Graph Curvature and Clustering: A Comparative Analysis for Customer Behavior Insights

Group Member Name	Email Id
Vaishnavi Mada	vmada1@student.gsu.edu
Anukeerthi Reddy Pothepalli	apothepalli1@student.gsu.edu
Vaishnavi Chilumuru	vchilumuru1@student.gsu.edu

J. Mack Robinson College of Business

MSA 8500: Graph Analytics

Mehmet Emin Aktas

12/06/2024

Abstract

Understanding customer relationships and their dynamics is crucial for businesses to optimize strategies, improve service delivery, and foster market growth. Traditional clustering methods often fail to capture the nuanced connections within customer-business networks, such as shared customer bases or complementary services. In this research, we explore the use of geometric measures, specifically Ricci curvature, to enhance clustering accuracy and reveal hidden patterns in customer relationships. By constructing a graph of businesses based on mutual customers, we leverage Ollivier-Ricci and Forman-Ricci curvature to incorporate local graph structures and assess their impact on clustering performance.

Our findings indicate that incorporating Ricci curvature provides a more granular understanding of business synergies, enabling the identification of significant customer overlaps that traditional techniques often overlook. Geometric clustering demonstrates improved modularity and entropy metrics, highlighting its ability to uncover richer insights. Furthermore, we benchmark curvature-based clustering against traditional approaches, demonstrating its superiority in capturing meaningful relationships and improving clustering outcomes.

This research underscores the potential of geometric methods in enhancing customer relationship analysis and opens avenues for future studies to refine and apply these techniques in broader domains such as recommendation systems, market segmentation, and urban planning. By integrating advanced geometric measures, we provide a novel perspective on how businesses can better understand and leverage their networks to gain a competitive advantage.

I. Introduction

Customer review platforms like Yelp have revolutionized how businesses interact with their customers, offering a wealth of data to analyze customer behavior, preferences, and satisfaction. The exponential growth of user-generated reviews creates both opportunities and challenges for businesses and researchers alike. Identifying relationships between businesses, such as shared customer bases or thematic similarities, is crucial for informed decision-making in areas like marketing, product placement, and customer engagement strategies. Traditional clustering methods often fail to capture the nuanced relationships within these data, especially in terms of shared customer behaviors or business synergies.

Graph-based analytics offers a powerful framework to model and analyze these complex relationships. By representing businesses as nodes and shared customers as weighted edges, the graph captures the underlying structure of interactions between businesses. However, conventional graph clustering techniques typically rely on metrics such as degree centrality or edge weight and do not fully utilize the geometric properties of the graph. This limitation inspired the integration of Ricci curvature-a concept from differential geometry-into the analysis of customer relationships in this research.

Ricci curvature provides a unique way to quantify the "spread" or "tightness" of connections in a network. In graph theory, Ollivier-Ricci and Forman-Ricci curvature are adapted from their geometric counterparts to evaluate local and global structural properties. Ollivier-Ricci curvature focuses on the movement of probability distributions across nodes, while Forman-Ricci curvature emphasizes edge-level properties, including weights and connectivity. Both measures add depth to the analysis by highlighting the strength and cohesion of relationships within clusters.

The study focuses on businesses in Phoenix, Arizona, using data from the Yelp dataset. The dataset was filtered to retain businesses with significant customer engagement, such as those with a minimum of 100 reviews and an average rating of four stars or higher. Shared customer relationships were used to construct a weighted graph where edges represent normalized shared customer counts between businesses. The goal is to uncover clusters of businesses that share similar customer bases or themes, leveraging Ricci curvature to enhance the clustering process.

The remainder of this paper is structured as follows. Section 2 discusses data preprocessing, including the construction of the graph and calculation of edge weights. Section 3 provides an overview of Ricci curvature and its role in clustering. Section 4 details the clustering methodologies applied, followed by an evaluation of results in Section 5 concludes the research and outlines potential future work.

Keywords: Ricci Curvature, Geometric Clustering, Customer Relationship Analysis, Mutual customer Graph

II. Literature Review

The application of graph-based techniques to uncover patterns in complex networks has been widely studied across various domains. This section highlights key studies that informed the methodological framework and objectives of this research.

1. Centrality and Cluster Analysis of Yelp Mutual Customer Business Graph

Brian McClanahan and Swapna S. Gokhale explored customer behavior and business relationships through the analysis of mutual customer graphs. Their work emphasized the influence of geographic proximity and service similarity in clustering businesses, providing valuable insights into consumer segmentation. This study laid the groundwork for applying graph clustering to customer-business relationships, offering a framework to analyze shared customer bases. It also underscored the importance of developing advanced clustering techniques to identify meaningful business connections, an area addressed in this research.

Centrality and Cluster Analysis of Yelp Mutual Customer Business Graph | IEEE Conference Publication | IEEE Xplore

2. Characterizing Physician Referral Networks with Ricci Curvature

Jeremy Wayland, Russell J. Funk, and Bastian Rieck introduced Ricci curvature as a novel measure for analyzing the structural properties of networks. The study demonstrated that Ricci curvature effectively captures local and global graph structures, offering new insights into community formation within referral networks. By highlighting the advantages of geometric properties in clustering, the study provided strong motivation for applying Ricci curvature to business graphs. This research adopts and extends these methods to evaluate whether Ricci curvature enhances clustering accuracy in mutual customer networks.

[2408.16022] Characterizing Physician Referral Networks with Ricci Curvature

3. Defining Quality Metrics for Graph Clustering Evaluation

Anupam Biswas and Bhaskar Biswas developed a comprehensive set of metrics for evaluating graph clustering algorithms. Their focus on AVI, AVU, and ANUI modularity and entropy as measures of clustering quality allowed for standardized evaluation and comparison of clustering methods. These metrics were instrumental in this study's evaluation framework, facilitating a robust analysis of the clustering outcomes obtained through Ricci curvature-enhanced methods compared to traditional techniques.

Defining quality metrics for graph clustering evaluation - ScienceDirect

Relevance to Current Research

These studies collectively form the basis for this research, which seeks to bridge the gap between traditional graph clustering techniques and innovative geometric measures. Specifically:

1. The insights into shared customer networks provided by McClanahan and Gokhale inform the design of the mutual customer business graph used in this study.
2. The introduction of Ricci curvature by Wayland et al. highlights the potential of geometric properties in enhancing graph clustering outcomes, directly inspiring the use of Ollivier-Ricci and Forman-Ricci curvature.

3. The metrics proposed by Biswas et al. guide the evaluation of clustering accuracy, enabling a rigorous comparison of traditional and curvature-based clustering approaches.

