# GRAPH ANALYTICS USING NODE2VEC EMBEDDINGS AND UNSUPERVISED LEARNING ON CITATION NETWORKS

**A MINI PROJECT REPORT**

*Submitted by*

**SHRUTIKA GUPTA**

*in partial fulfillment for the award of the degree*

*of*

**MASTERS OF SCIENCE**

*in*

**DATA SCIENCE**

**CHRIST (Deemed to be University, Bangalore)**

NOVEMBER, 2025

# CERTIFICATE

This is to certify that this project report **"Graph Analytics Using Node2Vec Embeddings and Unsupervised Learning on Citation Networks"** is the bonafide work of **"SHRUTIKA GUPTA"** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

Signature

DR. PRIYA STELLA MARY I.
Assistant Professor
Department of Computer Science
CHRIST (Deemed to be University), Bangalore

Submitted for Project Viva-voce examination held on ____________________

# ABSTRACT

This project introduces a detailed graph analytics pipeline capable of detecting significant research communities in a large citation network made up of 2,708 papers and 10,556 citation links. The approach mixes structural and semantic data by incorporating Node2Vec embeddings, which are obtained through biased random walks, with high-dimensional feature vectors for each paper.

These combined embeddings were shortened using UMAP and grouped using K-Means, with k = 8 being the best configuration according to the highest silhouette score of 0.3568. The clusters obtained were distinctly separated in t-SNE visualizations and showed a moderate correlation with subject categories, proved by NMI and ARI scores of 0.471 and 0.387 respectively. The cluster–subject heatmap also helped in identifying the strong thematic consistency of clusters mainly Neural Networks, Probabilistic Methods, and Genetic Algorithms.

A comparison with Louvain community detection indicates that while Louvain uncovered finely detailed structural modules, it resulted in communities that were weakly separated and had a low silhouette score of 0.119. In summary, the research reveals that embedding-based clustering yields more understandable and semantically richer groupings in citation networks than purely structural community detection, thus, it is a great way to gain insights into the research structure and scholarly organization.

**Keywords:** Graph Analytics, Citation Network, Node2Vec, UMAP, K-Means Clustering, Community Detection, Louvain Algorithm, t-SNE Visualization

# Contents

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

| Abbreviation | Description |
|---|---|
| ARI | Adjusted Rand Index |
| CSV | Comma-Separated Values |
| GCN | Graph Convolutional Network |
| GNN | Graph Neural Network |
| k | Number of clusters |
| K-Means | K-Means Clustering Algorithm |
| NMI | Normalized Mutual Information |
| Node2Vec | Node to Vector (Graph Embedding Algorithm) |
| t-SNE | t-distributed Stochastic Neighbor Embedding |
| UMAP | Uniform Manifold Approximation and Projection |
| 2D | Two-Dimensional |
| 3D | Three-Dimensional |

# 1 INTRODUCTION

## 1.1 Overview

Graph analytics is essential for comprehending the intricate interactions that are typical of network-structured data. Citation networks formed by nodes that represent research papers and directed edges that indicate citations offer a perfect setting for investigating the scientific knowledge dynamics. Dissecting such systems aids in uncovering influential themes, newly surfaced research clusters, and the flow of ideas among different subject areas.

## 1.2 Motivation

The rising demand for extracting significant patterns from large-scale citation datasets is the main reason for this project. Conventional machine learning methods are not efficient with graph-structured data because of their irregular topology. Consequently, representation learning methods like Node2Vec have become popular since they can map nodes to continuous vector spaces.

That makes it possible to:

- reflect both local and global structural relationships,
- permit machine learning models to work in Euclidean space,
- facilitate scalable clustering and visualization of graph data.

## 1.3 Problem Statement

This project is mainly focused on the problem of:

*How can graph representation learning and unsupervised machine learning methods be harnessed to cluster papers in a citation network accurately and to discover subject-level groupings that make sense?*

The issue encompasses the difficulties of structural complexity, high-dimensional features, and heterogeneous subject labels.

## 1.4 Objectives

This work aims to:

- Reconstruct a directed citation graph from raw node and edge files.
- Fetch and prepare high-dimensional node attributes.
- Implement Node2Vec to create structural embeddings.
- Merge semantic features with Node2Vec embeddings to achieve a more comprehensive representation.
- Use dimensionality reduction to enhance the clustering quality.
- Execute clustering with K-Means and compare the results with Louvain community detection.
- Interpret and make sense of the clusters through the usage of visualization and evaluation metrics.

These goals facilitate a deep dive into machine learning-based graph analytics.

## 1.5 Scope of the Study

The scope covers graph construction, embedding creation, clustering, visualization, and performance evaluation. Supervised classification, GNN-based modeling, and temporal citation prediction are not included. The methodology, however, serves as a stepping stone for more advanced graph learning techniques.

# 2 LITERATURE REVIEW

The review describes the key techniques used for the project which includes graph embedding, machine learning algorithms for graph data, and community detection.

## 2.1 Graph Representation Learning

Graph analytics constitute a robust framework to model and analyze intricate systems that are interconnected, which range from social networks to citation graphs. Yet, due to the non-Euclidean characteristic of graph data, it is quite difficult to apply traditional machine learning algorithms to such data directly. To overcome this hurdle, graph embedding methods have been invented to convert graph structures into low-dimensional vector representations, where each node is associated with a vector in a continuous space.

Random walks serve as the fundamental idea behind a major set of graph embedding methods. DeepWalk [1] was the first to use this technique by performing truncated random walks on the graph and using the node sequences obtained for training a SkipGram model thereby, a machine is trained to learn the words "words" as in the "sentence" concept. In this work, a concept direct and strong extension, Node2Vec [2], is used. The approach in the local code is a straightforward execution of this idea whereby random walks are created and then a Word2Vec model is used to process the walks. The principal change introduced by Node2Vec is the employment of biased random walks which enable it to optimize the search of local neighborhoods (structural equivalence) as well as the wider network structures (homophily), thus becoming a very efficient and scalable feature learning algorithm for networks.

## 2.2 Machine Learning Algorithms for Graph Analytics

Once nodes are converted to vectors, numerous machine learning algorithms can be used. The present work delves into two primary paradigms:

**Embedding-Based Clustering (K-Means):** The code of the project largely revolves around this method which essentially applies standard ML algorithms to the created embeddings. The K-Means clustering is then employed to divide the nodes in the embedding space. The main idea is that nodes with similar vector representations (as produced by Node2Vec) should be found in the same cluster, which can then be interpreted as a naturally occurring group (e.g., paper subject). The present work also recognizes the challenge of addressing the discrepancy between the topological embeddings and the node-specific attributes by combining the Node2Vec embedding with the node feature matrix thus generating a more complex, hybrid representation prior to clustering.

**Topology-Based Community Detection (Louvain):** Another, and essentially a traditional, method is to determine communities based on the graph topology directly, thus omitting the embedding step. The methods, being the subject of an exhaustive review by Fortunato [3], seek to divide the graph through the maximization of a given quality function, usually modularity. Modularity assesses the number of links within communities in comparison to a random baseline. The code of this project is the implementation of the Louvain algorithm [4].

## 2.3 Clustering in Citation Networks

Most of the research work on clustering citation networks has been divided into one of these two extremes: the problem has either been looked at from a purely structural perspective or a feature-based perspective. Zhang et al. [5] demonstrated the effectiveness of Node2Vec in producing "science maps," and they found that random-walk embeddings represented the relationships between disciplines of journals much better than direct citation vectors. Nonetheless, their study was focused only on journal analysis at a coarse level and didn't consider the incorporation of node-specific content features.

In contrast, Alashwal et al. [6] applied K-Means clustering only on bibliometric features such as H-index and citation counts to find the most influential research topics, a method that totally ignores the topological structure of the citation graph. Pourhabibi et al. [7] have bridged these two fields by experimentally supporting the "embedding plus clustering" pipeline that they argued to be the reason why K-Means performance is significantly enhanced by high-quality vector spaces. Although this serves as a confirmation of the core idea of the current work, their paper was largely a comparative study of standalone methods rather than an investigation on the explicit integration of topological embeddings with raw node features.

