Identifying Hierarchical Community Structures in Content-based Scholarly Social Networks

1st Md Asaduzzaman Noor
Gianforte School of Computing
Montana State University
Bozeman, Montana, USA
mdasaduzzamannoor@montana.edu

2nd John W. Sheppard
Gianforte School of Computing
Montana State University
Bozeman, Montana, USA
john.sheppard@montana.edu

3rd Jason A. Clark
Library
Montana State University
Bozeman, Montana, USA
jaclark@montana.edu

***Abstract*—Community detection plays a pivotal role in social network analysis by partitioning networks into cohesive groups of vertices with dense intra-group connections and sparse inter-group connections. In this paper, we utilized a scholarly social network based on researchers' topic similarity derived from their publication metadata to identify interdisciplinary research communities. As topics often form a hierarchy, we hypothesize that the constructed scholarly network will exhibit hierarchical community structures. Therefore, we explore the efficacy of two prominent community detection algorithms, Louvain and Spectral clustering, known for their capacity to detect hierarchical community structures within networks. While both algorithms demonstrate this capability, the original Louvain algorithm is susceptible to the resolution limit problem due to its reliance on the modularity measure. To address this limitation, we propose the nested hierarchical Louvain algorithm, which iteratively partitions the network based on previously identified subgraphs, and we find that the bias towards large communities is mitigated. To evaluate the hierarchy produced by each of the algorithms, we employ the Cophenetic Correlation Coefficient (CPCC), a metric commonly used in hierarchical clustering evaluations but less frequently utilized in hierarchical community analysis. We argue that CPCC can be a useful measure to identify the presence of implicit hierarchical community structure in social networks when it is not explicitly available from domain knowledge while also further mitigating the inherent bias present in using modularity as a metric. Experimental results, conducted on both synthetic networks and the scholarly social network, demonstrate that the nested hierarchical Louvain algorithm, as well as Spectral Clustering, successfully identifies more finely structured hierarchical communities, offering greater depth in the dendrogram compared to the basic Louvain algorithm.**

***Index Terms*—Social Networks, Hierarchical Community Detection, Clustering, Topic Models**

I. Introduction

Community detection is a fundamental task in social network analysis to understand the underlying organization and functional units within networks, offering insights into various complex systems such as social networks, biological networks, and technological networks [1]–[4]. One crucial aspect of community detection is the presence of hierarchical structures within networks where smaller communities form subsets of larger ones, leading to a multi-level hierarchy similar to the branches of a tree. While traditional community detection algorithms excel at identifying a single layer of communities, they often fail to capture the hierarchical structures inherent in many real-world networks. They are essential in network analysis for a deeper understanding of the network structure by exploring the hierarchical relationships from different granularity within complex systems.

For our analysis of real-world networks, we constructed a scholarly social network based on researchers' topic similarity derived from their publication metadata. We did not consider direct relationships like co-authorship and citations to promote the detection of interdisciplinary communities within the scholarly network. We assume that the topic-based network will exhibit hierarchical community structure, given that topics often follow a hierarchical pattern of specific terms nested under more abstract, related terms.

Modularity is a popular measure for assessing the quality of partitions or communities within networks [5]. However, modularity-based methods like Louvain often suffer from the resolution limit problem, where they tend to merge smaller communities into larger ones, which can obscure fine-grained hierarchical structures within networks. To mitigate this bias towards the modularity measure, we propose the nested hierarchical Louvain algorithm that follows a divisive clustering approach by iteratively partitioning the network based on previously identified subgraphs. This iterative partitioning on the subgraph changes the scale of the resolution limit resulting in finer-tuned, smaller communities and mitigating the bias towards large communities. Subsequently, we devised a Spectral clustering algorithm that utilizes the Hierarchical Agglomerative Clustering approach to identify hierarchical structures within networks to asses the hierarchical partitioning obtained from different algorithms.

To assess the hierarchical structures identified by different algorithms, we utilized the Cophenetic Correlation coefficient (CPCC), a statistical measure that evaluates how well the dendrogram produced by the clustering algorithm preserves the pairwise distances between original data points. Since CPCC is typically used to assess hierarchical clustering algorithms, we adapted its application to evaluate hierarchical community structures in our study. We hypothesize that the hierarchical community structure obtained from the nested hierarchical Louvain and Spectral Clustering will offer a more finely detailed hierarchical perspective compared to the results obtained from the original Louvain algorithm.

II. Related Work

Interdisciplinary community detection helps to determine potential connections and interactions between different fields of study. This knowledge is important for identifying potential collaborators, improving resource allocation, and enhancing the overall impact of research. However, this area has not gained much attention in the literature. Most community detection methods focus on using direct relationships, such as co-authorship and citations, to construct the network, limiting the detection of cross-disciplinary communities [6], [7]. Araki *et al.* [8] proposed a recommender system for interdisciplinary collaboration using research content similarity, claiming to be the first to address this issue. In addition, we hypothesize that content topics should follow a hierarchical structure, where hierarchical community detection on the *content-based* scholarly network would allow for an interdisciplinary community structure with different levels of granularity.