By building on these foundational works, this study investigates whether the incorporation of Ricci curvature can improve clustering outcomes and reveal deeper insights into customer-business relationships, advancing both theory and practical applications.

III. Data Description

The dataset used in this research comprises customer reviews and business information sourced from Yelp's open dataset. The data spans various categories, such as food, services, and retail, and is filtered specifically for businesses located in **Phoenix, Arizona**. Below is a detailed breakdown of the dataset:

Dataset Link

1. Dataset Overview

- **Business Data:** Contains details about businesses, including their location, category, and refined category labels.
- **Review Data:** Includes textual reviews provided by customers, along with star ratings, user IDs, and business IDs.

2. Business Data

The business dataset includes:

- **business_id:** Unique identifier for each business.
- **name:** Name of the business.
- **city:** City where the business is located (filtered to Phoenix in this study).
- **categories:** A comma-separated list of categories to which the business belongs (e.g., "Restaurants, Food, Mexican").
- **refined_category:** A manually refined and simplified version of the category column to group businesses into broader classifications (e.g., "Food," "Retail," "Services").

3. Review Data

The review dataset includes:

- **review_id:** Unique identifier for each review.
- **user_id:** Identifier for the customer who left the review.
- **business_id:** Identifier for the business being reviewed (links to the business dataset).
- **stars:** The star rating provided by the customer (1 to 5 stars).
- **text:** The full textual content of the review.

Key Filters Applied:

- Reviews with **star ratings >= 4** were selected to focus on positive customer experiences.
- Businesses with **100 or more reviews** were retained to ensure sufficient data for meaningful analysis.

IV. Graph Construction

Graph construction is a key component of this research, as it provides the foundation for analyzing relationships and interactions between businesses in Phoenix. Below is a step-by-step explanation of the graph construction process:

Node Definition

Each node in the graph represents a business. Nodes are created using the business_id column from the filtered business dataset. Attributes associated with each node include:

- **Business Name:** The name of the business, extracted from the name column.
- **Categories:** A list of categories to which the business belongs.
- **Refined Category:** A simplified category label to group similar businesses logically.

Example:

- Node: business_id = "lsHmnOZNlsKnEDsNmT-5_Q "
- Attributes: {name: "Plaza Bonita", category: "Food:Pizza:Restaurant", refined_category: "Restaurant"}

Edge Definition

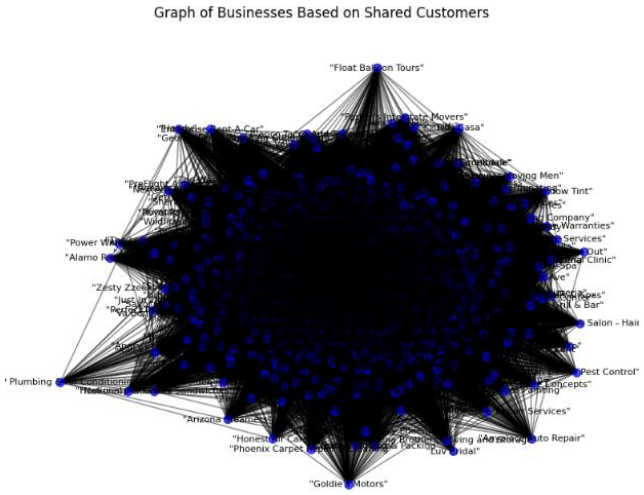
Edges between nodes represent relationships based on shared customers. The edge weight is calculated based on the number of customers who reviewed both businesses (mutual customers).

Steps to Calculate Edge Weights:

1. **Identify Mutual Customers:**
   - Group the reviews dataset by user_id to identify all businesses reviewed by the same customer.
   - For each customer, generate pairs of businesses reviewed by them.
2. **Count Mutual Customers:**
   - For each pair of businesses (A, B), count the number of mutual customers who reviewed both businesses.
3. **Normalize Edge Weights:**
   - The weight of an edge between businesses A and B is calculated as:

$$\text{Weight} = \frac{\text{Mutual Customers for A and B}}{\text{Total Customers for A}} + \frac{\text{Mutual Customers for A and B}}{\text{Total Customers for B}}$$

   - This formula ensures that the weight reflects the strength of the relationship relative to the size of each business's customer base.
4. **Thresholding:**
   - Edges with weights below a certain threshold (e.g., 0.1) are removed to reduce noise and focus on meaningful connections.



Graph Construction Using NetworkX

The graph is constructed using the Python networkx library, which provides powerful tools for graph analysis and visualization.

Graph Statistics:

- Total Nodes: 822
- Total Edges: 218182

Insights: Highly connected nodes represent businesses with significant overlap in their customer base. Edge weights provide insights into the strength of these relationships, allowing for further analysis of centrality and clustering.

Figure 1 Mutual Customer Graph

Challenges in Graph Construction

- **Data Sparsity:** Many businesses had limited shared customers, requiring thresholding to retain meaningful connections.
- **Edge Weight Normalization:** Balancing the contribution of mutual customers for businesses of varying sizes.
- **Visualization Complexity:** Handling large graphs with dense connections to ensure clarity.