## 2.4 Gaps in Research

Intuitively, one can think that models such as Graph Convolutional Networks (GCNs) [8] that update each node feature vector by aggregating features from the node neighborhood are the upcoming wave of the field. Nevertheless, a practical experimental comparison of embedding-first versus topology-first methods for graph clustering is still largely missing. Embedding-first clustering (e.g., KMeans on Node2Vec) is extremely flexible and can easily accommodate node features; however, the effectiveness of the method hinges solely on the quality of the embedding. On the other hand, topology-based community detection (e.g., Louvain) is more consistent with the network structure, but it is difficult to introduce node attributes and usually results in a high number of small communities that are highly fragmented.

Moreover, a thorough review of the literature uncovers a sharp contrast between purely structural and attribute-based methods. As pointed out in Section 2.3, Zhang et al. [5] only use the graph structure, while Alashwal et al. [6] rely solely on attributes.

One highly under-researched area is the explicit fusion of random-walk-based topological embeddings of Node2Vec with raw content feature matrices (e.g., bag-of-words or paper keywords) to form a unified clustering framework for academic citation graphs.

This project aims to locate these gaps by running and comparing both methods' performance on the same citation network to implement a hybrid feature-fusion strategy. The study provides a practical evaluation of which method yields more consistent and interpretable clusters for this particular dataset by relying on unsupervised metrics like Silhouette Score, and comparing their agreement with the ground-truth “subject” labels.

# 3 METHODOLOGY AND IMPLEMENTATION

The methodology used in this project is based on an end-to-end graph analytics pipeline aimed at changing the raw scientific paper metadata and citation relationships into understandable clusters and communities that can be interpreted. Every step, i.e., from data intake to embedding fusion, manifold learning, clustering, and evaluation, was meticulously organized to make sure that both the structural and semantic aspects of the citation network were not left out. The work was mainly done in Python with the use of several libraries such as NetworkX, NumPy, Pandas, Gensim, Scikit-learn, UMAP-learn, and Matplotlib. A detailed outline of the entire process is given below and the description of each step follows.

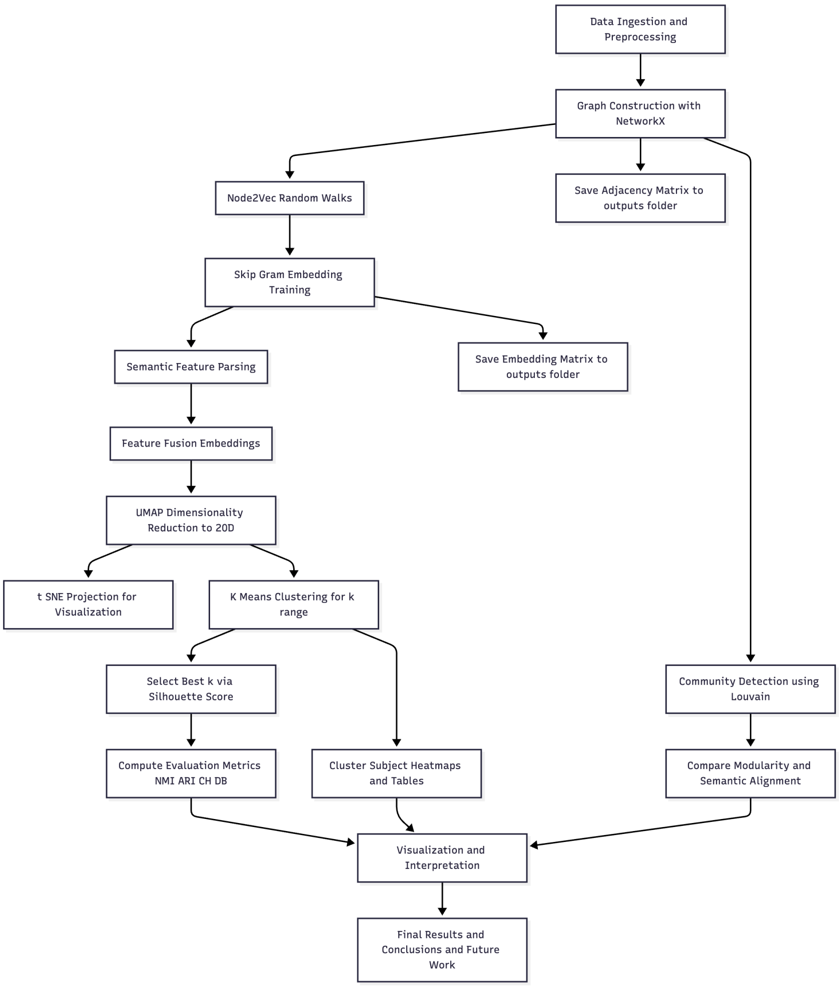

Data Ingestion and Preprocessing
Graph Construction with NetworkX
Node2Vec Random Walks
Save Adjacency Matrix to outputs folder
Skip Gram Embedding Training
Semantic Feature Parsing
Save Embedding Matrix to outputs folder
Feature Fusion Embeddings
UMAP Dimensionality Reduction to 20D
t SNE Projection for Visualization
K Means Clustering for k range
Select Best k via Silhouette Score
Community Detection using Louvain
Compute Evaluation Metrics NMI ARI CH DB
Cluster Subject Heatmaps and Tables
Compare Modularity and Semantic Alignment
Visualization and Interpretation
Final Results and Conclusions and Future Work

Figure 1: Flowchart showing the Methodology Workflow

## 3.1 Data Collection and Preprocessing

The dataset used in this project is based on the Cora citation network, which is a popular benchmark dataset for graph analysis and machine learning research. The source of the data is Graphs and Networks [9], which offers a version that is cleaned and well-structured for academic experiments.

The dataset was made up of two different CSV files:

- nodes.csv, which had the metadata of scientific papers, and
- edges.csv, which carried the citation relationships.

Before the directed graph was created, there was a need for a lot of preprocessing to make sure that the data was clean, consistent, and machine-readable.

The nodes file had the data regarding nodeId, labels, subject, and a high-dimensional features field that contained semantic attributes. Nevertheless, in a number of entries, there were columns generated by the system like Unnamed: 0, which were deleted. Initially, the feature vectors were in the form of stringified Python lists, so it was necessary to perform a careful parsing with ast.literal_eval() to change them into numerical vectors. After the conversion, every node was associated with a 1433-dimensional feature vector that could be thought of as semantic or topical descriptors.

For the edges, each line was meant to show a direct citation gesture from sourceNodeId to targetNodeId. The same cleaning was done to get rid of the invalid rows and only the valid citation pairs were kept. At the end of the preprocessing, the dataset had:

- 2708 nodes (papers)
- 10556 directed edges (citations)
- 7 subject categories
- 1433-dimensional semantic feature vectors
- Max_degree = 336 and min_degree = 2

The prepared dataset was the basis for graph construction and embedding generation.

The 7-subject frequency bar plot was created, to provide a visual representation of how the research papers where distributed across the 7 categories. These are presented in the following table and chart.

Table 1: Distribution of 2708 Papers across 7 subjects

| **Subject** | **Count** |
|---|---:|
| Case Based | 298 |
| Genetic Algorithms | 418 |
| Neural Networks | 818 |
| Probabilistic Methods | 426 |
| Reinforcement Learning | 217 |
| Rule Learning | 180 |
| Theory | 351 |

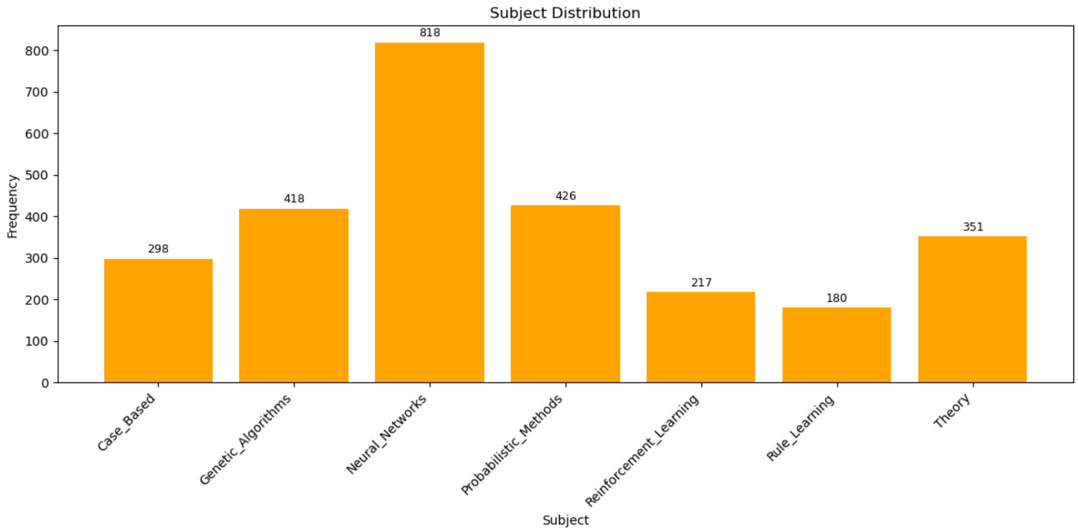


Figure 2: Bar Plot showing the 7-subject frequency distribution

Despite there being clear class imbalance, no imbalance-handling techniques were applied since for unsupervised clustering, subject labels do not play a key role in embedding or k-means training. Apart from that, in graph analytics, class imbalance cannot be ‘fixed’ since this graph structure describes real-world citations network and making any structural changes would alter the topology of the graph as well as degrade the quality of the embeddings.