Detecting hierarchical structures poses several challenges due to different factors such as network size, density, and the inherent ambiguity in defining hierarchical boundaries. The Louvain algorithm, proposed by Blondel *et al.* [9], is an agglomerative community detection algorithm capable of providing hierarchical structures. While effective for identifying communities in large graphs, Louvain's hierarchy often fails to capture deeper, multi-level hierarchies due to the resolution limit problem in optimizing modularity [10]. Another algorithm for detecting hierarchical structures is Hierarchical InfoMap, introduced by Rosvall *et al.* [11], which recursively partitions the network based on information flow. However, it relies on heuristics and may be sensitive to parameter choices.

Spectral clustering has been proposed as a method for identifying hierarchical structures. White *et al.* [12] introduced two spectral clustering approaches aimed at optimizing the modularity function. They achieved this by employing the k-means algorithm to select the best k value that maximizes the modularity. However, while these algorithms generate partitions based on varying k values, they do not offer a dendrogram. Subsequently, Wahl *et al.* [13] proposed an overlapping hierarchical spectral algorithm. They utilized fuzzy c-means for detecting overlapping communities and employed Jaccard similarity to map parent-child relationships in the dendrogram. Despite these advancements, the computational cost of finding network partitions for all potential k values can be challenging for large graphs.

Bhowmick *et al.* [14] proposed a hierarchical Louvain method for high-quality and scalable network embedding, where they utilize the Louvain algorithm recursively on the already detected communities to obtain sub-communities that are used later for efficient network embedding. While their method is similar to the nested hierarchical Louvain algorithm we present here, they used their approach only for network embedding and did not consider how well the network was partitioned across the nested levels. In this study, we adopted the Cophenetic Correlation Coefficient (CPCC) as a metric to evaluate hierarchical community structure.

III. Dataset

For the construction of a topic similarity-based scholarly social network, we accessed OpenAlex [1] [15], an open-source platform providing worldwide data on researchers and academics. Initially, we compiled information on 326 current researchers from our school database, including their names, departments, and college affiliations. Subsequently, we retrieved the publication histories of these researchers from OpenAlex spanning from 2004 to 2023, filtering them by specific researchers' names and affiliations. Although OpenAlex provides various publication details such as titles, abstracts, publication dates, venues, and citations, for our topic-based similarity network, we focused solely on the publication titles and abstracts, which were combined to create individual documents. Among the 326 researchers, we gathered a total of $9,659$ publication contents from OpenAlex, averaging 29 publications per researcher, with a maximum of 179 publications and a minimum of 4. Finally, we employed the publication contents to compute the topic-based similarity between researchers using topic modeling, as further detailed in Section V-A.

We hypothesize that building a network based on topic similarity will naturally lead to hierarchical structures, as topics tend to organize themselves into broader categories or concepts. For example, in a scholarly network, we expect to observe lower-level communities centered around specific topics like machine learning and artificial intelligence, which are likely subtopics within a broader community focusing on computer science, and subsequently, within even broader communities related to general scientific topics. Additionally, we created synthetic networks to simulate real-world scholarly networks (details in section V-B), allowing for in-depth analysis of hierarchical community detection algorithms.

IV. Methodology

To uncover hierarchical community structures within content-based social networks, we employed three distinct algorithms. The first is a method we developed for hierarchical community detection based on Spectral Clustering [16], followed by the standard Louvain algorithm, introduced by Blondel *et al.* [9]. Additionally, we developed a variant of the Louvain algorithm, called the nested hierarchical Louvain algorithm. This approach adopts a top-down recursive strategy to generate hierarchical community structures within networks. Subsequently, we utilized the Cophenetic Correlation Coefficient (CPCC) metric [17] to evaluate and compare the hierarchies produced by the various algorithms. We used CPCC rather than modularity since the latter would be biased in favor of Louvain because it is the objective function optimized by Louvain.

A. Spectral Clustering

Spectral clustering [16] is a popular method used in machine learning and data analytics to partition data points into distinct

[1] https://openalex.org/

clusters. The key idea behind spectral clustering is to represent the data points in a graph structure and analyze the graph's spectral properties. In spectral graph theory, the spectral properties of the graph are derived from the eigendecomposition of some representation of a matrix associated with the graph [18]. From a graph perspective, the matrix could be either the graph adjacency matrix or the graph Laplacian matrix, the latter being the most popular.

The graph Laplacian matrix $\mathcal{L}$ is computed as

$$\mathcal{L} = \mathbf{D} - \mathbf{A}$$

where $\mathbf{A}$ is the adjacency matrix and $\mathbf{D}$ represents the degree matrix. The degree matrix $\mathbf{D}$ is a diagonal matrix such that d_{ii} = the degree of vertex i and $d_{ij} = 0$ for $i \neq j$. The above Laplacian matrix is unnormalized, and vertices with large degrees may dominate, which is not always desirable with graphs of various sizes and densities. Therefore, a symmetric normalized Laplacian matrix, denoted as $\mathcal{L}_{\text{sym}}$ is often used instead:

$$\mathcal{L}_{\text{sym}} = \mathbf{D}^{-\frac{1}{2}}\mathcal{L}\mathbf{D}^{-\frac{1}{2}}$$

where $\mathcal{L}$ is the unnormalized Laplacian matrix, and again $\mathbf{D}$ is the degree matrix.

