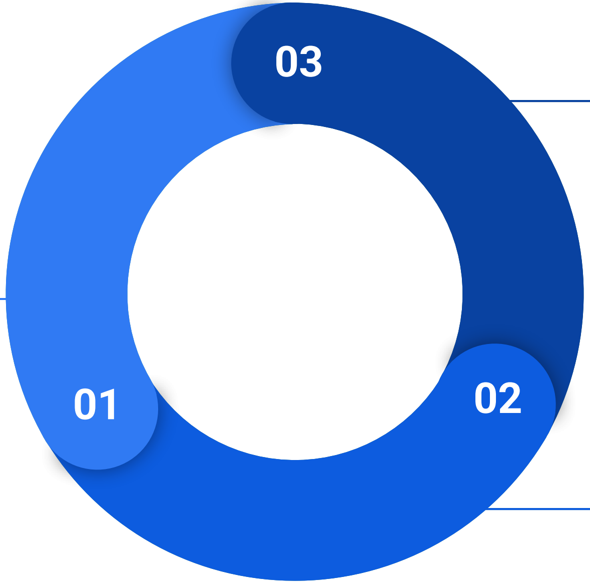
NETWORK ANALYSIS

ACSP - Senior Data Analyst - Stage Seven

INTRODUCTION

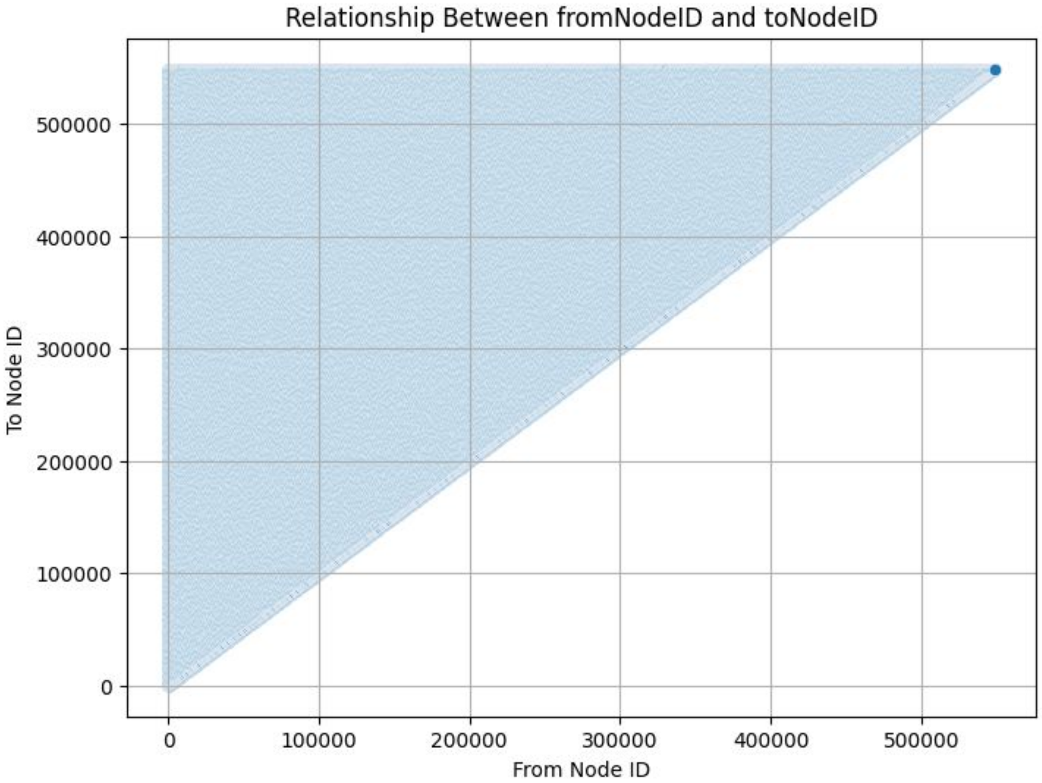
Dataset Overview
Analysis Objective
Initial Exploratory Data Analysis



Conclusion and Relevance

Research Questions and Findings

1. What are the largest connected components, and how fragmented is the network?
2. How does the degree centrality score vary across different nodes?
3. How does clustering coefficient distribution vary across nodes, and what does it reveal about local connectivity?
4. What is the redundancy of connections, and how does it affect robustness?
5. Is there a relationship between community size and density in a product co-purchase network?
6. how accurate is a machine learning model trained on network features for link prediction?



INTRODUCTION

- ★ **Dataset Overview**: The dataset is sourced from [Amazon]
- ★ **Analysis Objective**: The goal of the analysis is to explore the relationship between FromNodeId and ToNodeId and **uncover patterns in the data through a network analysis**.
- ★ The scatter plot indicates a **non-linear relationship between the two variables**, with some areas showing more concentrated node connections.

INTRODUCTION

- **Data Loading and Cleaning:**
  - The dataset is first uncompressed and loaded into a Pandas DataFrame.
  - No duplicate entries or missing values were found in the dataset.
- **Dataset Shape:**
  - The dataset consists of 925,872 rows and 2 columns.
- **Descriptive Statistics:** As shown in the table;
  - A mean value of approximately 185,663 for FromNodeId and 368,949 for ToNodeId
  - The range of node IDs spans from 1 to 548,411 for FromNodeId and 366 to 548,551 for ToNodeId
- **Unique Node Count: 265,933** unique FromNodeIds and **264,147** unique ToNodeIds.

	FromNodeId	ToNodeId
count	925872.000000	925872.000000
mean	185662.827571	368949.221348
std	133061.964633	132601.362414
min	1.000000	366.000000
25%	73349.000000	273732.000000
50%	162058.000000	392319.000000
75%	277243.000000	482703.000000
max	548411.000000	548551.000000

RESEARCH QUESTION 1

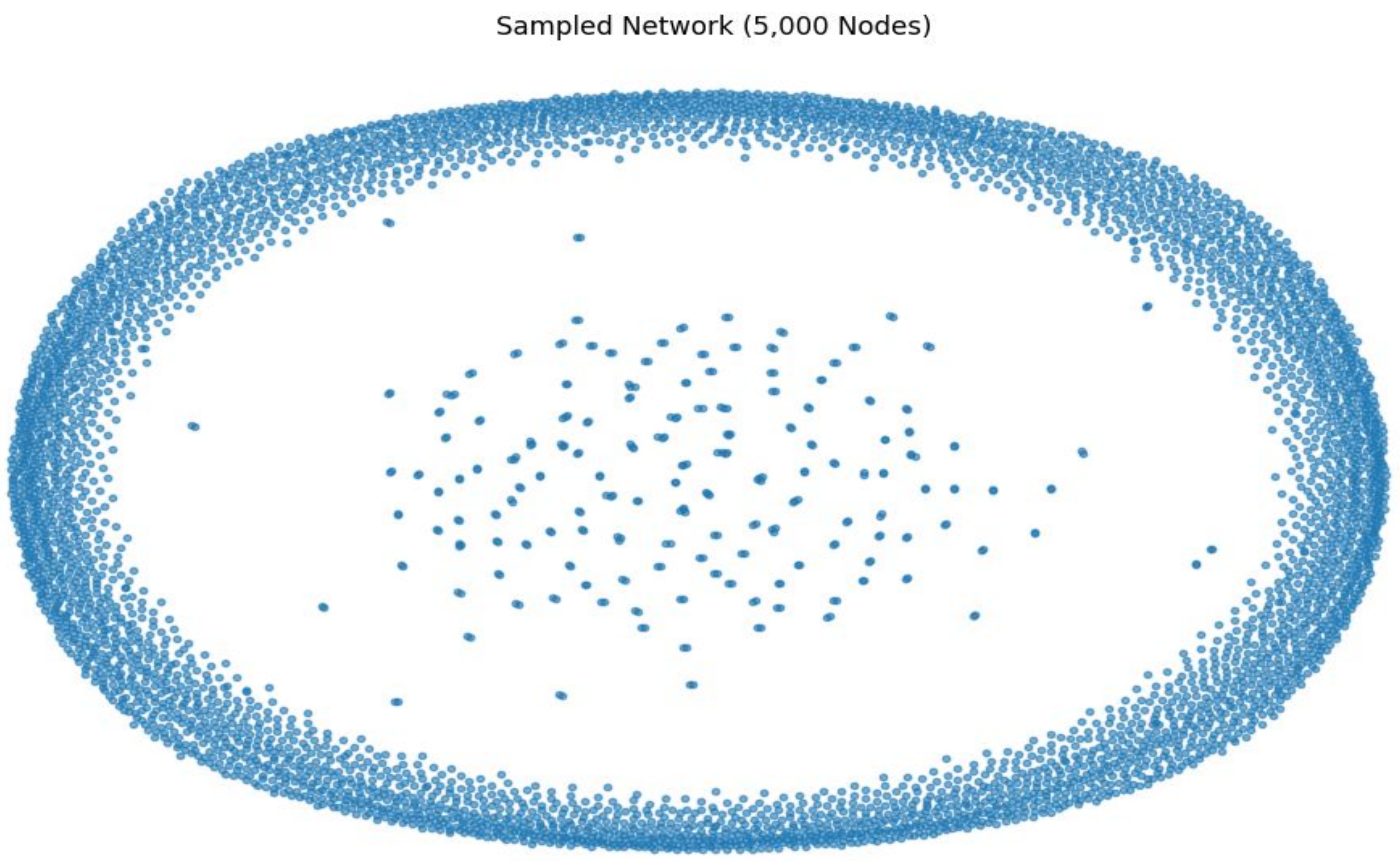
Number of connected components	1
Largest component size	334863
Network fragmentation index	0.00
The network is fully connected	True
Average component size	334863.00
Nodes	334863
Edges	925872
Network Density	0.000017
Average Clustering Coefficient	0.3967

