

ECE232E_Project2_606539982_106405230_106303745

April 28, 2025

#Question 1

[4]: pip install networkx matplotlib numpy

```
Traceback (most recent call last):
  File "/Users/tilboon/opt/anaconda3/envs/ece219/lib/python3.10/runpy.py", line
196, in _run_module_as_main
    return _run_code(code, main_globals, None,
  File "/Users/tilboon/opt/anaconda3/envs/ece219/lib/python3.10/runpy.py", line
86, in _run_code
    exec(code, run_globals)
  File "/Users/tilboon/opt/anaconda3/envs/ece219/lib/python3.10/site-
packages/pip/_\_main\_\_.py", line 8, in <module>
    if sys.path[0] in ("", os.getcwd()):
FileNotFoundError: [Errno 2] No such file or directory
Note: you may need to restart the kernel to use updated packages.
```

[111]:

```
import networkx as nx
import matplotlib.pyplot as plt
import numpy as np
import urllib.request
import math

# Step 1: Download the file from GitHub raw URL
url = "https://raw.githubusercontent.com/kbaseba/232E/main/Project_2/
facebook_combined.txt"
filename = "./facebook_combined.txt"
urllib.request.urlretrieve(url, filename)

# Step 2: Load the graph
G = nx.read_edgelist(filename, create_using=nx.Graph(), nodetype=int)

# Step 3: Basic info
num_nodes = G.number_of_nodes()
num_edges = G.number_of_edges()
print(f"Number of nodes: {num_nodes}")
print(f"Number of edges: {num_edges}")
```

Number of nodes: 4039

Number of edges: 88234

Q1.1: Number of Nodes and Edges

Number of nodes: 4039

Number of edges: 88234

```
[30]: # QUESTION 1.2: Is the network connected? Find GCC if not.  
is_connected = nx.is_connected(G)  
print(f"Is the network connected? {is_connected}")  
  
if not is_connected:  
    gcc = max(nx.connected_components(G), key=len)  
    G_gcc = G.subgraph(gcc).copy()  
else:  
    G_gcc = G  
  
print(f"Size of GCC - Nodes: {G_gcc.number_of_nodes()}, Edges: {G_gcc.  
    ↪number_of_edges()}")
```

Is the network connected? True

Size of GCC - Nodes: 4039, Edges: 88234

Q1.2: Is the network connected?

Yes, the Facebook network is connected.

Size of GCC: 4039 nodes, 88234 edges (same as the whole network).

1 Question 2

```
[31]: # QUESTION 2: Diameter of the GCC  
diameter = nx.diameter(G_gcc)  
print(f"Diameter of the network (or GCC): {diameter}")
```

Diameter of the network (or GCC): 8

Q2: Diameter of the network

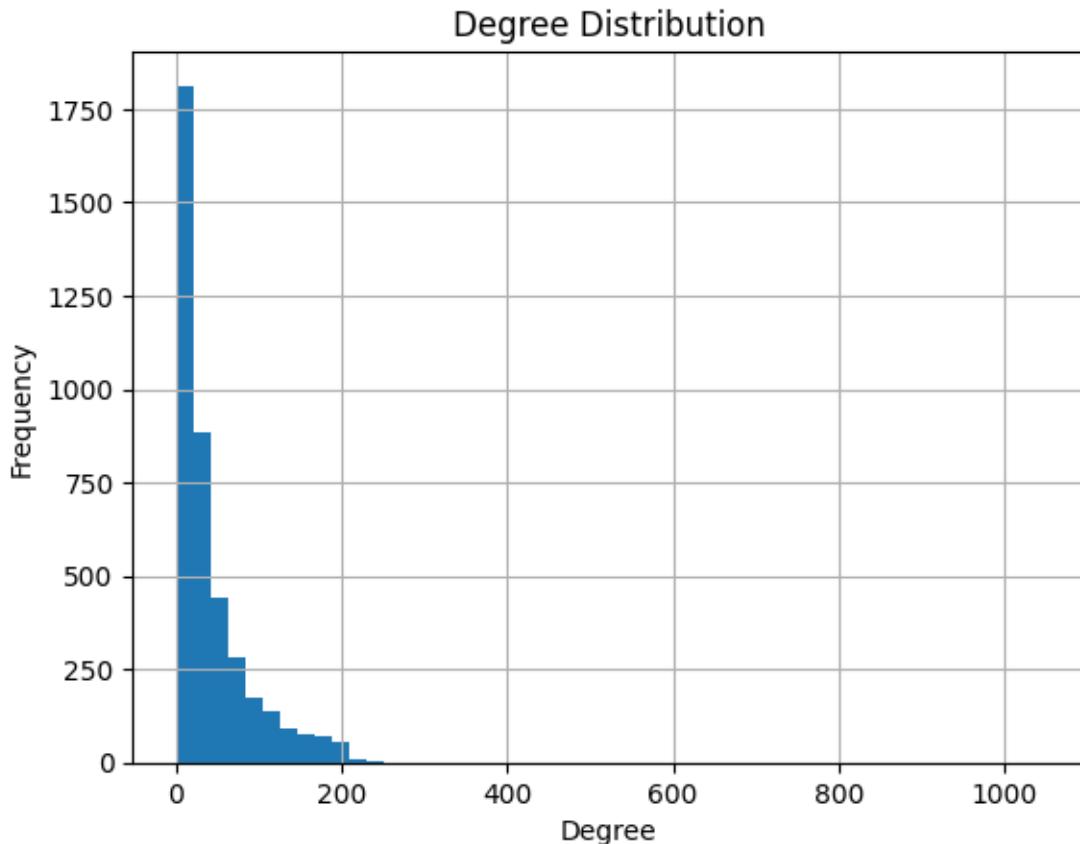
Diameter: 8

2 Question 3

```
[32]: # QUESTION 3: Degree distribution  
degrees = [d for n, d in G.degree()]  
avg_degree = np.mean(degrees)  
print(f"Average degree: {avg_degree:.2f}")  
  
# Plot degree distribution  
plt.figure()
```

```
plt.hist(degrees, bins=50)
plt.title("Degree Distribution")
plt.xlabel("Degree")
plt.ylabel("Frequency")
plt.grid(True)
plt.show()
```

Average degree: 43.69



Q3: Degree distribution

Average degree: 43.69

3 Question 4

```
[33]: # QUESTION 4: Log-log degree distribution
degree_counts = np.bincount(degrees)
degree_values = np.nonzero(degree_counts)[0]
counts = degree_counts[degree_values]
```

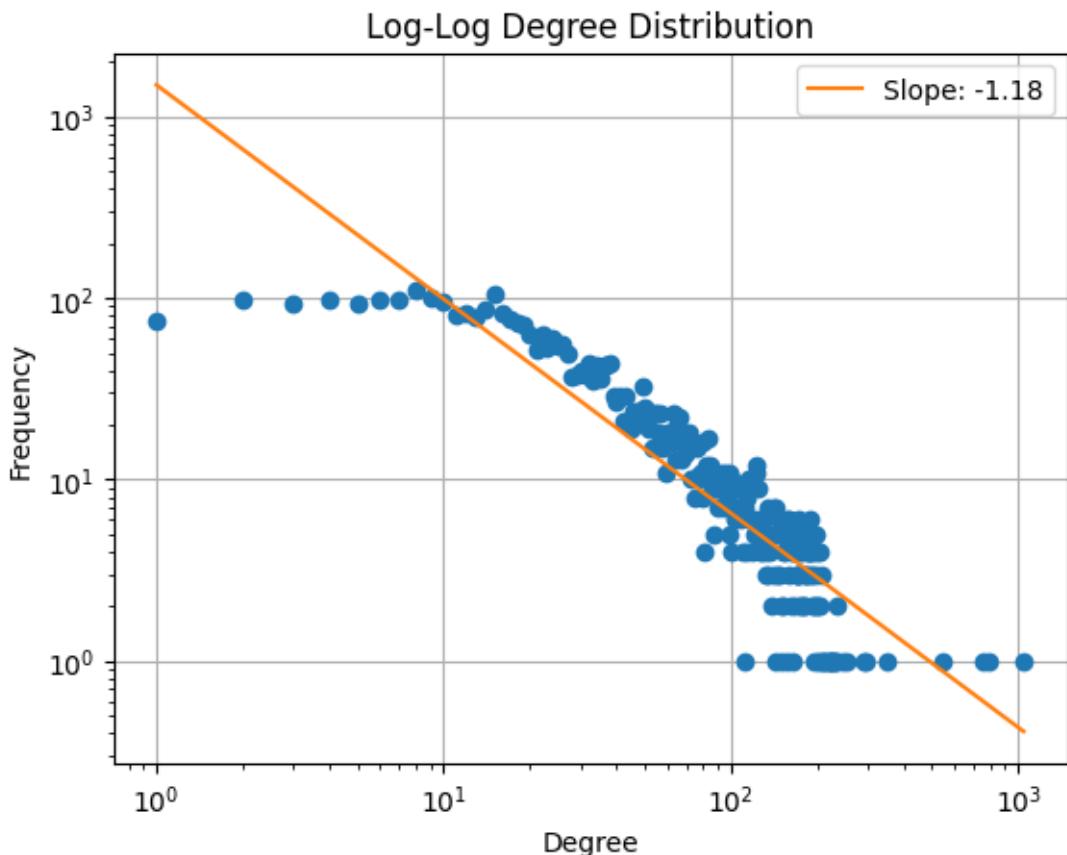
```

plt.figure()
plt.loglog(degree_values, counts, marker='o', linestyle='None')
plt.title("Log-Log Degree Distribution")
plt.xlabel("Degree")
plt.ylabel("Frequency")

# Fit line to log-log plot
log_degrees = np.log(degree_values)
log_counts = np.log(counts)
slope, intercept = np.polyfit(log_degrees, log_counts, 1)
plt.plot(degree_values, np.exp(intercept) * degree_values**slope, label=f"Slope: {slope:.2f}")
plt.legend()
plt.grid(True)
plt.show()

print(f"Estimated slope of the log-log plot: {slope:.2f}")

```



Estimated slope of the log-log plot: -1.18

Q4: Log-log degree distribution

Estimated slope: -1.18

The degree distribution of the Facebook network follows a power-law with a slope of approximately -1.18, indicating that the network is scale-free. This suggests the presence of a few highly connected hubs and many nodes with low degree, a characteristic structure commonly observed in real-world social networks.

4 Question 5

```
[34]: # ---- PERSONALIZED NETWORK ----
```

```
# In the edge list, node ID = 0 corresponds to graph node ID = 1
target_node = 0 # user whose ID is 1
```

```
[35]: # QUESTION 5: Create personalized network for node 0
```

```
personalized_net = nx.ego_graph(G, target_node) # ego_graph returns node +  
→neighbors
```

```
[36]: # Number of nodes and edges
```

```
pn_nodes = personalized_net.number_of_nodes()
pn_edges = personalized_net.number_of_edges()
print(f"Personalized Network Nodes: {pn_nodes}")
print(f"Personalized Network Edges: {pn_edges}")
```

Personalized Network Nodes: 348

Personalized Network Edges: 2866

Question 5: Personalized Network

Number of nodes: 348

Number of edges: 2866

5 Question 6

```
[46]: # QUESTION 6: Diameter
```

```
pn_diameter = nx.diameter(personalized_net)
print(f"Diameter of the Personalized Network: {pn_diameter}")
```

Diameter of the Personalized Network: 2

```
[38]: # Trivial bounds for diameter:
```

```
lower_bound = 1 # Minimum if everyone is directly connected
upper_bound = 2 # Max because all nodes are neighbors of center or each other
print(f"Lower Bound: {lower_bound}")
print(f"Upper Bound: {upper_bound}")
```

```
Lower Bound: 1  
Upper Bound: 2
```

Question 6: Diameter and Bounds

Diameter of the personalized network: 2

Lower bound: 1 (if everyone was connected to each other)

Upper bound: 2 (if friends are connected only through the center node)

6 Question 7

```
[39]: # QUESTION 7: Interpretations  
if pn_diameter == upper_bound:  
    print("Diameter equals upper bound: Some neighbors are not directly connected to each other.")  
elif pn_diameter == lower_bound:  
    print("Diameter equals lower bound: All neighbors are directly connected to each other (clique).")  
else:  
    print("Diameter is between bounds: Intermediate connectivity among neighbors.")
```

Diameter equals upper bound: Some neighbors are not directly connected to each other.

Question 7:

Since the diameter of the personalized network is equal to the upper bound (2), it means that some neighbors of the user are not directly connected to each other and must go through the central user to reach one another. This shows that the network is not fully connected like a clique, but instead has a star-like structure where most friends are connected mainly through the user rather than directly to each other.

7 Question 8

```
[40]: # --- Find Core Nodes (degree > 200) ---  
  
# Get degree of each node  
node_degrees = dict(G.degree())  
  
# Core nodes: nodes with degree > 200  
core_nodes = [node for node, deg in node_degrees.items() if deg > 200]  
  
# Number of core nodes  
num_core_nodes = len(core_nodes)  
print(f"Number of core nodes: {num_core_nodes}")
```

```
# Average degree of core nodes
avg_core_degree = np.mean([node_degrees[node] for node in core_nodes])
print(f"Average degree of core nodes: {avg_core_degree:.2f}")
```

Number of core nodes: 40
 Average degree of core nodes: 279.38

Question 8:

Number of core nodes: 40
 Average degree of the core nodes: 279.38

8 Question 9

```
[ ]: # Code for Q9
import networkx as nx
import igraph as ig
import matplotlib.pyplot as plt
import matplotlib.colors as mcolors
import random

# From assignment specification, core nodes are given as:
# core node = node ID - 1
core_nodes = [0, 107, 348, 483, 1086]

# Helper function to plot communities
def plot_communities(G_personal, membership, title):
    pos = nx.spring_layout(G_personal, seed=42)
    communities = {}
    for idx, cluster_id in enumerate(membership):
        communities.setdefault(cluster_id, []).append(idx)

    colors = list(mcolors.TABLEAU_COLORS.values()) + list(mcolors.CSS4_COLORS.
    ↪values())
    random.shuffle(colors)

    plt.figure(figsize=(6, 6))
    for cluster_id, nodes in communities.items():
        node_list = [list(G_personal.nodes())[node_idx] for node_idx in nodes]
        nx.draw_networkx_nodes(G_personal, pos, nodelist=node_list, ↪
        ↪node_color=colors[cluster_id % len(colors)], label=f'Community {cluster_id}')
    nx.draw_networkx_edges(G_personal, pos, alpha=0.5)
    plt.title(title)
    plt.axis('off')
    plt.legend()
    plt.show()
```

```

def fast_greedy(G_personal, core_node):
    G_ig = ig.Graph.TupleList(G_personal.edges(), directed=False)
    fg_clusters = G_ig.community_fastgreedy().as_clustering()
    modularity = G_ig.modularity(fg_clusters.membership)
    plot_communities(G_personal, fg_clusters.membership, f"Fast-Greedy (Core_{core_node})")
    return modularity

def edge_betweenness(G_personal, core_node):
    G_ig = ig.Graph.TupleList(G_personal.edges(), directed=False)
    eb_clusters = G_ig.community_edge_betweenness().as_clustering()
    modularity = G_ig.modularity(eb_clusters.membership)
    plot_communities(G_personal, eb_clusters.membership, f"Edge-Betweenness (Core_{core_node})")
    return modularity

def infomap(G_personal, core_node):
    G_ig = ig.Graph.TupleList(G_personal.edges(), directed=False)
    im_clusters = G_ig.community_infomap()
    modularity = G_ig.modularity(im_clusters.membership)
    plot_communities(G_personal, im_clusters.membership, f"Infomap (Core_{core_node})")
    return modularity

```

```

[66]: # Main loop
for core in core_nodes:
    print(f"== Core node {core} ==")
    G_personal = nx.ego_graph(G, core)

    mod_fg = fast_greedy(G_personal, core)
    print(f"Fast-Greedy modularity: {mod_fg:.4f}")

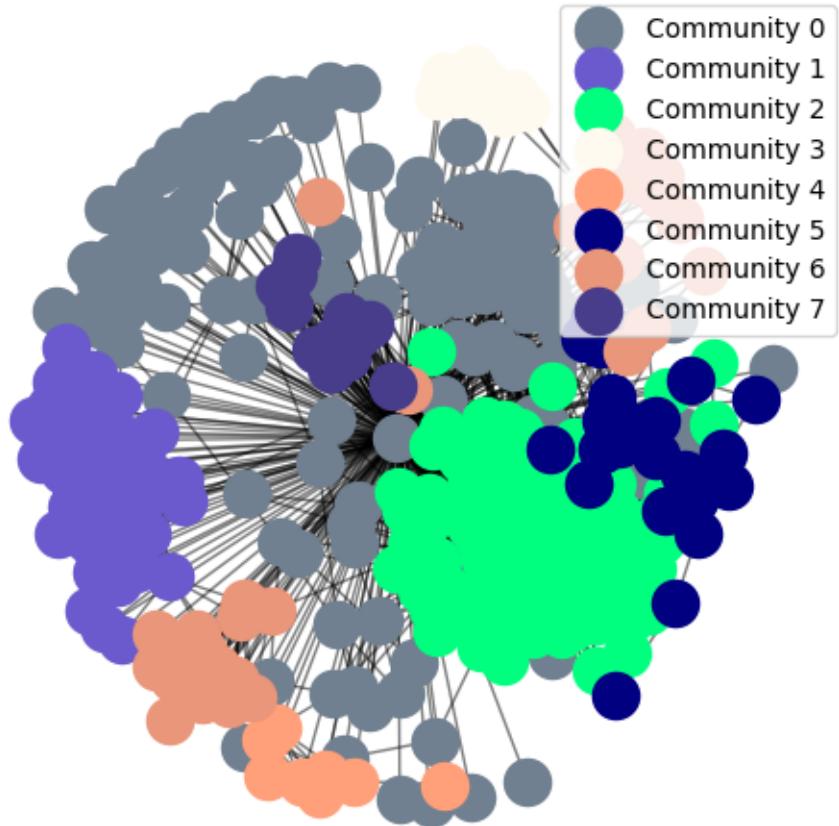
    mod_eb = edge_betweenness(G_personal, core)
    print(f"Edge-Betweenness modularity: {mod_eb:.4f}")

    mod_im = infomap(G_personal, core)
    print(f"Infomap modularity: {mod_im:.4f}")

