EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

High-Order Local Clustering on Hypergraphs

Jingtian Wei[1,*],Zhengyi Yang[1,*,§],Qi Luo[1,*], Yu Zhang[2,†],Lu Qin[3,‡],Wenjie Zhang[1,*]

[1]University of New South Wales, Sydney, Australia
[2]University of New South Wales Canberra, Canberra, Australia
[3]University of Technology Sydney, Sydney, Australia

Abstract

Graphs are a commonly used model in data mining to represent complex relationships, with nodes representing entities and edges representing relationships. However, graphs have limitations in modeling high-order relationships. In contrast, hypergraphs offer a more versatile representation, allowing edges to join any number of nodes. This capability empowers hypergraphs to model multiple relationships and capture high-order information present in real-world applications. We focus on the problem of local clustering in hypergraphs, which computes a cluster near a given seed node. Although extensively explored in the context of graphs, this problem has received less attention for hypergraphs. Current methods often directly extend graph-based local clustering to hypergraphs, overlooking their inherent high-order features and resulting in low-quality local clusters. To address this, we propose an effective hypergraph local clustering model. This model introduces a novel conductance measurement that leverages the high-order properties of hypergraphs to assess cluster quality. Based on this new definition of hypergraph conductance, we propose a greedy algorithm to find local clusters in real time. Experimental evaluations and case studies on real-world datasets demonstrate the effectiveness of the proposed methods.

Keywords: Local Clustering, Hypergraphs, Conductance

Received on 30 09 2024; accepted on 12 11 2024; published on 15 11 2024

doi:10.4108/eetsis.7431

1. Introduction

A graph is a commonly used data structure in data mining that organizes real-world entities and their relationships through nodes and edges. Graphs naturally facilitate the representation of complex relationships, making them particularly advantageous for managing highly interrelated data. However, an edge in a graph can only connect two nodes, which limits its application in representing higher-order relationships in the real world. Hypergraphs, on the other hand, are a generalization of graphs in which each hyperedge can connect an arbitrary number of nodes. This provides a more comprehensive abstraction for representing n-ary interactions, enabling the capture of more complex high-order information compared to traditional graphs [5, 9, 19, 21–24, 34, 37].

Graph theory offers profound applications across a variety of real-world scenarios, leveraging unique properties of graphs and hypergraphs. For instance, in collaboration networks, academic publications with multiple co-authors can be represented as hyperedges, linking all authors who contributed to the work. This captures the collective nature of scholarly collaborations. In biological networks, protein complexes, which comprise multiple interacting proteins, are aptly modeled as hyperedges to reflect their intricate interactions. Similarly, in e-commerce, the bundling of products in shopping carts can be effectively analyzed using hypergraph modeling, where each set of purchased products forms a hyperedge, providing insights into consumer purchasing patterns.

Additionally, graph theory underpins significant advancements in intelligent communication systems as

[*]Email: {jingtian.wei,zhengyi.yang,qi.luo1,wenjie.zhang}@unsw.edu.au
[†]Email: m.yuzhang@unsw.edu.au
[‡]Email: lu.qin@uts.edu.au
[§]Zhengyi Yang is the corresponding author.

evidenced by recent research [25]. Social media platforms employ graph-based analyses to monitor user-generated content, aiming to preemptively identify and mitigate risks such as potential user suicides [27]. Furthermore, graph theoretical approaches have found applications in health domains, such as analyzing patterns in eye care [26]. These applications demonstrate the versatile utility of graph-based models in analyzing complex, interconnected data across diverse domains.

Local clustering is a fundamental topic in graph analysis. Specifically, it identifies a cluster of a seed node within the graph where the nodes are more densely connected to each other than to the rest of the graph. This concept is crucial for understanding the underlying structure and organization of complex networks. While the problem has been extensively explored for graphs [18, 33, 36], the study of local clustering in hypergraphs remains rather limited. In this paper, we investigate the problem of local clustering in hypergraphs.

Applications. Hypergraph local clustering is a pivotal technique in various fields, capitalizing on its ability to encapsulate complex, high-order relationships that surpass traditional graph models. This method finds extensive utility in domains such as web ranking and community detection [7, 8]. In the sphere of academic research, it is instrumental in discerning collaborative clusters within co-authorship networks, thereby revealing concealed scientific communities. Within the realm of social network analysis, hypergraph local clustering elucidates groups connected by common activities, providing insights into the underlying social dynamics [10]. In the commercial sector, particularly in e-commerce, this approach is employed to scrutinize patterns of product co-purchases, thereby refining marketing strategies and enhancing recommendation systems. Moreover, the method holds significant relevance in bioinformatics [11], where it facilitates the identification of protein complexes within protein-protein interaction networks. Notably, in the health and medical domains, hypergraph local clustering serves critical functions, as exemplified by its application in identifying disease modules and predicting protein functions [31]. These applications underscore the versatility and efficacy of local clustering in hypergraphs, particularly in capturing and analyzing multifaceted relational data. Practically, we highlight the following two example applications in the health/medical domain.

