007. A tutorial on spectral clustering. *Stat. Comput.* 17, 4 (2007), 395–416. https://doi.org/10.1007/S11222-007-9033-Z

[63] Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing* 17 (2007), 395–416. https://doi.org/10.1007/s11222-007-9033-z

[64] Joyce Jiyoung Whang, Rundong Du, Sangwon Jung, Geon Lee, Barry L. Drake, Qingqing Liu, Seonggoo Kang, and Haesun Park. 2020. MEGA: Multi-View Semi-Supervised Clustering of Hypergraphs. *Proc. VLDB Endow.* 13, 5 (2020), 698–711. https://doi.org/10.14778/3377369.3377378

[65] Ming-Juan Wu, Ying-Lian Gao, Jin-Xing Liu, Chun-Hou Zheng, and Juan Wang. 2019. Integrative hypergraph regularization principal component analysis for sample clustering and co-expression genes network analysis on multi-omics data. *IEEE Journal of Biomedical and Health Informatics* 24, 6 (2019), 1823–1834.

[66] Xin Xia, Hongzhi Yin, Junliang Yu, Qinyong Wang, Lizhen Cui, and Xiangliang Zhang. 2021. Self-Supervised Hypergraph Convolutional Networks for Session-based Recommendation. In *AAAI*. AAAI Press, 4503–4511. https://doi.org/10.1609/AAAI.V35I5.16578

[67] Renchi Yang, Jieming Shi, Yin Yang, Keke Huang, Shiqi Zhang, and Xiaokui Xiao. 2021. Effective and Scalable Clustering on Massive Attributed Graphs. In *WWW*. ACM / IW3C2, 3675–3687. https://doi.org/10.1145/3442381.3449875

[68] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. 2018. Hierarchical Graph Representation Learning with Differentiable Pooling. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. 4805–4815. https://proceedings.neurips.cc/paper/2018/hash/e77dbaf6759253c7c6d0efc5690369c7-Abstract.html

[69] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph Contrastive Learning with Augmentations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/3fe230348e9a12c13120749e3f9fa4cd-Abstract.html

[70] Raphael Yuster and Uri Zwick. 2005. Fast sparse matrix multiplication. *ACM Trans. Algorithms* 1, 1 (2005), 2–13. https://doi.org/10.1145/1077464.1077466

[71] Lihi Zelnik-Manor and Pietro Perona. 2004. Self-Tuning Spectral Clustering. In *NIPS*. 1601–1608.

[72] Junwei Zhang, Min Gao, Junliang Yu, Lei Guo, Jundong Li, and Hongzhi Yin. 2021. Double-Scale Self-Supervised Hypergraph Learning for Group Recommendation. In *CIKM*. ACM, 2557–2567. https://doi.org/10.1145/3459637.3482426

[73] Shuqin Zhang and Hongyu Zhao. 2012. Community identification in networks with unbalanced structure. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 85, 6 (2012), 066114.

[74] Chen Zhe, Aixin Sun, and Xiaokui Xiao. 2019. Community Detection on Large Complex Attribute Network. In *KDD*. 2041–2049. https://doi.org/10.1145/3292500.3330721

[75] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. 2006. Learning with Hypergraphs: Clustering, Classification, and Embedding. In *NeurIPS*. 1601–1608. https://doi.org/10.7551/mitpress/7503.003.0205

[76] Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. 2009. Graph Clustering Based on Structural/Attribute Similarities. *Proc. VLDB Endow.* 2, 1 (2009), 718–729. https://doi.org/10.14778/1687627.1687709

[77] Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. 2010. Clustering Large Attributed Graphs: An Efficient Incremental Approach. In *2010 IEEE International Conference on Data Mining*. 689–698. https://doi.org/10.1109/ICDM.2010.41

[78] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2020. Deep Graph Contrastive Representation Learning. *CoRR* abs/2006.04131 (2020). arXiv:2006.04131 https://arxiv.org/abs/2006.04131


# A Hyperparameters

In this section, we perform sensitivity analysis on hyperparameters in TCL+ and GRC+. We conduct attributed hypergraph clustering on all datasets by varying hyperparameter values through a grid search, as commonly done in contrastive learning methods [42, 59, 78]. Clustering quality is evaluated using 6 metrics introduced in Section 6.1: F-measure, ARI, Jaccard Similarity, Purity, Balanced Accuracy, and NMI.