```

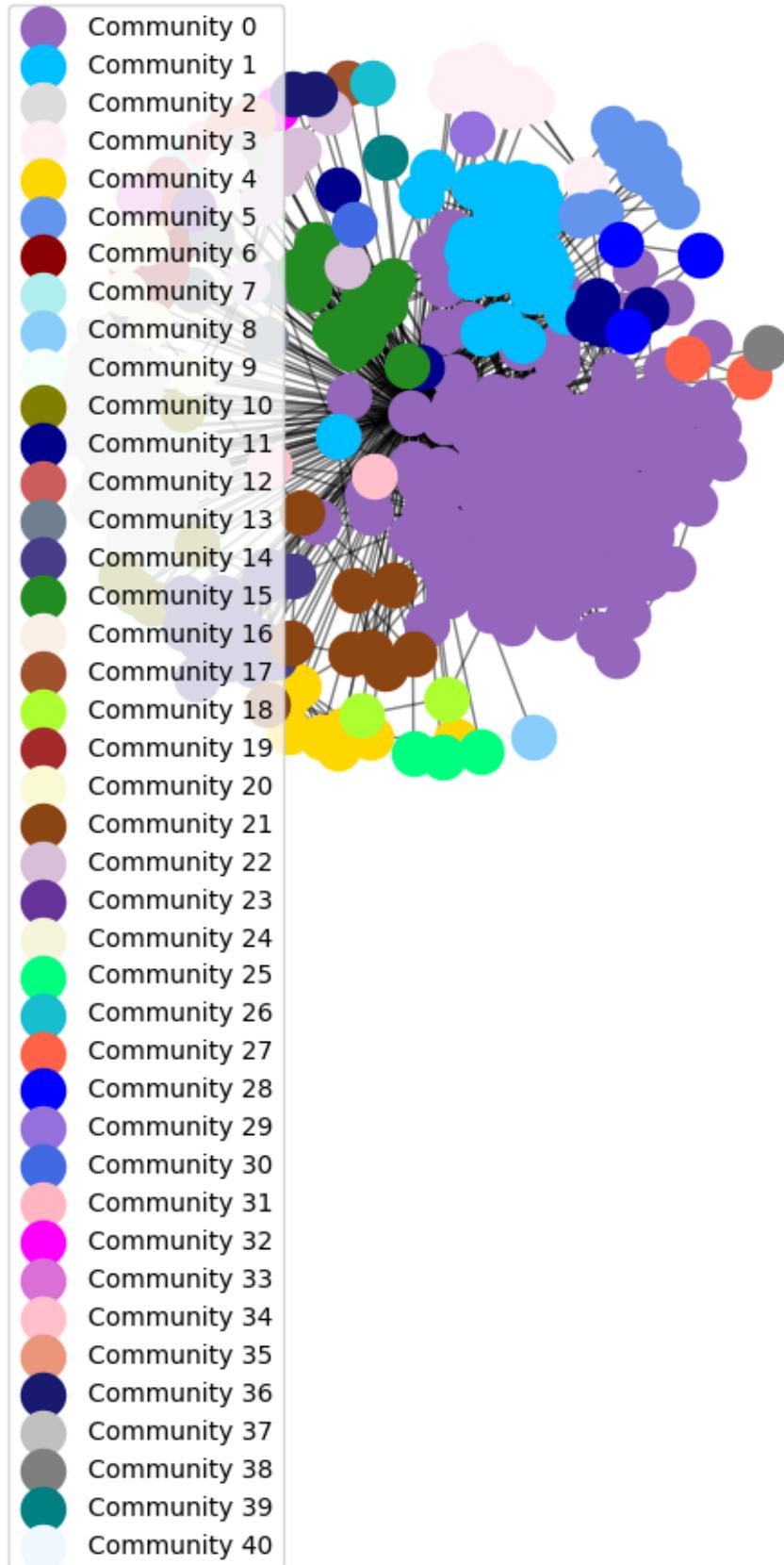
==== Core node 0 ===

Fast-Greedy (Core 0)

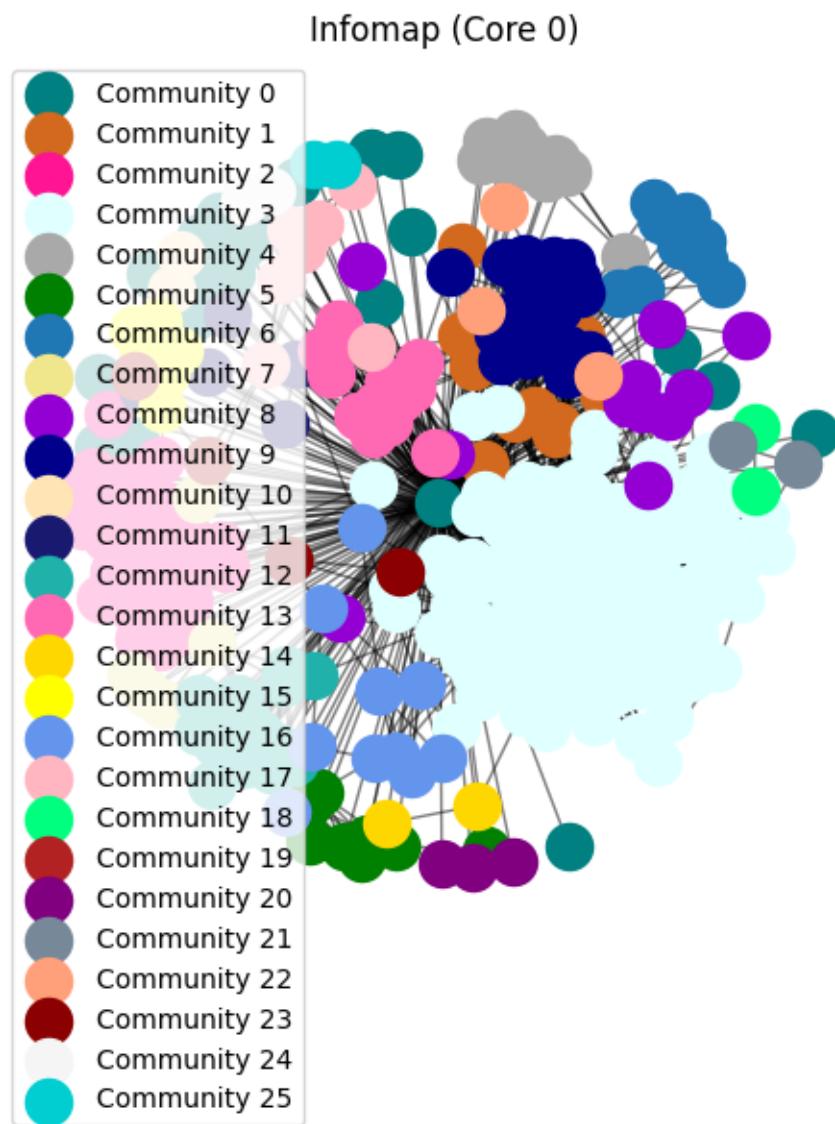


Fast-Greedy modularity: 0.4131

Edge-Betweenness (Core 0)



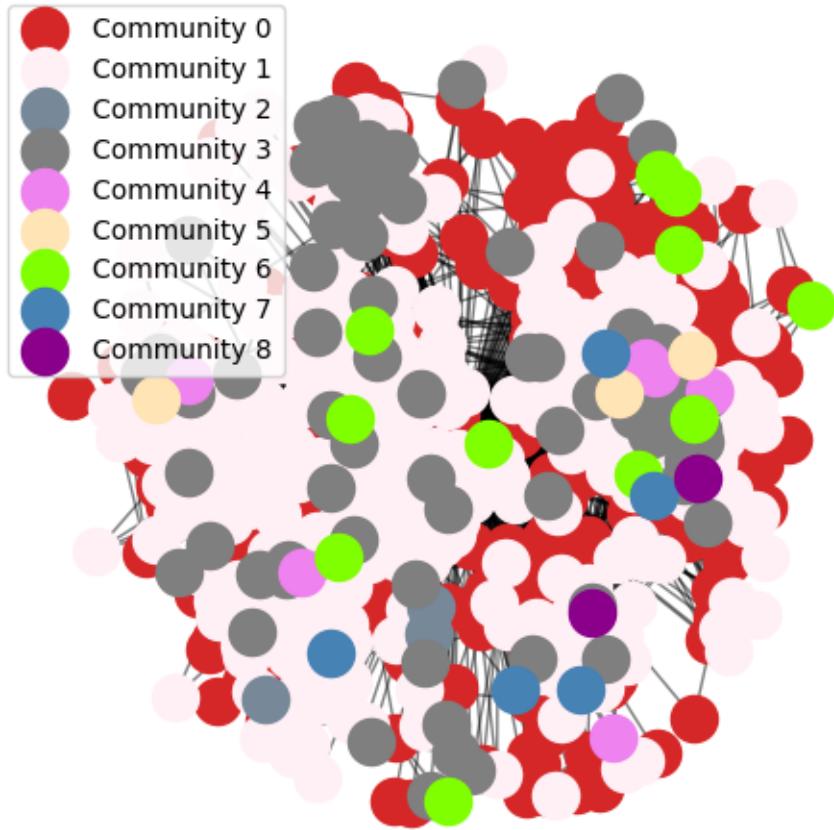
Edge-Betweenness modularity: 0.3533



Infomap modularity: 0.3891

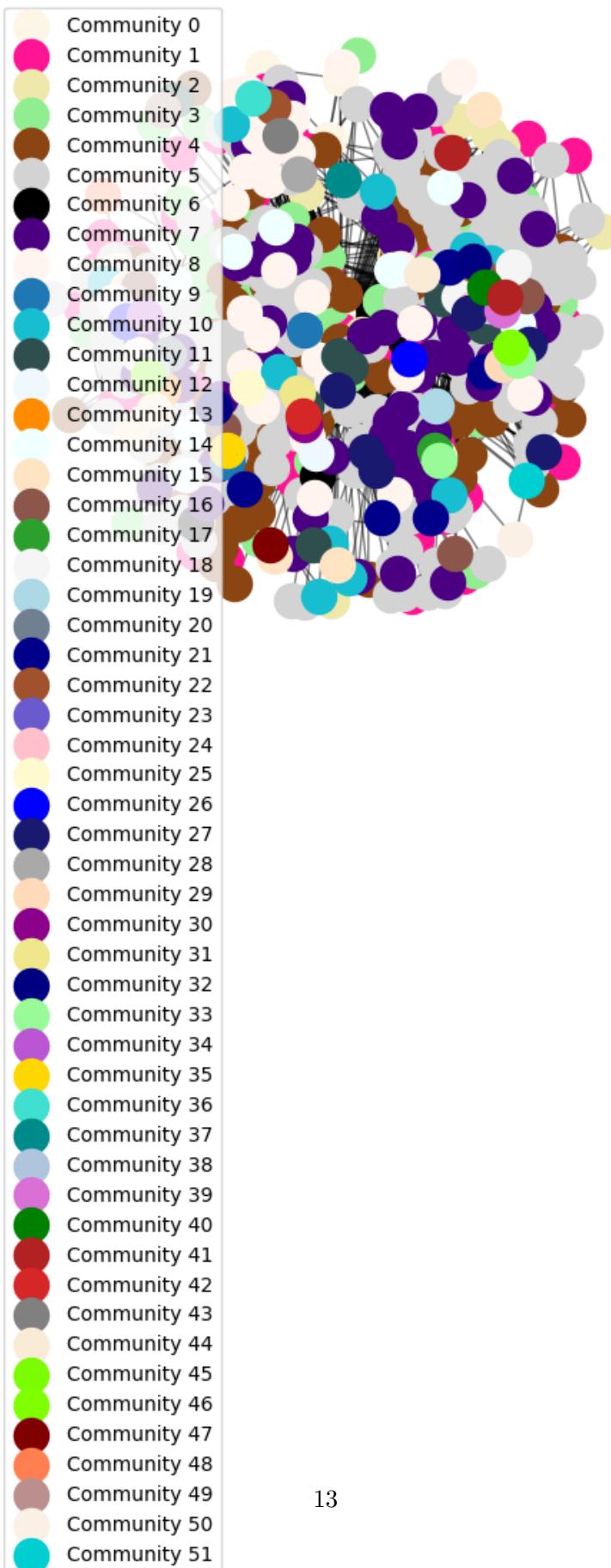
==== Core node 107 ===

Fast-Greedy (Core 107)

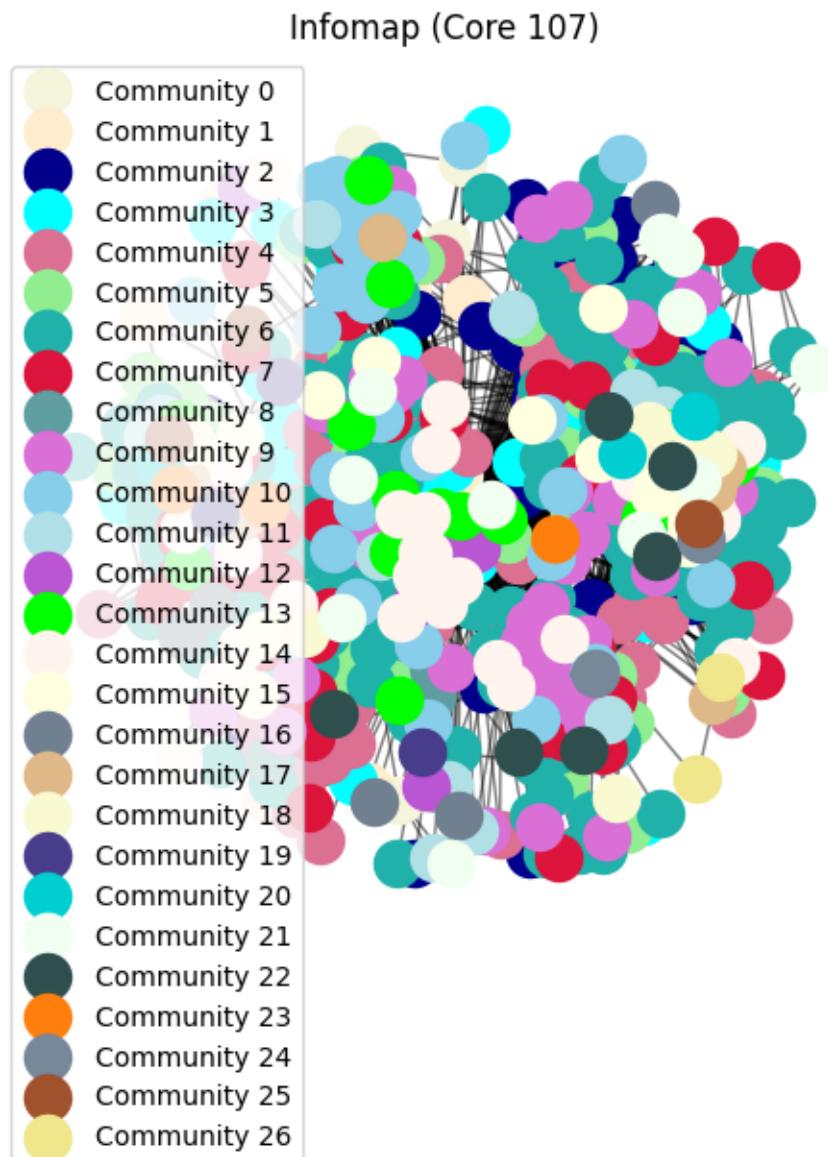


Fast-Greedy modularity: 0.4368

Edge-Betweenness (Core 107)



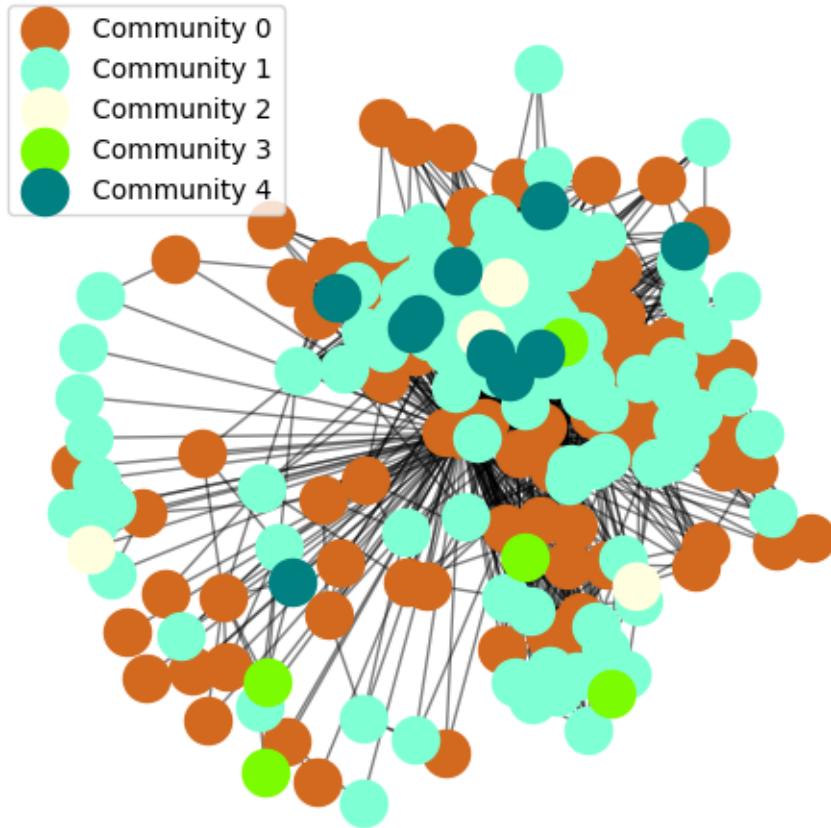
Edge-Betweenness modularity: 0.5068



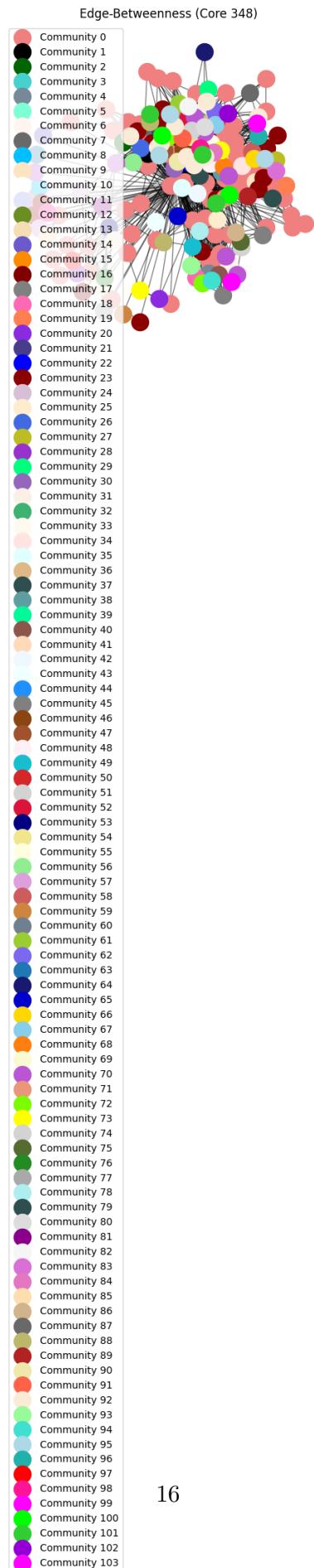
Infomap modularity: 0.5082

==== Core node 348 ===

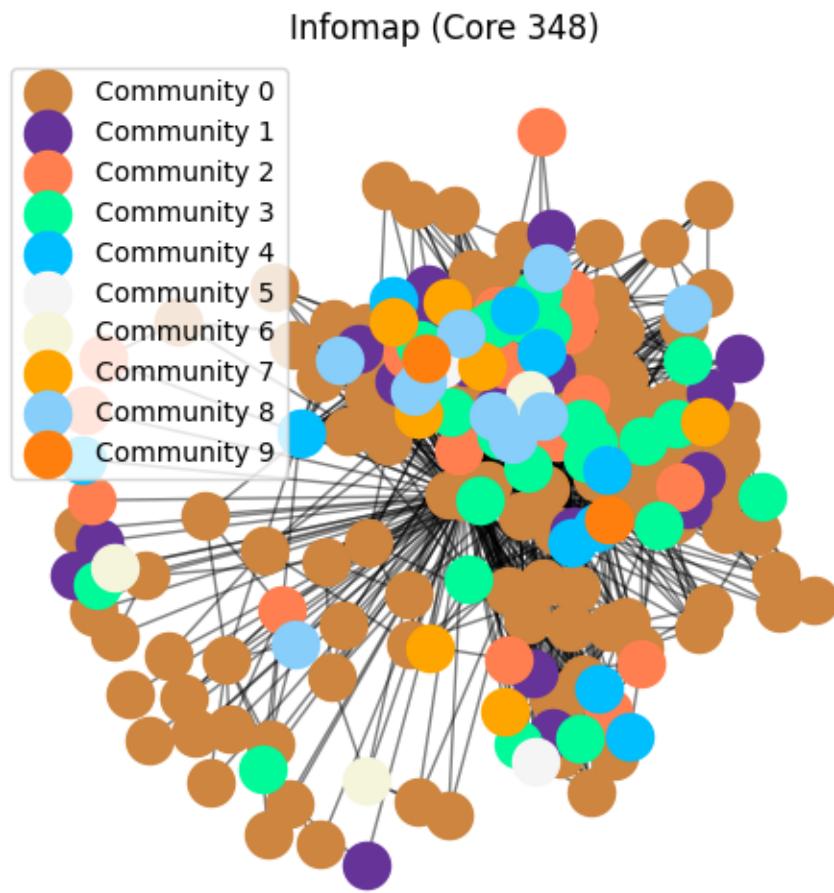
Fast-Greedy (Core 348)