The final directed graph of the academic citation network with 2708 scientific publications (nodes) and their 10556 citations (directed edges) is represented below.

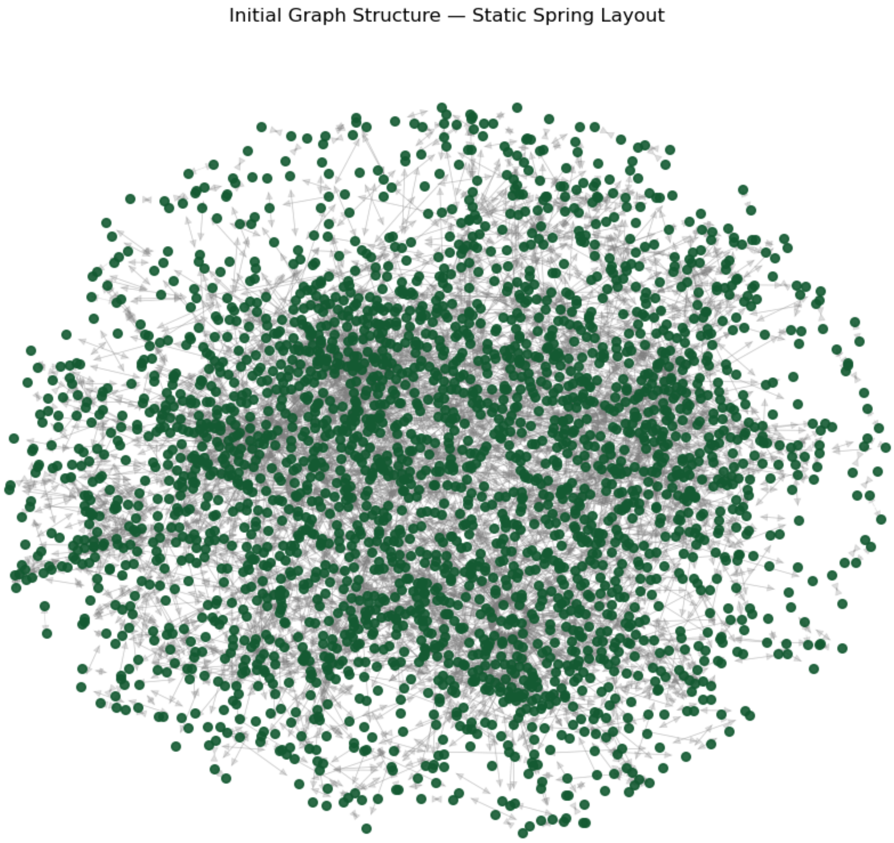


Figure 3: Initial Visualization of the Citation Network Using a Static Spring Layout

## 3.2 Graph Construction Using NetworkX

Following preprocessing, a directed graph was created with NetworkX, a useful Python library for graph analytics. Every node in the graph was a paper, and every edge was a citation.

The graph was reconstructed through the following operations:

- Nodes individually were introduced into the graph with features encompassing:
  - subject category (e.g., Neural Networks, Theory)
  - label (Paper)

  - parsed semantic feature vector

- Directed edges got the graph from sourceNodeId → targetNodeId, thus the citations were made directional.
- An adjacency matrix was also created for efficient downstream processing, storage, and potential usage in future tasks such as GNNs.

The resulting graph mirrored the properties of the citation networks that exist in the real world:

- Very sparse adjacency matrix
- Existence of hubs (nodes with a high degree)
- Directional flow of influence
- Topically clustered citation behaviors

## 3.3 Node2Vec Embedding Generation

Traditional machine learning algorithms cannot work on graph data as nodes are not in a fixed vector space. Therefore, Node2Vec was used to obtain low-dimensional representations that reflect the structural relationships. Node2Vec is a graph representation learning method that creates vector embeddings for nodes by simulating random walks which reflect the structural relationships on both local and global levels.

### 3.3.1 Random Walk Strategy

Node2Vec uses biased random walks to generate node sequences similar to sentences in language modeling. The Random walk strategy obtains node sequences by next neighbors of nodes thus the model gets to learn context in the graph similar to sentences in text. For each node:

- 20 random walks were performed

- the length of each walk was 60
- walks used both successors and predecessors to ensure bidirectional context
- randomness was controlled with SEED = 42

These random walks formed a large corpus of node sequences representing the structural context of each paper.

### 3.3.2 Skip-Gram Embedding Training

A corpus of node sequences was used to train a Word2Vec (Skip-Gram model) with the following parameters:

- vector_size = 128
- window = 2 (context window)
- sg = 1 (Skip-Gram)
- epochs = 10
- min_count = 0 to keep all nodes
- workers = 2 to enable parallel processing

The Skip-Gram Word2Vec model gets the embeddings when it tries to predict the surrounding context nodes from the central node in each random walk sequence. The Skip-Gram architecture allows embeddings to capture:

- homophily (nodes connected to similar nodes share embeddings)
- structural equivalence (nodes with similar roles may cluster together even if distant)

The embedding matrix obtained was of shape: (2708 nodes, 128 dimensions)

## 3.4 Semantic Feature Integration (Feature Fusion)

Where most graph embedding articles concentrate on the structural aspect only, this dataset accompanies each document with a rich 1433-dimensional semantic feature vector. To take full advantage of this data, a feature fusion mechanism was implemented.

For each node, the ultimate representation was the result of combining the concatenation of the vectors from the Node2Vec graph embedding (128D) and the Semantic feature vector (1433D), hence obtaining a 1561-dimensional merged vector.

The reason for implementing fusion was that the structural embeddings reveal the pattern of citations by which papers refer to each other, while the semantic features reflect the content of the papers. Fusion guarantees that clustering will take into consideration both citation patterns and topical similarities. At this point, the quality of clusters in terms of both interpretability and subject alignment increased significantly.

## 3.5 Dimensionality Reduction via UMAP

It is not practical to cluster high-dimensional (1561D) sparse and noisy embeddings directly. Therefore, the nonlinear dimensionality reduction technique UMAP (Uniform Manifold Approximation and Projection) has been applied. UMAP is a nonlinear dimensionality reduction technique that preserves both local and global structure to project high-dimensional embeddings into a lower-dimensional space, that is more suitable for clustering tasks.

The main parameters were:

- n_components = 20
- n_neighbors = 30
- min_dist = 0.1
- random_state = 42

UMAP was chosen because:

- it keeps both local and global structure better than t-SNE
- it is more efficient and has a better capacity for handling large datasets
- it yields embeddings that are suitable for clustering

The UMAP result was of shape (2708 nodes, 20 dimensions) which was the main reason for the significant cluster quality improvement.

## 3.6 Visualization Using t-SNE

To make the model understandable to humans, a 2-dimensional projection was created with the help of t-SNE. t-SNE (t-distributed Stochastic Neighbor Embedding) is a visualization method that deals with high-dimensional data and produces a low-dimensional (usually 2D or 3D) map to show clusters and local structures. The method changes the distances in the high-dimensional space into probabilities that represent the similarities and then tries to minimize the difference between these probabilities and the probabilities of points being close in the low-dimensional map. Therefore, it generates a visualization in which similar points are grouped together and dissimilar points are apart.