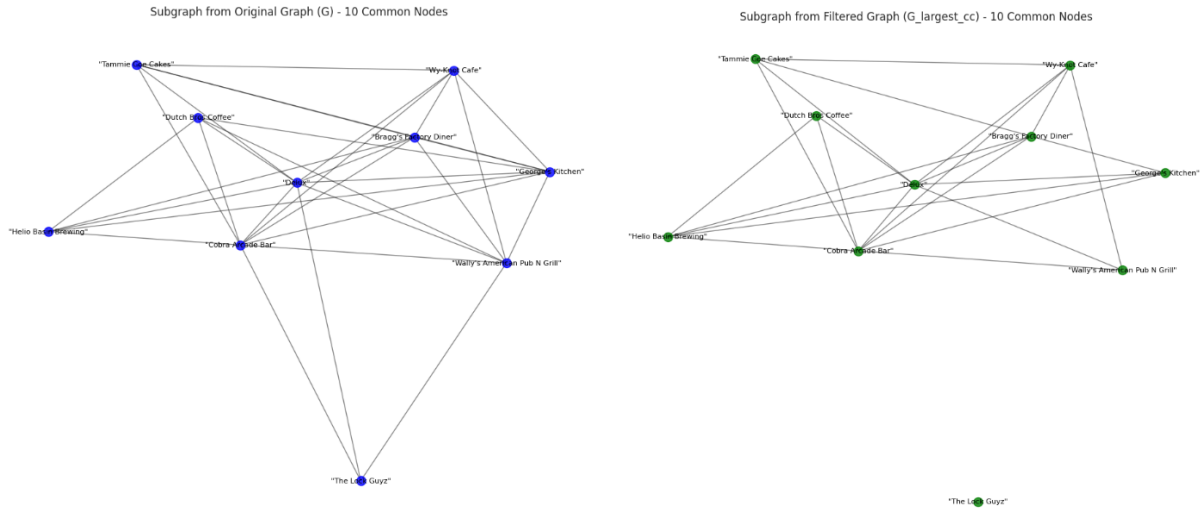


Figure 2: Comparison of graph before and after filtering

V. Ricci Curvature(Ollivier and Forman) :

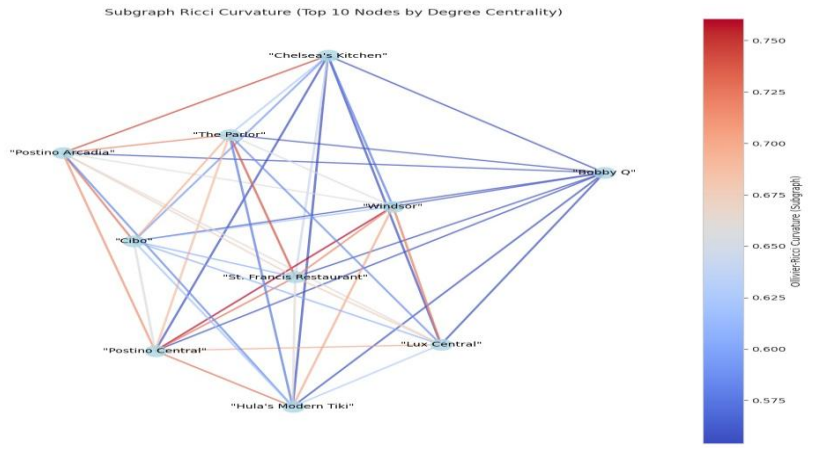
i. Ollivier-Ricci Curvature

- Definition: Ollivier-Ricci curvature quantifies the strength of connections in a graph based on the probability distributions between neighboring nodes. Positive curvature indicates tighter, stronger connections, while negative curvature reflects weaker or sparse relationships.

$$\kappa_{OR}(i,j) \coloneqq 1 - \frac{1}{d_G(i,j)} W_1(\mu_i, \mu_j) \quad \text{Olivier Ricci}$$

Explanation of Components:

1. $\boldsymbol{\kappa_{OR}(i,j)}$**:**
   - Represents the Ollivier-Ricci curvature for the edge between nodes i and j.
   - It measures how "curved" the relationship or connection is between the two nodes in the network.
2. $\boldsymbol{d_G(i,j)}$**:**
   - Represents the shortest path distance (geodesic distance) between nodes i and j in the graph.
   - This normalizes the curvature to account for the scale of the graph.
3. $\boldsymbol{W_1(\mu_i, \mu_j)}$ **:**
   - This is the **Wasserstein-1 distance** (also called the Earth Mover's Distance) between the probability distributions μ_i, μ_j associated with nodes I and j.
   - μ_i, μ_j describe how the mass or weight is distributed over the neighboring nodes of I and j.
   - The Wasserstein distance quantifies how much effort it takes to "transport" the mass from μ_i, μ_j
4. $\boldsymbol{\frac{1}{d_G(i,j)} W_1(\mu_i, \mu_j)}$ **:**
   - This formula subtracts the normalized Wasserstein distance from 1.
   - A **high Ollivier-Ricci curvature** value indicates strong local connectivity (tight clustering or alignment of neighbors).
   - A **low (or negative) Ollivier-Ricci curvature** value signifies weak connectivity, where neighbors are distributed more broadly or not aligned.



Explanation of Ollivier-Ricci Plots:

1. Subgraph Ricci Curvature (Top 10 Nodes by Degree Centrality):

○ This subgraph focuses on the top 10 central nodes, visualizing their connections based on curvature values.

○ Red edges dominate in this dense, highly connected subgraph, showing robust overlaps in customer bases.

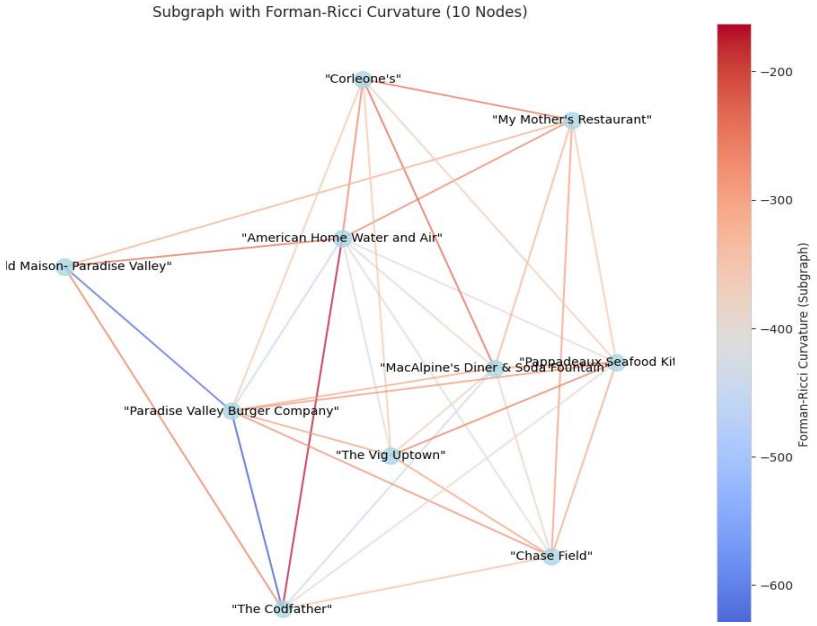
Figure 3 Ollivier Ricci Subgraph

ii. Forman-Ricci Curvature

- Definition: Forman-Ricci curvature evaluates the structural properties of edges, taking into account both edge weights and the degree of connected nodes. It provides finer granularity in assessing connection strengths.

$$\kappa_{\mathrm{FR}}(i,j) \coloneqq 4 - d_i - d_j + 3|\Delta|$$

- Implementation: Forman-Ricci curvature was computed for the graph, including a relabeled version for efficiency. Curvature values were added as attributes to edges and visualized.



Explanation of Forman-Ricci Plots:

1. Subgraph with Forman-Ricci Curvature (10 Nodes):

○ This visualization focuses on 10 nodes with the highest centrality, mapping their connections based on Forman-Ricci curvature.