Fast-Greedy modularity: 0.2503



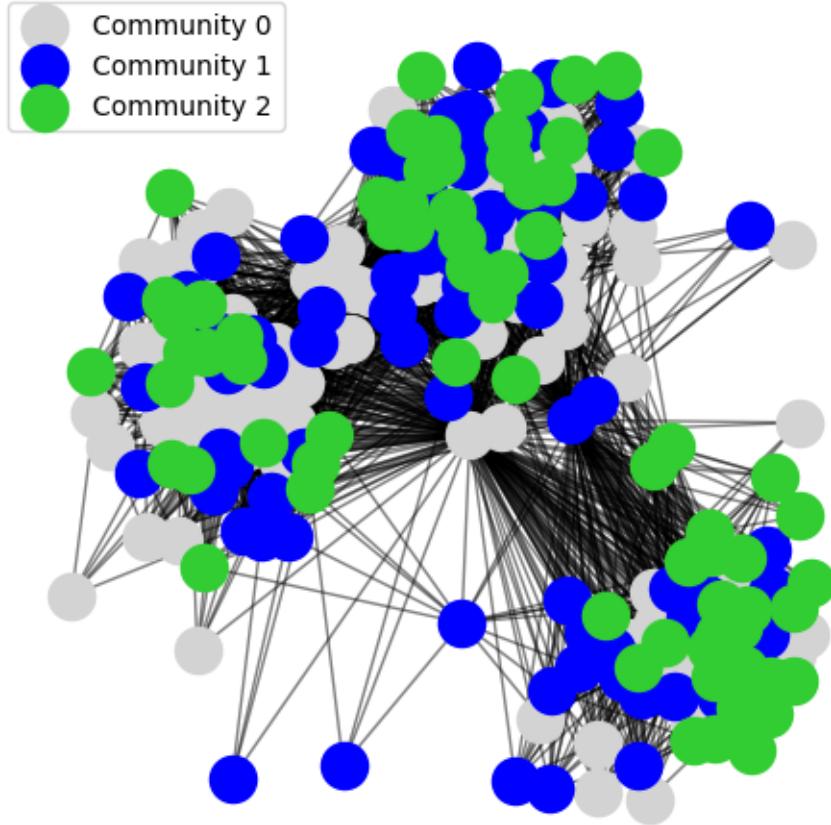
Edge-Betweenness modularity: 0.1335



Infomap modularity: 0.2038

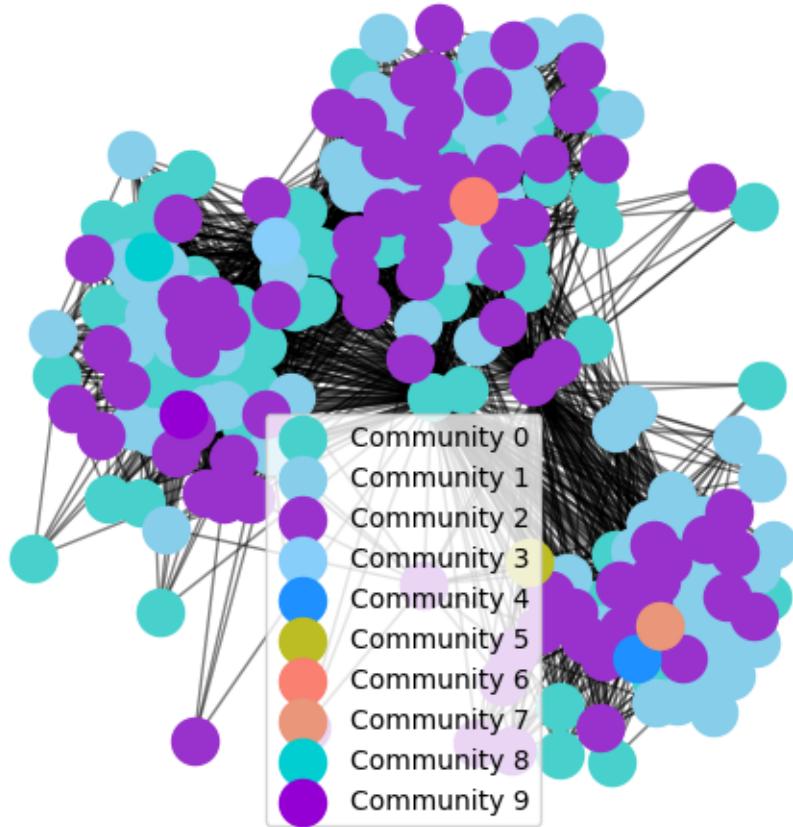
==== Core node 483 ===

Fast-Greedy (Core 483)



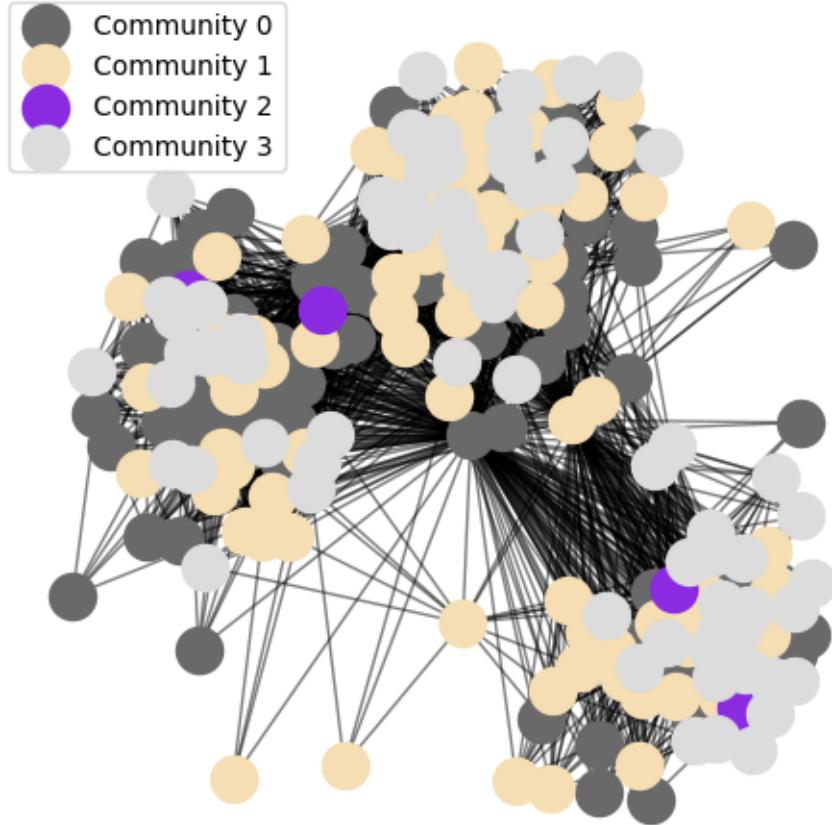
Fast-Greedy modularity: 0.5070

Edge-Betweenness (Core 483)



Edge-Betweenness modularity: 0.4891

Infomap (Core 483)

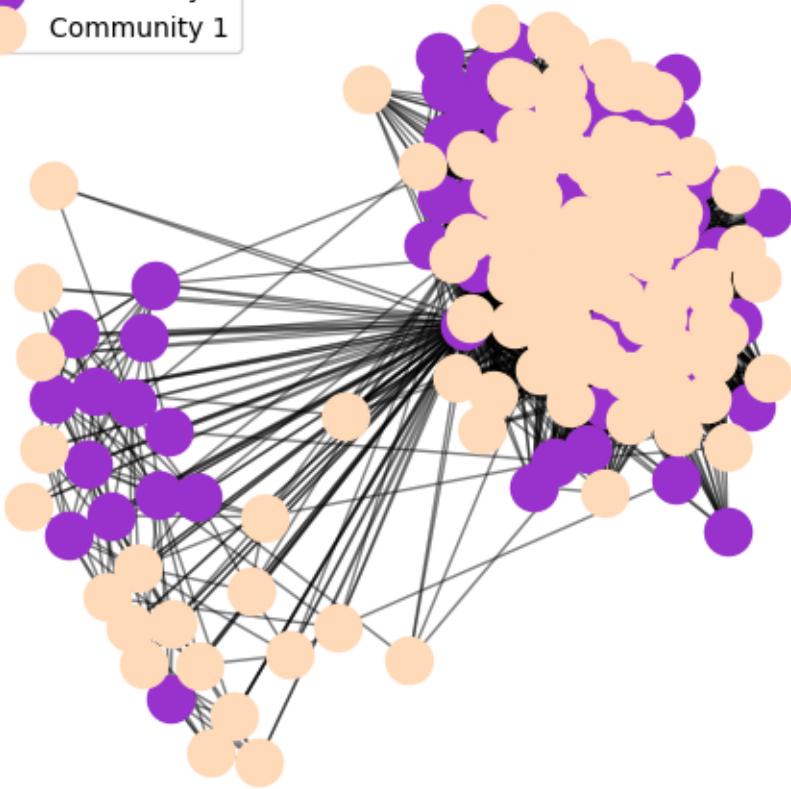


Infomap modularity: 0.5153

==== Core node 1086 ===

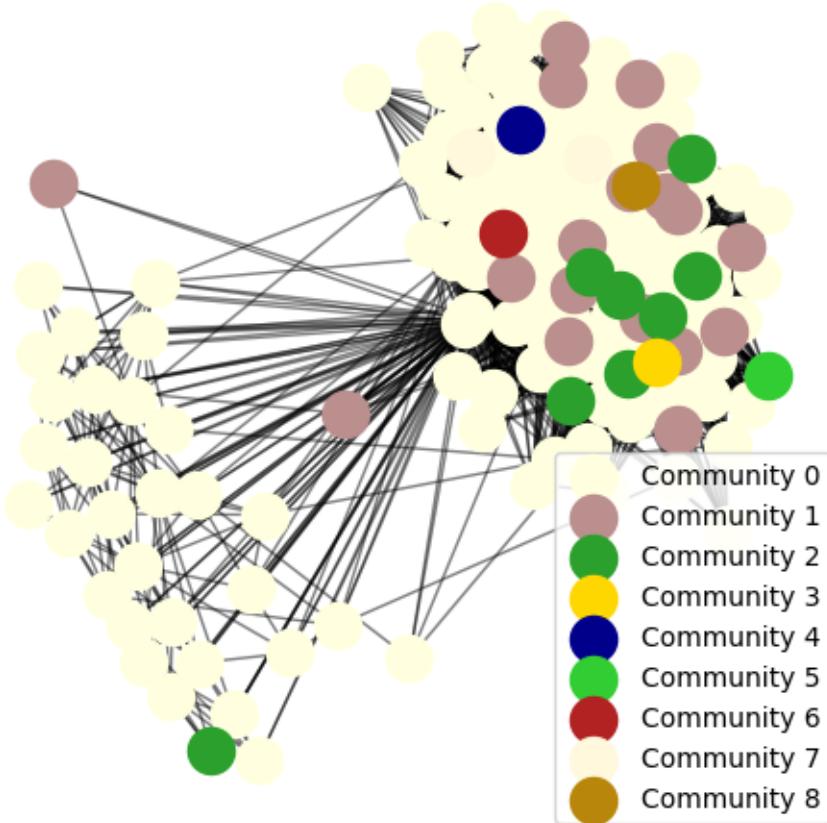
Fast-Greedy (Core 1086)

Community 0
Community 1

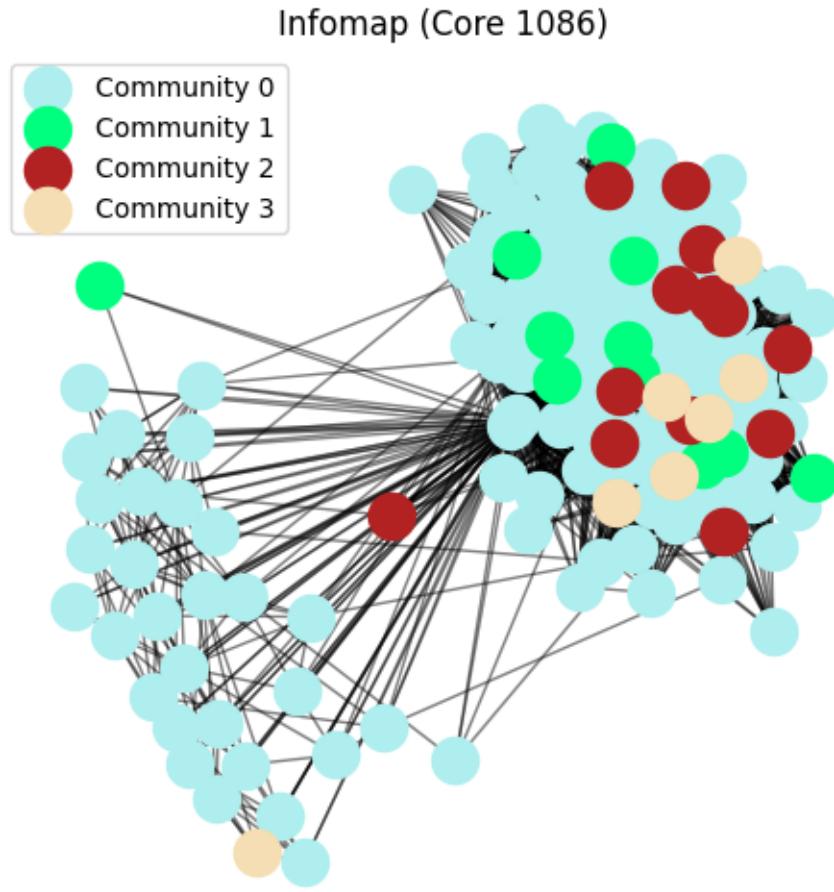


Fast-Greedy modularity: 0.1455

Edge-Betweenness (Core 1086)



Edge-Betweenness modularity: 0.0276



Infomap modularity: 0.0269

8.1 Results:

==== Core Node 1 ====

Fast-Greedy modularity: 0.4131

Edge-Betweenness modularity: 0.3533

Infomap modularity: 0.3891

==== Core Node 108 ====

Fast-Greedy modularity: 0.4368

Edge-Betweenness modularity: 0.5068

Infomap modularity: 0.5082

==== Core Node 349 ====

Fast-Greedy modularity: 0.2503
 Edge-Betweenness modularity: 0.1335
 Infomap modularity: 0.2038
 === Core Node 484 ===
 Fast-Greedy modularity: 0.5070
 Edge-Betweenness modularity: 0.4891
 Infomap modularity: 0.5153
 === Core Node 1087 ===
 Fast-Greedy modularity: 0.1455
 Edge-Betweenness modularity: 0.0276
 Infomap modularity: 0.0269