Spectral analysis on the graph Laplacian matrix reveals some interesting properties. Since the Laplacian is symmetric and positive-semidefinite, the eigenvalues can tell us about the connectedness of the graph structure. For example, the number of 0 eigenvalues corresponds to the number of connected components in the graph. Next, the second smallest eigenvalue, known as the algebraic connectivity, quantifies how well-connected the graph is. A higher algebraic connectivity indicates a graph with better connectivity. Following this, the eigenvectors corresponding to the k smallest eigenvalues of the Laplacian can be used for spectral encoding, representing the graph's structure in a lower-dimensional vector space. Subsequently, various clustering algorithms such as k-means clustering [19] can be applied to this low-dimensional vector space to unveil community structures.

However, as we are interested in detecting hierarchical community structure, we utilized Hierarchical Agglomerative Clustering (HAC) [20]. HAC starts with each data point forming its own cluster and then iteratively merges the closest clusters until either all the data points fall into a single cluster or a termination condition is met. For finding the distance between clusters, the algorithm typically uses a linkage criterion. For example, Single Linkage computes the distance between the closest points of the two clusters being merged, while Complete linkage computes the distance between the furthest points of the two clusters being merged. For community detection, these points represent vertices belonging to the same community. For a more in-depth exploration of graph spectral properties, please refer to [21].

B. Louvain Algorithm

The Louvain algorithm [9] (or just Louvain), is a greedy community detection algorithm that partitions the network resulting in a modularity score that is maximized. Modularity [5] is a widely used metric for assessing community quality without any ground truth labels. The score is based on the intuition that communities have more internal edges than one would expect when compared to a random network. Here, a random network would be one constructed with the same number of vertices and edges but connecting vertices randomly following the same degree distribution of the original network. To summarize, the modularity score compares the actual number of edges within communities to the expected number of such edges in a null model (random graph), defined as

$$Q = \frac{1}{2m}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\left(a_{ij} - \frac{d_i d_j}{2m}\right) \times \delta(c_i, c_j)\right]$$

where $a_{ij} \in \mathbf{A}$ represents an edge in the adjacency matrix, $m = |\mathbf{E}|$ is the number of edges, d_i is the degree of vertex v_i, c_i is the community assignment of vertex v_i, and $\delta(c_i, c_j)$ is the Kronecker delta function, which equals 1 if c_i and c_j represents the same community or 0 otherwise.

Q can range from -1 to $+1$. A value close to $+1$ indicates a strong community structure, while values close to 0 suggest a weak or non-existent community structure. A negative value, on the other hand, indicates that the partition is worse than a random assignment of vertices to a community.

The main idea behind Louvain is to find communities efficiently by locally optimizing the modularity. This local optimization process makes the algorithm computationally fast and well-suited for analyzing large-scale networks [22], [23]. Moreover, the algorithm follows an agglomerative (bottom-up) approach, enabling it to produce a hierarchical community structure, which is particularly useful for exploring community structures at different levels of detail.

Algorithm 1 shows the pseudocode for the Louvain algorithm where lines 2–4 initialize the vertices into their own communities and lines 6–12 show the local modularity optimization phase. At line 13 we record the locally formed

Algorithm 1 Louvain Algorithm

```
1:  Initialization:
2:  for each node i do
3:      Assign i to its own community
4:  end for
5:  while changes in modularity are significant do
6:      for each node i do
7:          for each neighbor community c of i do
8:              Move i to community c
9:              Calculate the change in modularity (ΔQ)
10:         end for
11:         Move i to the community yielding the max ΔQ
12:     end for
13:     Record Community history to a dendrogram
14:     Merge communities to form a new network
15: end while
16: Output: final aggregated network, dendrogram
```

communities into a dendrogram, and line 14 is the community aggregation phase. The **while** loop at line 5 repeatedly runs the two phases until no further gain in modularity is possible. Finally, the algorithm returns the community assignment of the last aggregated network as the best partition with the dendrogram structure.

Although Louvain works well for uncovering hidden structures in large and complex networks, it is not without limitations. One of the most important limitations is the resolution limit problem [10], [24]–[26]. The problem implies that, by prioritizing modularity optimization, the algorithm investigates possible network partitions at a broader level, potentially overlooking modules smaller than a certain scale. The resolution limit of modularity is not restricted to specific network architectures but arises solely from comparing the interconnectivity of communities to the total number of network links available.

The Modularity equation can be rewritten as follows:

$$Q = \sum_{c_k \in \mathbf{C}} \sum_{v_i \in c_k} \sum_{v_j \in c_k} \left[\left(\frac{a_{ij}}{2m} - \frac{d_i d_j}{4m^2} \right) \right]$$

where $\mathbf{C}$ represents the set of communities, and the modularity score is the sum of the number of detected communities and the value of each term. Therefore, finding the maximal modularity is equivalent to looking for the ideal tradeoff between the number of terms in the sum. However, an increase in the number of communities does not always guarantee an increase in the modularity score. Increasing the number of communities makes the community size smaller, resulting in fewer links inside each community. This reduction in internal links decreases the fraction of links within each community. Consequently, this can lead to a poorer modularity score overall. Therefore, for some number of communities, the modularity score has a "peak" value, and due to the organization of the mathematics, it often favors larger communities by merging smaller communities. However, in real-world networks, the actual community structures of the network may be heterogeneous in size, especially in large networks, so algorithms maximizing modularity may suffer from the resolution limit problem. According to Fortunato *et al.* [10], Louvain fails to detect communities of size $\sqrt{4m}$ or smaller where m is the total number of edges in the network.