- **Research Question**
  - **What are the largest connected components, and how fragmented is the network?**
  - This analysis aims to determine the extent of connectivity within the network by examining its largest connected components, fragmentation, and overall structure.
- **Findings**
  - **Network Connectivity:** The network is **fully connected** with **one single connected component**, meaning all nodes are reachable from any other node.
  - There is **no fragmentation**, as indicated by a **fragmentation index of 0.00.**

RESEARCH QUESTION 1

- **Findings**
  - **Largest Connected Component:**
    - The **largest component** consists of **334,863 nodes**, which represents the entire network.
    - The **average component size** is also **334,863 nodes**, confirming that all nodes belong to one major component.
  - **Network Structure & Properties:**
    - The network contains **334,863 nodes** and **925,872 edges**
    - The **network density** is **0.000017**, indicating a sparse graph where most nodes are not directly connected but still belong to the same network.
    - The **average clustering coefficient** is **0.3967**, suggesting a moderate level of local connectivity among nodes.
  - The **network sampled subgraph** of 5,000 nodes extracted **below**, shows clusters of nodes with varying connectivity, reflecting real-world patterns of interaction.

RESEARCH QUESTION 1



RESEARCH QUESTION 2

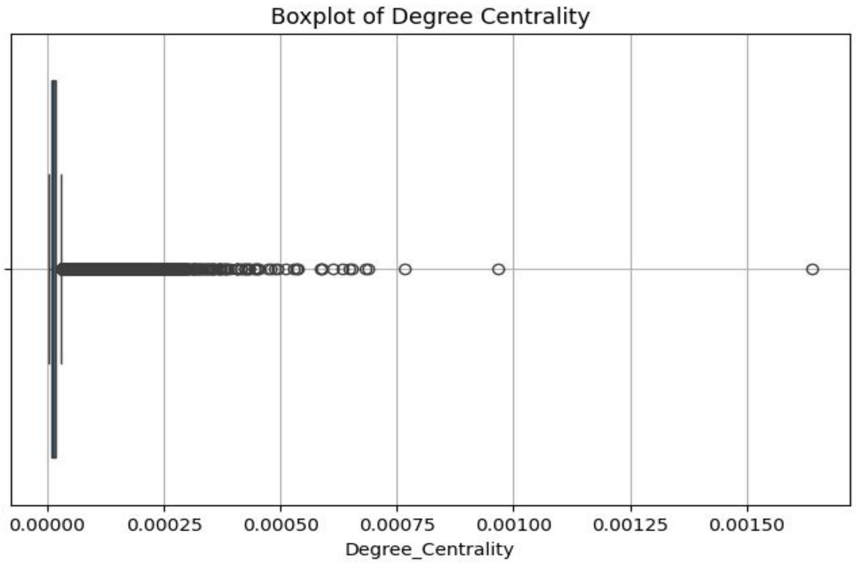
	NodeId	Degree_Centrality
count	334863.000000	334863.000000
mean	276768.565727	0.000017
std	159927.553896	0.000017
min	1.000000	0.000003
25%	138028.000000	0.000009
50%	276405.000000	0.000012
75%	415626.500000	0.000018
max	548551.000000	0.001639

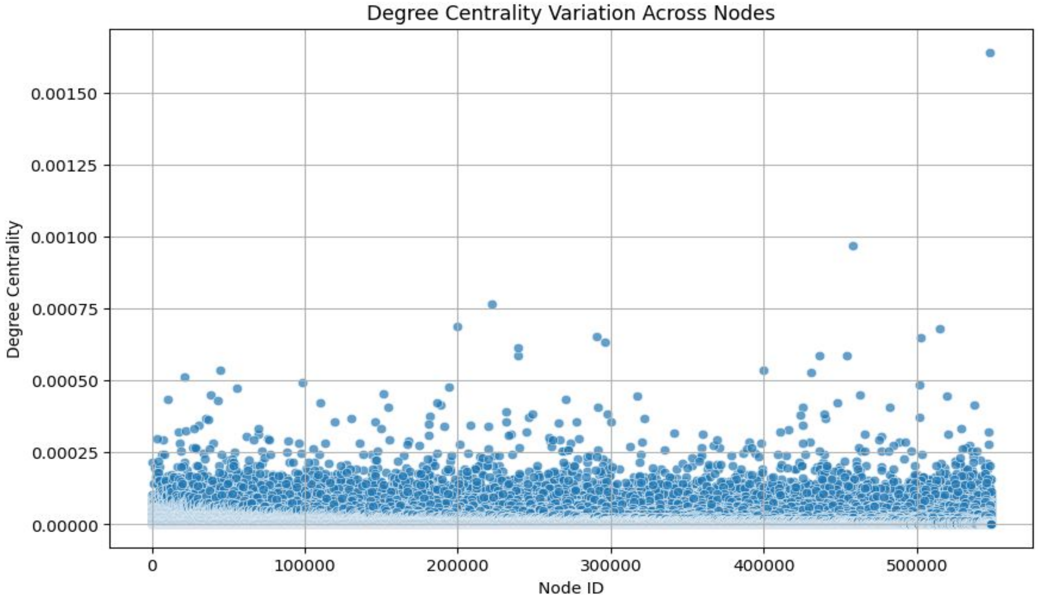
- ★ **Research Question Two**
  - **How does the degree centrality score vary across different nodes?**
  - This analysis examines how central various nodes are in the network based on their degree centrality, which measures the number of direct connections a node has relative to the total nodes.

- ★ **Findings**
  - **Degree Centrality Overview:**
    - The **mean degree centrality** is **0.000017**, indicating that most nodes have relatively low connectivity.
    - The **highest degree centrality** observed is **0.001639**, suggesting a few highly connected nodes

RESEARCH QUESTION 2

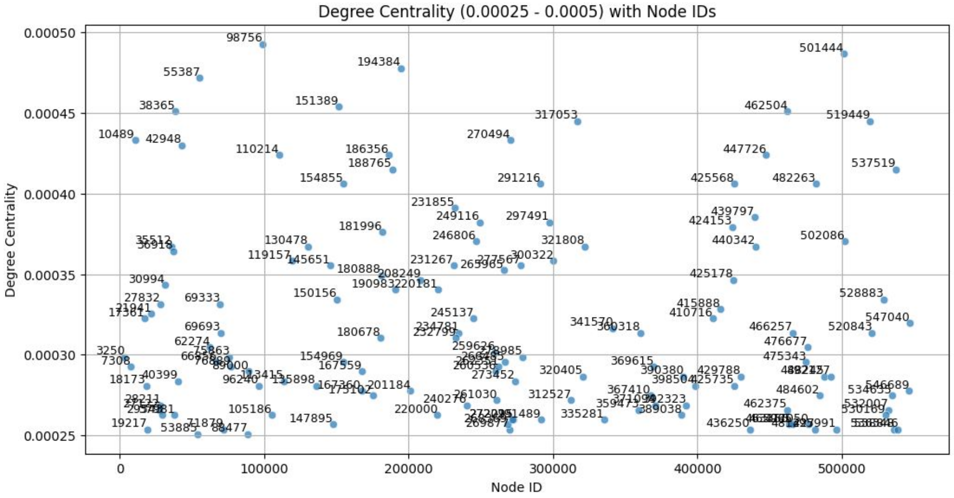
- **Findings**
  - **Degree Centrality Distribution:**
    - The **scatter plot** distribution below is **highly skewed**, with most nodes having low centrality and only a few nodes acting as hubs.
    - The **box plot** confirms that it is **positively skewed**, it suggests that while most nodes have low degree centrality, a few nodes (outliers) have much higher connectivity.