8.2 Algorithm Modularity Score Comparison:

9 Question 10

```
[68]: # Code for Q10
for core in core_nodes:
    print(f"== Core node {core} (Core node removed) ==")
    G_personal = nx.ego_graph(G, core)
    G_personal_removed = G_personal.copy()
    G_personal_removed.remove_node(core)

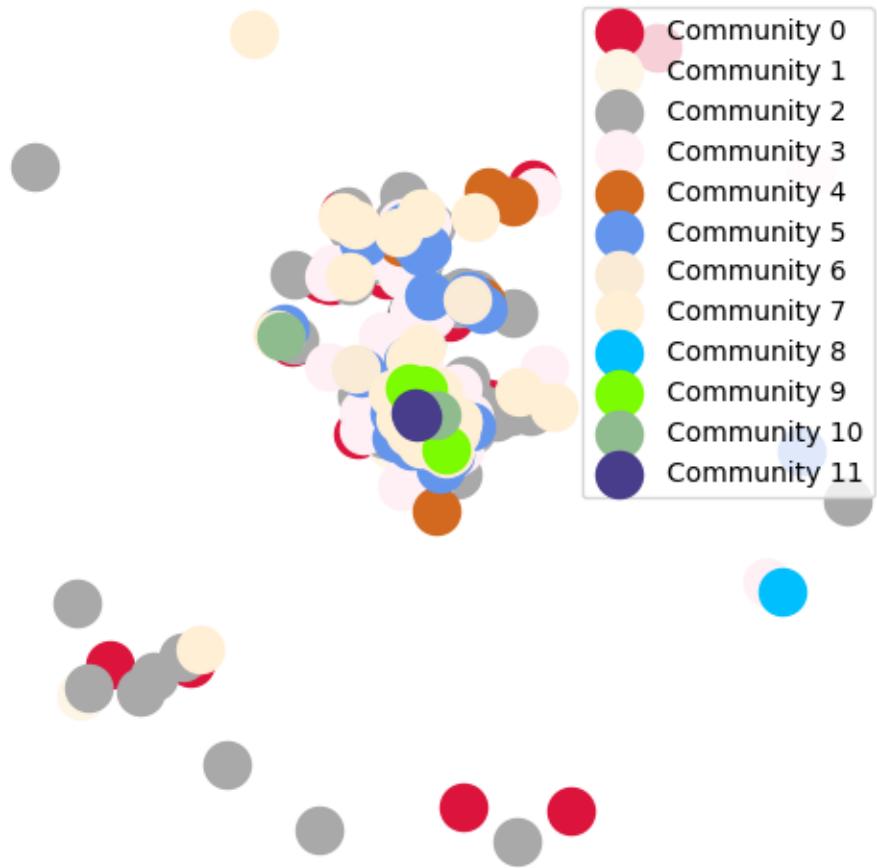
    # Now use the same functions as before
    mod_fg = fast_greedy(G_personal_removed, core)
    print(f"Fast-Greedy modularity (core removed): {mod_fg:.4f}")

    mod_eb = edge_betweenness(G_personal_removed, core)
    print(f"Edge-Betweenness modularity (core removed): {mod_eb:.4f}")

    mod_im = infomap(G_personal_removed, core)
    print(f"Infomap modularity (core removed): {mod_im:.4f}")

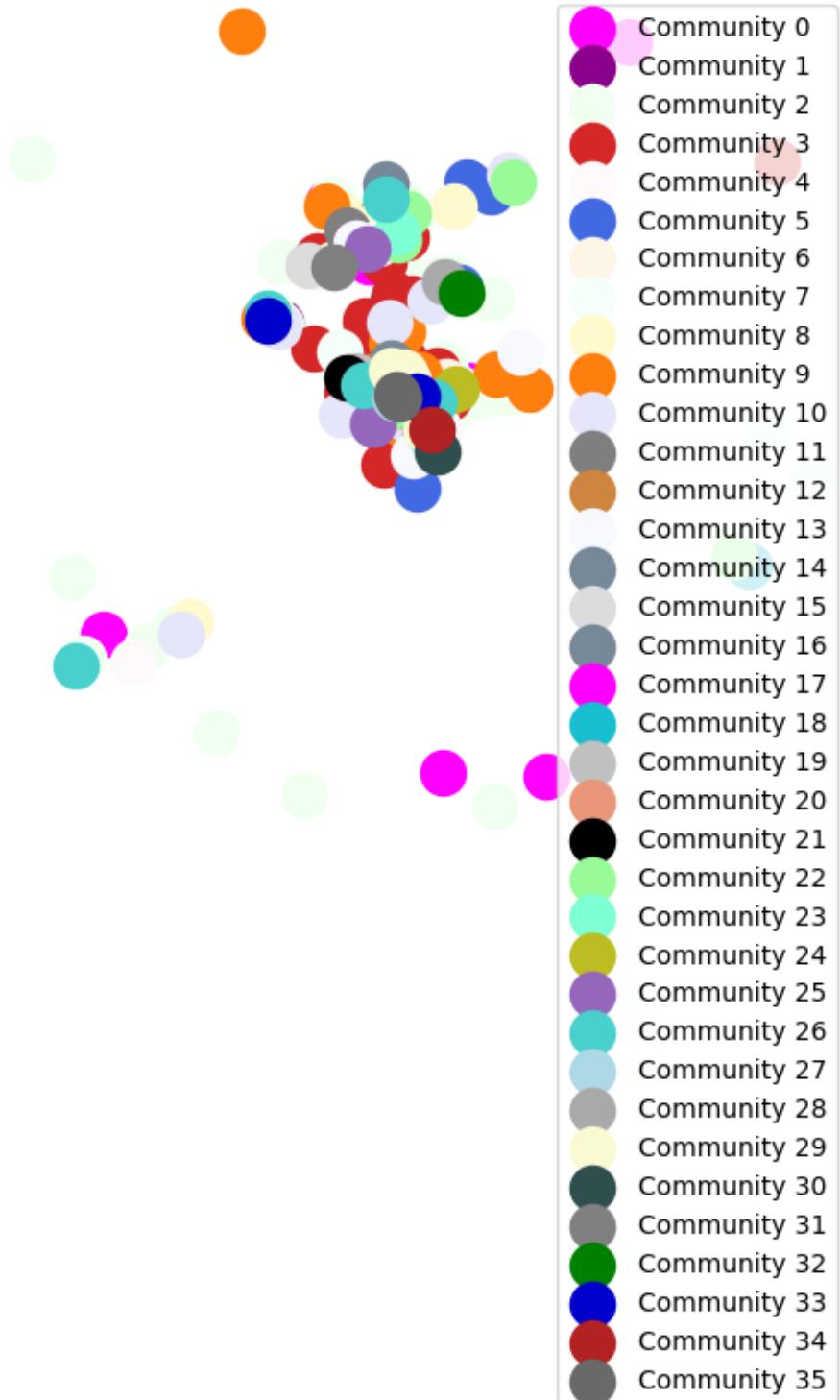
== Core node 0 (Core node removed) ==
```

Fast-Greedy (Core 0)



Fast-Greedy modularity (core removed): 0.4459

Edge-Betweenness (Core 0)



Edge-Betweenness modularity (core removed): 0.4161

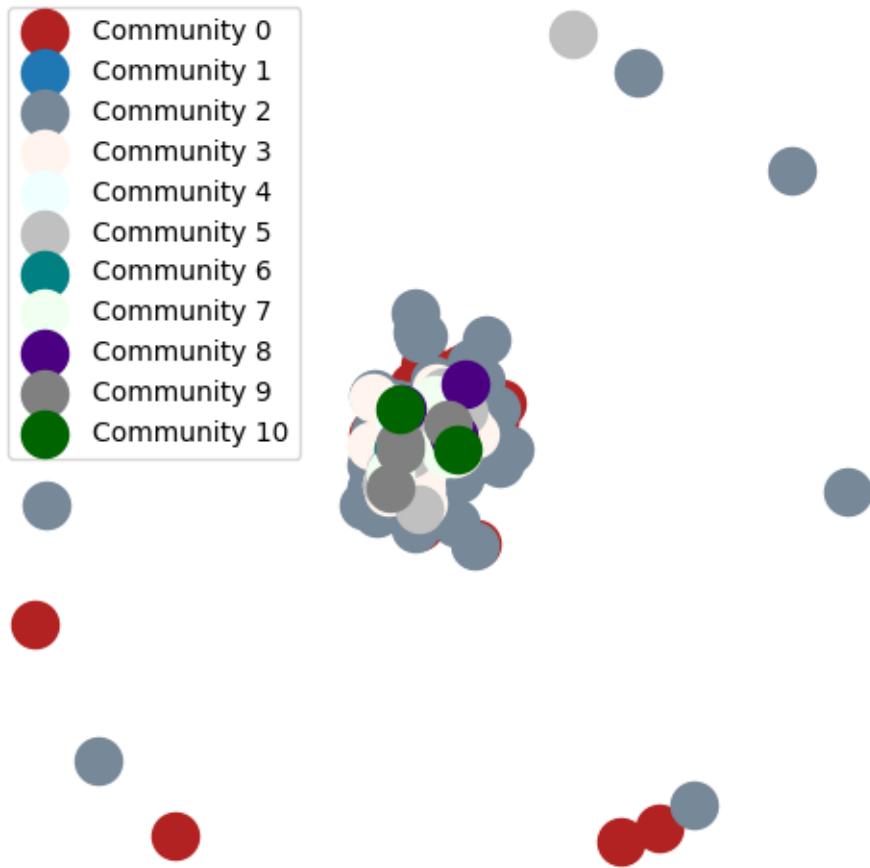
Infomap (Core 0)



Infomap modularity (core removed): 0.4180

==== Core node 107 (Core node removed) ===

Fast-Greedy (Core 107)

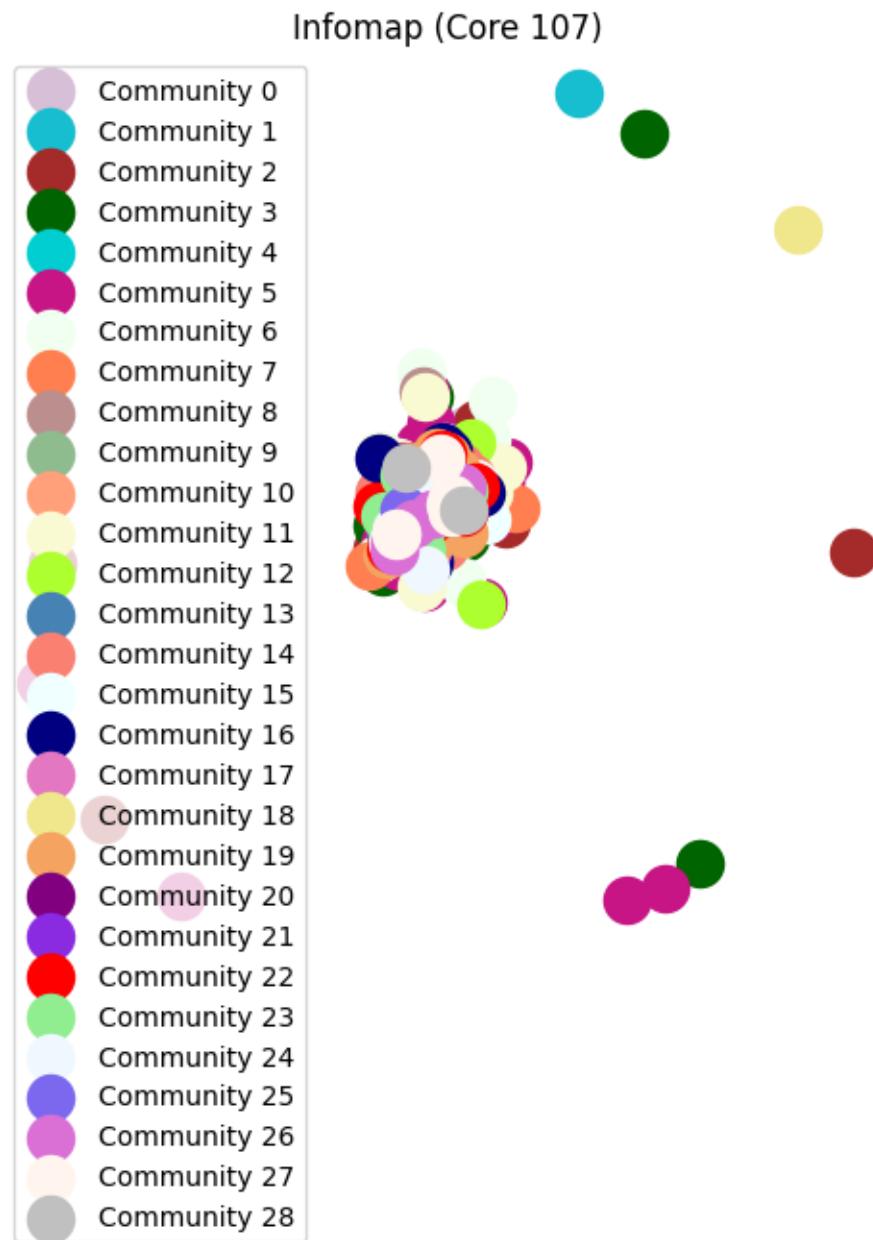


Fast-Greedy modularity (core removed): 0.4580

Edge-Betweenness (Core 107)

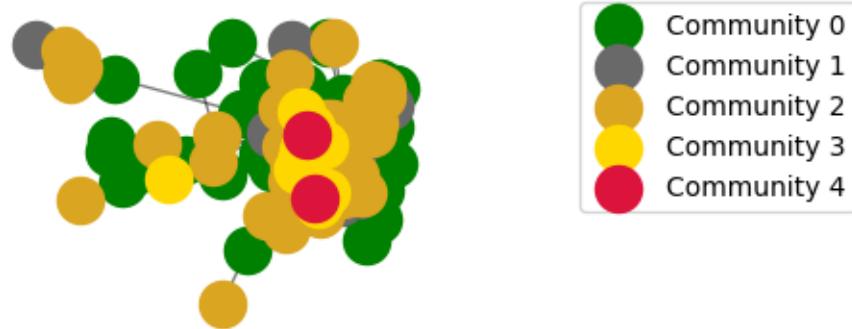


Edge-Betweenness modularity (core removed): 0.5213



Infomap modularity (core removed): 0.5186
== Core node 348 (Core node removed) ==

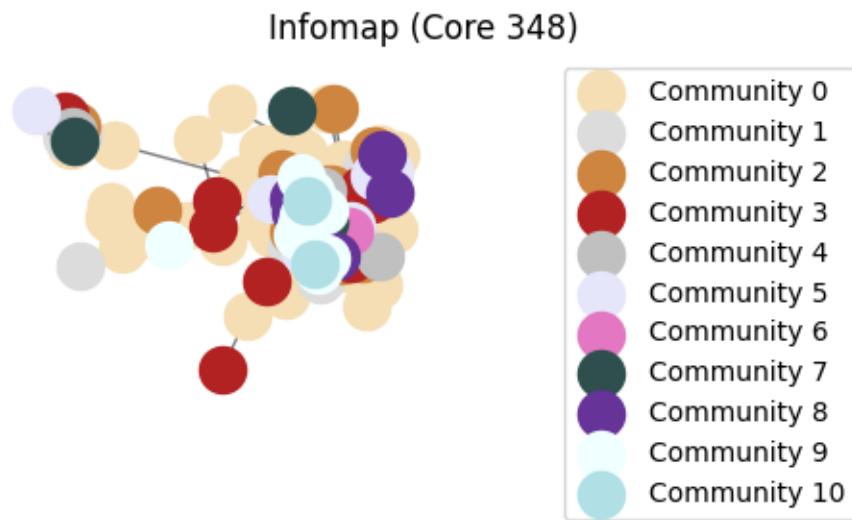
Fast-Greedy (Core 348)



Fast-Greedy modularity (core removed): 0.2457

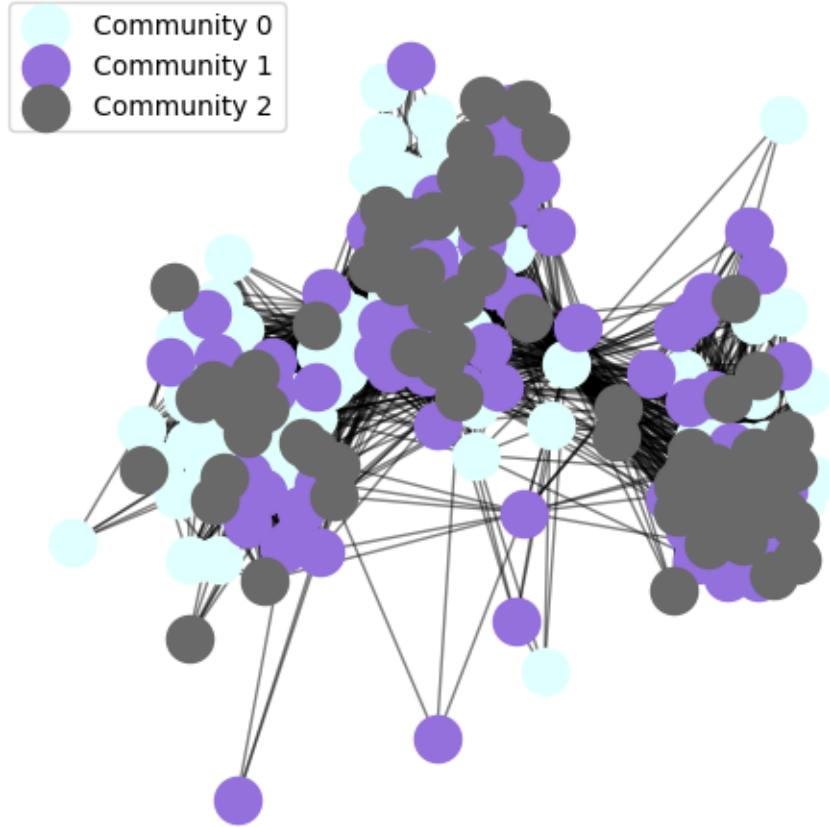


Edge-Betweenness modularity (core removed): 0.1506



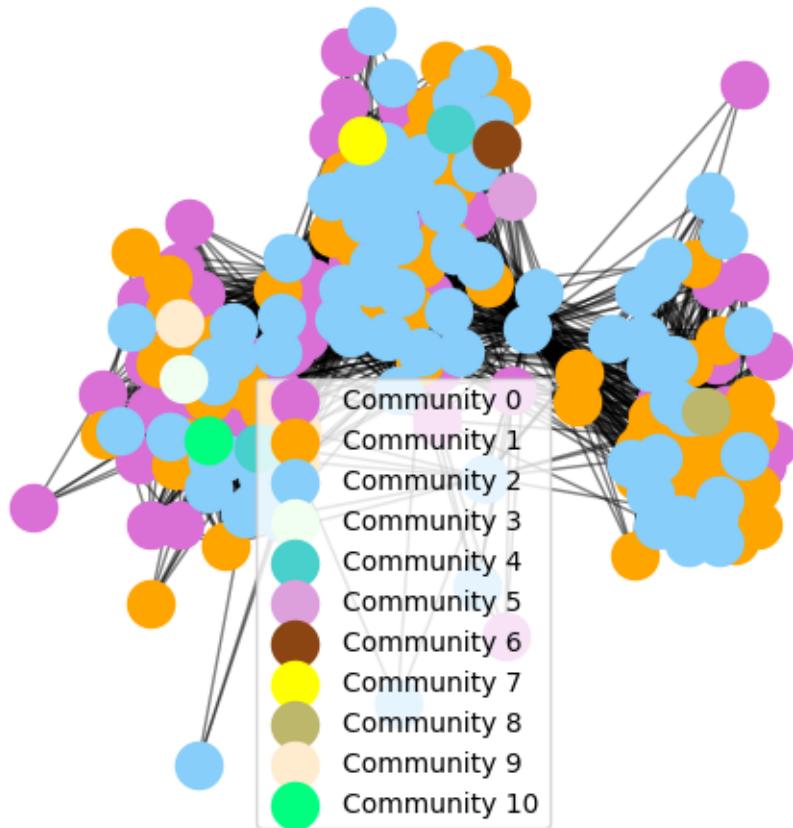
Infomap modularity (core removed): 0.2466
==== Core node 483 (Core node removed) ===

Fast-Greedy (Core 483)