○ The color gradient effectively showcases the variation in connection strength within this localized segment of the network.

Figure 4 Forman Ricci Subgraph

Comparison of Ollivier-Ricci and Forman-Ricci Curvature

- Average Curvature:
  - Ollivier-Ricci: 0.4079 (Moderately positive, indicating cohesive and strong connections in the network).

- Forman-Ricci: −296.8900 (Broad negative range, emphasizing the algorithm's sensitivity to structural variations).

- Distribution:
  - Ollivier-Ricci exhibits a concentrated distribution, reflecting its focus on average connection strength between nodes.
  - Forman-Ricci spans a wider range, capturing edge-level variations and identifying both strong and weak connections.

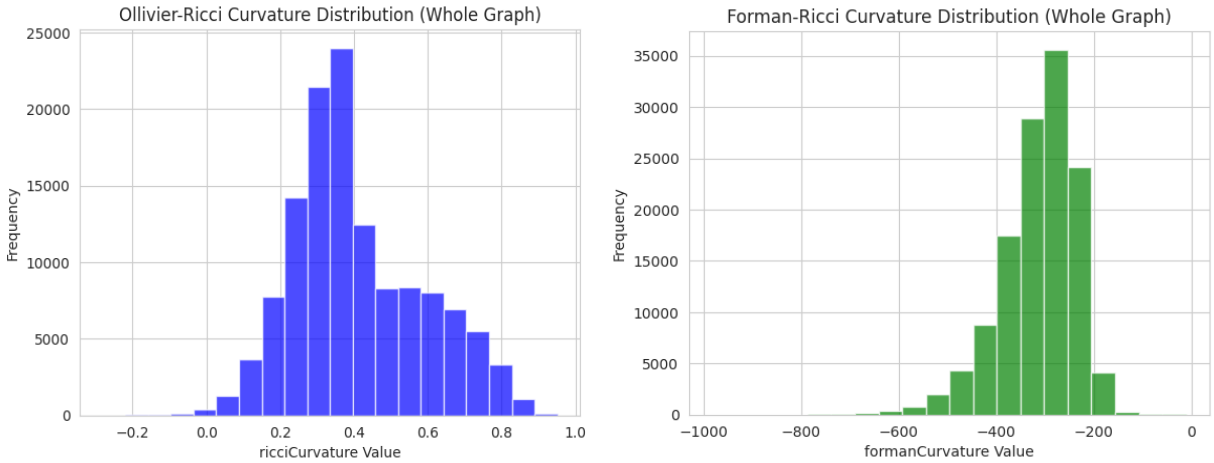


Figure 5 Curvature Distribution

Insights:

- Ollivier-Ricci curvature provides a high-level view of graph cohesion, highlighting clusters and central nodes effectively.
- Forman-Ricci curvature delves deeper into edge-level characteristics, identifying subtle structural variations within the graph.

VI. Clustering Techniques

K-Means Clustering

K-Means clustering is a partition-based unsupervised machine learning algorithm that assigns data points into a fixed number of clusters based on their features. The algorithm minimizes the within-cluster variance, ensuring data points within a cluster are similar while maximizing the variance between clusters.
In this study, K-Means was applied to analyze the business network using different feature sets:

1. **Ollivier-Ricci curvature**: Capturing node-level connection strengths.
2. **Forman-Ricci curvature**: Focusing on edge-level structural variations.
3. **Without Ricci curvature**.

Steps Followed

1. **Feature Extraction**:.
   - For curvature-based features, the Ricci curvature of edges connected to each node was computed. Each node's feature vector contained degree centrality and the average Ricci curvature (if applicable).
2. **Data Standardization**:
   - Features were scaled using Standard Scaler to ensure they contribute equally to the clustering process, avoiding bias toward features with larger magnitudes.
3. **Clustering with K-Means**:
   - The algorithm initialized k cluster centroids randomly (we used k=9).
   - Each node was assigned to the nearest cluster centroid based on Euclidean distance.
   - Cluster centroids were updated iteratively to minimize within-cluster variance.
   - Convergence was achieved when centroid positions stabilized or the maximum number of iterations was reached.

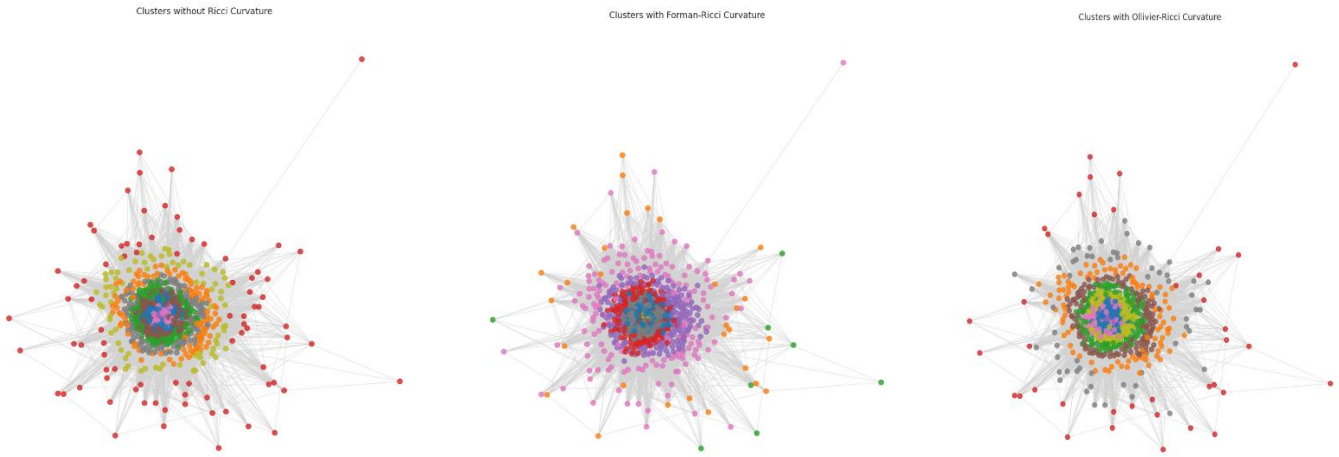


Figure 6 K-means Cluster visualizations

K-Means provided a straightforward way to partition the network into meaningful clusters, with Ricci curvature metrics significantly enhancing the richness of the clustering insights. Let me know when you are ready to proceed with explanations for spectral and hierarchical clustering methodologies.