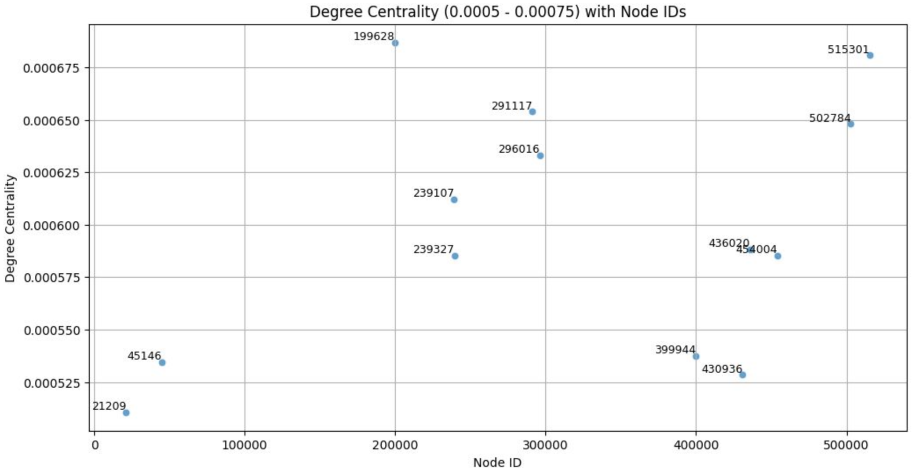


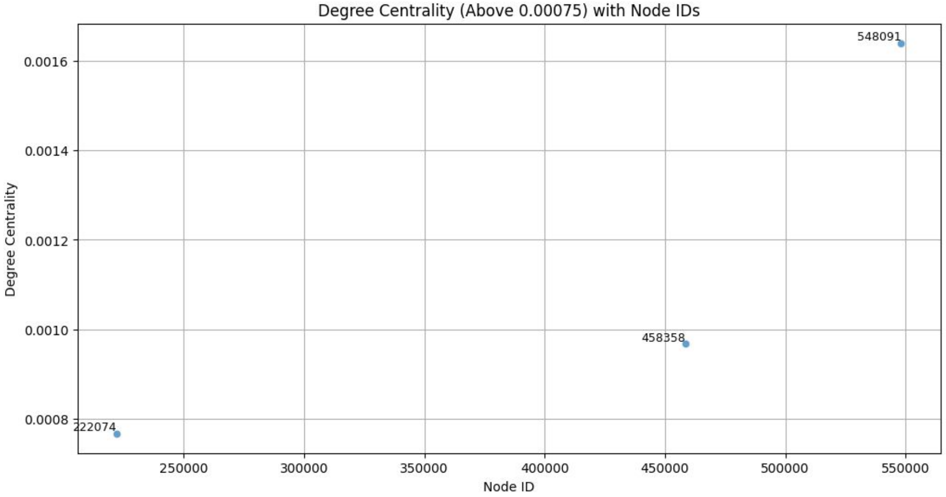


RESEARCH QUESTION 2

- ★ **Segmenting Nodes by Centrality:** Each category was visualized in separate scatter plots, with node IDs labeled.
- ★ Above 0.00075 - highly connected nodes, key hubs.
- ★ Between 0.0005 - 0.00075 - Moderately central nodes with above-average connectivity.
- ★ Between 0.00025 - 0.0005 - Lower-tier nodes with some level of connectivity.
- ★ The 0.00000 - 0.00025 - the lowest-tier nodes, characterized by very minimal connectivity

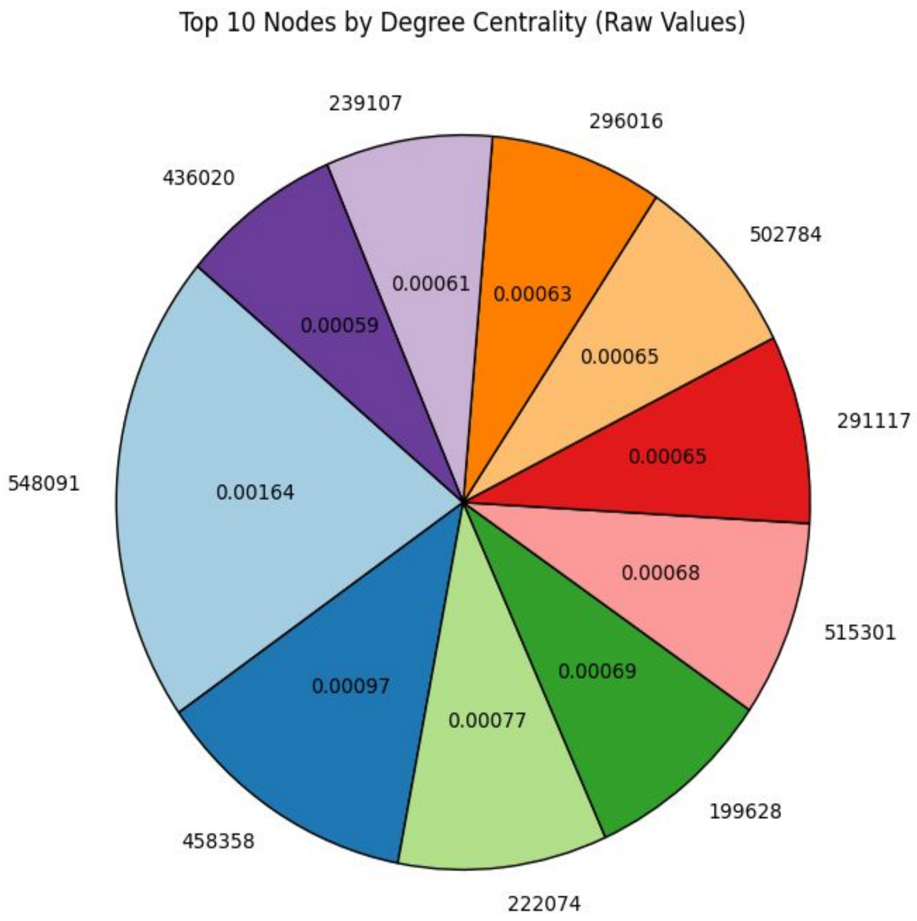






RESEARCH QUESTION 2

- ★ Highest Centrality Node: The **node 548091** has the **highest degree centrality** (0.00164)
  - it is the most connected node in the network.
  - It acts as a key hub in the structure, possibly bridging many other nodes.
- ★ Other High Centrality Nodes:
  - Nodes **458358 (0.00097)** and **222074 (0.00077)** are highly connected nodes, representing key hubs
- ★ If these high-centrality nodes were removed, it could significantly impact the connectivity of other nodes.



RESEARCH QUESTION 3

	NodeId	Clustering_Coefficient
count	334863.000000	334863.000000
mean	276768.565727	0.396746
std	159927.553896	0.329530
min	1.000000	0.000000
25%	138028.000000	0.100000
50%	276405.000000	0.333333
75%	415626.500000	0.666667
max	548551.000000	1.000000

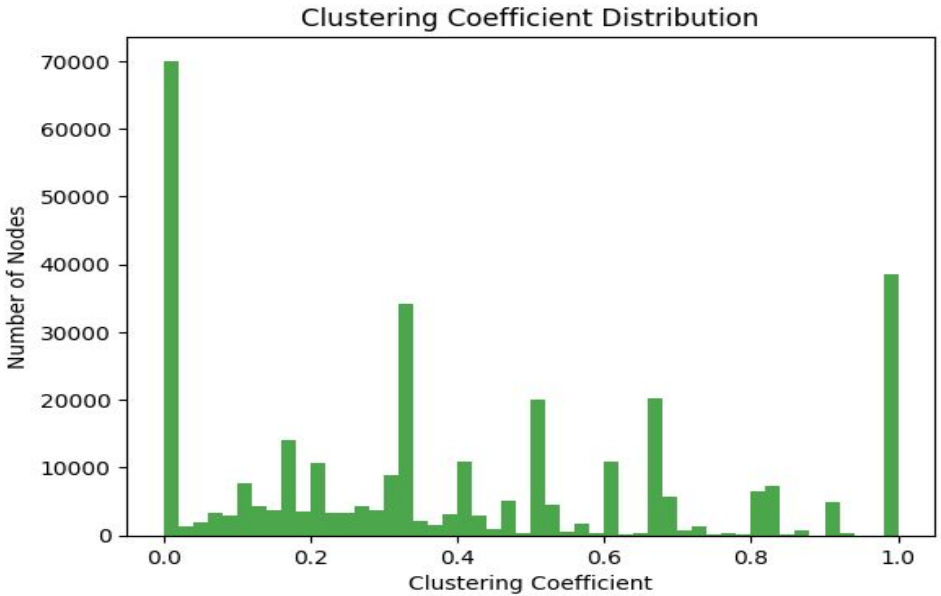
- **Research Question Three**
  - **How does clustering coefficient distribution vary across nodes, and what does it reveal about local connectivity?**
  - This analysis examines examine the distribution of clustering coefficients among nodes and reveals local connectivity patterns in the network