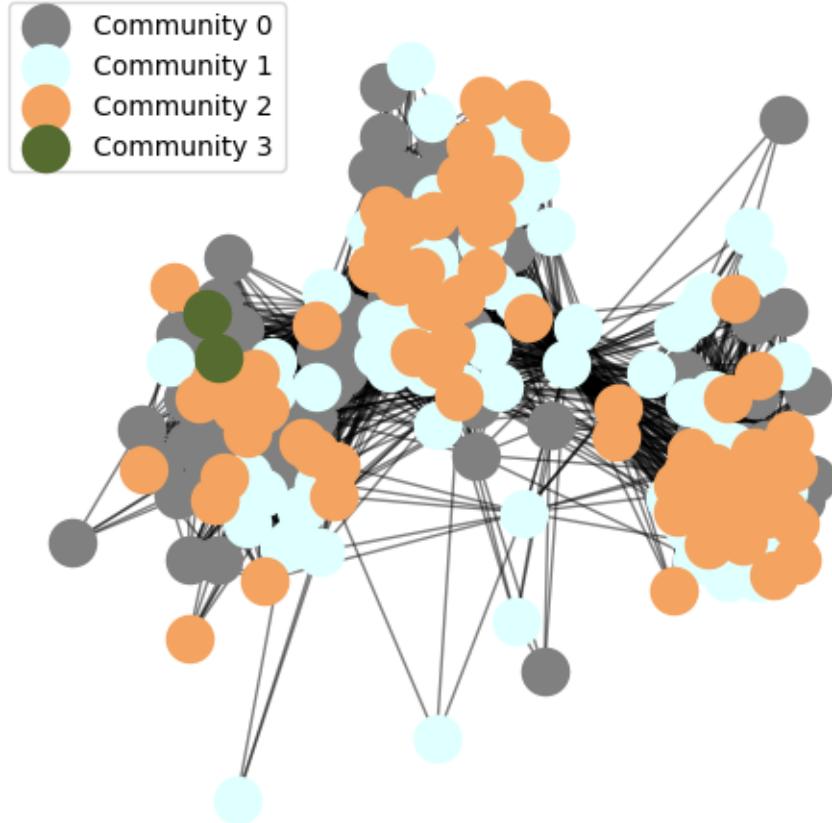
Fast-Greedy modularity (core removed): 0.5342

Edge-Betweenness (Core 483)



Edge-Betweenness modularity (core removed): 0.5154

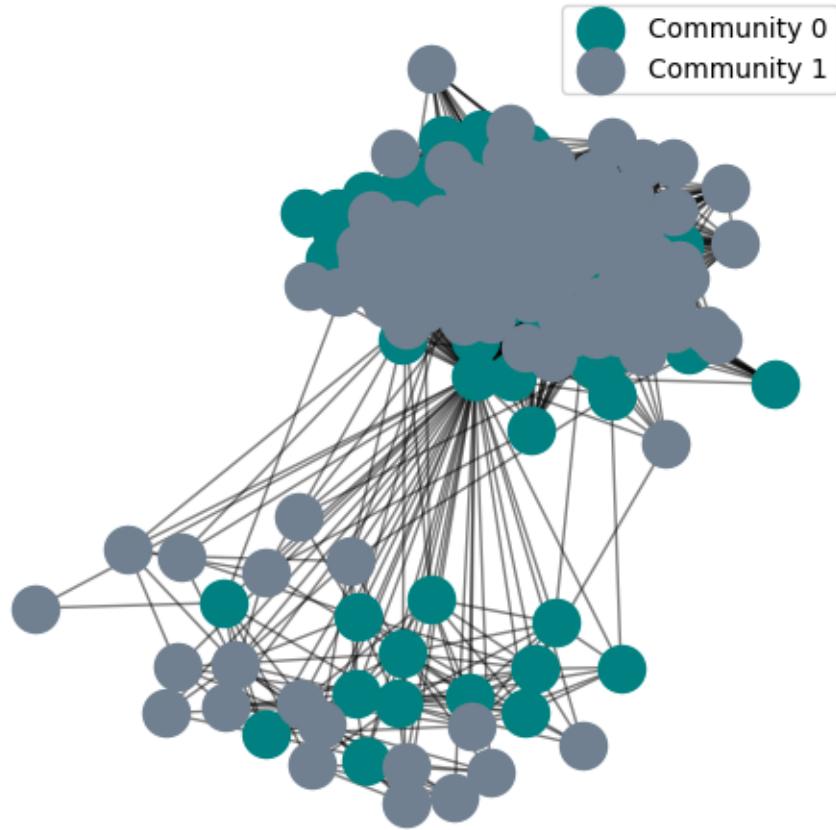
Infomap (Core 483)



Infomap modularity (core removed): 0.5434

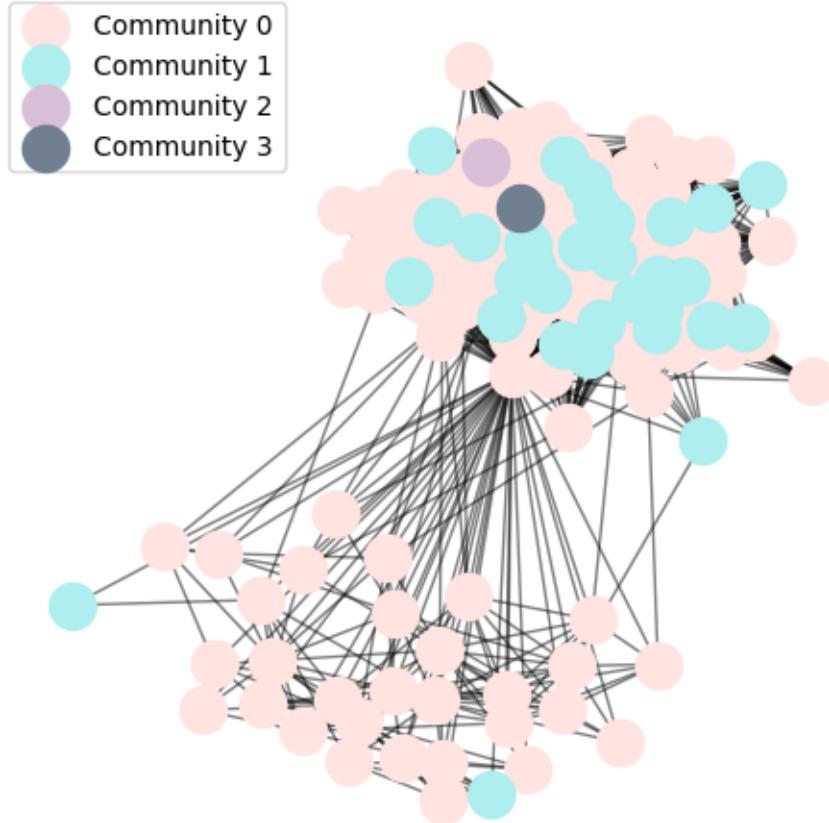
==== Core node 1086 (Core node removed) ===

Fast-Greedy (Core 1086)

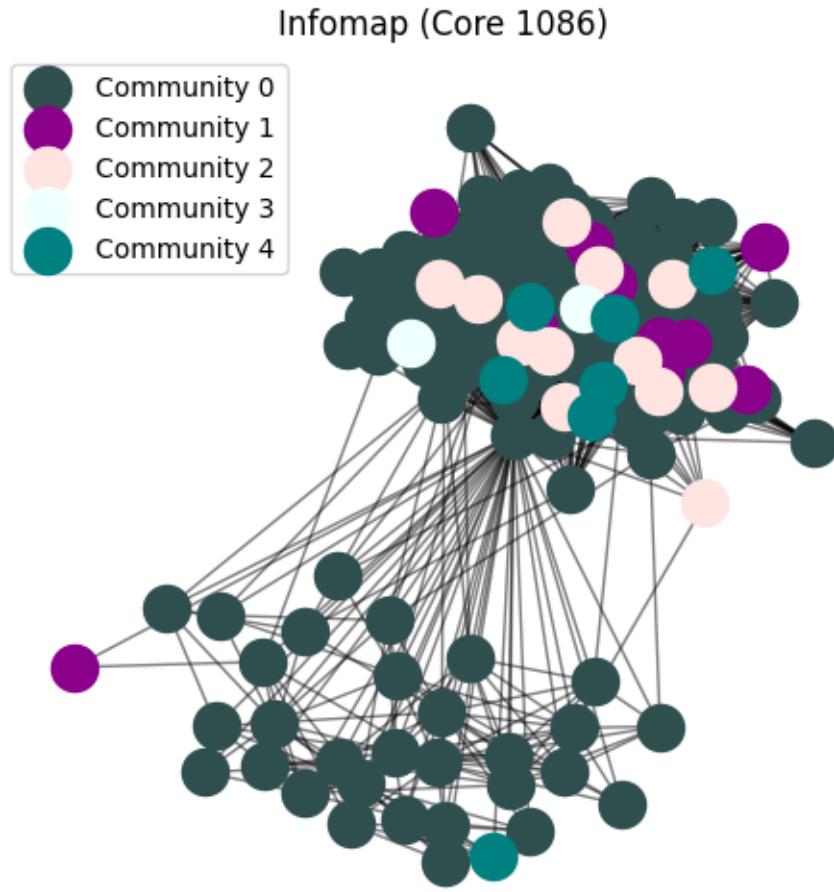


Fast-Greedy modularity (core removed): 0.1482

Edge-Betweenness (Core 1086)



Edge-Betweenness modularity (core removed): 0.0325



Infomap modularity (core removed): 0.0274

9.1 Results:

==== Core Node 1 ====

Fast-Greedy modularity (core removed): 0.4459

Edge-Betweenness modularity (core removed): 0.4161

Infomap modularity (core removed): 0.4180

==== Core Node 108 ====

Fast-Greedy modularity (core removed): 0.4580

Edge-Betweenness modularity (core removed): 0.5213

Infomap modularity (core removed): 0.5186

==== Core Node 349 ====

Fast-Greedy modularity (core removed): 0.2457
 Edge-Betweenness modularity (core removed): 0.1506
 Infomap modularity (core removed): 0.2466
 === Core Node 484 ===
 Fast-Greedy modularity (core removed): 0.5342
 Edge-Betweenness modularity (core removed): 0.5154
 Infomap modularity (core removed): 0.5434
 === Core Node 1087 ===
 Fast-Greedy modularity (core removed): 0.1482
 Edge-Betweenness modularity (core removed): 0.0325
 Infomap modularity (core removed): 0.0274

9.2 Modified Personalized Network Modularity Score Comparison:

10 Question 11

In general, the embeddedness between a core node v_{core} and a non-core node v_i is the number of mutual neighbors they share:

$$\text{Embeddedness}(v_i, v_{\text{core}}) = |\text{Neighbors}(v_i) \cap \text{Neighbors}(v_{\text{core}})|$$

In the personalized network (ego graph), since all non-core nodes are directly connected to the core node, the embeddedness simplifies to:

$$\text{Embeddedness}(v_i, v_{\text{core}}) = \text{Degree}(v_i) - 1$$

11 Question 12

```
[69]: # Code for Q12

# Compute embeddedness
def compute_embeddedness(G_personal, core_node, target_node):
    core_neighbors = set(G_personal.neighbors(core_node))
    target_neighbors = set(G_personal.neighbors(target_node))
    mutual_friends = core_neighbors.intersection(target_neighbors)
    return len(mutual_friends)

# Compute dispersion
def compute_dispersion(G_personal, core_node, target_node):
    # Find mutual friends
    core_neighbors = set(G_personal.neighbors(core_node))
    target_neighbors = set(G_personal.neighbors(target_node))
    mutual_friends = list(core_neighbors.intersection(target_neighbors))
```

```

if len(mutual_friends) < 2:
    return 0 # Dispersion is zero if fewer than 2 mutual friends

# Create modified graph
G_modified = G_personal.copy()
G_modified.remove_node(core_node)
G_modified.remove_node(target_node)

dispersion = 0
for i in range(len(mutual_friends)):
    for j in range(i+1, len(mutual_friends)):
        try:
            dist = nx.shortest_path_length(G_modified, mutual_friends[i], mutual_friends[j])
            dispersion += dist
        except nx.NetworkXNoPath:
            # If no path exists, skip
            pass

return dispersion

```

```

[71]: for core in core_nodes:
    G_personal = nx.ego_graph(G, core)
    neighbors = list(G_personal.neighbors(core))

    embeddedness_list = []
    dispersion_list = []

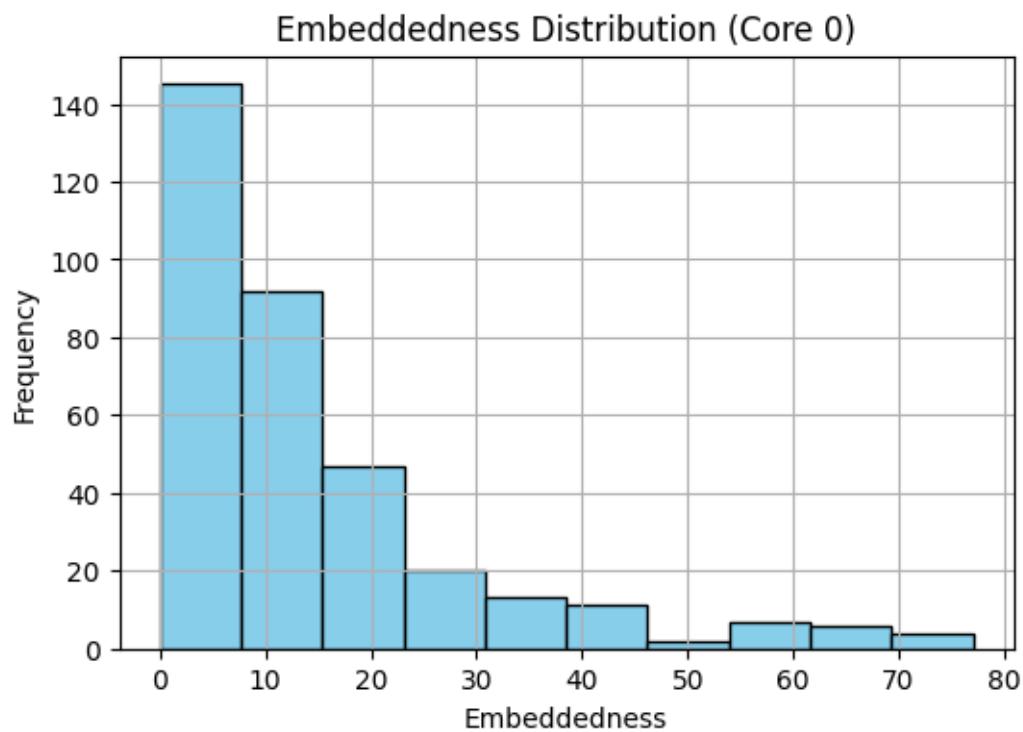
    for v in neighbors:
        emb = compute_embeddedness(G_personal, core, v)
        disp = compute_dispersion(G_personal, core, v)
        embeddedness_list.append(emb)
        dispersion_list.append(disp)

    # Plot Embeddedness Histogram
    plt.figure(figsize=(6, 4))
    plt.hist(embeddedness_list, bins=10, color='skyblue', edgecolor='black')
    plt.title(f"Embeddedness Distribution (Core {core})")
    plt.xlabel("Embeddedness")
    plt.ylabel("Frequency")
    plt.grid(True)
    plt.show()

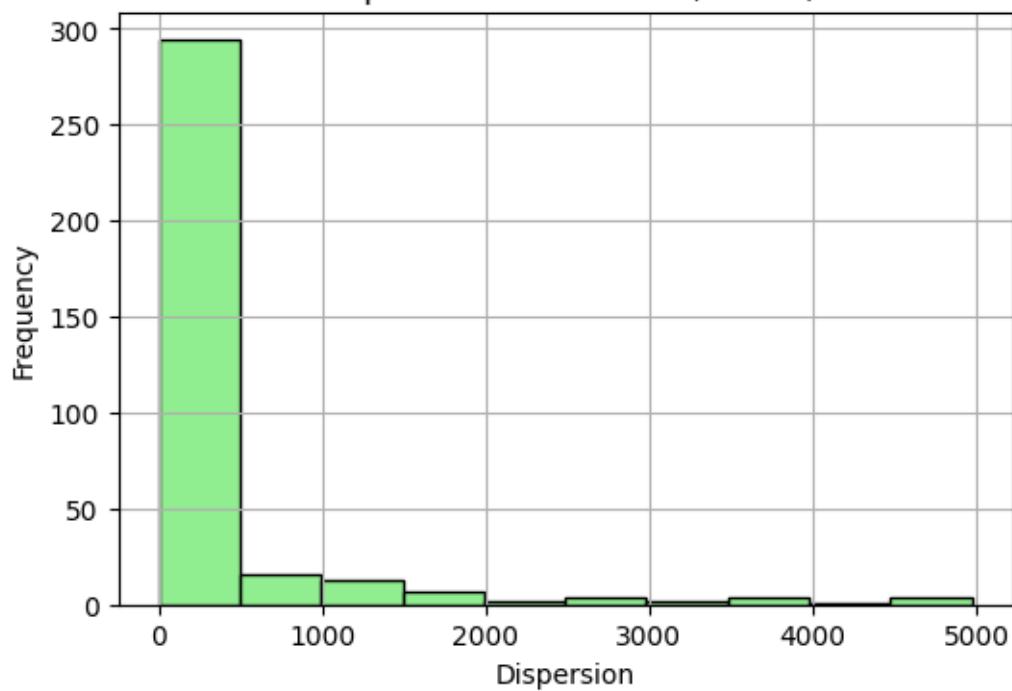
    # Plot Dispersion Histogram
    plt.figure(figsize=(6, 4))
    plt.hist(dispersion_list, bins=10, color='lightgreen', edgecolor='black')

```

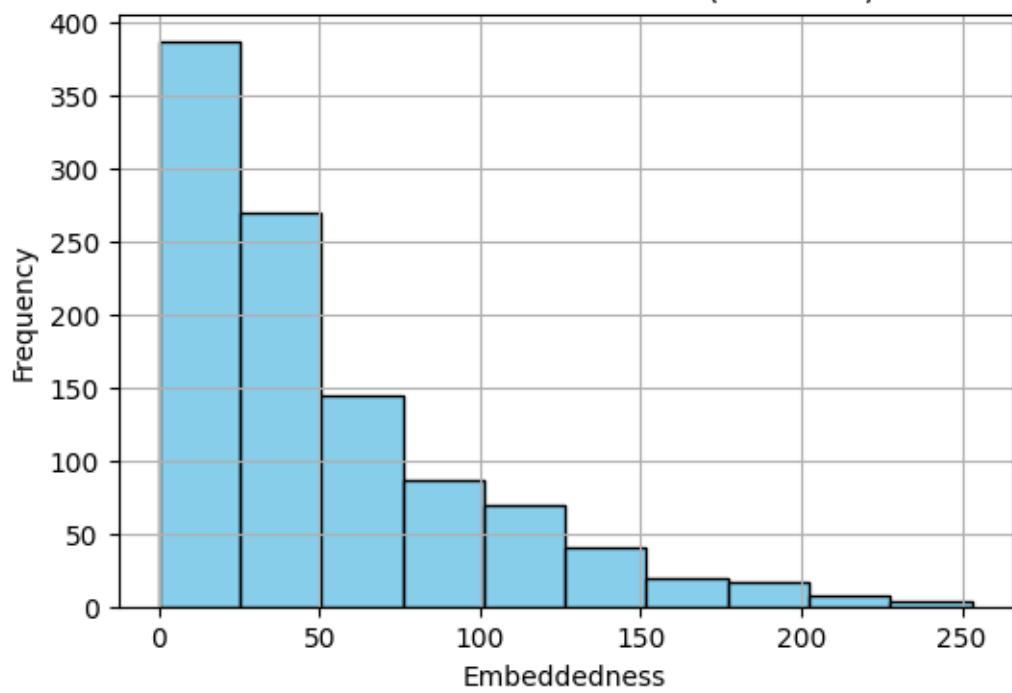
```
plt.title(f"Dispersion Distribution (Core {core})")
plt.xlabel("Dispersion")
plt.ylabel("Frequency")
plt.grid(True)
plt.show()
```