1. *Disease Gene/Drug Discovery.* Hypergraph local clustering enhances the identification of disease-associated genes and drugs by modeling relationships among genes, diseases, traits, and drug-target interactions. It uncovers that clusters provide insights into disease genetics and drug efficacy, identifying novel candidates for research and therapeutic development [29].

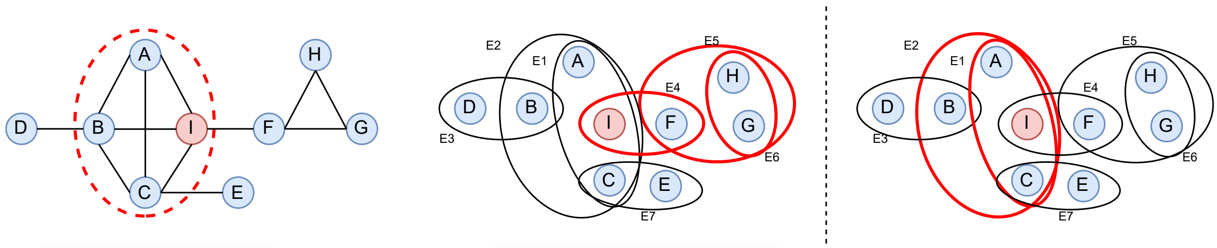
2. *Healthcare Analytics.* Hypergraph local clustering analyzes patient data across demographics, medical histories, and treatment outcomes. By identifying clusters of patients with similar profiles, it advances precision medicine by optimizing treatment strategies and discovering new patient subgroups for tailored interventions.

Motivations and Challenges. Traditional approaches to hypergraph local clustering often fall short due to their inability to effectively capture the high-order information in hypergraphs. Traditional approaches to hypergraph local clustering often struggle to effectively capture the complex, higher-order relationships present in hypergraphs. In local clustering, *conductance* is a key metric that assesses cluster quality by measuring the connectivity between a node subset and the rest of the graph. High conductance indicates a cluster has dense internal connections compared to external ones, which is crucial for identifying high-quality clusters. Commonly, hypergraphs are converted into clique graphs, where each hyperedge becomes a complete subgraph of pairwise edges, to apply graph-based clustering algorithms. This method, however, loses vital higher-order interaction data, reducing complex relationships to simple pairwise connections. As a result, conductance in these transformed graphs may not accurately reflect the true quality of the clusters in the original hypergraph.

Another method adapts graph-based clustering techniques directly to hypergraphs by applying conductance as is. Yet, this approach falls short because hyperedges connect multiple nodes at once, making simple edge counting insufficient for a true assessment of cluster quality. This highlights the need for hypergraph-specific metrics that can accurately reflect their distinct structures.

Example 1. In Figure 1: The three graphs show local clustering by node I as the seed node, which produces different clustering results based on different graph classes as well as metrics. Figure 1A is the clique graph of hypergraph in Figure 1B. Both use the ratio of the number of inner and outer edges as conductance for local clustering methods. Figure 1C then shows the local clustering performed based on our method.

In Figure 1 A, hyperedges e_1 and e_6 are ignored due to the inclusion of the hyperedge, leading to the loss of information inherent in the hypergraph. This results in a reduction in the perceived importance of these hyperedges. This phenomenon demonstrates that conductance calculation methods effective for graphs are not fully applicable to hypergraphs. Moreover, the process of converting hypergraphs to graphs affects the internal information of the hypergraphs, resulting in a degradation of clustering quality.



A. Edge Based Conductance

B. Hyperedge Based Conductance

C. Degree Based Conductance

Figure 1. An overview of local clustering on Clique Graph and Hypergraph.

The clustering differences observed in Figure 1 B and C underscore that varying definitions of conductance within hypergraphs can yield distinct clustering outcomes. The cluster in Figure 1 C is more closely related to the seed nodes. This highlights the importance of carefully designing metrics for local clustering that align with the intrinsic characteristics of hypergraphs. It is crucial to develop conductance metrics that accurately reflect the complex relationships and interactions within hypergraphs to ensure high-quality clustering results.

Contributions. For better quality hypergraph local clustering, we propose a novel method for computing high-order conductance specifically tailored for hypergraphs. Additionally, we introduce a greedy algorithm designed to identify local clusters within this framework. We summarize the contributions of this paper as follows.

1. *New Definition for Hypergraph Conductance.* We introduce a degree-based high-order conductance metric for hypergraphs, considering their unique properties and characteristics.

2. *Algorithms for Hypergraph Local Clustering.* We propose a greedy algorithm capable of identifying local clusters within hypergraphs, utilizing the newly devised conductance metric.

3. *Extensive Experiments and Case Studies.* We conduct extensive experiments and case studies to validate the effectiveness of our methods, demonstrating their practical applicability and superiority in various scenarios.