C. Nested Hierarchical Louvain

The hierarchical structure constructed by the base Louvain algorithm follows a bottom-up approach, which begins by placing all vertices in separate communities. Then, it merges vertices or communities with their neighboring communities as long as it improves modularity. Due to the resolution limit problem, smaller communities in this step are often merged together into larger communities. Therefore, the hierarchy or the dendrogram produced by the base Louvain algorithm is often shallow and fails to capture the fine-grained hierarchical structure of the network.

To address the resolution limit problem in the hierarchy captured by the base Louvain algorithm, we devised a top-down hierarchical approach that sits on top of the base Louvain. We call it the nested hierarchical Louvain algorithm where we follow a generalized top-down (divisive) approach for hierarchical clustering [27].

Algorithm 2 Hierarchical Louvain

```
1:  function HierarchicalLouvain(𝒢)
2:      partitions ← Louvain(𝒢)
3:      if sizeof(partitions) ≤ 1 then
4:          make partition leaf of parent partitions
5:      else
6:          for each community in partitions do
7:              make community child of parent partitions
8:              HierarchicalLouvain(𝒢[community])
9:          end for
10:     end if
11: end function
```

In the Nested Hierarchical Louvain (NH-Louvain) method, we begin with all the vertices in the same community. Then for the divisive step, we use the base Louvain algorithm to determine the first level of network partitions. Then, we recursively apply the Louvain algorithm only on the partitions obtained in the previous step until reaching a stopping criterion. For the stopping criterion, it could be when the partitions lead to a single vertex or a single community with possibly a clique. The process ultimately creates a dendrogram where the leaves are either single communities or single vertices.

The way NH-Louvain hierarchy addresses the resolution limit problem is by using base Louvain recursively on the partitions obtained in the previous level. As discussed before, modularity optimization relies on the total number of links or edges present in the network for which it fails to detect any communities of size equal to or smaller than $\sqrt{4m}$ (m being the total number of edges). However, in the nested Louvain hierarchy, as Louvain is applied to the previous partition, we are effectively reducing the total number of links in the subpartitions, thus reducing the resolution limit as compared to the total network. This forces the original Louvain method to provide smaller communities at that level. This way, the hierarchy produced by the nested Louvain would be able to provide more depth and fine-structured communities, increasing the granularity of the whole network partitions.

Algorithm 2 shows the pseudocode of the NH-Louvain algorithm. The algorithm takes the whole network $\mathcal{G}$ as input and runs base Louvain to find an initial partition at line 2. In line 3, it checks for the base condition or stopping criterion for the recursive call by checking if the partitions obtained from Louvain are single partitions or empty. If the condition is satisfied then it makes it a leaf node at the parent community and terminates at line 4. If the base condition is not satisfied, then for each community in the partition, it makes that community a child of the parent community and calls the recursive function with a subgraph induced by the vertices that belong to that community, effectively shrinking the size of the original network (lines 6–8). Finally, the dendrogram with the hierarchical relationships of the partitions is returned.

D. Cophenetic Correlation Coefficient

The Cophenetic Correlation Coefficient (CPCC) is a popular metric for evaluating hierarchical clustering in an unsupervised setting [17], [28]. It is a measure used to evaluate the goodness-of-fit of a hierarchical clustering solution to the original data by quantifying the similarity between the pairwise distances among data points and the distances in the dendrogram produced by the hierarchical clustering algorithm. However, it has not received much attention in evaluating hierarchical community structures in the context of community detection in social networks.

To compute CPCC, first, we need to obtain the Cophenetic Distance Matrix from the dendrogram obtained from a hierarchical community detection algorithm. Each vertex in the dendrogram represents a community at that level. The cophenetic distance between two vertices in the dendrogram is the distance or height of the lowest common ancestor vertex where the two vertices are first merged together into a community. Then, the cophenetic distance matrix contains the cophenetic distances of all pairwise vertices in the network obtained from the dendrogram.

Next, we need to compute the pairwise distance matrix for the original data points. As we are constructing networks based on content data, we get an undirected weighted graph where the weights represent the strength of the connection between two vertices. Therefore, for the similarity matrix of the original data points, we can utilize the adjacency matrix of the constructed graph. Let us assume, $\mathbf{A}$ denotes the adjacency matrix of the graph with a_{ij} denoting the edge weight between vertices v_i and v_j. Let $\mathbf{\Delta}$ denote the cophenetic distance matrix with $\delta_{i,j}$ being the cophenetic distance of vertices v_i and v_j. Then, the CPCC between these two matrices is calculated using the Pearson correlation coefficient as

$$CPCC = \frac{\sum_{i<j}(a_{ij} - \bar{\mathbf{A}})^2 \sum_{i<j}(\delta_{ij} - \bar{\mathbf{\Delta}})^2}{\sqrt{\sum_{i<j}(a_{ij} - \bar{\mathbf{A}})(\delta_{ij} - \bar{\mathbf{\Delta}})}}$$

where $\bar{\mathbf{A}}$ and $\bar{\mathbf{\Delta}}$ represent the mean of the adjacency matrix and the cophenetic distance matrix respectively. The CPCC values range from -1 to 1 with 1 indicating a perfect fit between the dendrogram and adjacency matrix, 0 indicating no correlation between the dendrogram and adjacency matrix, and -1 indicating perfect negative correlation.