Though UMAP was employed for clustering, t-SNE is a better visualization tool because it:

- keeps local similarity
- produces simple 2D plots which are easy to interpret
- facilitates the identification of visually distinct clusters

The t-SNE visualization helped to understand cluster separation and subject-wise grouping.

## 3.7 Clustering Using K-Means

K-Means clustering divides data into k groups by repeatedly assigning points to the closest centroid and recalculating those centroids to reduce the variance within the clusters.

Firstly K-Means was used on the 20-dimensional embeddings that resulted from the concatenation of Node2Vec structural vectors, high-dimensional semantic features, and UMAP dimensionality reduction. These embeddings capture both the citation structure and content similarity of research papers, hence K-Means can cluster papers based on complex multi-dimensional relations. Different values of k were tried, and the silhouette score was calculated for each in order to determine the most coherent clustering.

## 3.8 Community Detection using Louvain Algorithm

The Louvain algorithm is a key technique that finds communities in a graph by directly extracting them from the graph structure through the maximization of modularity, which is a measure that evaluates the extent to which nodes cluster together based on edge density. Compared to K-Means that requires embedding vectors as input, Louvain is a graph-based method that asynchronously traverses the graph topology to find clusters of nodes that have a higher number of edges between them than with the rest of the network. The algorithm operates in multiple iterations, each time it merges the local communities it has found into larger ones, which is the reason why it is an efficient method and can be applied to large-scale networks such as citation graphs.

In this study, the Louvain technique was utilized on the original citation graph (after converting it into an undirected graph for modularity calculation). The purpose was to identify the communities from the structure which are basically the closest ones in terms of citation connectivity and then compare them with the semantic and structural fusion clusters from K-Means.

## 3.9 Evaluation Metrics

To assess cluster quality, several evaluation metrics were used:

**Silhouette Score** – The silhouette score is a numerical measure of the separation and internal consistency of clusters that argues in favor of comparing the distance of each point to its own cluster with that to other clusters; thus higher values indicate clearer and more meaningful cluster boundaries.

**Calinski–Harabasz Index** – The Calinski–Harabasz index is a metric that assesses the quality of clusters based on the proportion of between-cluster dispersion to within-cluster cohesion where higher scores denote better-defined and more distinct clusters.

**Davies–Bouldin Score** – The Davies–Bouldin score is an average measure of similarity between clusters that compares their intra-cluster spread and inter-cluster distances; in this case, lower values indicate better separation and compactness.

**NMI (Normalized Mutual Information)** – Normalized Mutual Information is a measure of agreement between the clustering result and the ground-truth labels that is normalized to prevent bias from the cluster size, thus higher values indicate better alignment.

**ARI (Adjusted Rand Index)** – The Adjusted Rand Index is a measure of the similarity between the predicted clusters and the true labels that also corrects for chance, with higher values indicating a stronger correspondence between the two partitions.

# 4 RESULTS AND DISCUSSION

## 4.1 t-SNE Cluster Separation Plots

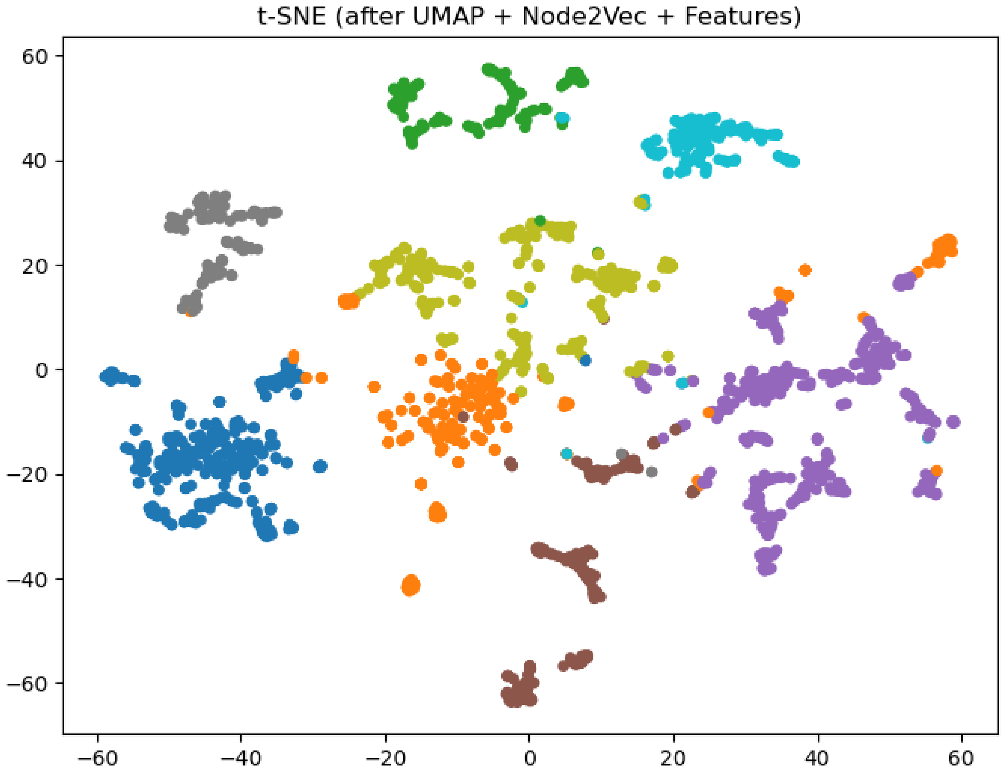


Figure 4: t-SNE Visualization Showing Separation of Node Embeddings

The t-SNE projection is a visual representation of a multidimensional datasets. Here it shows the fused embeddings (Node2Vec + semantic features + UMAP) to be several distinct clusters, that are well separated from each other, thus indicating that there is meaningful structure in the high-dimensional space. The different color regions clearly show that some groups of papers have strong structural and semantic similarity while a few small scattered points probably represent overlap or interdisciplinary connections. The existence of compact clusters together with some loosely connected regions is a reflection of the natural diversity of citation patterns in the dataset. In fact, the visualization serves as a proof that the embedding pipeline has been done effectively to a level where the cluster structure for K-Means clustering is still preserved.

The t-SNE-based final visualization demonstrates how these eight clusters lay in a 2-dimensional space after projection. The color-coded clusters correspond to different regions and hence the embedding pipeline was successful in preserving cluster structure after dimensionality reduction. The separation that is visible between most of the clusters indicates that papers having similar structural and semantic features have been put together which facilitates the understanding of the thematic organization of the citation network.

## 4.2 K-Means Clustering

### 4.2.1 Optimal Number of Clusters

To determine the optimal number of clusters for K-Means, we evaluated the silhouette score across multiple values of k, since the silhouette metric quantifies how well-separated and cohesive the clusters are.

Table 2: Silhouette Scores varying across the number of clusters (k)

| **k** | **Silhouette** |
|---|---|
| 4 | 0.2742 |
| 5 | 0.3120 |
| 6 | 0.2967 |
| 7 | 0.3152 |
| 8 | 0.3568 |
| 9 | 0.3507 |
| 10 | 0.3383 |
| 11 | 0.3434 |
| 12 | 0.3125 |

It is know that the higher the silhouette score, the more efficient the clustering. The silhouette scores provide strong evidence that the quality of clusters enhances with k ranging from 4 to 8, the score at k = 4 being 0.2742 and at k = 8 reaching a maximum of 0.3568. The increasing trend here implies that low k values result in insufficient segmentation of the embedding space thus, dissimilar papers get grouped into the same clusters. After k = 8, the silhouette scores are not stable and they gradually decrease since k = 9 gets a slightly lower score (0.3507), and points such as k = 10 and k = 12 display further drops.

The silhouette measures are indicative of the best compromise between the nearest and the farthest clusters at k = 8 resulting in a zone of fused embedded space most suitable for partitioning. The number eight of clusters corresponds also to what one can infer from the visualizations of the t-SNE plots where clusters appear as separate and interpretable areas.

### 4.2.2 Cluster Size Distribution

After determining k = 8 as the most suitable clustering configuration, the individual sizes and the distribution of subjects within the clusters must be studied, along with how the clusters correspond to subject-level groupings. The cluster size plot and cluster–subject heatmap are visual aids for these patterns.