Organization. The paper is structured as follows. Section 2 reviews related works. Section 3 introduces the background. Our proposed method for hypergraph local clustering is detailed in Section 4. Section 5 presents experimental results, while Section 6 concludes the paper.

2. Related Works

Clustering on Graphs. Graphs are extensively used in data mining to represent complex relationships, with nodes representing entities and edges representing their interactions. Clustering involves classifying nodes in a graph into cohesive clusters based on common characteristics. Several clustering methods have been developed based on various features in the graph, such as Distance-Based Structural Graph Clustering [20], Possibilistic Fuzzy C-means Clustering [2] and Conductance-Based Graph Clustering [18]. Additionally, methods like SA-Cluster [36] and ACMin [33] address attributed graph clustering problems. Local clustering, in contrast, focuses on finding clusters closely related to a given seed node. One notable algorithm for local clustering in graphs is NIBBLE [28]. This algorithm leverages Personalized PageRank (PPR) values to identify small sets of nodes (clusters) with low conductance. NIBBLE performs a series of random walks and uses the resulting PPR values to find a set of nodes that form a cluster around a seed node.

Clustering on Hypergraphs. Hypergraph-based models have become increasingly prominent in research. Unlike traditional graphs, hypergraphs can capture higher-order information and represent more complex relationships. Numerous hypergraph-based clustering methods have been developed in recent years. For instance, JNMF [6] and AHCKA [17] are both attributed hypergraph clustering methods. Addressing the problem of local clustering in hypergraphs, a sweep cut method has been proposed [30] to identify high-quality clusters containing a given node. This method uses hypergraph Personalized PageRank (PPR) values to effectively determine the clusters. Additionally, the *HyperGo* algorithm [16] adapts the *NIBBLE* algorithm [28]—a truncated random walk-based local algorithm for graph partitioning—to hypergraphs. Some earlier methods [1, 38] converted hypergraphs to graphs before processing them, which resulted in the loss of higher-order information inherent in the hypergraph. Current approaches often either extend graph-based local clustering directly to hypergraphs or transform hypergraphs into graphs, overlooking their inherent higher-order features, leading to low-quality local clustering.

3. Preliminaries

In this paper, we address the problem of local clustering in hypergraphs. This section outlines the preliminaries, beginning with an introduction to the notations employed throughout the paper. Subsequently, we provide the definitions of hypergraphs and local clustering, and discuss the concept of conductance in the context of graphs.

Definition 1 (**Hypergraph**). Let $H = \{V, E\}$ denote an unweighted hypergraph, where V is the set of nodes consisting of n nodes, and E is the set of hyperedges consisting of m hyperedges. Each hyperedge $e \in E$ is a subset of V.

If a node v exists in hyperedge e, then the hyperedge e is *incident* with the node v, and vice versa. For each node $v \in V$, the *degree* of node v is represented as $\delta(v)$, which is defined as the number of hyperedges incident to node v. Similarly, for each hyperedge $e \in E$, the degree of hyperedge e is represented as $\delta(e)$, which is defined as the number of nodes incident to hyperedge e.

Definition 2 (**Dual-Hypergraph**). The dual-hypergraph of a hypergraph $H = \{V, E\}$ is hypergraph $H^* = \{V^*, E^*\}$. The nodes of the dual hypergraph V^* correspond to the hyperedges E of the original hypergraph H. The hyperedges E^* of the dual hypergraph correspond to nodes V of the original hypergraph H.

Definition 3 (**Sub-Hypergraph**). Given a hypergraph $H = \{V, E\}$, where V is the set of nodes and E is the set of hyperedges, a sub-hypergraph H' is defined as $H' = \{V', E'\}$. $V' \subseteq V$ is a subset of nodes in V. $E' \subseteq E$ is a subset of hyperedges from E.

Definition 4 (**Local Clustering**). Given an unweighted hypergraph H and a seed node s, this paper focuses on the methodology for local clustering in hypergraphs. The objective is to identify a cluster S within the hypergraph, originating from the given seed node or seed hyperedge s.

Remark. The resulting cluster S must satisfy the following criteria:

1. *Relevance to the Seed.* The nodes and hyperedges within the cluster should exhibit a high degree of relevance to the initial seed node or seed hyperedge. This ensures that the cluster is contextually meaningful and closely related to the origin node.
2. *Locally Optimal Quality.* The cluster S should achieve a locally optimal quality, defined by specific metrics or criteria pertinent to the structure and properties of the hypergraph. This optimally ensures that the cluster is well-formed and robust within its local context, even if it is not globally optimal.