V. Experimental Design

To evaluate and compare our two hierarchical community detection methods with the original Louvain method, we examined the scholarly social network with several synthetically generated networks. The synthetic networks are constructed using known block structures to simulate networks similar to the content-based scholarly network.

A. Scholarly Network

To construct the network, we followed a procedure similar to the one described in [29]. First, we used title and abstract data obtained from OpenAlex to build a topic model, specifically Latent Dirichlet Allocation (LDA) [30] with Gibbs sampling for parameter estimation [31], to obtain the latent topics from the publication corpus. Next, we calculated the topic probability distribution of each researcher from the trained LDA model. Then, to connect each researcher in the constructed network, we applied the Jensen-Shannon divergence (JSD) metric [32] to estimate the topic similarity between the researchers. The JSD value ranges between 0 and 1, with 0 indicating maximum similarities and 1 indicating complete dissimilarities. Therefore, we used $(1 - JSD)$ as edge weights, where a weight close to 1 signifies strong topic-based similarity between the two researchers.

Constructing the scholarly social network based on topic-based similarity results in a fully connected network, wherein every vertex connects to every other vertex, and consistently yields a value greater than zero. Despite the network being weighted, the dense graph formed poses challenges for community detection. To address this, we pruned edges whose weights fell below a given threshold Θ while ensuring the resulting network remained connected (i.e., with only one connected component). In the fully connected graph with a threshold value of 0 (i.e., no edges pruned), the number of edges amounts to $52,975$. With a threshold of 0.2, the edge count reduces to $5,535$ without causing any disconnected components in the initial graph. Consequently, we considered two versions of the network: one with a threshold of 0 (i.e., without edge pruning) and another with a threshold of 0.2 (i.e., with edge pruning).

Spectral clustering for hierarchical community detection requires some hyperparameters to be defined, such as the similarity metric, the linkage criterion for merging the clusters, and the number of eigenvectors corresponding to the k smallest eigenvalues. After experimenting with several different options, we found the *cosine* similarity metric to work best. For the linkage criterion, we tested *single, complete, average,* and *Ward* linkage, and *complete* linkage was found to work the best. For the number of eigenvectors, we used the CPCC metric to select the k eigenvectors.

When computing CPCC for Spectral hierarchical community detection, we can get two baseline similarity matrices, the first based on the original pairwise topic-based similarity stored in the graph adjacency matrix, and the second being the pairwise distance matrix from the eigendecomposition of the normalized graph Laplacian used to encode the graph structure. Figure 1 shows the CPCC values for different numbers of eigenvectors obtained from both the graph adjacency matrix (depicted with a red line) and the eigendecomposition on the normalized Laplacian matrix (depicted with a blue line) using edge threshold of 0.2.

As shown, the CPCC values when using the spectral encoded matrix are consistently higher than the CPCC values obtained with the original adjacency matrix. This is expected as the algorithm uses the spectral encoded matrix for producing hierarchical clusters, and the cophenetic distances obtained from the dendrogram are much more aligned with the encoded matrix. Nevertheless, since we are comparing the spectral

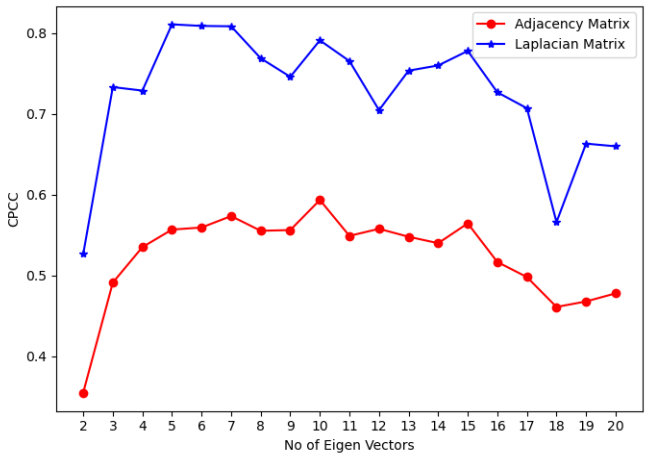


Fig. 1: CPCC vs. number of eigenvectors

clustering hierarchy with the Louvain algorithms, which use the original adjacency matrix for network partitioning, we opted to use the CPCC values based on the adjacency matrix to select the best k for a fair comparison. Based on Figure 1, this resulted in selecting 10 eigenvectors.

B. Synthetic Networks

To evaluate the algorithms' performance further, we employ a hierarchical Stochastic Block Model (SBM) [33] to generate synthetic networks that capture nested community structures. This model extends the classical SBM by introducing multiple levels of community organization, where vertices are grouped into blocks at each hierarchical level. The parameters for the hierarchical SBM include the number of vertices (N), the number of hierarchical levels (L), a branching factor (B) where each block is divided into B child blocks, an intra-block edge probability (P_{intra}) to connect vertices within a block, and an inter-block edge probability (P_{inter}) to connect vertices between blocks.