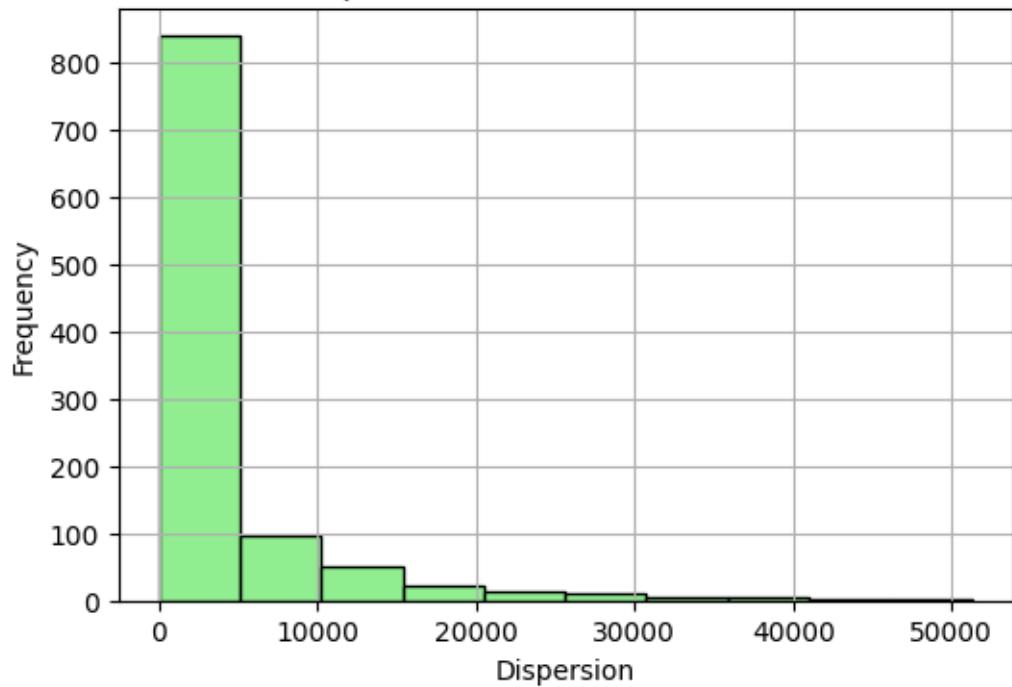
Dispersion Distribution (Core 0)



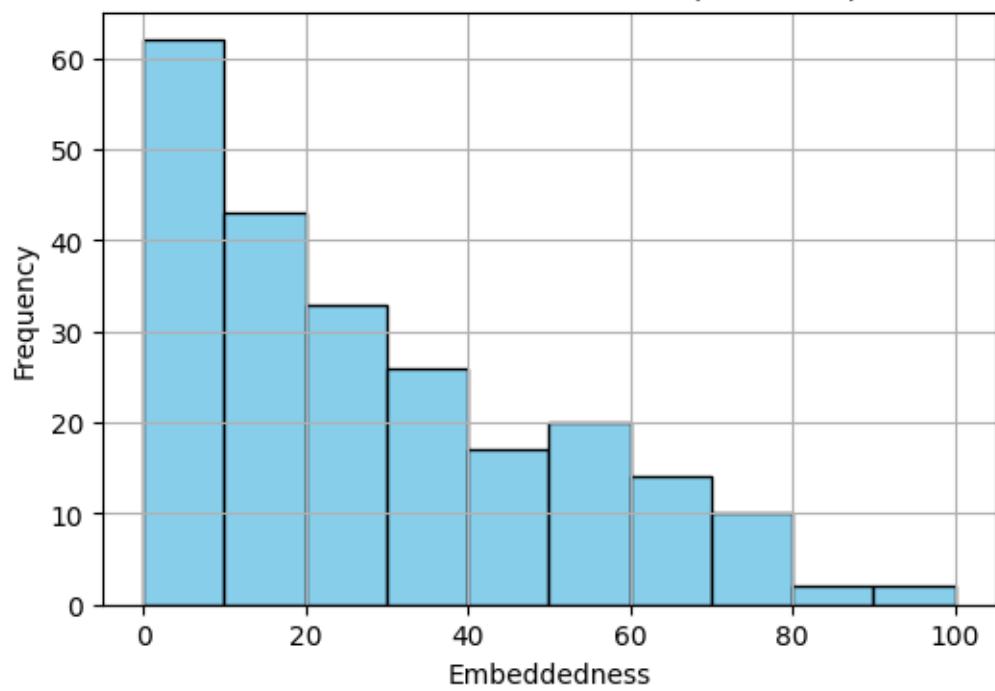
Embeddedness Distribution (Core 107)



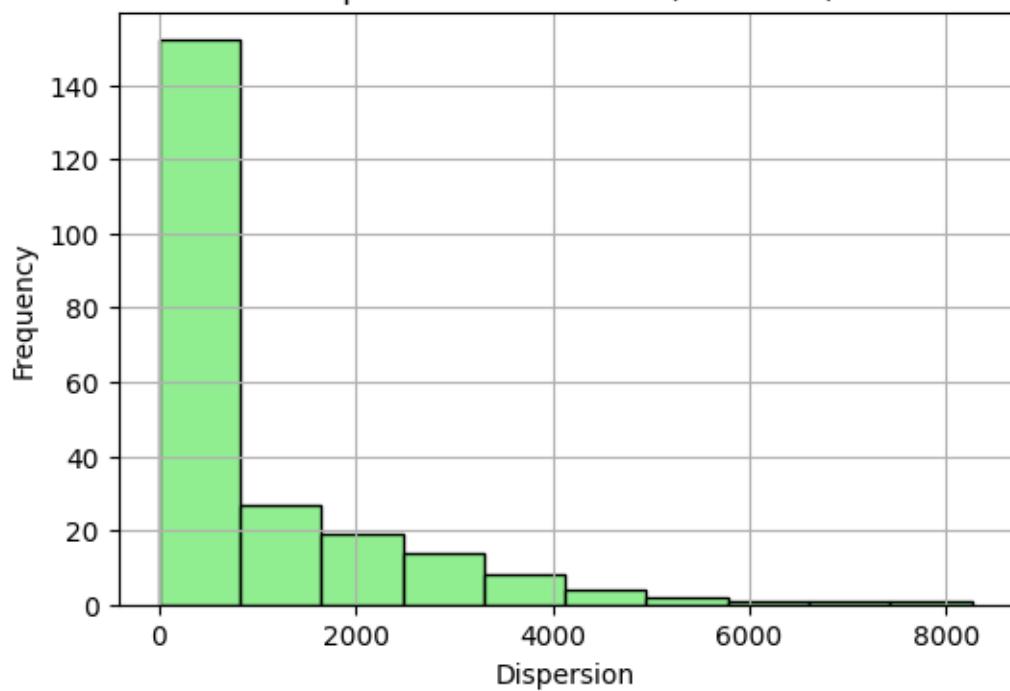
Dispersion Distribution (Core 107)



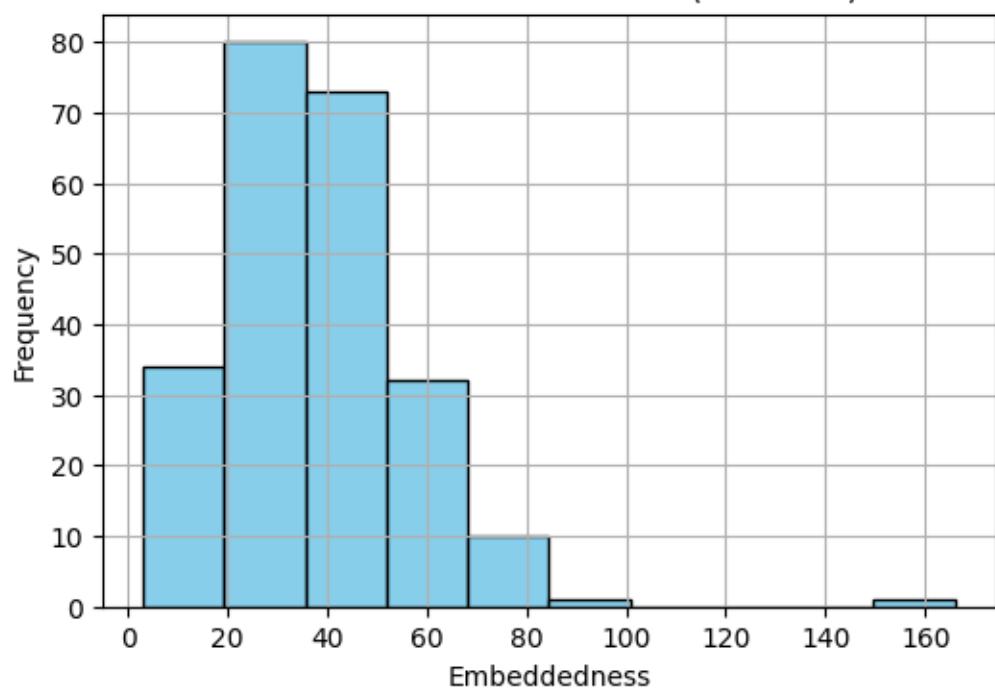
Embeddedness Distribution (Core 348)



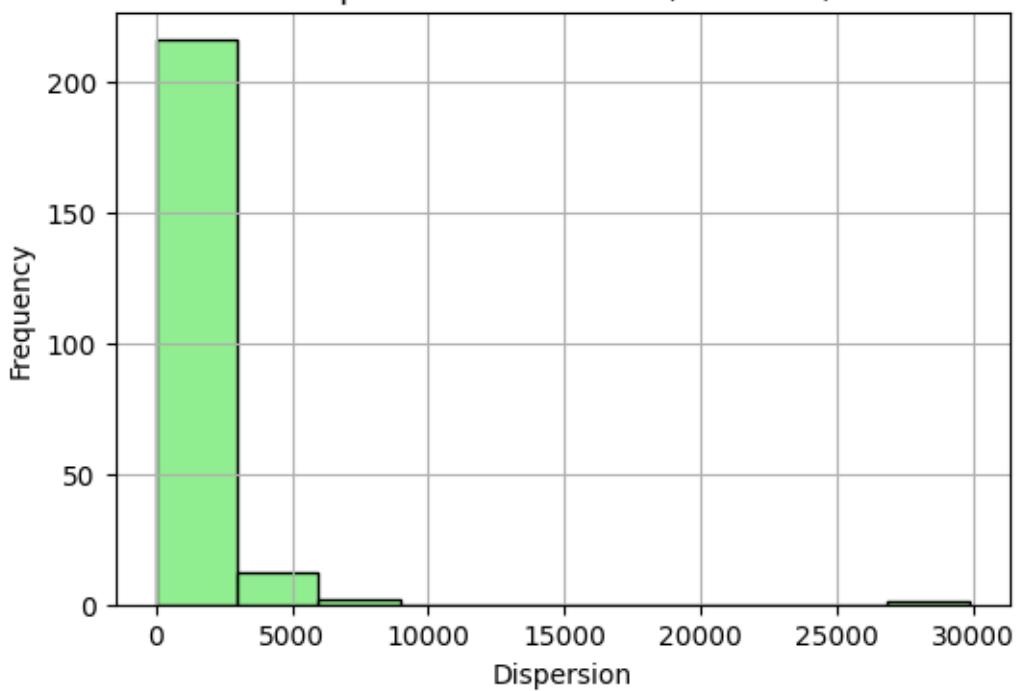
Dispersion Distribution (Core 348)



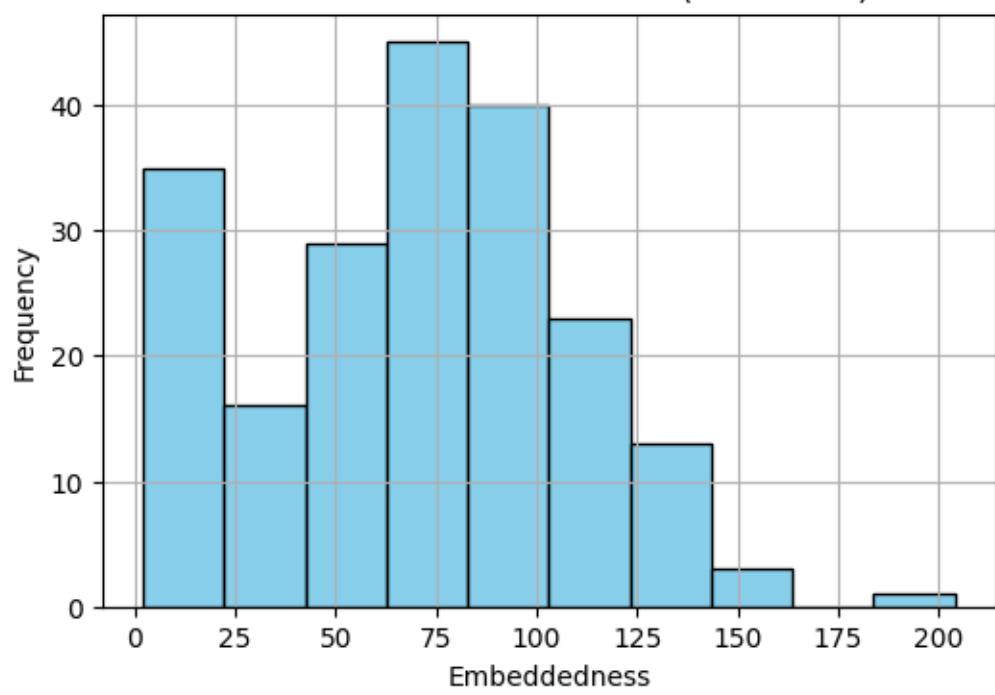
Embeddedness Distribution (Core 483)

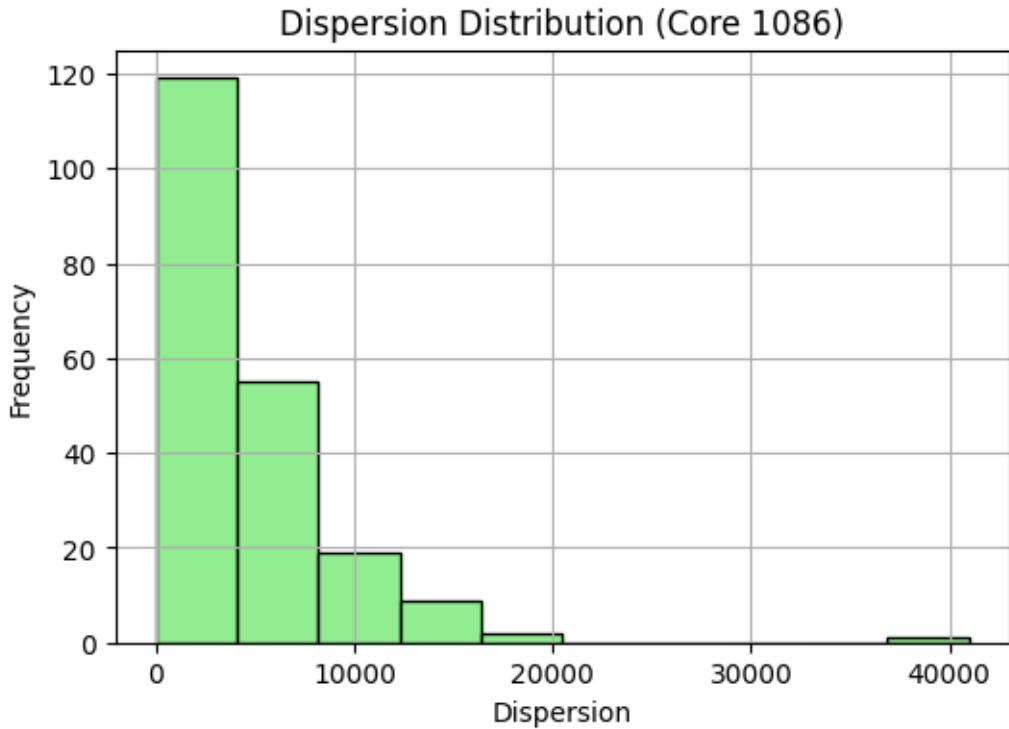


Dispersion Distribution (Core 483)



Embeddedness Distribution (Core 1086)





[92]: # Code for Q13

```
# Fast Greedy Membership Function
def fast_greedy_no_plot(G_personal):
    G_ig = ig.Graph.TupleList(G_personal.edges(), directed=False)
    fg_clusters = G_ig.community_fastgreedy().as_clustering()
    return fg_clusters.membership
```

[94]: def plot_communities_highlight_two_nodes(G_personal, membership, ↴max_embeddedness_node, max_ratio_node):
 pos = nx.spring_layout(G_personal, seed=42, k=0.3, iterations=100)

 communities = {}
 for idx, cluster_id in enumerate(membership):
 node = list(G_personal.nodes())[idx]
 communities.setdefault(cluster_id, []).append(node)

 colors = list(mcolors.TABLEAU_COLORS.values()) + list(mcolors.CSS4_COLORS. ↴values())
 random.shuffle(colors)

 plt.figure(figsize=(7, 7))

```

# Draw normal edges first
all_edges = set(G_personal.edges())
incident_edges_embed = set(G_personal.edges(max_embeddedness_node))
incident_edges_ratio = set(G_personal.edges(max_ratio_node))
all_highlight_edges = incident_edges_embed.union(incident_edges_ratio)
normal_edges = list(all_edges - all_highlight_edges)

nx.draw_networkx_edges(G_personal, pos, edgelist=normal_edges, alpha=0.4)

# Draw normal nodes
for cluster_id, nodes in communities.items():
    nx.draw_networkx_nodes(G_personal, pos, nodelist=nodes,
                           node_color=colors[cluster_id % len(colors)], ▾
                           alpha=0.8, node_size=80)

# Highlight embeddedness node
nx.draw_networkx_nodes(G_personal, pos, nodelist=[max_embeddedness_node],
                       node_color='blue', node_size=300, ▾
                       edgecolors='black', linewidths=2, label="Max Embeddedness Node")

# Highlight ratio node
nx.draw_networkx_nodes(G_personal, pos, nodelist=[max_ratio_node],
                       node_color='red', node_size=300, edgecolors='black', ▾
                       linewidths=2, label="Max Dispersion/Embeddedness Node")

# Highlight incident edges
nx.draw_networkx_edges(G_personal, pos, ▾
                       edgelist=list(incident_edges_embed), width=3, edge_color='blue')
nx.draw_networkx_edges(G_personal, pos, ▾
                       edgelist=list(incident_edges_ratio), width=3, edge_color='red')

plt.title("Community Structure: Max Embeddedness (Blue) and Max Dispersion/
          Embeddedness (Red)")
plt.axis('off')
plt.legend()
plt.show()

```

[95]: # Main loop

```

for core in core_nodes:
    G_personal = nx.ego_graph(G, core)
    membership = fast_greedy_no_plot(G_personal)

    embeddednesses = []
    dispersions = []
    ratios = []