Definition 5 (**Conductance on Graphs**). Conductance is a measure of the proportion of relationships that connect nodes within a cluster S relative to those that connect nodes in S to nodes outside S. To compute conductance, count the total number of edges that lie entirely within the cluster S (internal edges) and the edges that link nodes in S to nodes outside S (external edges). The conductance is then defined as the ratio of internal edges to external edges. This ratio reflects the cluster's connectivity relative to its separation from the rest of the graph.

$$\phi_c(S) = \frac{\sum_{e \in E_{\text{int}}} |e|}{\sum_{e \in E_{\text{ext}}} |e|}. \tag{1}$$

Conductance is a widely recognized metric for assessing cluster quality, particularly in graph-based community detection [32, 35]. Studies have demonstrated that conductance is effective in evaluating the authenticity of community structures within real-world graphs. As outlined in Definition 5, conductance quantifies the balance between internal and external connectivity of a cluster [4, 14]. Generally, clusters with higher conductance are indicative of more cohesive and well-defined communities. This metric is especially valuable in applications where preserving community integrity is essential, as it provides insight into the cluster's connectivity profile relative to its surroundings [12].

Example 2. Figure 2 depicts a cluster within a graph, outlined by a red dashed boundary. The edges that connect nodes entirely within the cluster (internal edges) and those that link nodes in the cluster to nodes outside of it (external edges) are shown with green dashed arrows, respectively. The conductance of the cluster is calculated using Equation 1, based on the ratio of internal to external connections. This illustration helps visualize how conductance quantifies the balance between internal cohesion and external connectivity for the cluster.

Example 3. Figure 3 presents an example of a cluster within a hypergraph, delineated by a black dashed line. The green dashed arrow indicates the hyperedges that are fully contained within the cluster, while the two red hyperedges represent those that are only partially contained. The conductance of the hypergraph cluster can be calculated using Equation 2 and 3.

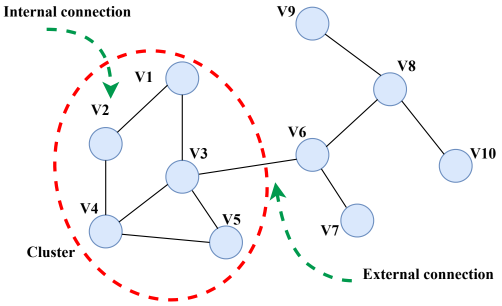


Figure 2. Graph Conductance

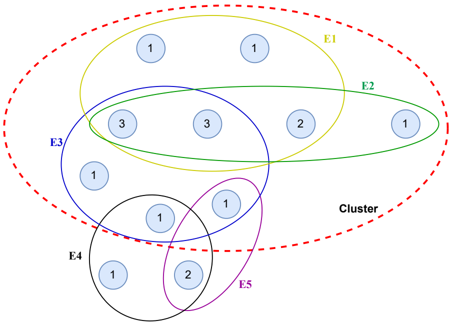


Figure 3. Hypergraph Conductance

4. Our Approach

In our proposed method for calculating conductance, we consider both the degrees of nodes whose hyperedges are fully contained within the cluster S and nodes whose hyperedges are only partially contained within S.

Definition 6 (**High-Order Conductance Calculation**). We define the conductance as the ratio between the sum of the degrees of nodes which in fully contained hyperedges and the sum of the degrees of nodes in partially connected hyperedges of the cluster. Formally,

$$\phi_{hc}(S) = \frac{\sum_{v \in S} \delta(v)}{\sum_{v \in \overline{S}} \delta(v)}. \tag{2}$$

Remark. The conductance can also be computed from the degree of the hyperedge:

$$\phi_{hc}(S) = \frac{\sum_{v \in S} \delta(e)}{\sum_{v \in \overline{S}} \delta(e)}. \tag{3}$$

In conventional graph theory, the method of calculating conductance does not account for the higher-order relationships inherent in hyperedges within a hypergraph. To address this limitation, a novel hypergraph-based conductance calculation method is proposed, as outlined in Definition 6. Instead of directly computing the ratio of the number of edges, this method employs the ratio of the degrees of internal and external nodes within the cluster S to determine conductance. This approach is designed to better capture the complexity of hypergraph structures. Consequently, clusters exhibiting higher conductance values are indicative of superior quality, reflecting a more robust and meaningful grouping of nodes within the hypergraph context.

Algorithm 1 calculates the conductance of a cluster within a hypergraph. The conductance metric is used to evaluate the quality of the cluster by examining the proportion of hyperedges that are fully contained within the cluster versus those that are only partially connected. The hypergraph H for which the conductance is being calculated. It consists of a set of hyperedges, each connecting a subset of nodes. The cluster S of nodes within the hypergraph for which the conductance is to be calculated. A numerical value conductance ϕ_{hc} representing the conductance of the cluster within the hypergraph. The algorithm effectively evaluates the quality of clusters within a hypergraph by analyzing the distribution of the sum of degrees of fully contained and partially connected hyperedges. A higher conductance value signifies a higher quality cluster, as it indicates that hyperedges entirely within the cluster exhibit higher degrees. In contrast, hyperedges with lower degrees are either neighboring hyperedges or located outside the cluster. This approach ensures that clusters accurately capture the intrinsic structure of the hypergraph by emphasizing the connectivity and degree of hyperedges within the cluster.