Initially, vertices are assigned to a single root block at the topmost level (Level 0). At each subsequent level, each block is subdivided into child blocks according to the branching factor B, and edges are added based on the specified edge probabilities P_{intra} and P_{inter}, forming a dendrogram structure. Additionally, since the scholarly network is weighted and the adjacency matrix represents the similarity matrix, we assign weights to the edges in the synthetic networks. For intra-block edges, weights are randomly assigned in the range $[0.5, 1.0]$, and for inter-block edges, weights are randomly assigned in the range $[0.0, 0.5]$.

The block sizes at each hierarchical level are determined by the total number of vertices, the branching factor, and the hierarchy level. For the synthetic networks, we used $N = 500, 1000$ and $L = 3, 4, 5, 6$ with a branching factor $B = 3$, an intra-block edge probability of $P_{\text{intra}} = 0.7$, and an inter-block edge probability of $P_{\text{inter}} = 0.7$. The reason for choosing similar and high intra- and inter-edge probabilities is to ensure that the edge weights dictate the hierarchical structure, similar to the scholarly social network. As a result, we generated 8 synthetic networks, with four different hierarchy levels for networks with 500 and 1000 vertices.

TABLE I: CPCC comparison

Dataset	Louvain	NH-Louvian	Spectral
Scholar_0.0	0.53	0.55	0.53
Scholar_0.2	0.54	0.61	0.59
A_3	0.40	0.50	0.37
A_4	0.36	0.49	0.40
A_5	0.36	0.47	0.36
A_6	0.37	0.50	0.33
B_3	0.40	0.50	0.40
B_4	0.40	0.52	0.39
B_5	0.37	0.50	0.37
B_6	0.34	0.45	0.33

TABLE II: Maximum dendrogram depth

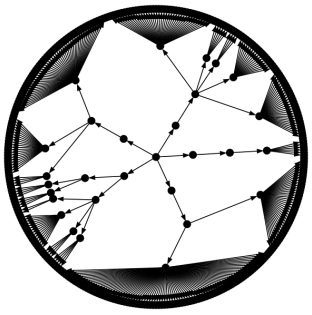
Dataset	Louvain	NH-Louvian	Spectral
Scholar_0.0	3	6	12
Scholar_0.2	4	6	18
A_3	3	7	14
A_4	4	6	14
A_5	3	6	18
A_6	4	6	16
B_3	3	7	16
B_4	4	7	18
B_5	4	7	18
B_6	4	7	18

VI. Results & Discussion

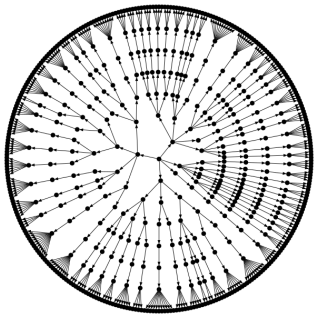
Table I shows the Cophenetic Correlation Coefficient (CPCC) values obtained by the three different algorithms. In the table, the column labeled **NH-Louvain** refers to the nested hierarchical Louvain algorithm, while **Spectral** corresponds to spectral clustering with HAC. The first two rows correspond to the scholarly social networks constructed from topic-based similarity. The label "Scholar-0.0" indicates the entire network without edge pruning, while "Scholar-0.2" indicates the removal of edges below $\Theta = 0.2$. Subsequent entries labeled "A-(ℓ)' correspond to four synthetic networks using 500 vertices and ℓ hierarchical levels, and "B-(ℓ)" represents synthetic networks using 1000 vertices, again with ℓ levels.

Analysis of the results reveals that, for the scholarly network without edge pruning, all three algorithms yielded similar CPCC values. However, upon edge pruning to reduce network density, notable improvements in CPCC values were observed, particularly for NH-Louvain and Spectral clustering. This suggests that the removal of low-weight edges enhanced the detection of hierarchical structures by these algorithms. For the synthetic networks, NH-Louvain consistently provided higher CPCC values compared to the original Louvain and Spectral clustering. When comparing Louvain with Spectral, neither method demonstrates a clear advantage.

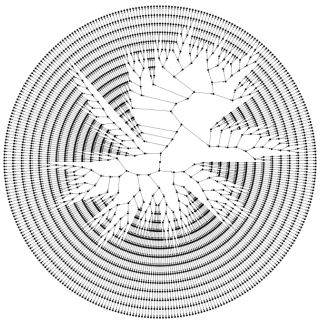
Table II shows the maximum depth of dendrograms generated by the three algorithms for each of the networks. As shown, the Louvain algorithm yields relatively shallow trees compared to NH-Louvain and Spectral. In particular, Spectral demonstrates the greatest tree depth compared to



(a) Dendrogram of Louvain



(b) Dendrogram of NH-Louvain



(c) Dendrogram of Spectral

Fig. 2: Dendrograms of "Scholor-0.2"

both of the Louvain algorithms. This difference is expected, given the distinct clustering methodologies employed by each algorithm. Spectral employs an agglomerative approach within a hierarchical framework, where each iteration merges only a single pair of clusters with minimal distance. On the other hand, Louvain, while also following an agglomerative strategy, can merge multiple communities during each iteration, driven by modularity score improvement. In contrast, NH-Louvain adopts a top-down, divisive approach, systematically subdividing previous partitions to obtain finer granularity in subsequent partitions. But since NH-Louvain still uses Louvain at each of the levels, multiple communities may still result as children of a previous community.