- **Findings**
  - **Clustering Coefficient Distribution:**
    - Mean: 0.3967 → Nodes, on average, exhibit moderate clustering.
    - Std Dev: 0.3295 → High variability in local connectivity.
    - Min: 0.0000 → Some nodes have no local clustering
    - Max: 1.0000 → Some nodes are part of fully interconnected local groups

RESEARCH QUESTION 3

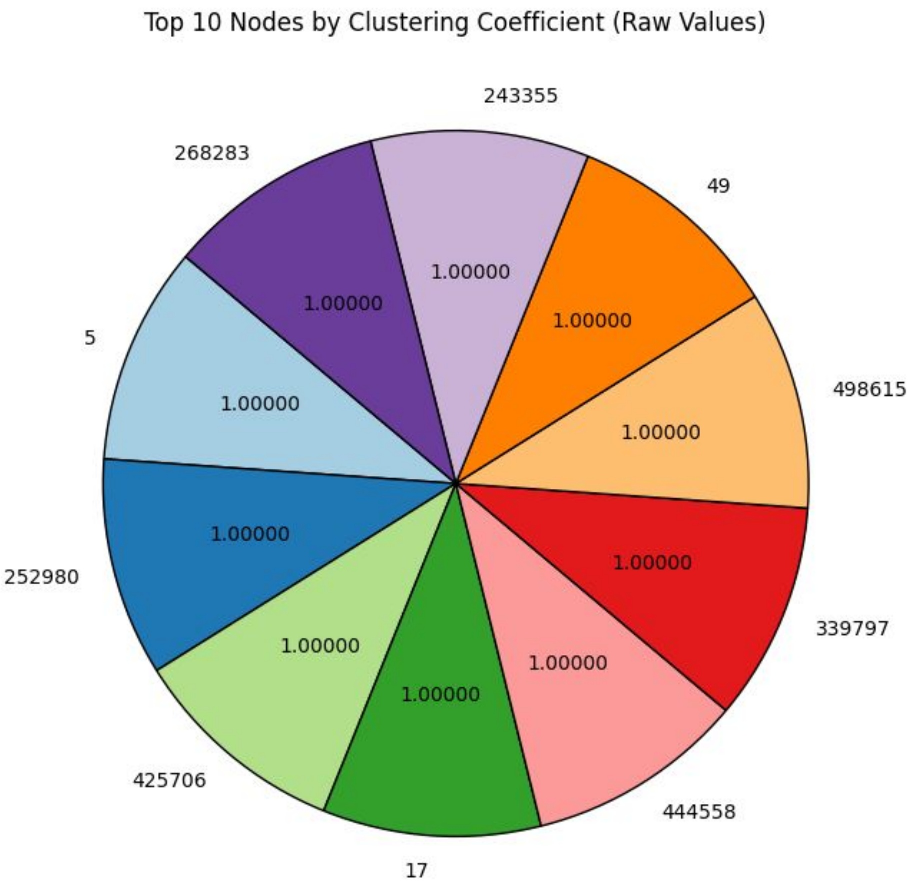
- **Skewed Distribution with Distinct Peaks**
  - A large proportion of nodes (~70,000) have a clustering coefficient of 0, meaning they are poorly connected locally and likely serve as bridges between different communities
  - Another significant group has a clustering coefficient of 1, indicating they belong to highly cohesive clusters.
  - Distinct peaks around 0.33, 0.5, and 0.67 suggest structured connectivity patterns, possibly reflecting hierarchical or modular organization within the network.

- **Implications on Local Connectivity**
- Nodes with Clustering Coefficient = 0: Act as **bridges** between communities, linking otherwise disconnected regions
- Nodes with High Clustering (0.67 - 1.0): Form **highly interconnected local groups**, likely representing **dense sub-networks** or **strongly connected communities.**



RESEARCH QUESTION 3

- ★ The equal distribution in the pie chart confirms that multiple nodes share the same maximum clustering coefficient.
- ★ All top 10 nodes have a clustering coefficient of 1.00000 (fully clustered).
- ★ These nodes represent products that are frequently purchased together within tightly knit groups.
- ★ The highly clustered nodes indicate strong co-purchasing behavior within niche product groups.



RESEARCH QUESTION 4

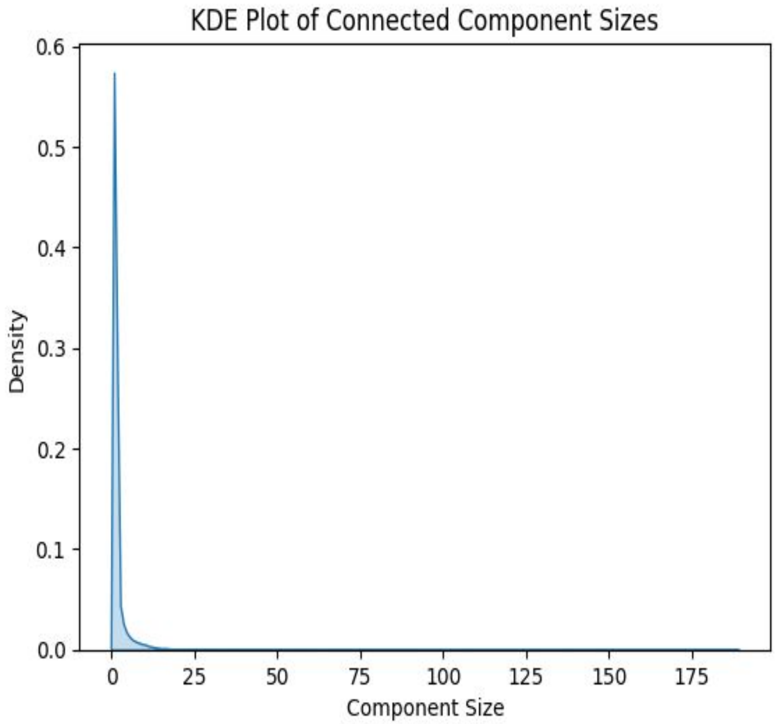
10% of Highest-Degree Nodes removal

Number of connected components	47453
Largest component size	210170
Network fragmentation index	0.30
The network is fully connected	False
Average component size	6.35
Nodes	301377
Edges	409930
Network Density	0.000009
Average Clustering Coefficient	0.2660

- **Research Question**
  - **What is the redundancy of connections, and how does it affect robustness?**
  - This analysis aims to analyze the impact of connection redundancy on the structural robustness of Amazon's co-purchase network
- **Attack Scenarios One Key Insights**
  - **Removing 10% of Highest-Degree Nodes (Hub Nodes)**
  - **Significantly disrupts the network** – The number of connected components increases dramatically **(47,453 components)**.
  - **The largest connected component shrinks** – Drops to **210,170 nodes**, meaning **30% of the network is fragmented.**
  - **Average component size is very small (6.35 nodes per component)**, indicating **severe network disintegration**.
  - **Network density drops to 0.000009**, showing **reduced connectivity.**
  - **Average clustering coefficient decreases to 0.2660, suggesting weaker local structure.**
  - **Hub nodes are crucial for connectivity; their removal severely fragments the network.**

RESEARCH QUESTION 4

10% of Highest-Degree Nodes removal component size kde plot without the largest component



- ★ **Attack Scenarios One component sizes Key Insights**
  - **Removing 10% of Highest-Degree Nodes (Hub Nodes)**
  - The KDE plot shows that most components are very small, clustered near 1.
  - The component size ranges from 1 to less than 200, but there is a major jump to 210,170 in size.
  - This suggests that the majority of nodes become isolated single-node components after the attack.
  - The network loses global connectivity, but a large connected subgraph still survives.
  - The remaining giant component suggests that some parts of the network were resilient due to redundancy.
  - The Network graph of the medium component size (btw 100 to 200 nodes) indicate that;
    - Moderate Complexity in Connectivity: The network structure is fairly spread out, with clear branching patterns.
    - Some sections are highly connected, while others appear more linear or tree-like.
    - Few densely packed clusters suggest localized connectivity rather than a centralized core.