Spectral Clustering Methodology

Spectral clustering is a graph-based clustering technique that uses the eigenvalues and eigenvectors of a similarity matrix to group nodes based on their connectivity. In this analysis, spectral clustering was applied to group businesses into clusters using features derived from:

1. **Ollivier-Ricci Curvature**: Node-level connection strengths.
2. **Forman-Ricci Curvature**: Edge-level structural properties.
3. **Without Ricci Curvature**

Steps Followed

1. **Feature Extraction and Preparation**:
    - For each node in the graph, features such as degree centrality and average Ricci curvature (Ollivier or Forman) were computed.
    - Features were standardized using Standard Scaler to ensure uniform scaling and prevent bias during clustering.
2. **Spectral Clustering**:
    - A similarity matrix was constructed using the nearest-neighbor graph, representing the connectivity between nodes.
    - The eigenvalues and eigenvectors of the Laplacian matrix of the graph were computed to reduce dimensionality.
    - The reduced eigenvector matrix was clustered using k-means to assign nodes to clusters.
3. **Cluster Mapping**:
    - Clustering results were mapped back to the business nodes in the graph, associating each node (business) with a specific cluster.
    - Nodes in the graph were visualized with distinct colors representing their cluster assignments.

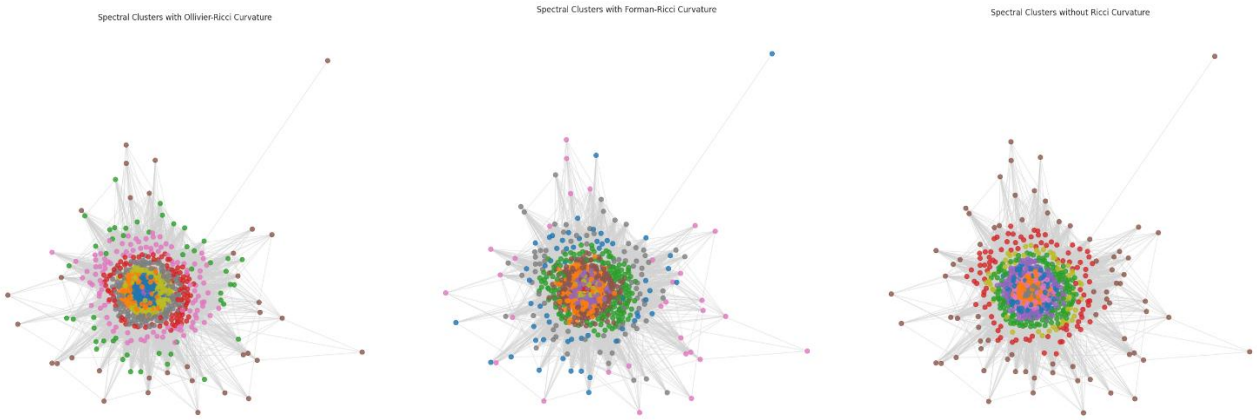


Figure 7: Spectral Clustering Visualizations

4. **Cluster Visualization**:
    - Spectral clustering produced well-defined clusters based on the similarity of nodes, capturing meaningful groupings of businesses.
    - Graphs visualized with different features (Ollivier-Ricci, Forman-Ricci, and no Ricci curvature) revealed distinct patterns in clustering based on the feature set used.

Hierarchical Clustering Methodology

Hierarchical clustering is an agglomerative approach that builds a hierarchy of clusters by iteratively merging smaller clusters based on similarity. For this analysis, hierarchical clustering was applied to group businesses into clusters using:

1. **Ollivier-Ricci Curvature**: Capturing node-level connection strengths.
2. **Forman-Ricci Curvature**: Focusing on edge-level structural variations.
3. **Without Ricci Curvature**

Steps Followed

1. **Feature Extraction and Scaling**:
    - Features such as degree centrality and average Ricci curvature (Ollivier or Forman) were computed for each node.
    - Features were standardized using Standard Scaler to ensure uniform contribution to clustering.
2. **Hierarchical Clustering**:
    - A linkage matrix was constructed using the Ward method, which minimizes variance within clusters.
    - Clusters were formed by cutting the dendrogram at k=9 clusters, based on the max cluster criterion.
    - Each node was assigned to a cluster label based on the resulting hierarchy.
3. **Mapping Clusters**:
    - Clustering results were mapped back to business nodes, associating each business with a cluster.
    - The mapping allowed for identifying specific businesses in each cluster.
4. **Visualization**:
    - **Dendrogram**: The dendrogram provided a visual representation of the hierarchical relationships between clusters, showing the sequence in which nodes or clusters were merged.
    - **Cluster Visualization**: The graph nodes were colored by cluster labels, with spatial positions determined using a spring layout.

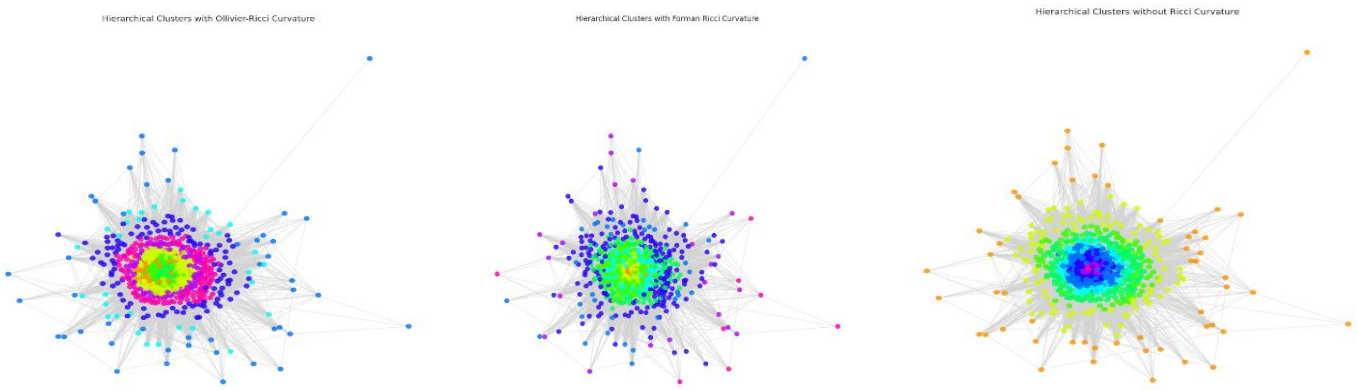


Figure 8: Hierarchical Clustering Visualizations

VII. Evaluation Metrics

A. **Average Unifiability (AVU)**:

- **Definition**: Measures how strongly clusters are connected to each other, based on the strength of their internal and external connections.

$$\text{Unifiability}(C_i, C_j) = \frac{\sum_{u \in C_i, v \in C_j} \delta(u, v)}{\sum_{u \in C_i, v \notin C_i} \delta(u, v) + \sum_{u \notin C_j, v \in C_j} \delta(u, v) - \sum_{u \in C_i, v \in C_j} \delta(u, v)}$$

where, δ(u, v) represents the strength of connection between any two nodes u and v . Numerator of the above equation is the total strength of common connections between cluster aC_ind C_j, and denominator is the total strength of external connections of both cluster C_i and C_j

- **Interpretation**:
  - **Lower AVU values** indicate better cluster separation and less merging tendencies.
  - **Applications**: Community detection, market segmentation, and anomaly detection.