For the synthetic networks, we observed differences between the original hierarchy levels and the tree depths obtained from different algorithms. Although the network structure was created using a predefined hierarchy, the introduction of random edge weights might cause the discovered hierarchy to deviate from the constructed one due to a corresponding relaxation of edge weights. Additionally, the Louvain algorithm produces maximum tree depths ranging from 3 to 4 levels, whereas NH-Louvain produces depths ranging from 6 to 7 levels. NH-Louvain uses the base Louvain algorithm for generating subpartitions, which may still be locally susceptible to the resolution limit problem. Although this effect is less pronounced than in the original Louvain, it can still impact the subpartitions, potentially explaining why NH-Louvain does not provide hierarchies deeper than 7 levels. In contrast, Spectral clustering shows a variety of tree depths ranging from 12 to 18 levels. The greater tree depth in Spectral clustering is due to the pairwise merging of nodes in each iteration.

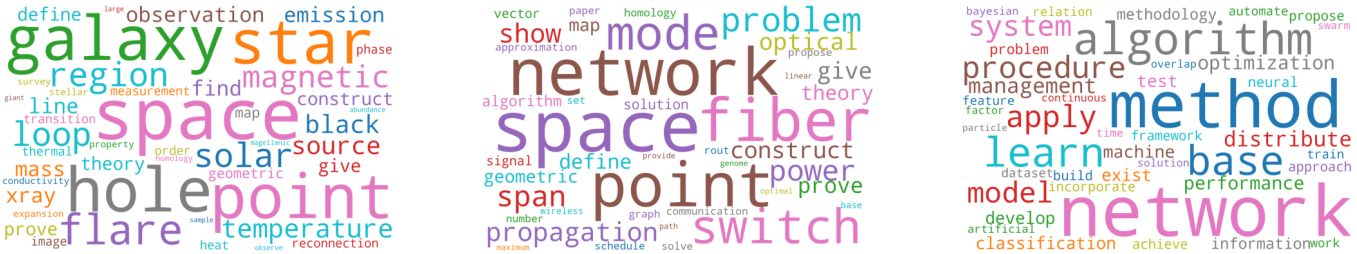
Further analysis of the dendrograms for the "Scholar-0.2" network, produced by the Louvain algorithms, is shown in Figure 2. The base Louvain algorithm, after the first iteration, forms communities that include large groups of members, almost revealing the final partition immediately. In subsequent iterations, only a few communities merge, resulting in a shallow hierarchy, as depicted in Figure 2a. In contrast, the dendrogram from NH-Louvain, shown in Figure 2b, and the dendrogram from Spectral clustering, shown in Figure 2c, present a more detailed and fine-grained view of the evolving hierarchy with more levels.

Figure 3 presents hierarchical community WordCloud examples across different levels of the hierarchy, again for the "Scholar-0.2" network. To generate these WordClouds, we utilized metadata from the publications of community members to extract latent topics using the trained topic model. In Figure 3c, the WordCloud represents a community at the bottom of the hierarchy, primarily associated with topics related to algorithms and machine learning. As we ascend the hierarchy for this particular community, we observe a shift towards computer science-related topics at level 3. Figure 3b displays the WordCloud at hierarchy level 2, which includes additional topics such as optics and mathematics. Finally, at hierarchy level 1, we encounter general science topics (Figure 3a), with level 0 representing the entire network.

VII. Conclusion & Future Work

Our analysis of content-based scholarly networks highlights the effectiveness of the proposed NH-Louvain algorithm and Spectral Clustering in identifying finely structured hierarchical communities, providing enhanced depth in the dendrogram compared to the basic Louvain algorithm. These results also supported our first hypothesis with respect to NH-Louvain and Spectral Clustering leading to finer-grained identification of hierarchical structure. When examining the results on "Scholar-0.0" and "Scholar-0.2", we also found support for our second hypothesis that the topic model-based social network contained hierarchical structure.

In this study, we did not focus on producing the best single-layer partition from the dendrogram. Although modularity is commonly used for this purpose, we argue that it may not always yield the optimal partition due to the resolution limit. Therefore, in future research, we plan to explore alternative measures to evaluate network partitioning, including leveraging large language models (LLMs) for advanced topic modeling. By incorporating LLM-based approaches, we aim to construct scholarly networks with improved detection of hierarchical community structures. Lastly, we did not consider overlapping memberships in the hierarchy, which are prevalent in real-world networks. Investigating this aspect further presents another exciting avenue for future exploration.