For GRC+, we investigate the impact of $p_e$ and $p_a$, which determine the portions of edges and node attributes to be removed, respectively. Figure 8 reports the metric scores with $p_e$ and $p_a$ values ranging from 0.1 to 0.9 in increments of 0.1. Due to space limits and the generally similar trends observed across metrics, we present the scores for F-measure and ARI. Figure 9 shows the scores for 6 metrics with the learning rate set to values from [1e−5, 5e−5, 1e−4, 5e−4, 1e−3, 5e−3].

For TCL+, we investigate the impact of $p_d$, which determines the probability that an entry in $\mathbf{S}$ is set to zero, and $w_s$ in Equation 7. Figure 10 reports the F-measure and ARI scores on all datasets with $p_d$ values ranging from 0.1 to 0.9 in increments of 0.2, and $w_s$ values chosen from $[2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3, 2^4]$. Figure 11 shows the scores for 6 metrics with the learning rate set to values from [1e−05, 5e−05, 1e−04, 5e−04, 1e−03, 5e−03].

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

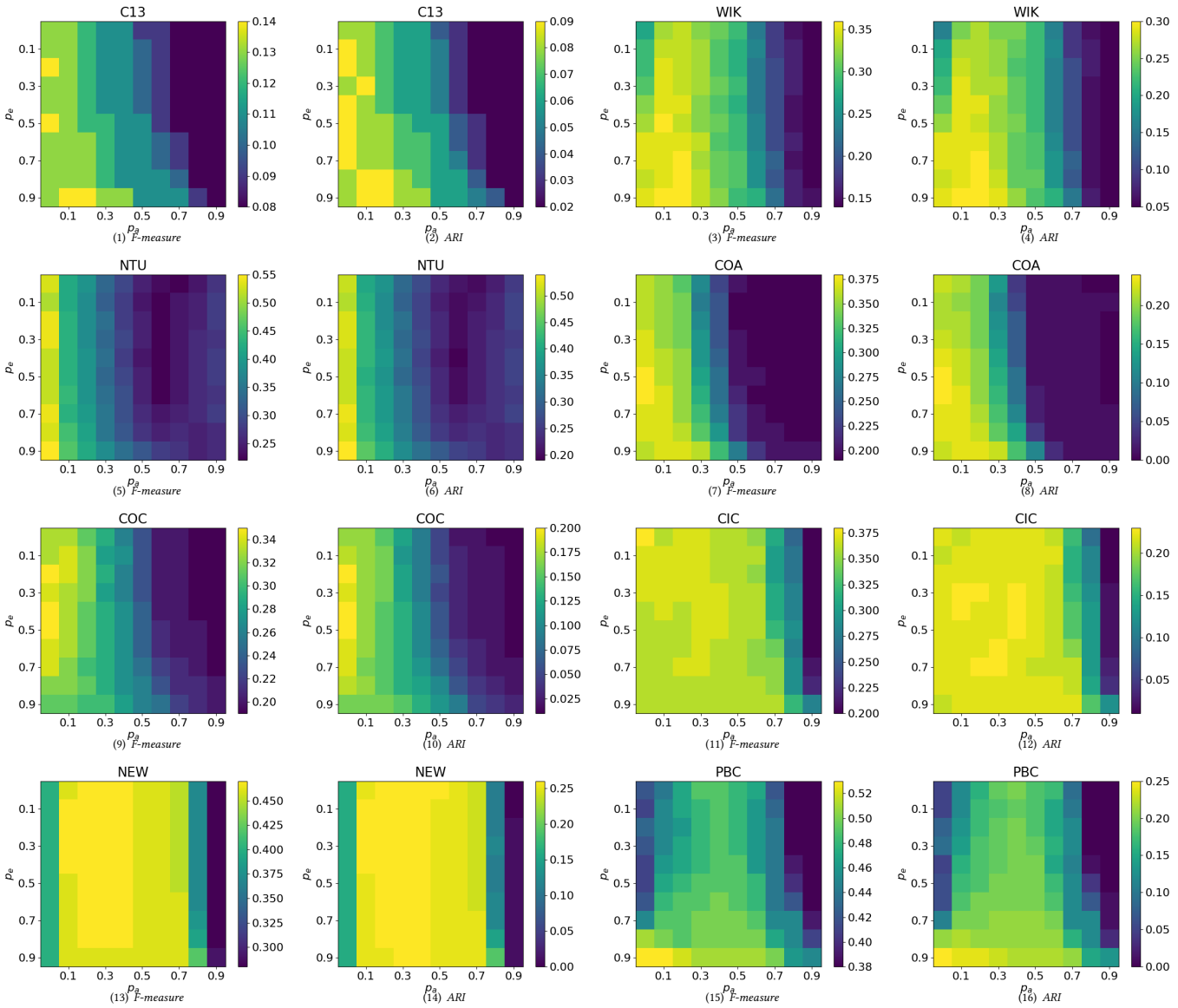



**Figure 8: Sensitivity: Clustering Quality of GRC+ on Varying Hyperparameters $p_e$ and $p_a$ in terms of F-measure and ARI**