```

```

for neighbor in G_personal.neighbors(core):
    emb = compute_embeddedness(G_personal, core, neighbor)
    disp = compute_dispersion(G_personal, core, neighbor)

    embeddednesses[neighbor] = emb
    dispersions[neighbor] = disp

    if emb > 0:
        ratios[neighbor] = disp / emb # Only define ratio if embeddedness
        ↪> 0

# Find max embeddedness node
valid_embeds = {node: val for node, val in embeddednesses.items() if val >
    ↪0}
max_embed_node = max(valid_embeds, key=valid_embeds.get)

# Find max dispersion/embeddedness node
max_ratio_node = max(ratios, key=ratios.get)

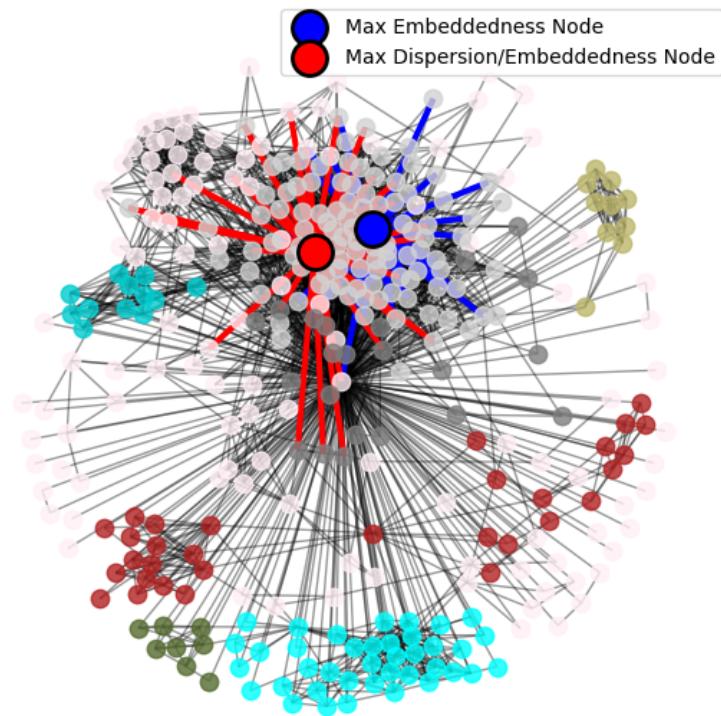
print(f"Core Node {core+1}: Max Embeddedness Node = {max_embed_node+1}, Max
    ↪Dispersion/Embeddedness Node = {max_ratio_node+1}")

plot_communities_highlight_two_nodes(G_personal, membership,
    ↪max_embed_node, max_ratio_node)

```

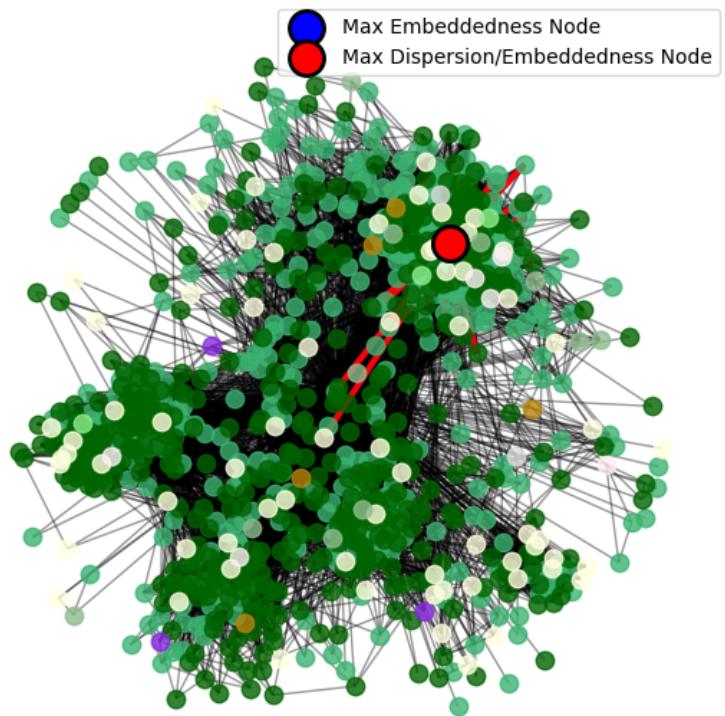
Core Node 1: Max Embeddedness Node = 57, Max Dispersion/Embeddedness Node = 26

Community Structure: Max Embeddedness (Blue) and Max Dispersion/Embeddedness (Red)



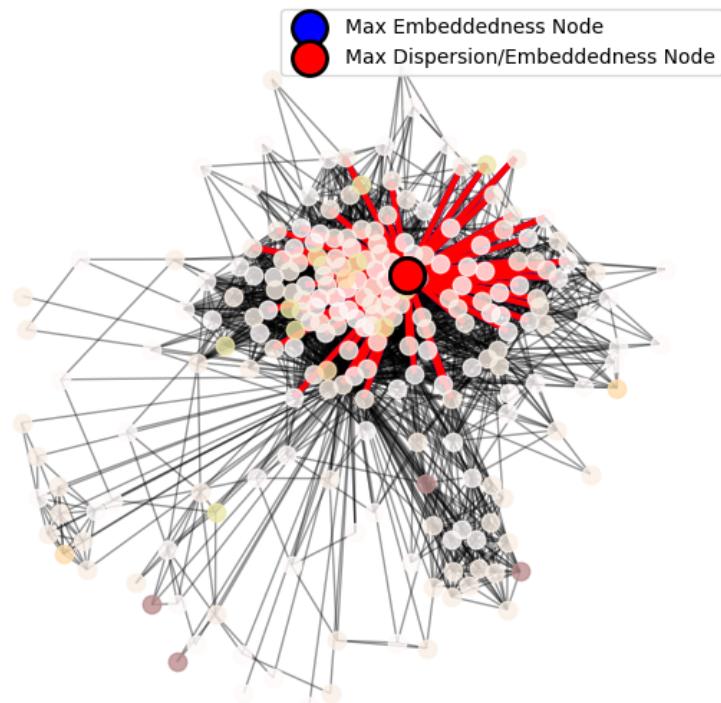
Core Node 108: Max Embeddedness Node = 1889, Max Dispersion/Embeddedness Node = 1889

Community Structure: Max Embeddedness (Blue) and Max Dispersion/Embeddedness (Red)



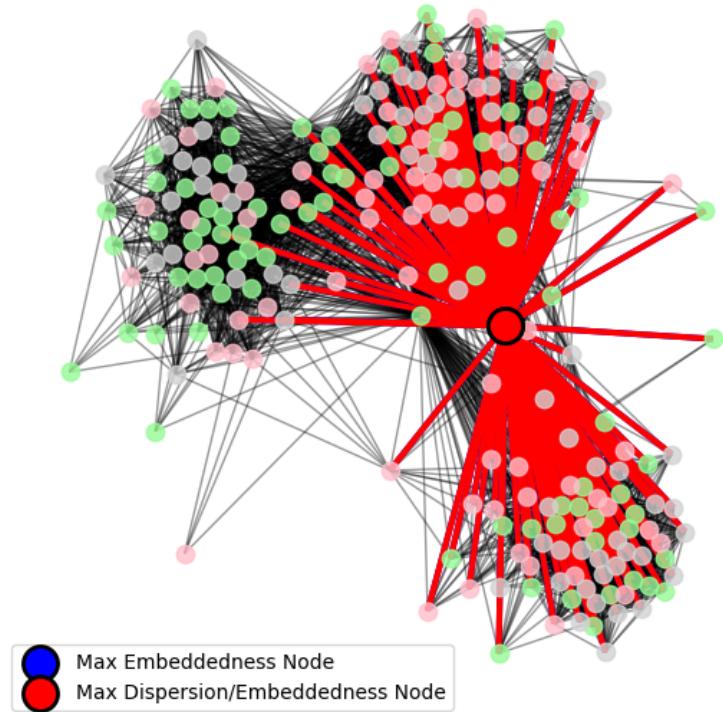
Core Node 349: Max Embeddedness Node = 377, Max Dispersion/Embeddedness Node = 377

Community Structure: Max Embeddedness (Blue) and Max Dispersion/Embeddedness (Red)



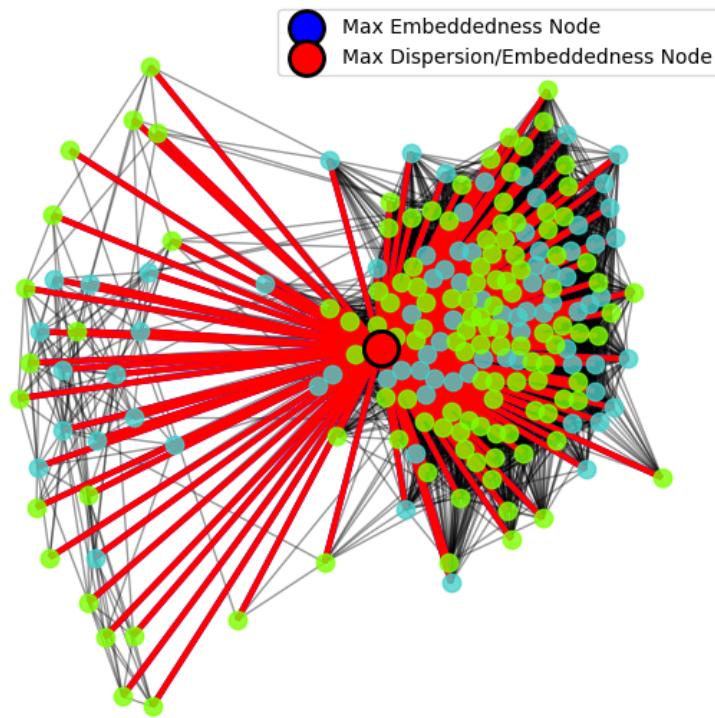
Core Node 484: Max Embeddedness Node = 108, Max Dispersion/Embeddedness Node = 108

Community Structure: Max Embeddedness (Blue) and Max Dispersion/Embeddedness (Red)



Core Node 1087: Max Embeddedness Node = 108, Max Dispersion/Embeddedness Node = 108

Community Structure: Max Embeddedness (Blue) and Max Dispersion/Embeddedness (Red)



12 Question 13

```
[ ]: # Code for Q13

# Fast Greedy Membership Function
def fast_greedy_no_plot(G_personal):
    G_ig = ig.Graph.TupleList(G_personal.edges(), directed=False)
    fg_clusters = G_ig.community_fastgreedy().as_clustering()
    return fg_clusters.membership

[98]: def plot_communities(G_personal, membership, highlight_nodes_colors, title):
    """
    highlight_nodes_colors: list of tuples (node, color)
    e.g., [(node1, "red"), (node2, "blue")]
    """

    pos = nx.spring_layout(G_personal, seed=42, k=0.3, iterations=100)

    communities = {}
    for idx, cluster_id in enumerate(membership):
        node = list(G_personal.nodes())[idx]
```

```

        communities.setdefault(cluster_id, []).append(node)

    colors = list(mcolors.TABLEAU_COLORS.values()) + list(mcolors.CSS4_COLORS.
    ↪values())
    random.shuffle(colors)

    plt.figure(figsize=(7, 7))

    # Draw normal edges first
    all_edges = set(G_personal.edges())
    highlighted_edges = set()
    for node, _ in highlight_nodes_colors:
        highlighted_edges.update(G_personal.edges(node))

    normal_edges = list(all_edges - highlighted_edges)
    nx.draw_networkx_edges(G_personal, pos, edgelist=normal_edges, alpha=0.4)

    # Draw normal nodes
    for cluster_id, nodes in communities.items():
        nx.draw_networkx_nodes(G_personal, pos, nodelist=nodes,
                               node_color=colors[cluster_id % len(colors)], ↪
    ↪alpha=0.8, node_size=80)

    # Draw highlighted nodes
    for node, color in highlight_nodes_colors:
        nx.draw_networkx_nodes(G_personal, pos, nodelist=[node],
                               node_color=color, node_size=300, ↪
    ↪edgecolors='black', linewidths=2, label=f'{color} node')

    # Draw highlighted edges last
    for node, color in highlight_nodes_colors:
        incident_edges = list(G_personal.edges(node))
        nx.draw_networkx_edges(G_personal, pos, edgelist=incident_edges, ↪
    ↪width=3, edge_color=color)

    plt.title(title)
    plt.axis('off')
    plt.legend()
    plt.show()

```

```

[101]: for core in core_nodes:
    G_personal = nx.ego_graph(G, core)
    membership = fast_greedy_no_plot(G_personal)

    dispersions = {}

    for neighbor in G_personal.neighbors(core):

```

```

dispersions[neighbor] = compute_dispersion(G_personal, core, neighbor)

# Find node with maximum dispersion
max_disp_node = max(dispersions, key=dispersions.get)

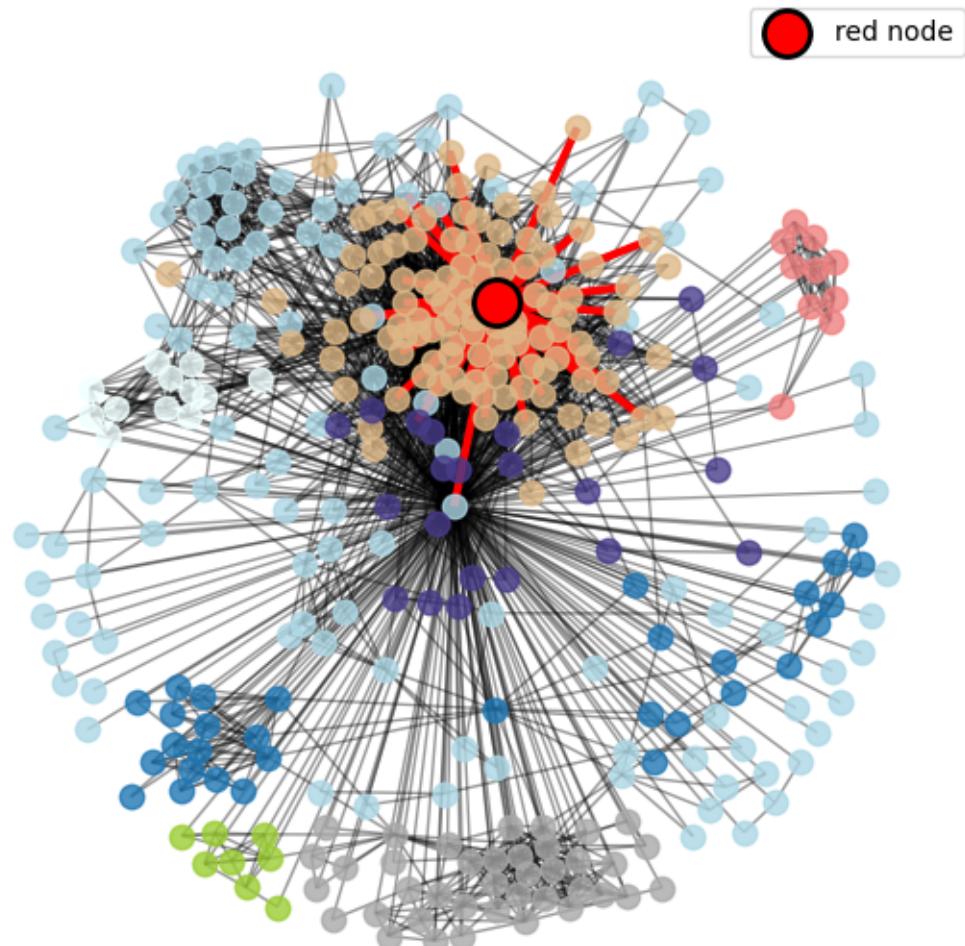
# Print node ID
print(f"Core Node {core+1}: Max Dispersion Node = {max_disp_node+1}")

# Plot
plot_communities(G_personal, membership,
                  highlight_nodes_colors=[(max_disp_node, "red")],
                  title=f"Q13 - Core {core+1}: Max Dispersion Node")