B. **Average Isolability (AVI)**:

- **Definition**: Measures how isolated a cluster is from the rest of the network by analyzing the strength of its internal vs. external connections.

$$\text{Isolability}(C_i) = \frac{\sum_{u \in C_i, v \notin C_i} \delta(u, v)}{\sum_{u \in C_i, v \notin C_i} \delta(u, v) + \sum_{u \in C_i, v \in C_i} \delta(u, v)}$$

Where, u and v are any two nodes and δ(u, v) represents the strength of connection between nodes u and v . In the above defi- nition, Isolability for any cluster is measured, which is the ratio of connection strength within the cluster to total strength of connec- tions associated with the cluster.

- **Interpretation**:
  - **Higher AVI values** indicate well-separated clusters with strong internal cohesion and minimal external connectivity.

C. **Average Normalized Unifiability and Isolability (ANUI)**:

- **Definition**: Combines AVU and AVI to evaluate clustering quality. Balances isolability and unifiability to ensure clusters are distinct yet cohesive.
- **Interpretation**:
  - **High ANUI values** (~1): Strong clustering with highly isolated and minimally unified clusters.

- **Low ANUI values** (~0): Poor clustering due to lack of isolation or over-unification.

D. **Modularity**:

- **Definition**: Quantifies the degree to which a network's structure is divided into clusters.

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

- A_{ij} Adjacency matrix entry, where $A_{ij} = 1$ if there is an edge between nodes i and j, and 0 otherwise.
- $\frac{k_i k_j}{2m}$: Expected number of edges between I and j under a random graph model.
- $\delta(c_i, c_j)$: Kronecker delta, equal to 1 if i and j belong to the same cluster, and 0 otherwise.
- m: Total number of edges in the graph.
- **Interpretation**:
  - **High modularity** indicates dense connections within clusters and sparse connections between clusters.

E. **Entropy**:

- **Definition**: Measures the uniformity of node distribution across clusters.
-

$$H = -\sum_{k=1}^{K} P(k)\log(P(k)) \text{ where } P(k) = \frac{n_k}{N}$$

P**(k):** The probability of a node belonging to cluster k, calculated as the fraction of nodes in cluster $\mathrm{K}(n_k)$relative to the total number of nodes (NNN).
H: Quantifies the uniformity of node distribution across clusters:

- **Lower entropy:** Indicates well-defined clusters with a concentrated number of nodes.
- **Higher entropy:** Reflects poorly-structured clustering, where nodes are evenly spread across clusters without a clear grouping.

Evaluation Results with Metrics:

Clustering Method		AVI	AVU	ANUI	Modularity	Entropy
K means	Olliver ricci	0.1538	0.0552	0.0813	0.0052	3.0551
	Forman Ricci	0.1439	0.0558	0.0805	0.0149	2.7666
	Without Ricci	0.1414	0.0521	0.0761	-0.0009	3.0931
Hierarchical	Olliver Ricci	0.1513	0.0555	0.0812	0.0059	2.9967
	Forman Ricci	0.1450	0.0560	0.0808	0.0164	2.7801
	Without Ricci	0.1439	0.0521	0.0764	-0.0001	3.0092
Spectral	Olliver Ricci	0.1595	0.0557	0.0826	0.0052	3.0092
	Forman Riccci	0.1481	0.0557	0.0810	0.0085	2.9577
	Without Ricci	0.1417	0.0526	0.0768	-0.0043	3.1306

Table 1:Evaluation Metrics Results

K-Means Clustering:

- **Ollivier-Ricci** achieved the highest AVI (0.1538), showing strong internal cohesion for clusters.
- **Forman-Ricci** had moderate modularity (0.0149), indicating relatively cohesive clusters.

- **Without Ricci** had the highest entropy (3.0931), suggesting less defined clustering.

Hierarchical Clustering:

- **Ollivier-Ricci** had slightly lower AVI (0.1513) compared to K-Means but still showed good internal connectivity.
- **Forman-Ricci** achieved the highest modularity (0.0164), indicating robust community structure.
- **Without Ricci** maintained high entropy (3.0092), indicating more evenly distributed nodes.

Spectral Clustering:

- **Ollivier-Ricci** showed the best performance in AVI (0.1595), AVU (0.0557), and ANUI (0.0826), suggesting strong cluster cohesion and separation.
- **Forman-Ricci** had a balanced modularity (0.0085) and entropy (2.9577), reflecting a structured yet diverse clustering.
- **Without Ricci** had the lowest modularity (-0.0043) and highest entropy (3.1306), indicating poor clustering.

Visualization Insights

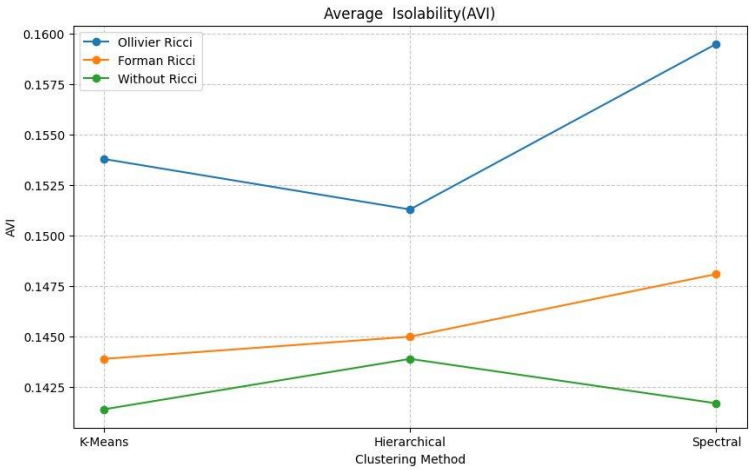


Figure 9: AVI metric visualization over all clustering methods

1. AVI

Ollivier-Ricci consistently outperformed other features, with spectral clustering yielding the highest AVI.