Algorithm 1 Calculate Hypergraph Conductance

```
Require: Hypergraph H, Cluster S
Ensure: Conductance $\phi_{hc}$
function CALCULATE_HYPERGRAPH_CONDUCTANCE(H, S)
    for each (he_id, he) in H.hyperedges do
        if he $\cap$ S $\neq \emptyset$ then
            if he $\subseteq$ S then
                Fully.add(he_id)
            else
                Partially.add(he_id)
    numerator $\leftarrow \sum_{he\_id \in F} \mathrm{len}(H.hyperedges[he\_id])$
    denominator $\leftarrow \sum_{he\_id \in P} \mathrm{len}(H.hyperedges[he\_id])$
    if denominator = 0 then
        return $\infty$
    else
        return numerator / denominator
```

Greedy Local Clustering. Algorithm 2 identifies the next hyperedge to be added to a cluster to maximize the cluster's conductance. Conductance is used as a metric to assess the quality of the cluster, and the

Algorithm 2 Find Next Hyperedge to Maximize Conductance

```
Require: Hypergraph $H$, Cluster $S$
Ensure: next_hyperedge_id, max_conductance
function FIND_NEXT_HYPEREDGE(H, S)
    for each (he_id, he) in $H$.hyperedges do
        if $he \cap S \neq \emptyset$ and $he \not\subseteq S$ then
            potential_cluster ← $S \cup he$
            conductance ← cal_con($H$, potential_cluster)
            if conductance > max_conductance then
                max_conductance ← conductance
                next_hyperedge_id ← he_id
    return next_hyperedge_id, max_conductance
```

algorithm aims to iteratively improve this quality by adding the optimal hyperedge. The hypergraph H within which the clustering is performed. It consists of a set of hyperedges, each connecting multiple nodes. The current cluster S of nodes within the hypergraph. The output of Algorithm 2 is the hyperedge id and the maximum conductance of the cluster S. The hyperedge identifier that, when added to the cluster, maximizes the conductance. The maximum conductance value achieved by adding the identified hyperedge to the cluster.

Algorithm 3 identifies a cluster with maximal conductance within a hypergraph, starting from a given seed node s. Conductance ϕ_{hc} is a metric used to evaluate the quality of the cluster, with a higher conductance indicating a higher quality cluster. The outputs are the final cluster of nodes identified as having maximal conductance and the conductance value of the final cluster S.

This two algorithms iteratively assess the potential benefits of incorporating each hyperedge into the current cluster, ensuring that the selected hyperedge is the one that optimally enhances the cluster's conductance. This iterative process incrementally improves the cluster's quality by continuously optimizing its conductance at each step until the conductance of the cluster cannot be increased. By leveraging this approach, the algorithm can efficiently identify clusters with locally optimal quality.

5. Experiments

Hardware. The algorithms are all implemented in Python. All of the experiments are conducted on a Linux machine with Apple M1 Pro chip with 3.22GHz and 32GB unified memory. In this section, we first introduce the datasets, parameters and baseline methods used in this experiment. Then the experimental settings will also be given. Finally, we will analyze the results of the experiment and give case studies.

Datasets. We use two real-world network datasets: Amazon [15], and DBLP [3]. Amazon is a product co-purchasing network dataset, where the nodes of the dataset represent the co-purchase scenarios for commodities and hyperedges represent products. Also in this dataset, each product belongs to one or more hierarchically organized product categories, and products of the same category define a group, which is considered a ground-truth community [32]. Same as DBLP scientific collaboration network dataset. Nodes represent the paper that authors co-authored, and hyperedges represent the authors. Authors publishing in the same research area (conference or journal types) can form a ground-truth communities. Dataset *Amazon-5000* consists 60,320 nodes and 15,845 hyperedges, average degree of hyperedges is 32.19, min and max hyperedge degree are 1 and 101. *DBLP-5000* consists 197,891 nodes and 87,391 hyperedges, average degree of hyperedges is 47.09, min and max hyperedge degree are 1 and 204.

Baselines. Our proposed hypergraph-based high-order local clustering method is compared with two baselines:

1. *Edge-Based Method:* Using an edge-based conductance calculation method shown in equation 1 as a judgment condition applied to the same greedy algorithm allows for a better validation of the effectiveness of the new conductance calculation method.