RESEARCH QUESTION 4

- Lack of a Strong Core: The structure indicates that these medium-sized components lack a dominant hub.
- Vulnerability to Further Fragmentation: Many of the nodes seem to be connected in a tree-like fashion, meaning removing a key link could easily isolate parts of the network.
- Resilience within Subgroups: The presence of localized dense clusters within the structure suggests some redundancy in connections.

RESEARCH QUESTION 4

10% Ordinary Nodes removal

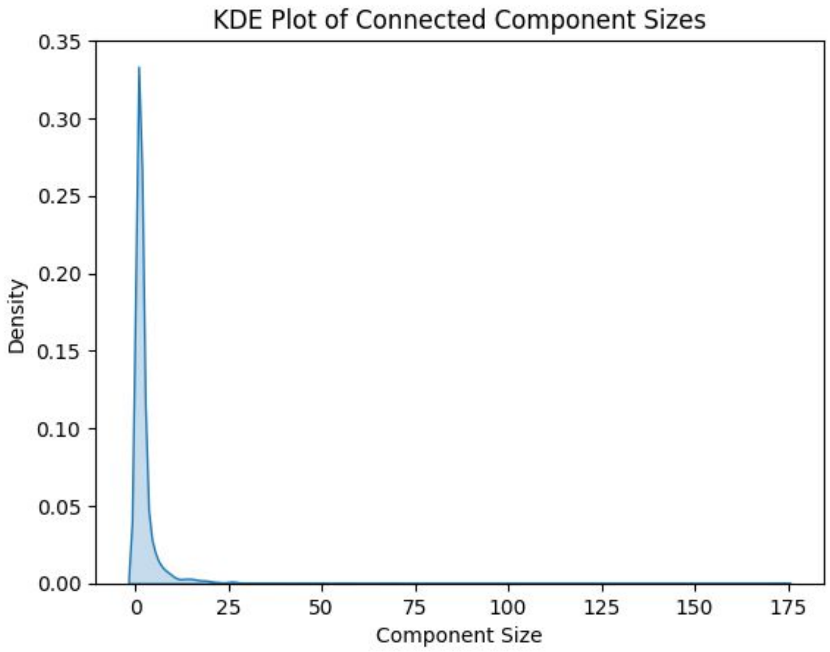
Number of connected components	3824
Largest component size	292413
Network fragmentation index	0.03
The network is fully connected	False
Average component size	78.81
Nodes	301377
Edges	749882
Network Density	0.000017
Average Clustering Coefficient	0.3852

★ **Attack Scenarios Two Key Insights**

- **Removing 10% of Random Nodes (Ordinary Nodes).**
- **Network remains much more intact** – Only **3,824 components**, compared to 47,453 in the first scenario..
- **Largest connected component remains large (292,413 nodes),** with only **3% fragmentation**.
- **Average component size is significantly higher (78.81 nodes per component),** indicating **minimal disruption.**
- **Network density drops to 0.000009**, showing **reduced connectivity.**
- **Network density is higher (0.000017)** compared to the first case.
- **Average clustering coefficient increases to 0.3852**, meaning **local connectivity remains strong**.
- Ordinary nodes contribute less to overall network structure; their removal has minimal impact on connectivity.

RESEARCH QUESTION 4

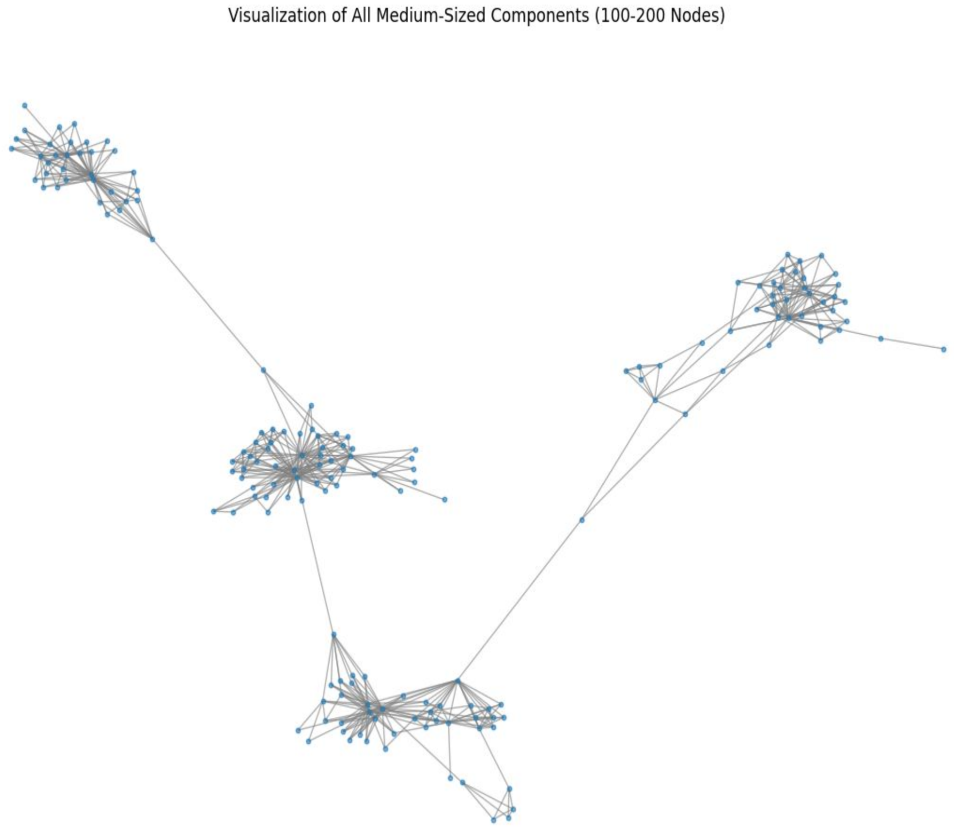
10% Ordinary Nodes removal component size kde plot without the largest component