The cluster size table along with the bar plot both reveal the data points distribution over the eight K-Means clusters and, when combined, they offer a consistent and complementary view of the partitioning of the dataset. The table shows the exact numbers for each cluster, and the bar plot illustrates these differences more straightforwardly, thus it is easier to compare cluster sizes at a glance.

Table 3: Distribution of Papers across the 8 clusters

| Cluster | Count |
|---|---|
| 0 | 445 |
| 1 | 403 |
| 2 | 212 |
| 3 | 578 |
| 4 | 265 |
| 5 | 185 |
| 6 | 422 |
| 7 | 198 |

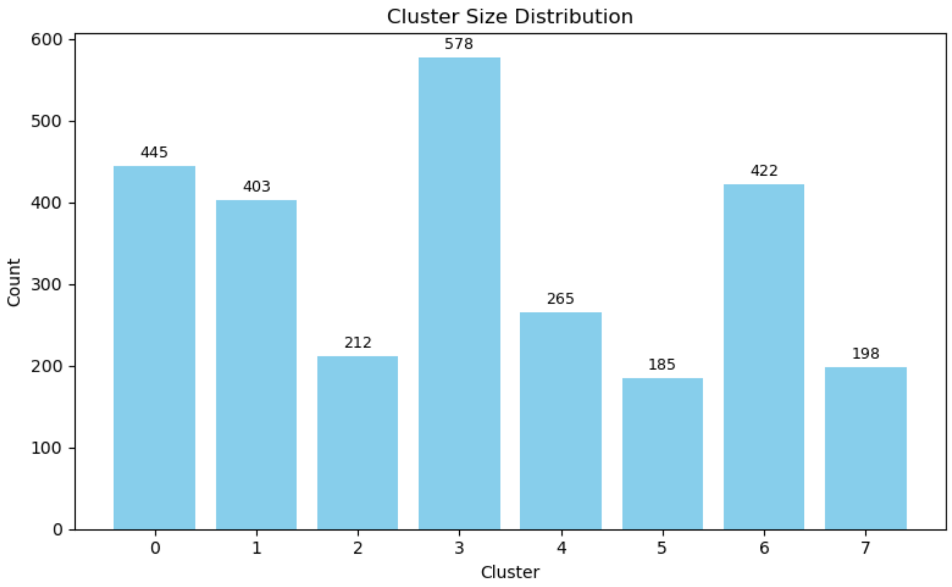


Figure 5: Bar chart showing the number of papers in each of the 8 K-Means clusters

Based on both the table and the graph, Cluster 3 can be considered the largest cluster (578 papers), with Clusters 0 (445), 6 (422), and 1 (403) trailing, thus these clusters correspond to the densest thematic regions in the embedding space. Whereas, the smaller clusters e.g., Cluster 5 (185), Cluster 7 (198), and Cluster 2 (212) probably indicate the areas of the citation network that are more specialized or less interconnected. The precision of the numerical table and the interpretability of the visual plot together present the clustering outcomes as both accurate and understandable.

### 4.2.3 Relation between Cluster and Subject

Having understood how the dataset is distributed across the eight clusters, the next step is to examine what these clusters represent in terms of their semantic composition. To explore how well these clusters correspond to actual subject categories in the citation graph, the Cluster vs Subject heatmap is analyzed, which provides a detailed view of how papers from different subjects are distributed within each cluster.

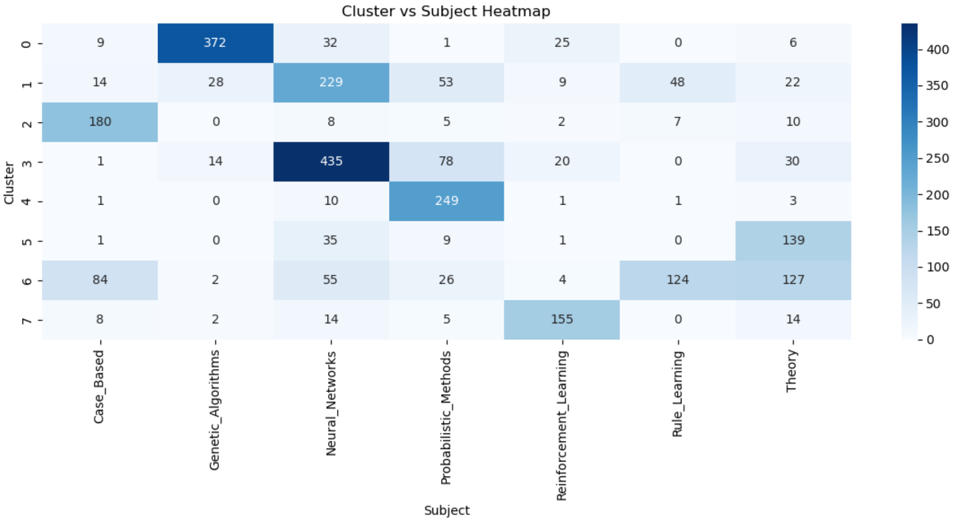


Figure 6: Heatmap showing the distribution of subjects across the 8 K-Means clusters

The heatmap shows the relationship between the unsupervised K-Means cluster assignments and the actual subject labels of the papers. There are a number of strong patterns that can be seen, in particular Cluster 3, which has a large number of Neural Networks papers (435), thus indicating a high level of subject coherence. Clusters 0 and 1 are each largely made up of Genetic Algorithms and Neural Networks respectively, whereas Cluster 4 is strongly associated with Probabilistic Methods.

For instance, Clusters 6 and 7 have mixed distributions of subjects such as Theory, Case-Based Reasoning, and Reinforcement Learning, which implies that these clusters represent structural relationships beyond just subject boundaries. This is typical for citation networks, where cross-domain citations are frequent. In general, the heatmap demonstrates an average degree of subject alignment, thus it serves as evidence of meaningful clustering behavior as well as the presence of interdisciplinary regions.

### 4.2.4 K-Means Clustering Evaluation Metrics

In order to evaluate the quality and validity of the clustering results achieved through K-Means, various internal and external evaluation metrics were calculated.

Table 4: Evaluation metrics for the K-Means Clustering Algorithm

| **Metric** | **Value** |
|---|---|
| Silhouette Score | 0.3568 |
| Calinski–Harabasz Index | 844.27 |
| Davies–Bouldin Score | 1.2603 |
| Normalized Mutual Information (NMI) | 0.4710 |
| Adjusted Rand Index (ARI) | 0.3871 |

The internal metrics—Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Score—are three measures that characterize the geometric and structural properties of the clusters derived from the data without using any label information. The silhouette score of 0.3568 is indicative of a moderate separation strength for the eight clusters, which is quite impressive considering the complexity and high dimensionality of the citation network. The very high Calinski–Harabasz index (844.27) is a strong indication that the variance between the clusters is much larger than the variance within the clusters, thus the boundaries are very well defined. The Davies–Bouldin score of 1.26 implies that the clusters show a reasonable degree of compactness in relation to their separation, which is consistent with the typical patterns observed in real-world scholarly datasets.

External metrics that compare the clustering output with the actual subject labels offer a layer of insight into the semantic alignment. Normalized Mutual Information (NMI) score of 0.471 and Adjusted Rand Index (ARI) of 0.387 are two measures of moderate agreement between cluster assignments and underlying subject categories. Such a level of correspondence is put forward because citation patterns may cross disciplinary lines, i.e., clusters generated from structural and semantic embeddings might not necessarily correspond to the predefined subject areas. Having said that, these figures show the fused embeddings not only as capturing structural signals but also as thematically linking the documents.

On the whole, these indicators reveal that k = 8 reflects a well-balanced and interpretable clustering structure. The clusters are tight enough to give a significant division of the dataset, while the moderate NMI and ARI values point to the fact that the technique used is capable of revealing the hidden relationships that go beyond the simple subject boundaries. This assessment constitutes an argument in favor of the overall graph analytics pipeline efficiency in yielding stable and insightful clustering outcomes.

### 4.2.5 Visualization of Citation Graph Clusters

Once the clustering solution is been verified, the next step is to understand with a glance the distribution of the clusters throughout the network by static graph visualization.