2. *NIBBLE:* The parameters of baseline method are set as: $\alpha = 0.15$, $\epsilon = 1e - 5$, $window = 3$. The experimental hypergraph data is transformed into clique graph data and local clustering is carried out using the NIBBLE algorithm, and the results can be analyzed to determine whether transforming the hypergraph into a graph will

Algorithm 3 Greedy Walk Algorithm

```
Require: Hypergraph H, Seed Node s
Ensure: Cluster S, Conductance $\phi_{hc}$
function FIND_CLUSTER_WITH_MAX_CONDUCTANCE(H, s)
    while True do
        next_hyperedge_id, max_conductance ← FIND_NEXT_HE(H, S)
        if max_conductance ≤ current_conductance then
            break
        current_conductance ← max_conductance
        S ← S ∪ H.hyperedges[next_hyperedge_id]
    return S, current_conductance
```

Table 1. Comparison Results

Datasets	Amazon-5000		DBLP-5000	
Methods	Avg Prec.	Size	Avg Prec.	Size
Deg-Con Greedy	1.0	13.42	0.8	18.84
Edge-Con Greedy	1.0	12.04	0.8	14.68
NIBBLE on Clique	1.0	9.36	0.5	6.88

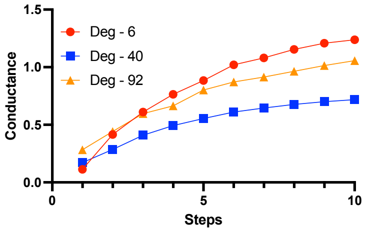


Figure 4. Conductance

result in the loss of higher information and thus make the clustering effect decrease.

Settings. We pre-process the datasets to generate hypergraph datasets. Specifically, we select the top 5000 communities with the highest quality from each dataset to construct the hypergraph datasets. We then sort all seed IDs and divide them equally into 100 portions, randomly selecting one ID from each portion. This procedure yields a randomized set of 100 seed input IDs for our experiments. Additionally, we set a maximum of 10 steps for all methods, allowing a maximum of 10 IDs to be output in the cluster.

Exp 1 - Evaluation of Effectiveness. The effectiveness of the three methods applied to the two datasets is presented in Table 1. This analysis primarily focuses on the precision of the methods' results and the sizes of the resulting data clusters. The average precision is determined by comparing the experimental results to the ground truth datasets. The cluster size is calculated using the harmonic mean of the number of hyperedges and nodes.

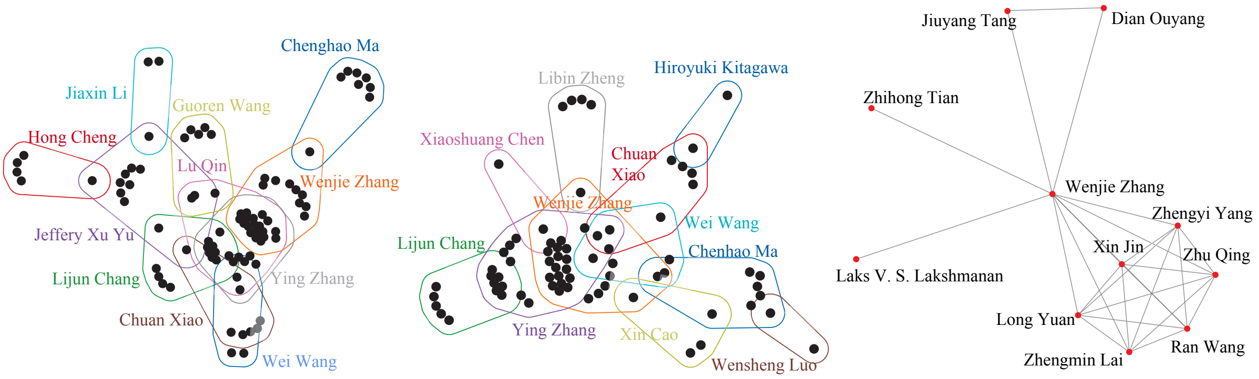
The results indicate that our method performs optimally in the same experimental settings. Despite ensuring precision, our method consistently produces the largest cluster size within the constraint of a maximum of 10 steps. Our method effectively captures nodes related to seed IDs that contain a higher information content within the hypergraph structure. This demonstrates that our method can identify a substantial number of high-quality IDs related to the seed ID while maintaining a certain level of precision.

Exp 2 - Conductance with Varying Seeds. We sort the hyperedges by their degree size, divide them equally into three parts, and randomly select one hyperedge from each part as seed input. Figure 4 illustrates the impact of the degree magnitude of the seed hyperedge on the growth curve of conductance using our method. It is observed that the conductance growth curve for the ID with a median degree is not as rapid as those for the other two IDs. This slower growth is attributed to the average degrees of its neighboring IDs, which lack significantly high-degree IDs that could quickly enhance the cluster conductance.

Exp 3 - Case study. We conducted a case study using the DBLP dataset to evaluate our clustering approach, using the author “Wenjie Zhang” as the seed author. Cluster A is the result of our method, while Clusters B and C are generated using the two baseline methods, respectively. In Clusters A and B, hyperedges represent authors and nodes represent co-authorships. Conversely, in Cluster C, nodes represent authors and edges represent co-authorships. Figure 5 illustrates the results, each contains the top-10 authors in the local cluster.