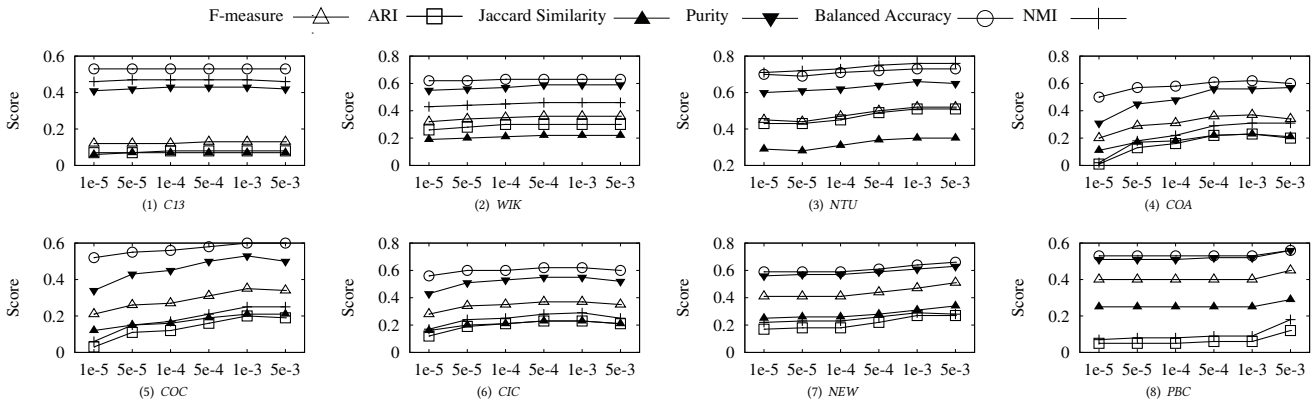



**Figure 9: Sensitivity: Clustering Quality of GRC+ on Varying Learning Rate**

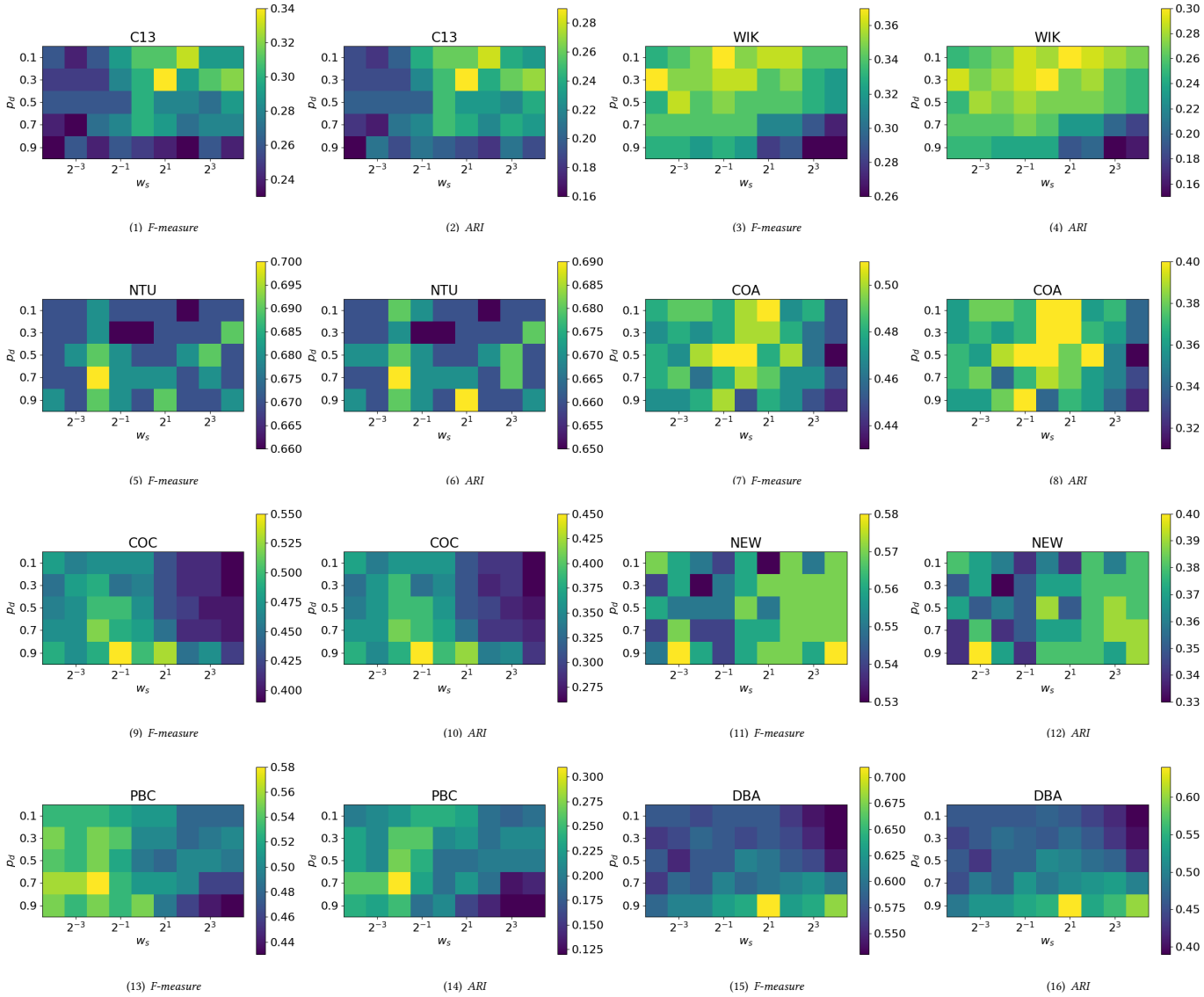



**Figure 10: Sensitivity: Clustering Quality