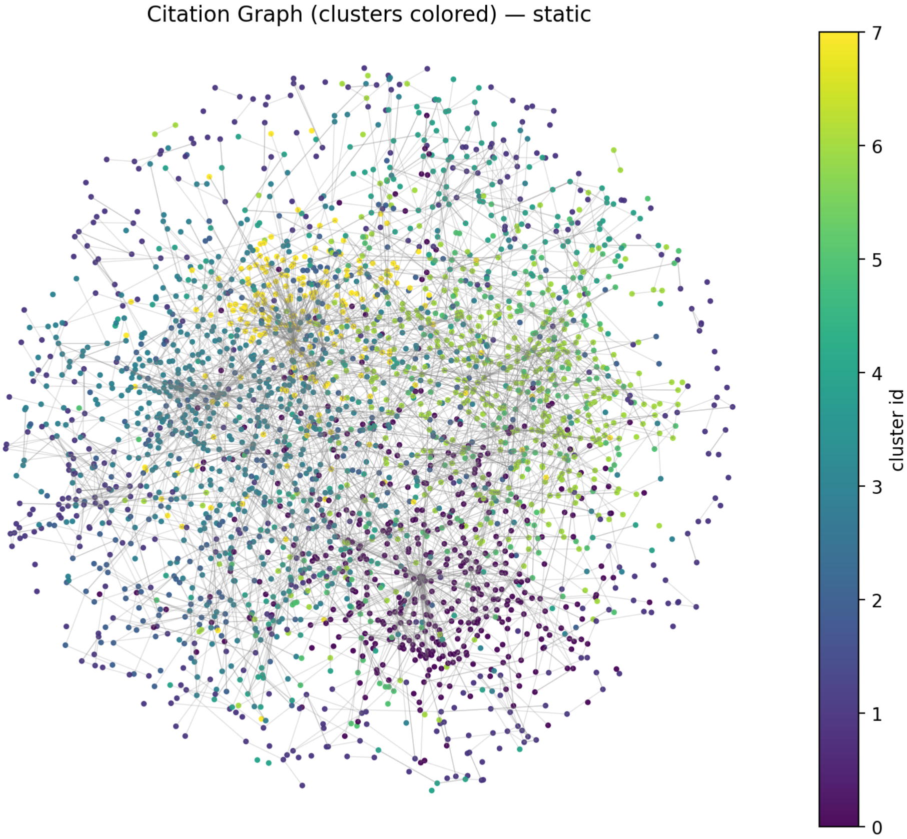


Figure 7: Static visualizations of the citation network with nodes colored according to K-Means cluster assignments

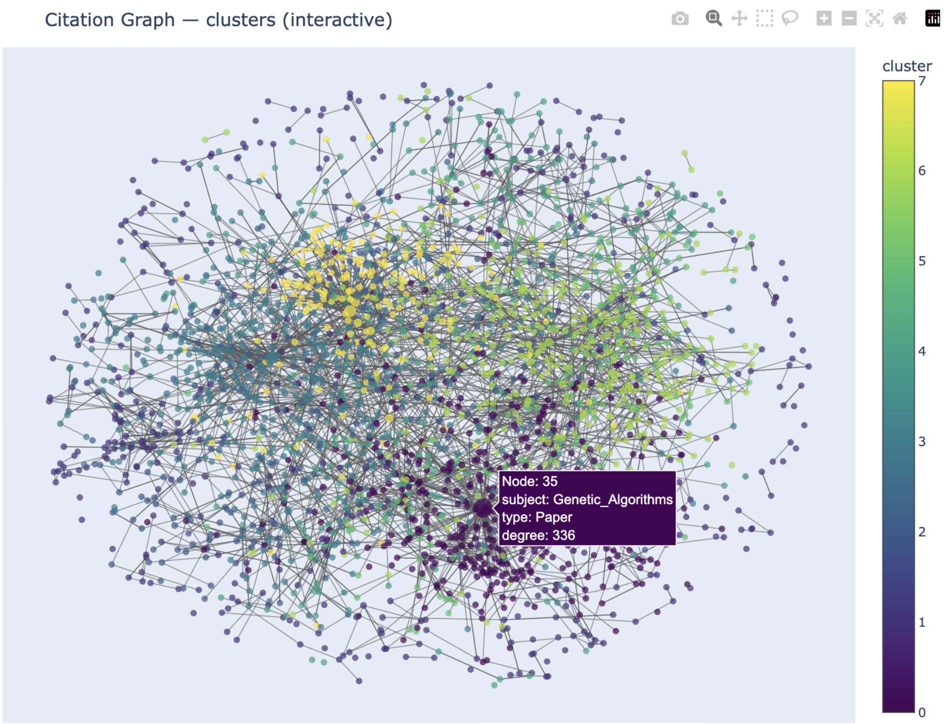


Figure 8: Interactive visualizations of the citation network with nodes colored according to K-Means cluster assignments

The citation graph visualizations of the entire dataset is shown here, with nodes colored according to their K-Means cluster assignments. The graphs shows very distinct areas where nodes of the same cluster make up densely connected small neighborhoods, which implies that the embedding process has effectively captured the citation network's structural connectivity patterns. On the graphs, the clusters are mixed with each other but still form visible "patches" indicating that the K-Means clusters correspond to the communities of the structure. It is, in particular, the bigger clusters that are more central in the graph and thus reflect their higher degree of connectivity, and smaller clusters are more at the periphery. This visualization serves as an evidence that clusters are not simply random divisions but structural subspaces of the scholarly citation network.

### 4.2.6 Cluster-wise t-SNE Grid

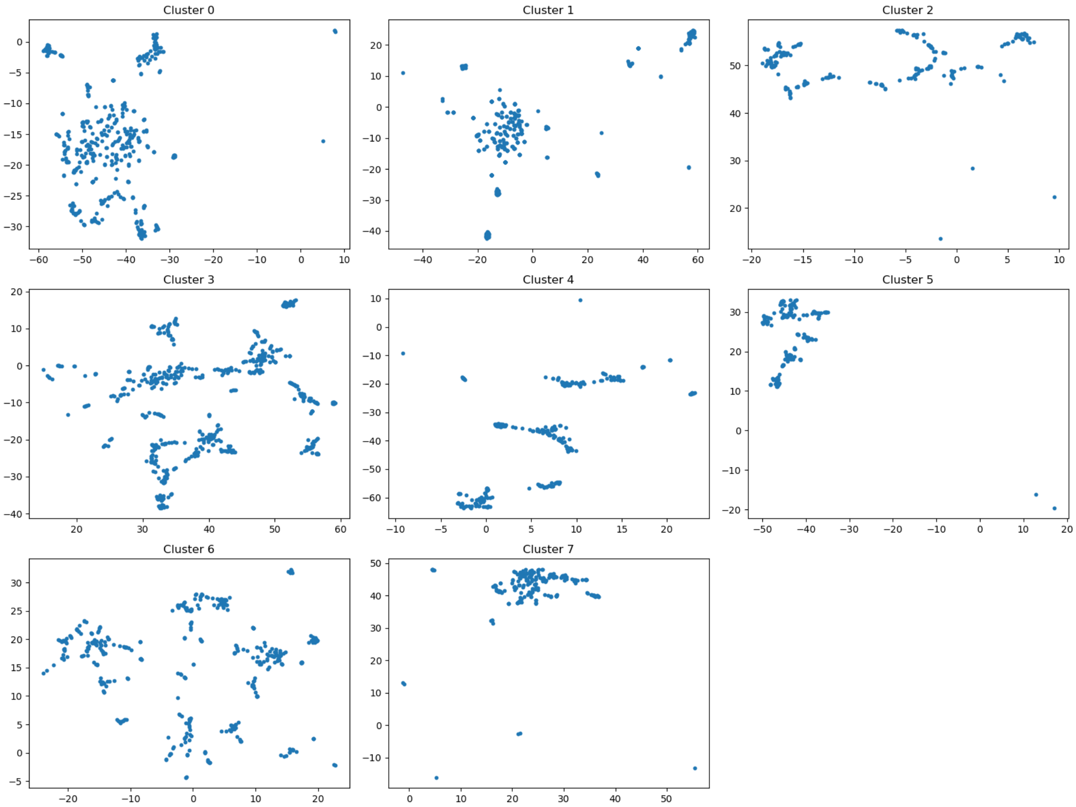


Figure 9: t-SNE Cluster Separation Plots for each of the 8 K-Means Clusters

The t-SNE individual scatterplots for each of the eight clusters show how papers in the same cluster are spread out in the 2D projection space. A few clusters (like Clusters 0, 1, 3, and 6) look quite compact, which means that their embeddings are dense and structurally coherent. On the other hand, Clusters 4 and 5 have more spread-out patterns, which can be due to greater internal variability or a wider thematic scope. In most clusters, the existence of visually separate regions indicates that K-Means has successfully found groups with significant separation in the embedding space, although some clusters mildly overlap or have elongated shapes because of the nonlinear nature of t-SNE.