Cluster A includes authors who are highly correlated with the seed author, such as "Ying Zhang", "Lu Qin", "Chuan Xiao", and "Jeffery Xu Yu". These individuals are prominent researchers in the field of graph mining. As illustrated in the cluster, they have co-authored many papers. Additionally, several of these authors have worked at the same institution and completed their PhD studies under the same supervisor. This close academic and research relationship underscores the effectiveness of our method in identifying and clustering authors with significant collaborative and academic connections. In Cluster B, however, some authors with strong relationships to the seed author in Cluster A, such as "Lu Qin" and "Jeffery Xu Yu", are not present. Many of the authors in Cluster B, such as "Hiroyuki Kitagawa" and "Wensheng Luo", do not have deep relationships with the seed author in either their field of study or their academic journey. This disparity highlights the limitations of the baseline method used for Cluster B, which fails to capture the significant academic and collaborative connections that our method effectively identifies in Cluster A. Several co-authors in Cluster C appear in the same article [13] as the seed author. By transforming a hypergraph into a clique graph, nodes that belong to a single hyperedge become densely connected. This transformation causes local clustering methods in clique graphs to prioritize the complete identification of all nodes within the same hyperedge, especially hyperedges with high degree.

The distributions of margins from each author to the seed author are presented in Figure 6. In Cluster C, all authors have a distance of 1 from the seed author, indicating that this baseline method consistently selects



Cluster A: Deg-Based

Cluster B: Edge-Based

Cluster C: Nibble-Based

Figure 5. Local Clusters for Wenjie Zhang (ID 177)

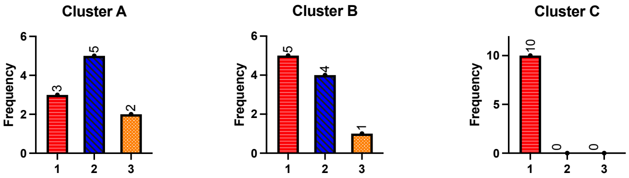


Figure 6. Edge distance from seed id

nodes with a direct connection to the seed. When comparing the margin distributions of Clusters A and B, it is evident that the overall margin of Cluster A is larger than that of Cluster B. This suggests that our method prioritizes authors who are farther from the seed author, incorporating these important but less directly connected authors into the cluster. This broader consideration highlights the effectiveness of our approach in identifying and clustering significant authors in the dataset.

6. Conclusion

In this paper, we study the problem of local clustering in hypergraphs. We observe that existing methods for hypergraph local clustering often overlook high-order information of entities, and propose a novel concept of high-order conductance for clusters, along with a greedy algorithm that utilizes this concept. This improved method leverages the complex structural details unique to hypergraphs, thereby identifying higher-quality clusters. The experimental results prove that our approach effectively captures this higher-order information.

References

[1] Agarwal, S., Lim, J., Zelnik-Manor, L., Perona, P., Kriegman, D.J., Belongie, S.J.: Beyond pairwise clustering. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society (2005)

[2] Arora, J., Tushir, M., Dadhwal, S.K.: A new suppression-based possibilistic fuzzy c-means clustering algorithm. EAI Endorsed Transactions on Scalable Information Systems (2023)

[3] Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X.: Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD (2006)

[4] Chierichetti, F., Lattanzi, S., Panconesi, A.: Rumour spreading and graph conductance. In: Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms. SIAM (2010)

[5] Deng, N., Wang, Y., Huang, G., Zhou, Y., Li, Y.: Semantic coherence analysis of english texts based on sentence semantic graphs. EAI Endorsed Transactions on Scalable Information Systems (2023)

[6] Du, R., Drake, B.L., Park, H.: Hybrid clustering based on content and connection structure using joint nonnegative matrix factorization. J. Glob. Optim. (2019)

[7] Epasto, A., Feldman, J., Lattanzi, S., Leonardi, S., Mirrokni, V.: Reduce and aggregate: similarity ranking in multi-categorical bipartite graphs. In: Proceedings of the 23rd international conference on World wide web (2014)

[8] Gargi, U., Lu, W., Mirrokni, V., Yoon, S.: Large-scale community detection on youtube for topic discovery and exploration. In: Proceedings of the International AAAI Conference on Web and Social Media (2011)

[9] Gong, X., Wang, H., Wang, X., Chen, C., Zhang, W., Zhang, Y.: Influence maximization on hypergraphs via multi-hop influence estimation. Information Processing & Management (2024)

[10] Jeub, L.G., Balachandran, P., Porter, M.A., Mucha, P.J., Mahoney, M.W.: Think locally, act locally: Detection of small, medium-sized, and large communities in large networks. Physical Review E (2015)

[11] Jiang, B.B., Wang, J.G., Wang, Y., Xiao, J.: Gene prioritization for type 2 diabetes in tissue-specific protein interaction networks. Systems Biology (2009)

[12] Koutis, I., Miller, G.L.: Graph partitioning into isolated, high conductance clusters: Theory, computation and applications to preconditioning. In: Proceedings of the twentieth annual symposium on Parallelism in algorithms and architectures (2008)