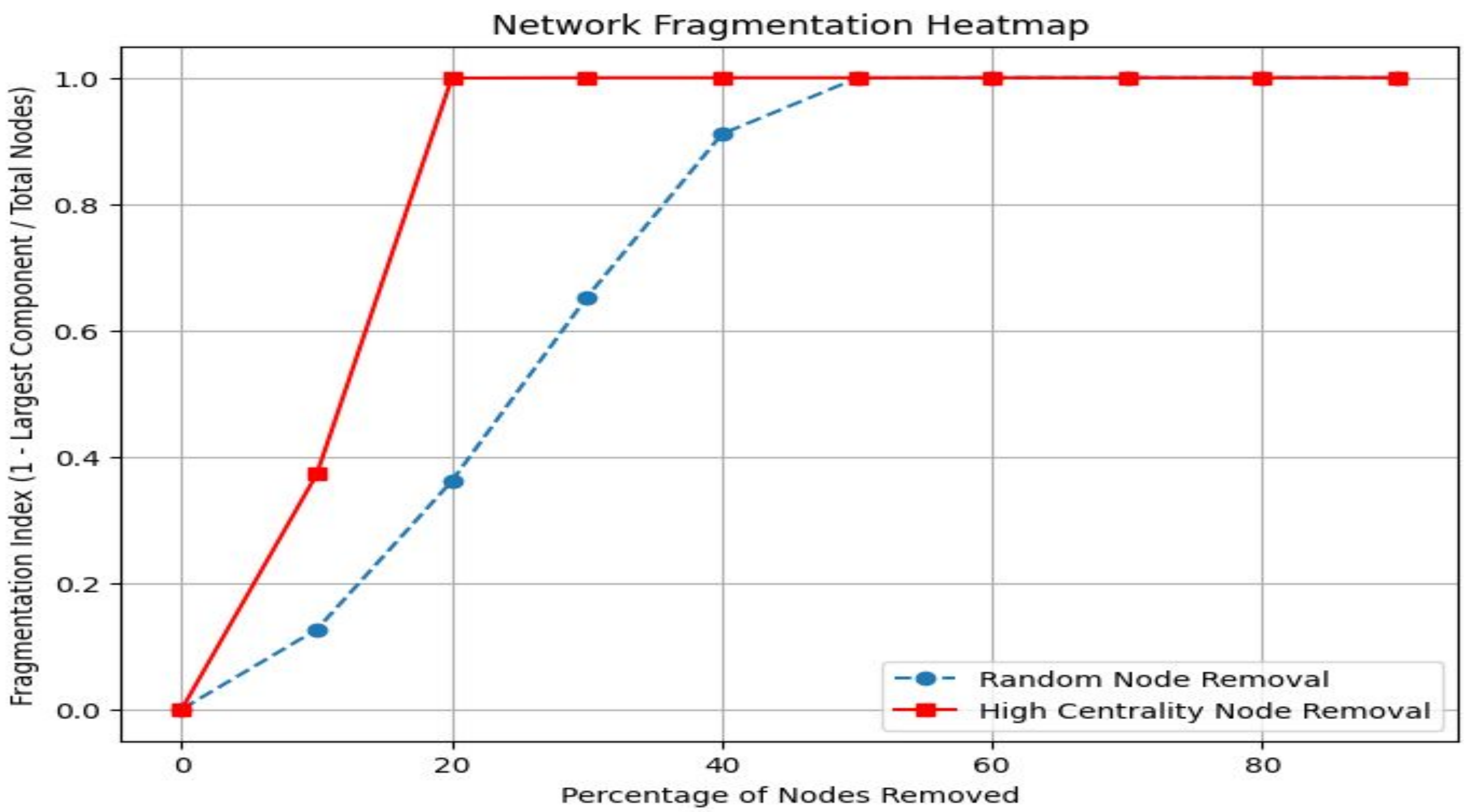


- **Attack Scenarios One component sizes Key Insights**
  - **Removing 10% Ordinary Nodes**
  - The majority of components are very small, mostly between 1 and 5 nodes.
  - The density declines rapidly as component size increases, showing that larger components are rare.
  - while most of the network fragments into tiny components after the attack, a single massive connected component size of 292413 remains intact, likely representing a core resilient structure of the network.
  - The remaining giant component suggests that some parts of the network were resilient due to redundancy.
  - The Network graph of the medium component size (btw 100 to 200 nodes) indicate that;
    - Fragmentation into Distinct Clusters: removing 10% of ordinary nodes successfully broke larger structures into smaller sub-networks but did not fully disintegrate the network

RESEARCH QUESTION 4

- Presence of Bridges: Some components are loosely connected by thin links, meaning a few critical nodes still serve as bridges between sub-communities.
- Localized Connectivity Within Components: removing only ordinary nodes does not disrupt the network's core connectivity, but only isolates cluster.
- Implications for Network Resilience: The main network structure still persists, albeit in fragmented pieces.

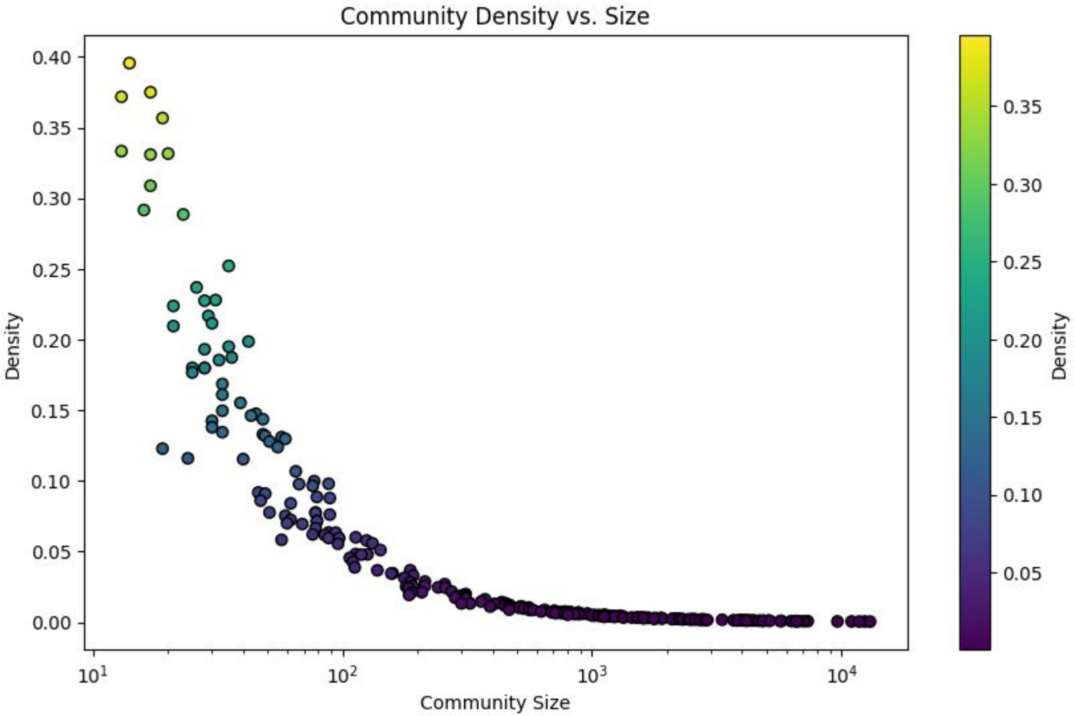



Network Fragmentation Heatmap
Fragmentation Index (1 - Largest Component / Total Nodes)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Percentage of Nodes Removed
Random Node Removal
High Centrality Node Removal

RESEARCH QUESTION 5

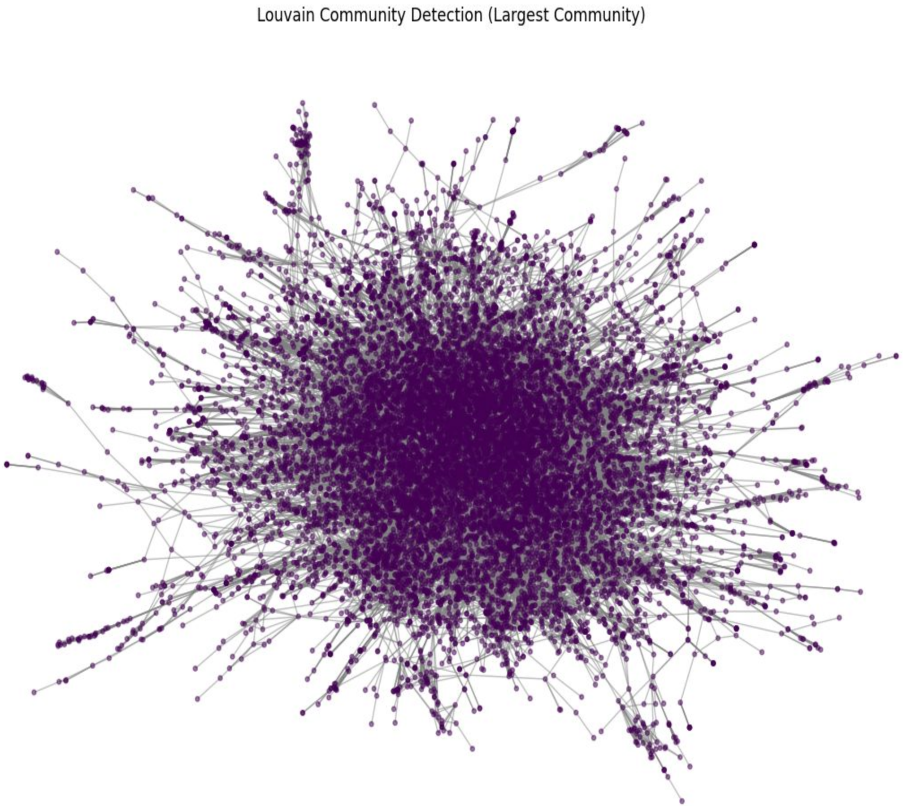
- **Research Question:** Is there a relationship between community size and density in a product co-purchase network?
  - This research aims analyze the relationship between community size and density in a product co-purchase network using community detection techniques

- Louvain Community Detection was applied to identify 252 communities in the network.
- Density vs. Size Relationship:
  - Smaller communities have higher density (up to 0.4).
  - Larger communities have significantly lower density (approaching 0.01)