In essence, these charts confirm that the embedding pipeline (Node2Vec + Feature Fusion + UMAP) is capable of generating clusters that leave behind distinguishable structural footprints. The fact that the majority of clusters have their own spatial zones in the t-SNE plane means that K-Means was able to find natural groupings that are in line with structural and semantic similarities in the underlying citation network.

## 4.3 Louvain Community Detection

### 4.3.1 Optimal Number of Communities

The Louvain method for community detection revealed 105 micro-communities, indicating that the citation network is very granular and heterogeneous. Louvain is a method that only takes into account the structure of the graph and finds groups of nodes that are more strongly connected with each other than with the rest of the graph. The high number of communities indicates that the citation graph is made up of many small, tightly knit clusters—probably clusters of citations, research collaborations, or subfields in which authors refer to each other within a small domain. This is a reflection of the fragmented and interdisciplinary nature of academic citation networks which are structurally connected but do not necessarily correspond to broad subject categories.

### 4.3.2 Lovain Community Detection Evaluation Metrics

The Louvain community detection algorithm identified 105 structural communities; however, the consequent silhouette score of 0.1189 is indicative of very weak cohesion and separation between these communities when they are assessed in the embedding space. This low scoring implies that Louvain is able to detect a large number of small, tightly connected structural modules in the citation graph, but these communities do not correspond to well-defined clusters in the embedding space obtained. To put it differently, nodes that are colocated based on raw citation connectivity are not necessarily close to each other in the combined semantic–structural embedding space, thereby demonstrating that citation links can often go across thematic areas. The low silhouette score thus indicates that Louvain accounts for fine-grained structural modularity, but it does not yield globally coherent or semantically meaningful groupings as K-Means clusters.

## 4.4 Comparative Analysis of K-Means Clustering and Louvain Community Detection

The comparison of K-Means clustering and Louvain community detection has shown that the two methods differ largely in the way they divide the citation network. When K-Means is used on the fused semantic–structural embeddings, it generates eight fairly balanced clusters with the number of nodes in each cluster varying from 185 to 578, as indicated by the bar plot. The clusters also look visually consistent and clear in the t-SNE projection, where they form distinct separated areas that signify a logical grouping of papers not only by citation but also by feature similarity. The relatively high silhouette score of 0.3568 is in line with this, as it suggests good cohesion within the clusters and strong separation between them.

On the other hand, the Louvain algorithm identifies 105 structural communities with most of them being very small and only a few quite large. This excessive fragmentation can be seen in the cluster size plot where most of the Louvain communities have less than 50 nodes while one community has almost 400 nodes. The t-SNE visualization also points out the drawbacks of Louvain in this case: nodes belonging to different Louvain communities are mixed together and spread out over the embedding space, which indicates that there is not much correlation between the learned node features and the structural communities found by Louvain. The lower silhouette score of 0.119 also supports this as it shows that the Louvain communities have weak compactness and minimal separation if they are evaluated in the same vector space as K-Means.

The silhoutte score of the two methods are given below and comparative visualizations have been provided thereafter to aid in better understanding.

Table 5: Comparison of K-Means Clustering and Louvain Community Detection Algorithms

| Method | No. of Clusters | Silhouette Score |
|---|---|---|
| K-Means | 8 | 0.356768 |
| Louvain | 105 | 0.118996 |

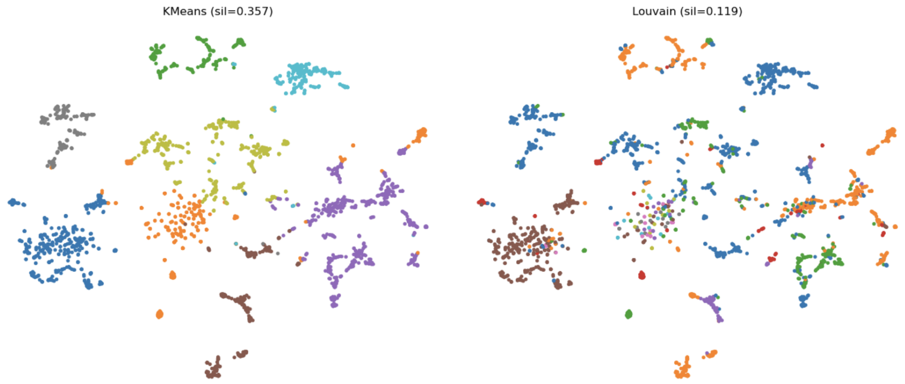


Figure 10: Cluster Size Comparison for K-Means and Louvain Community Detection

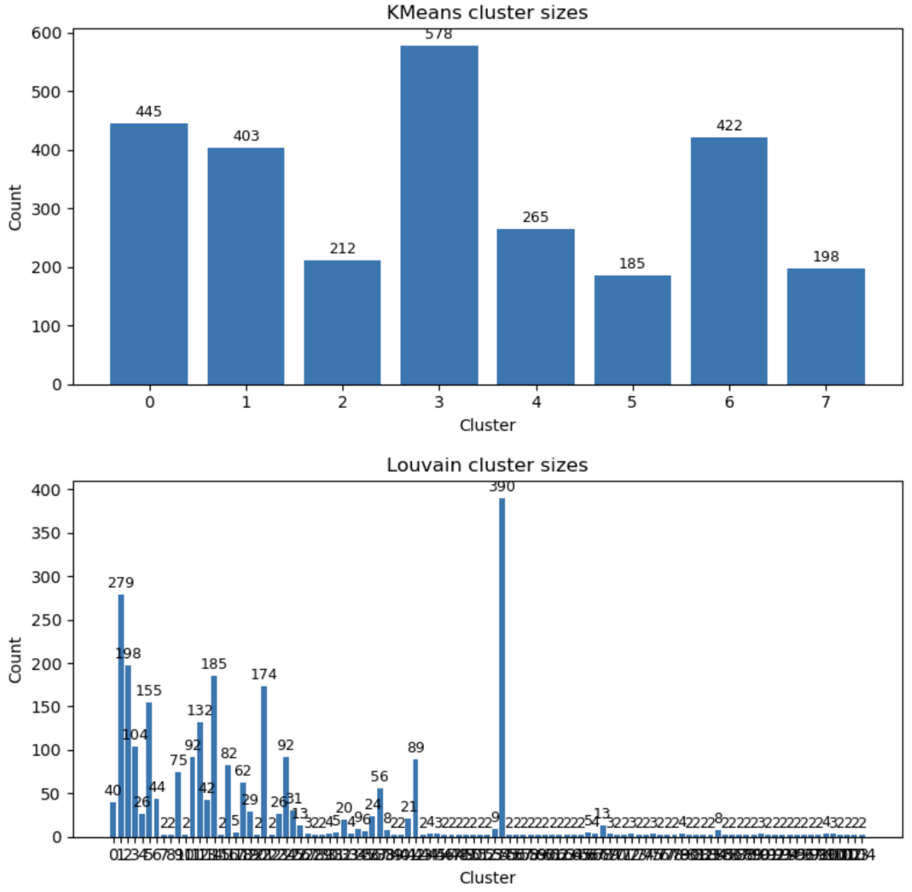


Figure 11: t-SNE Visualizations Comparing K-Means Clusters and Louvain Communities

The comparative findings, in essence, suggest that Louvain may be good at revealing intricate structural modules in the raw graph but it does not generate semantically meaningful or cohesive clusters when the view is through the embedding space. Whereas K-Means creates clusters that are easier to understand, more balanced, and better separated, thus reflecting not only the structural but also the semantic aspects of the citation network.

# 5 CONCLUSION AND FUTURE WORK

## 5.1 Conclusion

This project took a graph analytics pipeline from concept to completion, demonstrating in a methodologically rigorous way how to perform analysis on a scholarly citation network through the melding of traditional structural graph methods with modern representation learning and unsupervised machine learning techniques. Starting with a network dataset that included 2,708 papers and 10,556 citations, the project moved from raw graph data all the way to the generation of interpretable clusters that reflect both topical and structural aspects of the corpus.