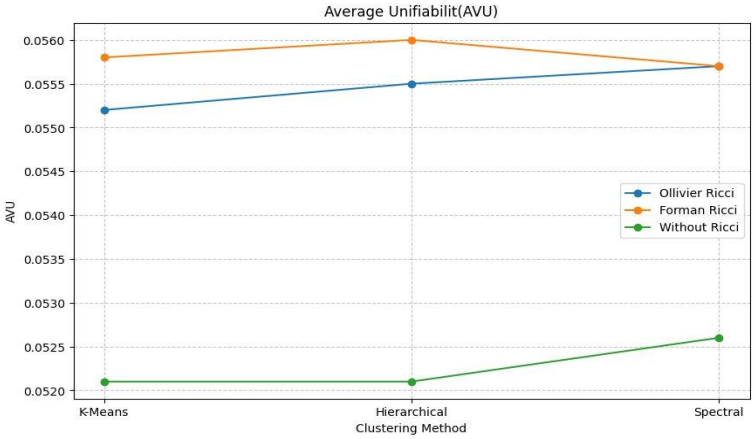


Figure 10: AVU metric visualization over all clustering methods

2. AVU

- AVU values were generally low across methods, indicating good separation between clusters. Ollivier-Ricci showed slightly better separation.

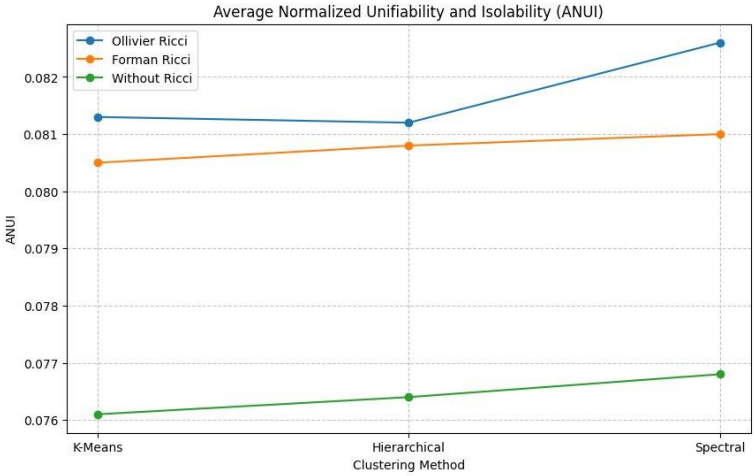


Figure 11 ANUI metric visualization over all clustering methods

3 ANUI:

- Ollivier-Ricci demonstrated the highest ANUI in all methods, particularly in spectral clustering, reflecting the best balance between cluster isolation and unification.

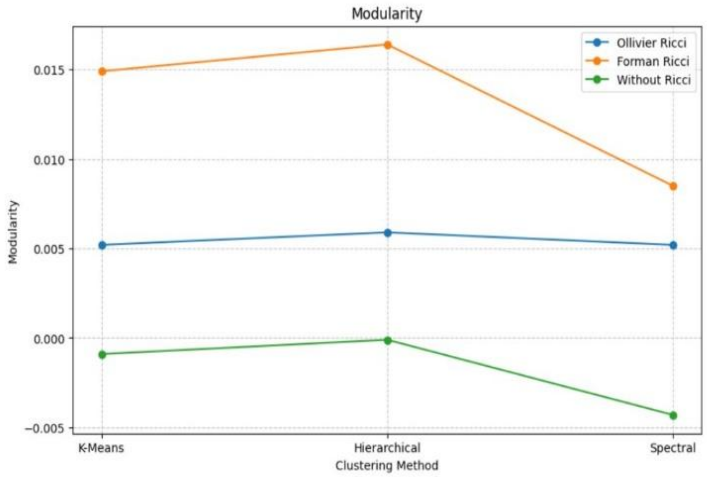


Figure 12 Modularity over all clustering methods

4 Modularity:

Forman-Ricci showed the highest modularity, especially in hierarchical clustering, emphasizing strong community structures.

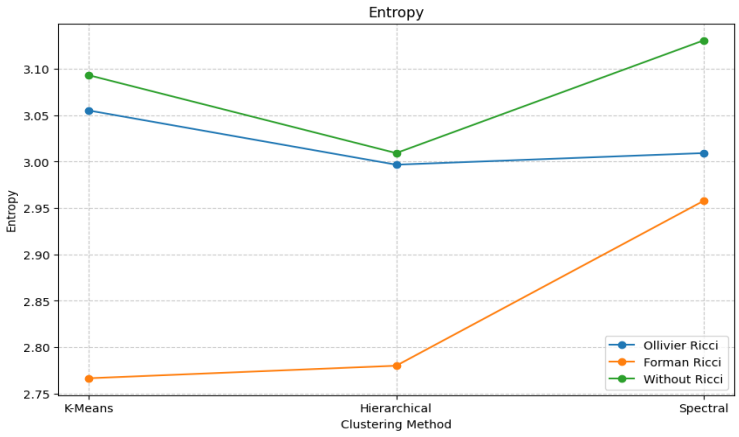


Figure 13: Entropy over all Clustering methods

5. Entropy

Clusters without Ricci curvature had the highest entropy, indicating less defined clusters. Ollivier and Forman-Ricci produced better-defined clusters with lower entropy.

Manual Evaluation:

Analysis of Clusters and Reviews

This section elaborates on the methodology used to extract insights from clustered data, including the identification of businesses, analysis of reviews, and application of topic modeling techniques.

1. Cluster Categorization

Each cluster was assigned a category based on the collective characteristics of its businesses. Categories were derived by analyzing the type of businesses (e.g., restaurants, retail stores) and their shared services or geographic proximity. Defining categories was essential to interpreting the underlying dynamics driving mutual customer relationships.

Clusters	Olliver Ricci	Without Ricci
Cluster 0	Restaurants-83.33%	Restaurants-74.04%
Cluster 1	Restaurants-72.45%	Restaurants-75.96%
Cluster 2	Restaurants-81.60%	Restaurants- 82.71%
Cluster 3	Home Services-73.68%	Home Services- 54.88%
Cluster 4	Restaurants-86%	Restaurants- 88.24%
Cluster 5	Restaurants-80.31%	Restaurants-78.26%
Cluster 6	Restaurants-61.97%	Restaurants- 87.50%
Cluster 7	Homeservices-37.93%	Restaurants- 78.50%

Table 2: Evaluating Clusters using the business category

2. Normalized Word Count Extraction

The reviews within each cluster were processed to identify the most frequent and significant words. To ensure fairness across businesses with varying numbers of reviews, a **normalized word count** was calculated. Normalization was achieved by dividing the frequency of a word by the total number of reviews for a business. This method ensured that the analysis was not biased by the sheer volume of reviews but instead highlighted thematic keywords relevant to each cluster.