(a) "Scholar-0.2" WordCloud at level 1 (b) "Scholar-0.2" WordCloud at level 2 (c) "Scholar-0.2" WordCloud at level 4

Fig. 3: Hierarchical WordCloud examples for "Scholar-0.2"

ACKNOWLEDGMENTS

This paper is based on work supported, in part, by NSF EPSCoR Cooperative Agreement OIA-2242802. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] C. Cortes, D. Pregibon, and C. Volinsky, "Communities of interest," in *Advances in Intelligent Data Analysis*. Springer Berlin Heidelberg, 2001, pp. 105–114.

[2] R. Guimera and L. A. Nunes Amaral, "Functional cartography of complex metabolic networks," *nature*, vol. 433, no. 7028, pp. 895–900, 2005.

[3] L. S. Haggerty, P.-A. Jachiet, W. P. Hanage, D. A. Fitzpatrick, P. Lopez, M. J. O'Connell, D. Pisani, M. Wilkinson, E. Bapteste, and J. O. McInerney, "A pluralistic account of homology: adapting the models to the data," *Molecular biology and evolution*, vol. 31, no. 3, pp. 501–516, 2014.

[4] M. Shrestha, S. V. Scarpino, E. M. Edwards, L. T. Greenberg, and J. D. Horbar, "The interhospital transfer network for very low birth weight infants in the united states," *EPJ Data Science*, vol. 7, pp. 1–14, 2018.

[5] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[6] J. Li, F. Xia, W. Wang, Z. Chen, N. Y. Asabere, and H. Jiang, "Acrec: a co-authorship based random walk model for academic collaboration recommendation," in *Proceedings of the 23rd international conference on world wide web*, 2014, pp. 1209–1214.

[7] T. Huynh, A. Takasu, T. Masada, and K. Hoang, "Collaborator recommendation for isolated researchers," in *2014 28th international conference on advanced information networking and applications workshops*. IEEE, 2014, pp. 639–644.

[8] M. Araki, M. Katsurai, I. Ohmukai, and H. Takeda, "Interdisciplinary collaborator recommendation based on research content similarity," *IEICE TRANSACTIONS on Information and Systems*, vol. 100, no. 4, pp. 785–792, 2017.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[10] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, pp. 36–41, 2007.

[11] M. Rosvall and C. T. Bergstrom, "Detecting hierarchical community structures in networks using infomap," *Inf. Sci.*, vol. 179, no. 15, pp. 3080–3091, 2009.

[12] S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," in *Proceedings of the 2005 SIAM international conference on data mining*. SIAM, 2005, pp. 274–285.

[13] S. Wahl and J. Sheppard, "Fuzzy spectral hierarchical communities in evolving political contribution networks," in *Proceedings of the 13 International Florida Artificial Intelligence Research Society Conference*, 2017, pp. 371–376.

[14] A. K. Bhowmick, K. Meneni, M. Danisch, J.-L. Guillaume, and B. Mitra, "Louvainne: Hierarchical louvain method for high quality and scalable network embedding," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, p. 43–51.

[15] J. Priem, H. Piwowar, and R. Orr, "Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," 2022. [Online]. Available: https://arxiv.org/abs/2205.01833

[16] A. Y. Ng, M. I. Jordan, and Y. Weiss, "Spectral clustering for image segmentation," in *Proceedings of the 2002 conference on Advances in neural information processing systems*, 2002, pp. 849–856.

[17] R. R. Sokal and F. J. Rohlf, "The comparison of dendrograms by objective methods," *Taxon*, vol. 11, no. 2, pp. 33–40, 1962.

[18] F. R. K. Chung, *Spectral Graph Theory*. American Mathematical Society, 1997.

[19] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. University of California Press, 1967, pp. 281–297.

[20] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[21] D. A. Spielman, "Spectral graph theory and its applications," *Linear Algebra and its Applications*, vol. 438, no. 11, pp. 2415–2445, 2013.

[22] X. Que, F. Checconi, F. Petrini, and J. A. Gunnels, "Scalable community detection with the louvain algorithm," in *2015 IEEE international parallel and distributed processing symposium*. IEEE, 2015, pp. 28–37.

[23] N. S. Sattar and S. Arifuzzaman, "Scalable distributed louvain algorithm for community detection in large graphs," *The Journal of Supercomputing*, vol. 78, no. 7, pp. 10 275–10 309, 2022.

[24] V. A. Traag, P. Van Dooren, and Y. Nesterov, "Narrow scope for resolution-limit-free community detection," *Physical Review E*, vol. 84, no. 1, p. 016114, 2011.

[25] J. W. Berry, B. Hendrickson, R. A. LaViolette, and C. A. Phillips, "Tolerating the community detection resolution limit with edge weighting," *Physical Review E*, vol. 83, no. 5, p. 056119, 2011.

[26] J. Guo, P. Singh, and K. E. Bassler, "Resolution limit revisited: community detection using generalized modularity density," *Journal of Physics: Complexity*, vol. 4, no. 2, p. 025001, 2023.

[27] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[28] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Pearson, 2005.

[29] M. A. Noor, J. Sheppard, and J. Clark, "Finding potential research collaborations from social networks derived from topic models," in *10th International Conference on Behavioural and Social Computing*, 2023, pp. 1–7.

[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[31] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, 1984, pp. 721–741.

[32] S. M. Ross, *Introduction to Probability Models*, 6th ed. San Diego, CA, USA: Academic Press, 1997.

[33] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.