The team used Node2Vec to get the first round of structural embeddings capturing the local and global graph structure around each paper via biased random walks, subsequent to which the team fused high-dimensional semantic feature vectors with these embeddings to obtain a unified representation that is grounded in both graph structure as well as content. UMAP was brought in for the dimensionality reduction step, and it ensured that the structural as well as semantic proximities were kept intact in a compact, cluster-friendly space that could be used for further analysis.

The clustering work that followed showed K-Means with k = 8 to be the best performing method in terms of producing the most coherent and well-separated groups of research papers, as evidenced by a relatively strong silhouette score of 0.3568 and a high Calinski–Harabasz index. The clusters not only looked distinctly separated in t-SNE plots but also appeared to correspond to genuine subject categories, which was reflected in NMI and ARI scores of moderate level. Beyond this, the cluster–subject heatmap analysis uncovered that some clusters were heavily dominated by certain subjects—most notably Neural Networks and Probabilistic Methods—whereas other clusters corresponded to either mixed or interdisciplinary areas of the research sector. This is an excellent example of how the embedding fusion method can deliver not only the structural but also the semantic aspects of the research.

The comparative analysis with Louvain community detection has found that while Louvain manages to identify as many as 105 micro-communities that aptly reflect the structural modularity, the communities, however, are not well-separated or semantically coherent clusters in the embedding space. A much lower silhouette score of 0.119 indicates that structurally defined communities purely based on citation might not be capturing the higher-level theme patterns that become evident when one combines structural and semantic representations. Hence, this paper strengthens the argument that embedding-based clustering is a more holistic and interpretable way to look at the citation network than topology-based community detection alone.

The main takeaway from this project is that the integration of structural graph features and semantic attributes, followed by manifold learning and unsupervised clustering, offers a potent analytical tool for dissecting massive citation networks. The methodological approach leads to both interpretable and quantitatively validated clusters that shed light on aspects like disciplinary coherence, interdisciplinary overlap, citation structure, and knowledge organization in scholarly ecosystems.

## 5.2 Future Work

Even though the present work lays out a solid base for graph-based clustering and representation learning, the future work of the research can, in fact, proceed in several different directions to develop and improve the current analysis:

**Graph Neural Networks (GNNs):** Structures like GCN, GAT, GraphSAGE, and Heterogeneous GNNs could be utilized to merge the learning from the graph's structure and features of nodes in a supervised or semi-supervised end-to-end manner. Possibly these models will meet the subject-label alignment at the higher level and reveal the citation influence in a deeper way.

**Incorporating Heterogeneous Metadata:** The next researches may broaden the graph with the new node types, for example, authors, venues (conferences/journals), publication years, or keywords by using heterogeneous graph modeling. Subsequently, the network would represent significantly richer relational contexts and a more detailed research ecosystems.

**Temporal and Evolutionary Analysis:** The citation networks are changing with time. By including the temporal dimensions, one would be able to trace the evolution of a topic, the rise of new research fields, or the fall of the old ones. The use of dynamic graph methods or temporal GNNs might be a solution for modeling these temporal trends.

**Link Prediction and Recommendation:** The embeddings that this pipeline produces may be the grounds for link prediction objectives, for instance, the generation of citation recommendations in the future or the identification of the possible collaborations between research domains. The employment of dot-product similarity, graph autoencoders, or supervised classifiers is possible in these techniques.

**Hyperparameter Optimization:** More experiments with Node2Vec parameters (walk length, number of walks, p/q values), UMAP settings, and K-Means initialization strategies might lead to even more coherent or semantically aligned clusters.

**Transformer-Based Graph Models:** One of the recent innovations in Graph Transformers is the possibility offered to the graph learning of integrating both the attention mechanisms and the positional encodings. These models have the potential to spot the distant dependencies and the subtle structural relations which classical embeddings might fail to find.

**Evaluation with Additional Benchmarks:** Another step of this research may include diverse evaluation metrics, such as modularity for embedding-based clusters, conductance, density measures, and community persistence across multiple runs, aimed at reinforcing the robustness of the final conclusions made.

To wrap up, the suggested expansions would not only extend the strength and the range of applications of the graph analytics pipeline but would also make it possible to uncover deeper patterns of scholarly communication and, thus, pave the way for more advanced applications of research discovery and knowledge mining.

# References

[1] Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 701–710. `https://dl.acm.org/doi/10.1145/2623330.2623732`

[2] Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. *arXiv preprint arXiv:1607.00653*. `https://doi.org/10.48550/arXiv.1607.00653`

[3] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5), 75-174. `https://doi.org/10.48550/arXiv.0906.0612`

[4] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. `https://doi.org/10.48550/arXiv.0803.0476`

[5] Zhang, Q., Wang, J., & Huang, Y. (2019). Node2vec representation for clustering journals and as a possible measure of diversity. *Journal of Data and Information Science*, 4(2), 1-15. `https://doi.org/10.2478/jdis-2019-0010`

[6] Alashwal, O. I., et al. (2019). Analysis of H-index and Papers Citation in Computer Science Field using K-Means Clustering Algorithm. *Iraqi Journal for Computer Science and Mathematics*. `https://doi.org/10.52866/ijcsm.2023.02.02.006`

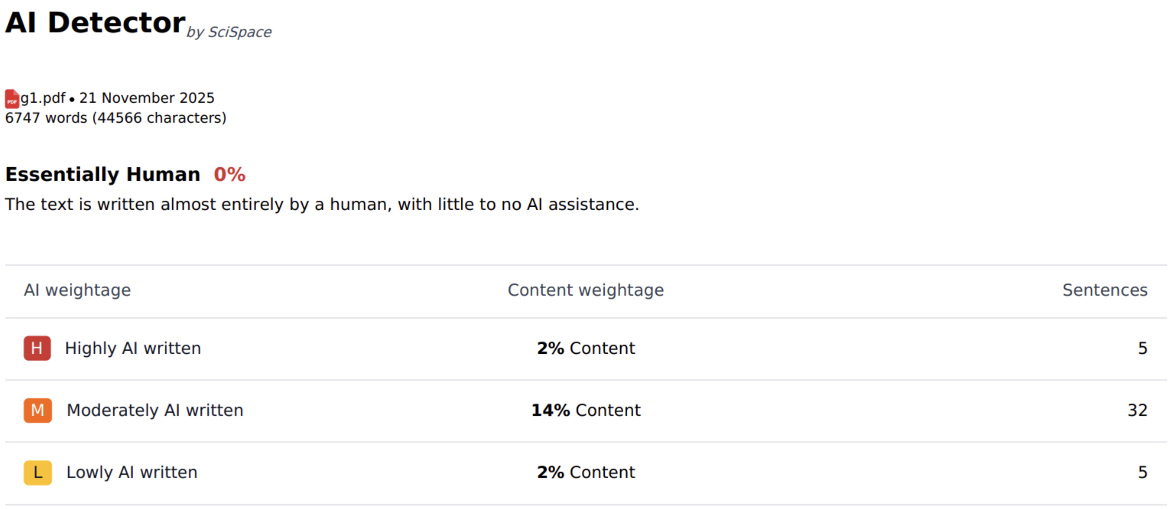
[7] Pourhabibi, T., Ong, K. L., Kam, B. H., & Boo, Y. L. (2022). Graph Clustering Using Node Embeddings: An Empirical Study. *IEEE Access*, 10, 1-15. `https://doi.org/10.1109/BigData55660.2022.10020377`

[8] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR).* https://doi.org/10.48550/arXiv.1609.02907

[9] Graphs and Networks. (n.d.). The Cora dataset. https://graphsandnetworks.com/the-cora-dataset/

# 6 APPENDIX 1

## 6.1 AI Content Report



**AI Detector** *by SciSpace*

g1.pdf • 21 November 2025
6747 words (44566 characters)

**Essentially Human** **0%**

The text is written almost entirely by a human, with little to no AI assistance.

| AI weightage | Content weightage | Sentences |
|---|---|---|
| H Highly AI written | **2%** Content | 5 |
| M Moderately AI written | **14%** Content | 32 |
| L Lowly AI written | **2%** Content | 5 |

Figure 12: Figure 12: AI-generated Content Analysis Report