of TCL+ on Varying Hyperparameters $p_d$ and $w_s$ in terms of F-measure and ARI**

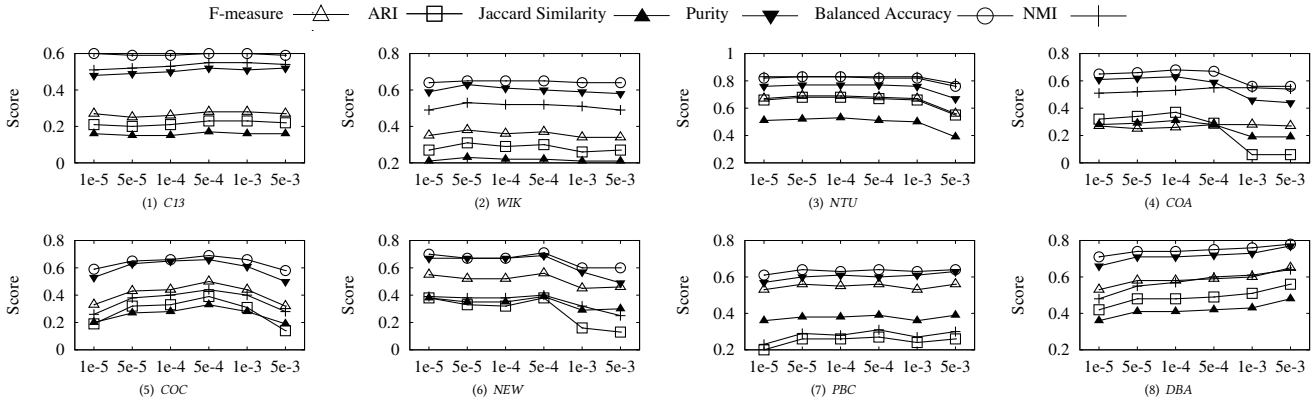



**Figure 11: Sensitivity: Clustering Quality of TCL+ on Varying Learning Rate**