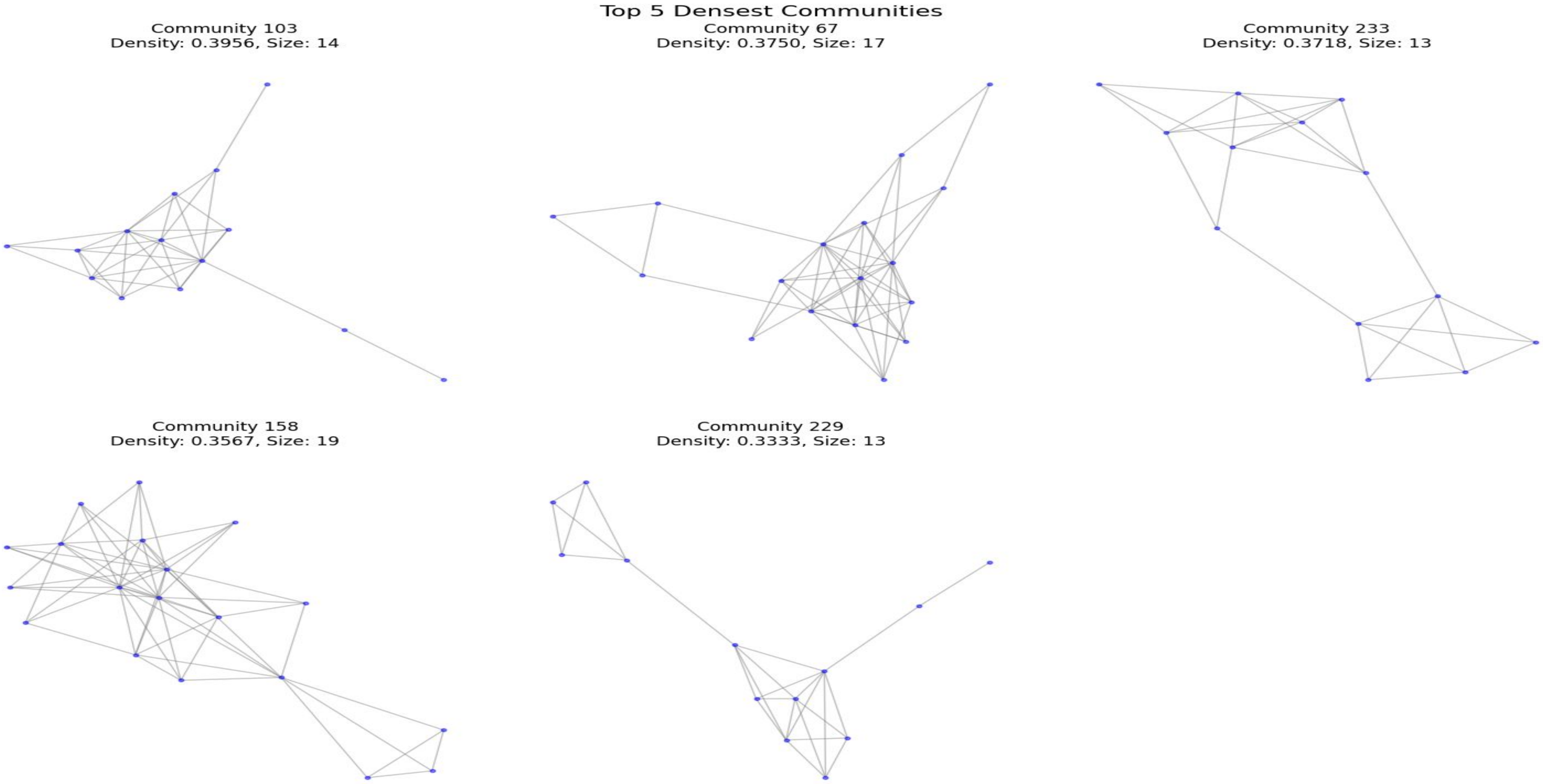


RESEARCH QUESTION 5

- ★ A negative correlation is observed: As community size increases, density decreases, indicating that larger communities are sparser in terms of connections.
- ★ The log-log scale satter plot visualization above confirms this trend.
- ★ The Network graph represent largest community contains 13,041 nodes, showing the presence of dominant clusters.
- ★ The top-five dense community are given in the plot below



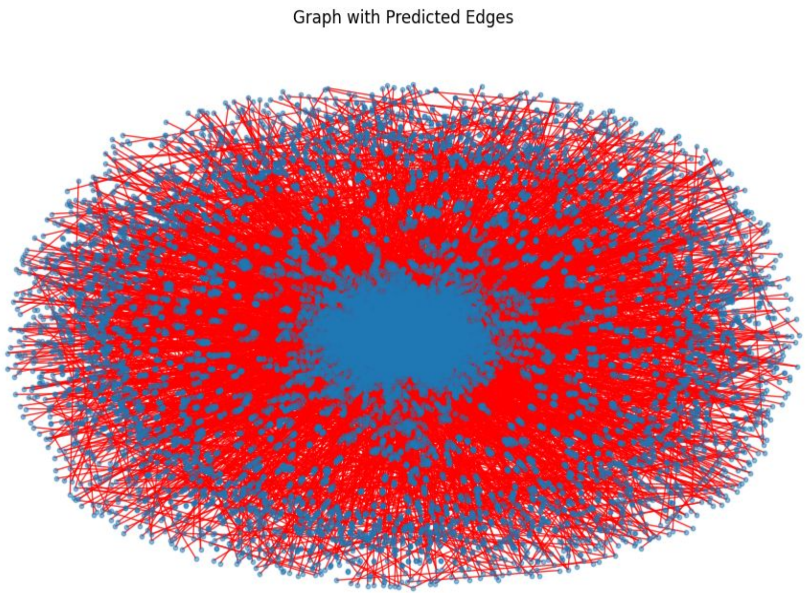
Research Question 5



RESEARCH QUESTION 6

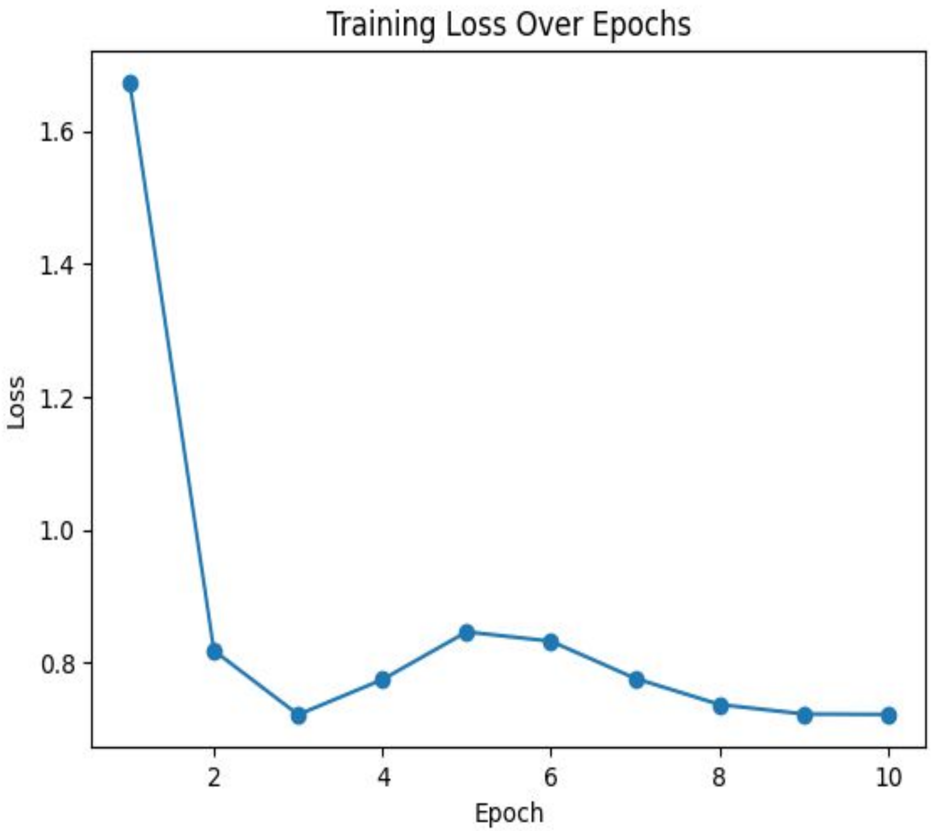
- **Research Question:** HOW ACCURATE IS A MACHINE LEARNING MODEL TRAINED ON NETWORK FEATURES FOR LINK PREDICTION?
  - The aim of your research question is to evaluate the effectiveness of a machine learning model trained on network features for predicting links (connections) in a graph.