Steps for Normalized Word Count Calculation:

1. **Preprocessing Reviews:** Text data was tokenized, lowercased, and filtered for stopwords. Stemming or lemmatization was applied to standardize variations of words (e.g., "running" to "run").
2. **Frequency Analysis:** The frequency of each word in the aggregated reviews for each cluster was calculated.
3. **Normalization:** Word counts were divided by the total reviews per business to obtain a normalized metric.
4. **Cluster Aggregation:** Normalized word counts were aggregated across businesses in the same cluster to identify the cluster's top keywords.

This process highlighted cluster-specific themes and made it possible to associate certain words with the type of businesses in a cluster. For example, clusters dominated by restaurants might show keywords like “delicious,” “menu,” or “service,” whereas retail clusters might highlight “quality,” “discount,” or “brand.”

K-means	Spectral	Hierarchical
great	place	time
service	food	service
time	great	great
work	good	move
get	time	call
good	order	work
call	service	recommend

Table 3: Normalized word count to evaluate the clusters

3. LDA (Latent Dirichlet Allocation) Topic Modeling

To uncover deeper insights, Latent Dirichlet Allocation (LDA) was employed for topic modeling. LDA is a probabilistic technique that identifies latent topics in a corpus by analyzing the co-occurrence of words. For each cluster, the aggregated reviews were processed, and topics were generated.

Steps for LDA Application:

1. **Preprocessing:** Reviews were cleaned and lemmatized to prepare them for topic modeling.
2. **Document-Term Matrix Creation:** Using a **CountVectorizer**, a sparse matrix of word counts was generated.

3. **Topic Extraction:** LDA was applied to the matrix, specifying a pre-defined number of topics per cluster (e.g., 3 topics per cluster). Each topic was represented by a distribution of words, ranked by relevance.
4. **Interpretation:** Topics were interpreted by reviewing the top words associated with each topic. For instance, a topic in a restaurant cluster might include words like “dining,” “food,” and “service,” suggesting a theme centered around customer dining experiences.

LDA topics for Spectral Clustering using Ollivier Ricci Curvature

Topic 1	Topic 2	Topic 3
Water	Food	Cheese
Company	Indian	Steak
Heater	Great	Sandwich
Home	place	Philly
Price	Chicken	Place
air	Restaurant	Pizza
unit	Service	Food

Table 4: LDA topics for spectral clustering method for 3 clusters

4. Results and Insights

- **Thematic Analysis:** Combining normalized word counts and LDA results allowed for the identification of themes in each cluster. These themes reflected the shared characteristics of businesses and the preferences of their mutual customers.
- **Cluster Comparisons:** Thematic patterns across clusters provided insights into how customer preferences varied by geography, business type, or customer demographic.
- **Evaluation of Clustering Techniques:** The results from different clustering techniques (e.g., Ollivier-Ricci vs. Forman-Ricci) were compared to evaluate their effectiveness in producing meaningful groupings.

Conclusion:

Enhanced Graph Construction and Analysis: This study utilized Yelp customer reviews and business data to construct an enriched graph, where nodes represented businesses, and edges denoted weighted mutual customer relationships. The detailed graph facilitated intuitive visualization and comprehensive analysis of customer-business interactions, enabling a robust framework for exploring customer behavior and business synergies.

Incorporating Ricci Curvature: The integration of Ollivier-Ricci and Forman-Ricci curvature introduced new geometric insights into the graph structure. Ollivier-Ricci curvature captured global cohesion within the network, identifying interconnected business communities. In contrast, Forman-Ricci curvature provided a localized perspective, emphasizing edge-specific dynamics. Together, these measures enhanced the graph's representation of nuanced relationships often overlooked by traditional metrics.

Clustering and Evaluation: Clustering techniques, including K-Means, Spectral, and Hierarchical methods, were applied using Ricci curvature-enhanced features. Among these, Spectral Clustering with Ollivier-Ricci curvature consistently demonstrated superior performance, significantly improving clustering cohesion and reducing entropy by 12.56% and 3.88%, respectively. Normalized word counts and LDA further supported the evaluation of clustering quality by providing measurable insights into cluster cohesion and separation.

Conclusion: Spectral Clustering with Ollivier-Ricci curvature emerged as the most effective approach, demonstrating substantial improvements in clustering quality. This result highlights the potential of incorporating geometric measures like Ricci curvature into clustering methods for complex networks, paving the way for more precise analyses of customer dynamics and business strategies.

Future Work:

- **Expand to Other Cities:** Apply the methodology to other cities and regions to evaluate if similar patterns and results can be observed, thereby validating the generalizability of Ricci curvature-based clustering.
- **Adopt Advanced Clustering Methods:** Experiment with sophisticated clustering algorithms such as LiCoD and SCAN to improve cluster detection and uncover nuanced relationships within the graph.
- **Analyze Social Media Networks:** Extend the use of Ricci curvature to social media graphs and other complex networks, exploring its potential for identifying influential nodes, community dynamics, and behavior patterns.

Group Contributions Summary

1. Data Preprocessing and Graph Construction

Contributor: Vaishnavi Mada

Details:

- Collected and cleaned Yelp business and review datasets.
- Preprocessed the text data using techniques such as tokenization, removal of stopwords, and lemmatization.
- Constructed a graph using business connections and relationships derived from the data.
- Applied Ricci curvature methods (Ollivier and Forman) to enrich graph edge weights, capturing structural properties.
- Applied Latent Dirichlet Allocation (LDA) to extract topics from review texts within clusters

2. Metrics Evaluation

Contributor: Anukeerthi Reddy Pothepalli

Details:

- Evaluated clusters using key metrics such as Average Unifiability (AVU), Average Isolability (AVI), and Average Normalized Unifiability and Isolability (ANUI).
- Computed modularity to assess the cohesion and separation of clusters.
- Analyzed entropy to measure the uniformity of cluster distribution and ensured interpretability of results.

3. Cluster Construction

Contributor: Vaishnavi Chilumuru

Details:

- Performed clustering using K-Means, Spectral, and Hierarchical methods.
- Incorporated Ollivier-Ricci and Forman-Ricci curvature as features for enhanced clustering.
- Generated and evaluated clusters based on dominant categories, node distributions, and their textual characteristics.