[13] Lai, L., Qing, Z., Yang, Z., Jin, X., Lai, Z., Wang, R., Hao, K., Lin, X., Qin, L., Zhang, W., et al.: Distributed subgraph matching on timely dataflow. Proceedings of the VLDB Endowment **12**(10) (2019)

[14] Lang, K., Rao, S.: A flow-based method for improving the expansion or conductance of graph cuts. In: International Conference on Integer Programming and Combinatorial Optimization. Springer (2004)

[15] Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Transactions on the Web (TWEB) **1**(1), 5–es (2007)

[16] Li, J., He, J., Zhu, Y.: E-tail product return prediction via hypergraph-based local graph cut. In: SIGKDD. pp. 519–527 (2018)

[17] Li, Y., Yang, R., Shi, J.: Efficient and effective attributed hypergraph clustering via k-nearest neighbor augmentation. Proc. ACM Manag. Data **1**(2) (2023)

[18] Lin, L., Li, R., Jia, T.: Scalable and effective conductance-based graph clustering. In: IAAI. pp. 4471–4478 (2023)

[19] Liu, B., Zhang, F., Zhang, W., Lin, X., Zhang, Y.: Efficient community search with size constraint. In: 37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021. IEEE (2021)

[20] Liu, K., Wang, S., Zhang, Y., Xing, C.: An efficient algorithm for distance-based structural graph clustering. Proc. ACM Manag. Data **1**(1), 45:1–45:25 (2023)

[21] Luo, L., Fang, Y., Cao, X., Zhang, X., Zhang, W.: Detecting communities from heterogeneous graphs: A context path-based graph neural network model. In: Proceedings of the 30th ACM international conference on information & knowledge management (2021)

[22] Luo, Q., Yu, D., Cai, Z., Lin, X., Wang, G., Cheng, X.: Toward maintenance of hypercores in large-scale dynamic hypergraphs. The VLDB Journal **32**(3) (2023)

[23] Luo, Q., Yu, D., Liu, Y., Zheng, Y., Cheng, X., Lin, X.: Finer-grained engagement in hypergraphs. In: (ICDE). IEEE (2023)

[24] Luo, Q., Zhang, W., Yang, Z., Wen, D., Wang, X., Yu, D., Lin, X.: Hierarchical structure construction on hypergraphs. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (2024)

[25] Nahi, H.A., Al-dolaimy, F., Abbas, F.H., Almohamadi, M., Hasan, M.A., Alkhafaji, M.A., Guneser, M.T.: A multi-objective optimization for enhancing the efficiency of service in flying ad-hoc network environment. EAI Endorsed Transactions on Scalable Information Systems (2023)

[26] Sarki, R., Ahmed, K., Wang, H., Zhang, Y., Wang, K.: Convolutional neural network for multi-class classification of diabetic eye disease. EAI Endorsed Transactions on Scalable Information Systems (2021)

[27] Singh, R., Subramani, S., Du, J., Zhang, Y., Wang, H., Miao, Y., Ahmed, K.: Antisocial behavior identification from twitter feeds using traditional machine learning algorithms and deep learning. EAI Endorsed Transactions on Scalable Information Systems (2023)

[28] Spielman, D.A., Teng, S.: A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. SIAM J. Comput. (2013)

[29] Su, G., Wang, H., Zhang, Y., Coster, A.C., Wilkins, M., Canete, P.F., Yu, D., Yang, Y., zhang, w.: Inferring gene regulatory networks by hypergraph variational autoencoder. bioRxiv (2024)

[30] Takai, Y., Miyauchi, A., Ikeda, M., Yoshida, Y.: Hypergraph clustering based on pagerank. In: KDD. pp. 1970–1978. ACM (2020)

[31] Voevodski, K., Teng, S.H., Xia, Y.: Spectral affinity in protein networks. BMC systems biology (2009)

[32] Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. In: ACM SIGKDD Workshop (2012)

[33] Yang, R., Shi, J., Yang, Y., Huang, K., Zhang, S., Xiao, X.: Effective and scalable clustering on massive attributed graphs. In: WWW. ACM / IW3C2 (2021)

[34] Yang, Z., Zhang, W., Lin, X., Zhang, Y., Li, S.: Hgmatch: A match-by-hyperedge approach for subgraph matching on hypergraphs. In: (ICDE) (2023)

[35] Zhang, Y., Rohe, K.: Understanding regularized spectral clustering via graph conductance. Advances in Neural Information Processing Systems (2018)

[36] Zhou, Y., Cheng, H., Yu, J.X.: Clustering large attributed graphs: An efficient incremental approach. In: ICDM 2010, The 10th IEEE. IEEE (2010)

[37] Zhou, Z., Zhang, F., Lin, X., Zhang, W., Chen, C.: K-core maximization: An edge addition approach. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019. ijcai.org (2019)

[38] Zien, J.Y., Schlag, M.D.F., Chan, P.K.: Multilevel spectral hypergraph partitioning with arbitrary vertex sizes. IEEE **18**(9), 1389–1399 (1999)