- Dataset and Graph Construction:
  - Used a real-world graph dataset with 20,000 high-degree sample nodes for better structure representation.
  - Created an induced subgraph to retain high-connectivity nodes
- Feature Engineering & Preprocessing:
  - Degree centrality was used to filter high-degree nodes.
  - Dummy node features were initialized to ensure compatibility with PyTorch Geometric (PyG).



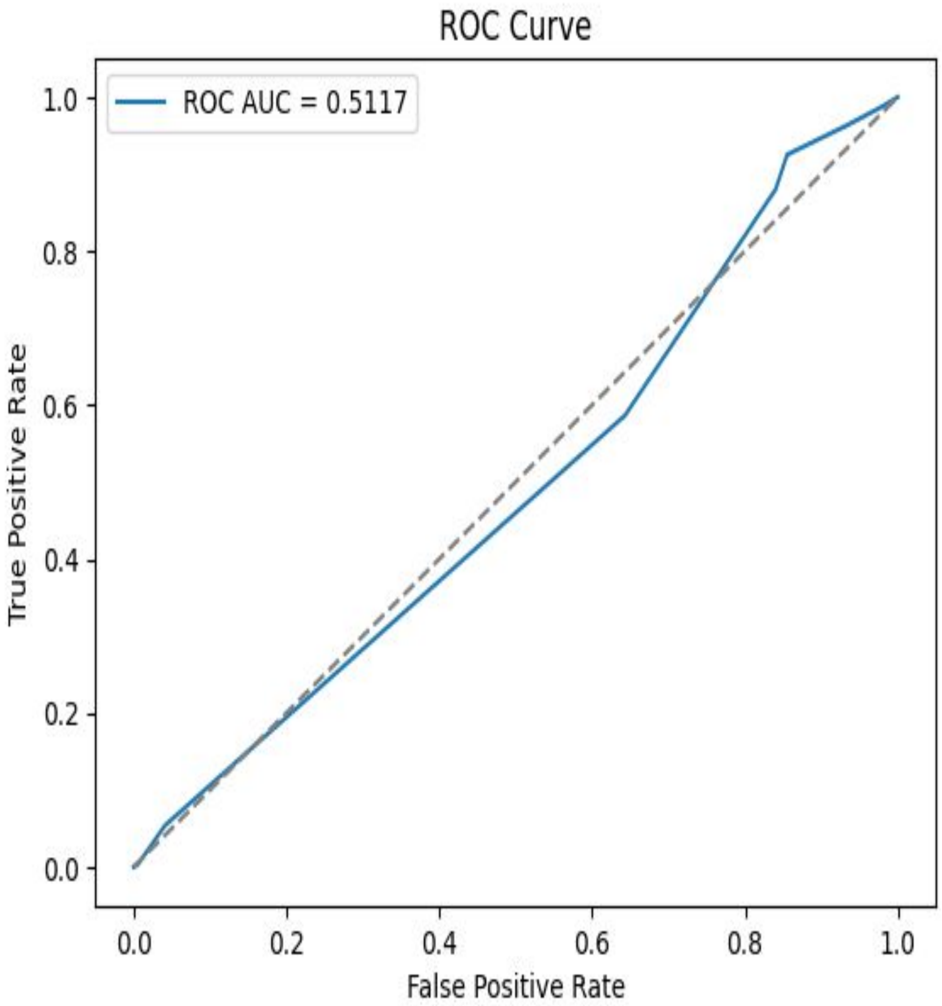
RESEARCH QUESTION 6

- Model & Training Process:
  - Implemented GraphSAGE, a graph neural network model, to learn node embeddings.
  - Split dataset into train (85%), validation (5%), and test (10%) using RandomLinkSplit
  - Binary cross-entropy loss was used to optimize link prediction.
  - Adam optimizer with a learning rate of 0.01 was applied.
- Training Results:
  - As shown in the line plot Training loss decreased significantly, from 1.67 in the first epoch to 0.72 in the final epoch.



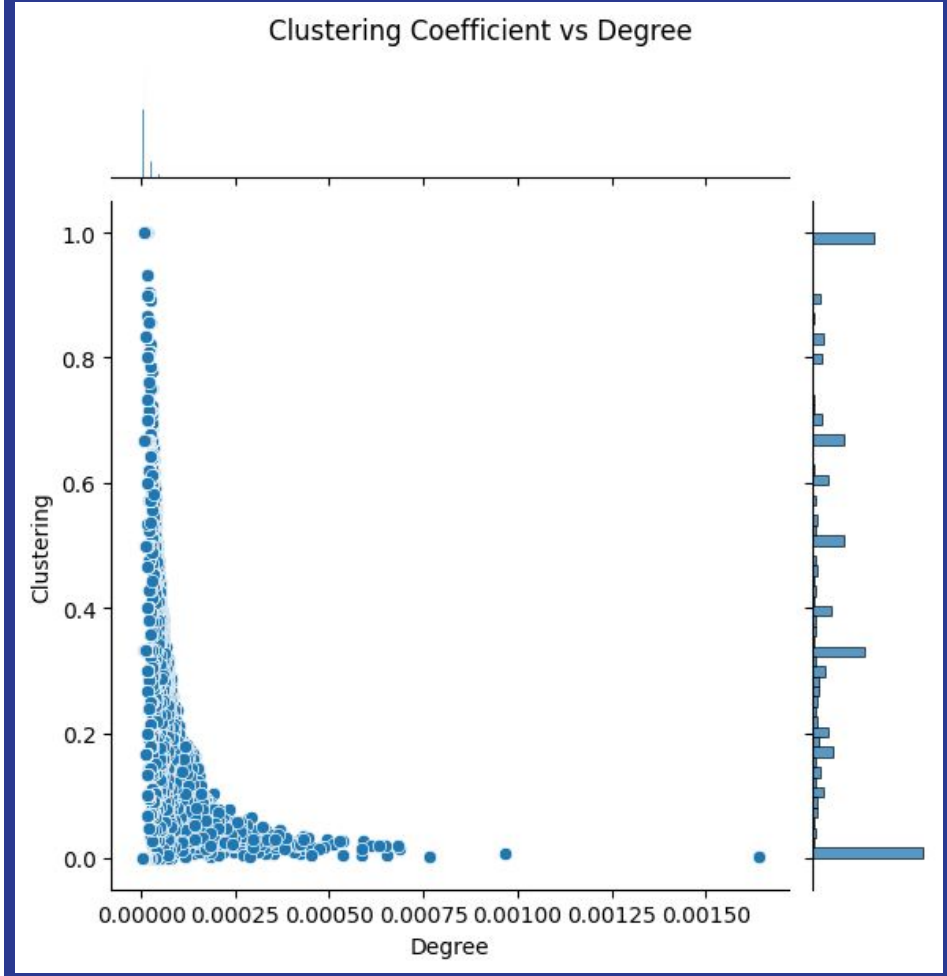
RESEARCH QUESTION 6

- Evaluation Metrics::
  - Test Loss: 1.06, indicating room for improvement in generalization.
  - Test Accuracy: 0.5000, meaning the model performs similarly to random guessing.
  - ROC AUC Score: 0.5117, suggesting weak discriminative ability between positive and negative links..
  - Adam optimizer with a learning rate of 0.01 was applied.
  - Average Precision Score: 0.5088, indicating that the ranking of predicted links is close to random.



CONCLUSION

- ★ The network structure is highly dependent on central hub nodes.
- ★ Smaller, denser communities indicate stronger product relationships, useful for recommendations.
- ★ Machine learning models for link prediction need additional features to improve accuracy.



RELEVANCE OF THE RESEARCH TO SHOP DESK

1) Identifying High-Impact Products (Hubs in the Network)
   a) Just like in the Amazon network, certain products act as key hubs with high centrality (e.g., Node 548091).
   b) Shop Desk can use sales data to identify best-selling or frequently co-purchased products, helping businesses prioritize inventory and marketing efforts.
2) Understanding Product Bundling & Cross-Selling Opportunities
   a) The clustering coefficient analysis showed strong co-purchasing behavior within niche product groups.
   b) By analyzing which products are commonly bought together, Shop Desk can suggest bundling strategies or personalized recommendations to boost sales to Users.

3) Assessing Business Resilience & Inventory Redundancy: Shop Desk can help businesses assess which products are critical to overall sales stability, ensuring they don’t go out of stock or disrupt the supply chain

4)Community-Based Demand Forecasting: Shop Desk Business owners can use community-based demand forecasting to identify seasonal trends, regional preferences, and category-specific demand, optimizing inventory stocking.