```

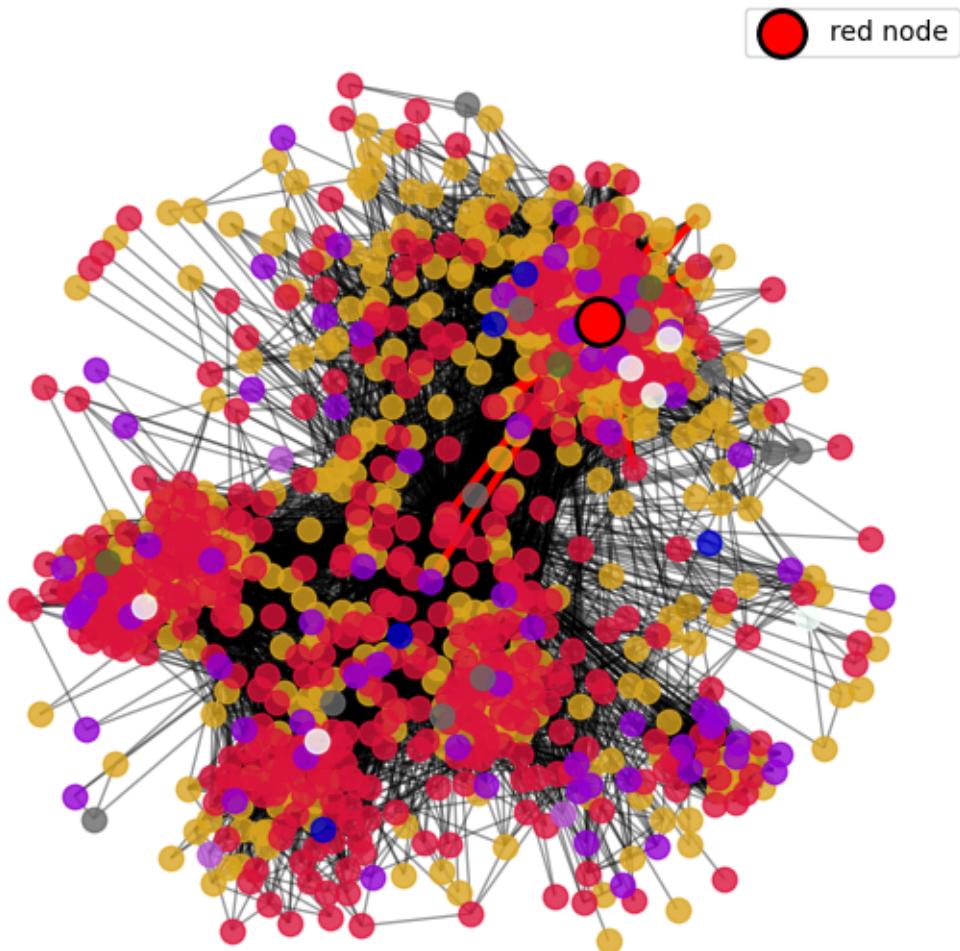
Core Node 1: Max Dispersion Node = 57

Q13 - Core 1: Max Dispersion Node



Core Node 108: Max Dispersion Node = 1889

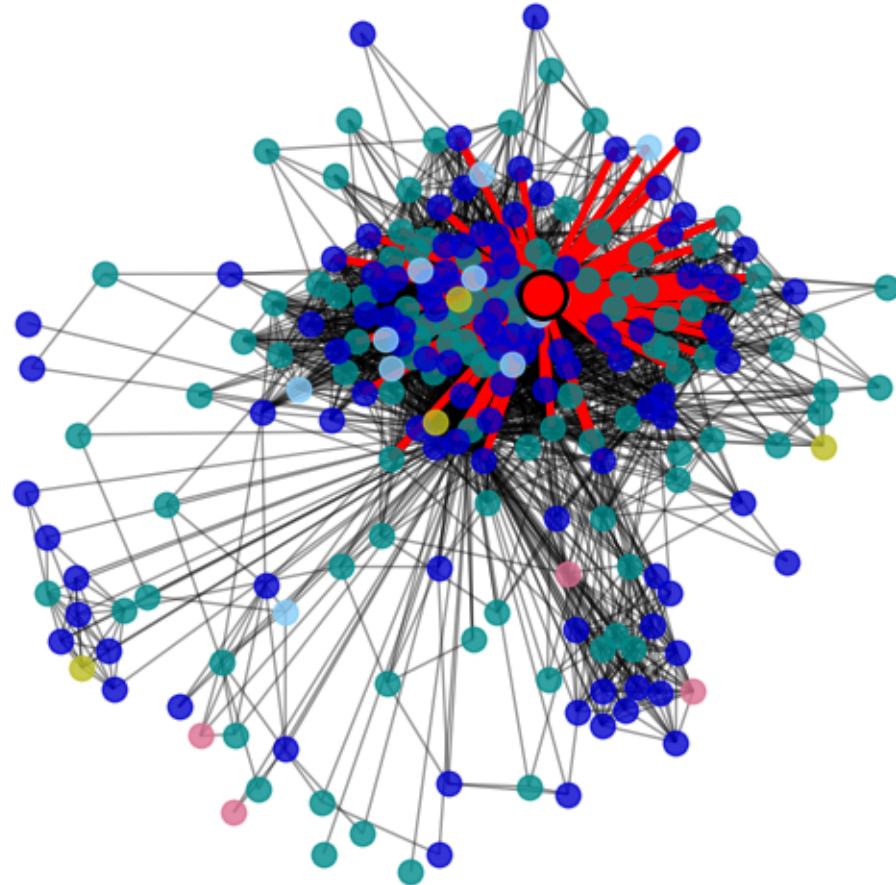
Q13 - Core 108: Max Dispersion Node



Core Node 349: Max Dispersion Node = 377

Q13 - Core 349: Max Dispersion Node

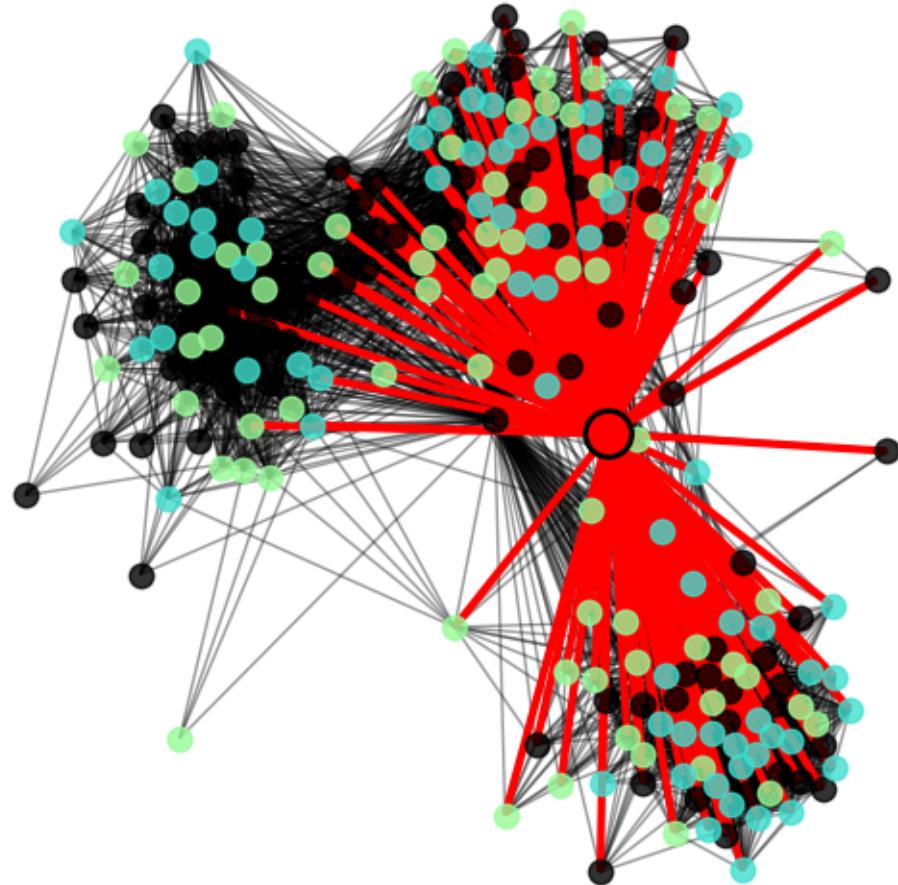
 red node



Core Node 484: Max Dispersion Node = 108

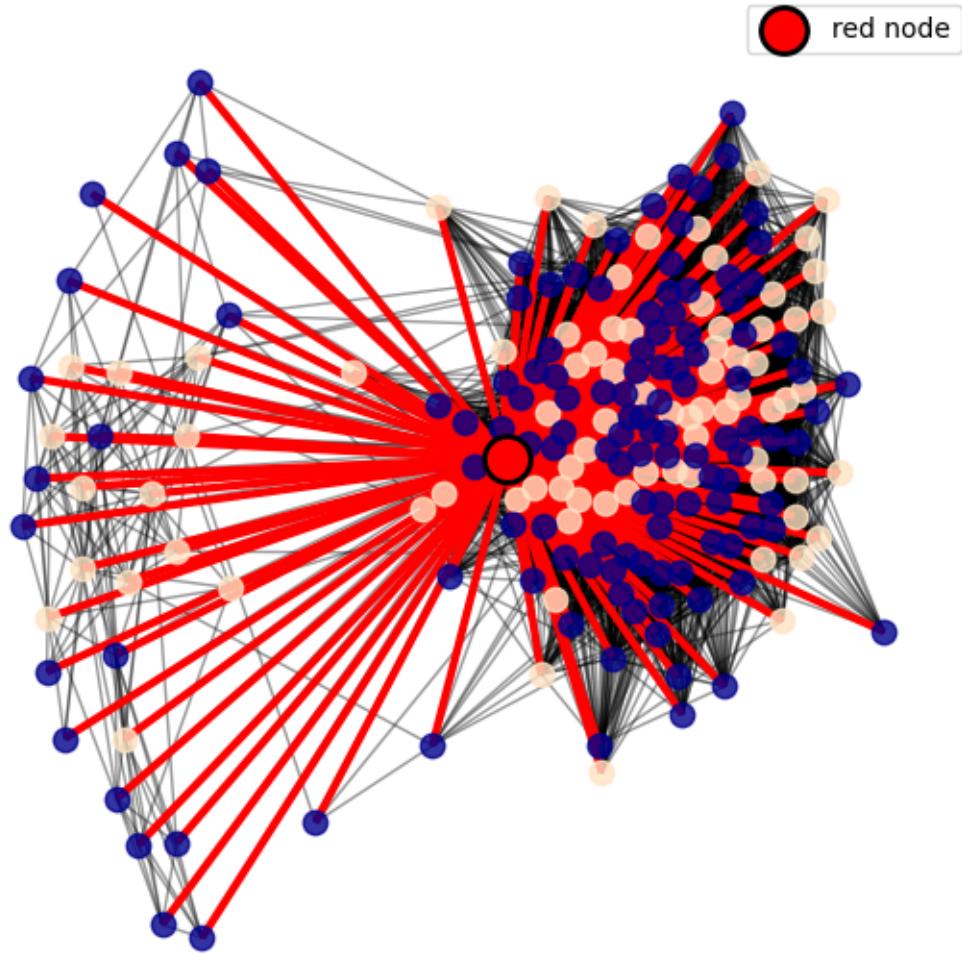
Q13 - Core 484: Max Dispersion Node

 red node



Core Node 1087: Max Dispersion Node = 108

Q13 - Core 1087: Max Dispersion Node



13 Question 14

```
[100]: for core in core_nodes:  
    G_personal = nx.ego_graph(G, core)  
    membership = fast_greedy_no_plot(G_personal)  
  
    embeddednesses = []  
    dispersions = []  
    ratios = []  
  
    for neighbor in G_personal.neighbors(core):  
        emb = compute_embeddedness(G_personal, core, neighbor)
```

```

disp = compute_dispersion(G_personal, core, neighbor)

embeddednesses[neighbor] = emb
dispersions[neighbor] = disp

if emb > 0:
    ratios[neighbor] = disp / emb

# Find node with maximum embeddedness
valid_embeds = {node: val for node, val in embeddednesses.items() if val > 0}
max_embed_node = max(valid_embeds, key=valid_embeds.get)

# Find node with maximum (dispersion / embeddedness)
max_ratio_node = max(ratios, key=ratios.get)

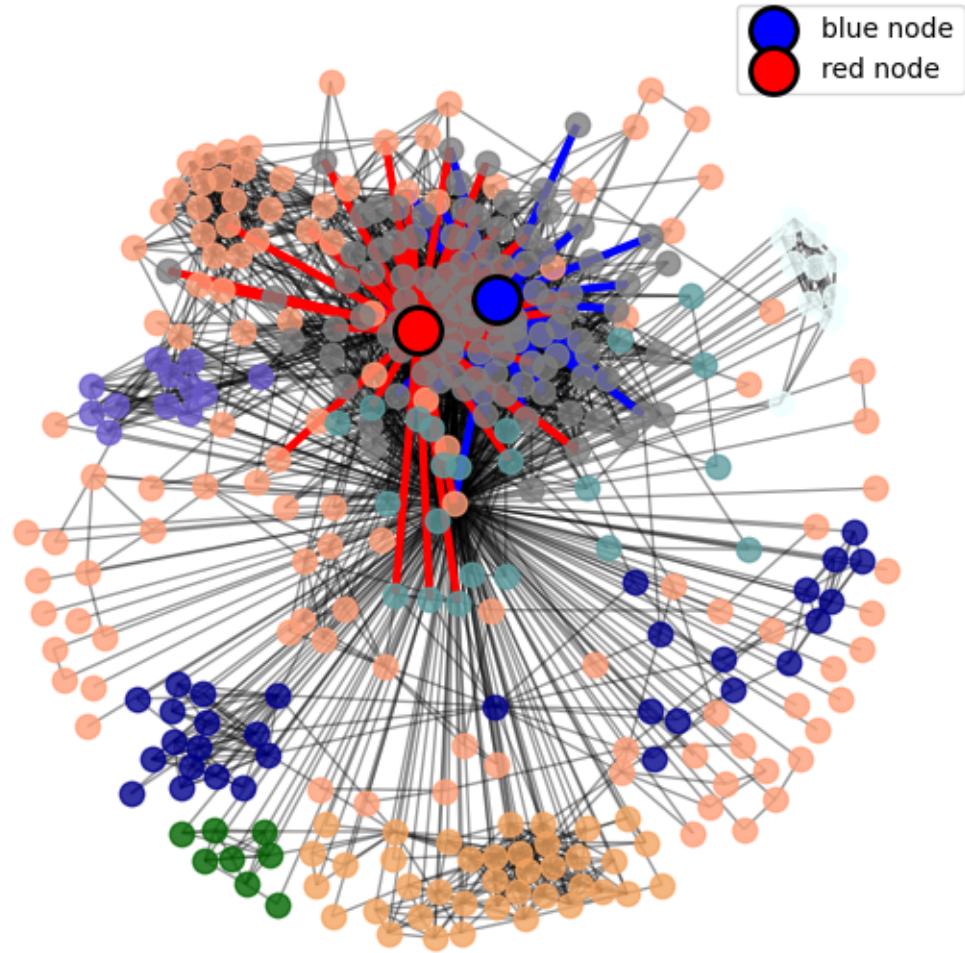
print(f"Core Node {core+1}: Max Embeddedness Node = {max_embed_node+1}, Max Disp/Embed Node = {max_ratio_node+1}")

# Plot
plot_communities(G_personal, membership,
                  highlight_nodes_colors=[(max_embed_node, "blue"), (max_ratio_node, "red")],
                  title=f"Q14 - Core {core+1}: Max Embed (Blue), Max Disp/Embed (Red)")

```

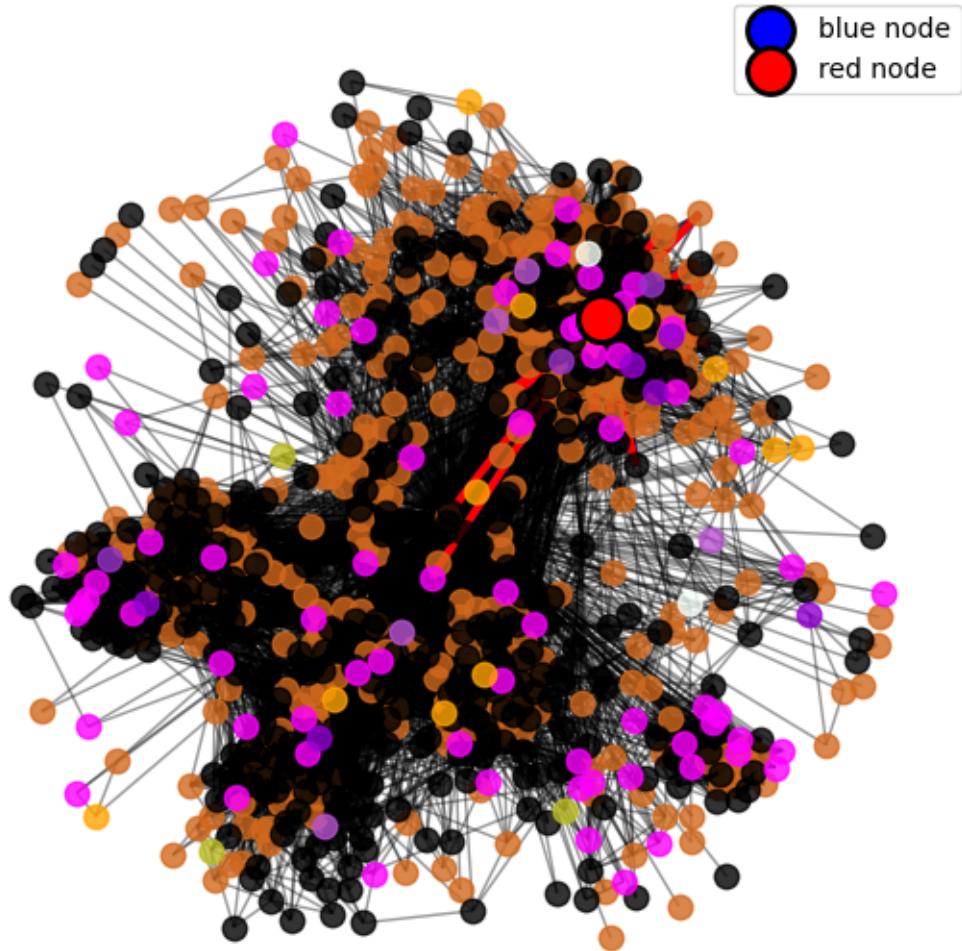
Core Node 1: Max Embeddedness Node = 57, Max Dispersion/Embeddedness Node = 26

Q14 - Core 1: Max Embed (Blue), Max Disp/Embed (Red)



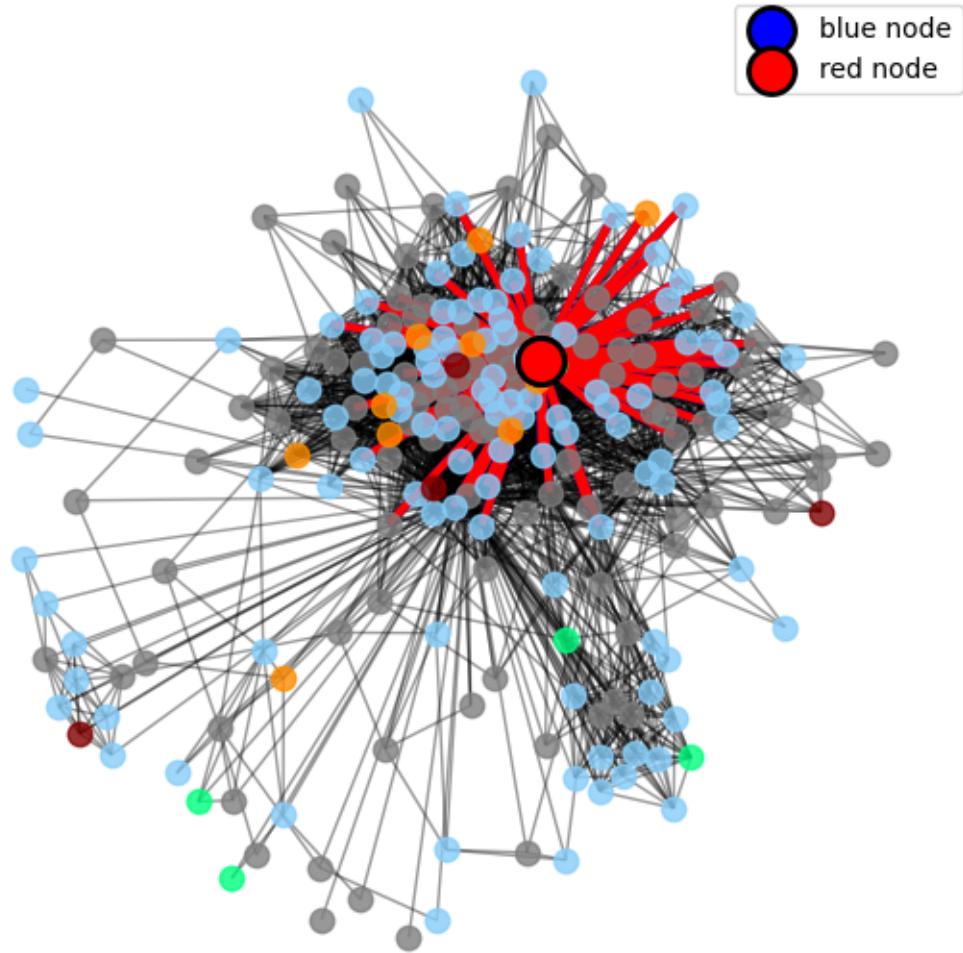
Core Node 108: Max Embeddedness Node = 1889, Max Dispersion/Embeddedness Node = 1889

Q14 - Core 108: Max Embed (Blue), Max Disp/Embed (Red)



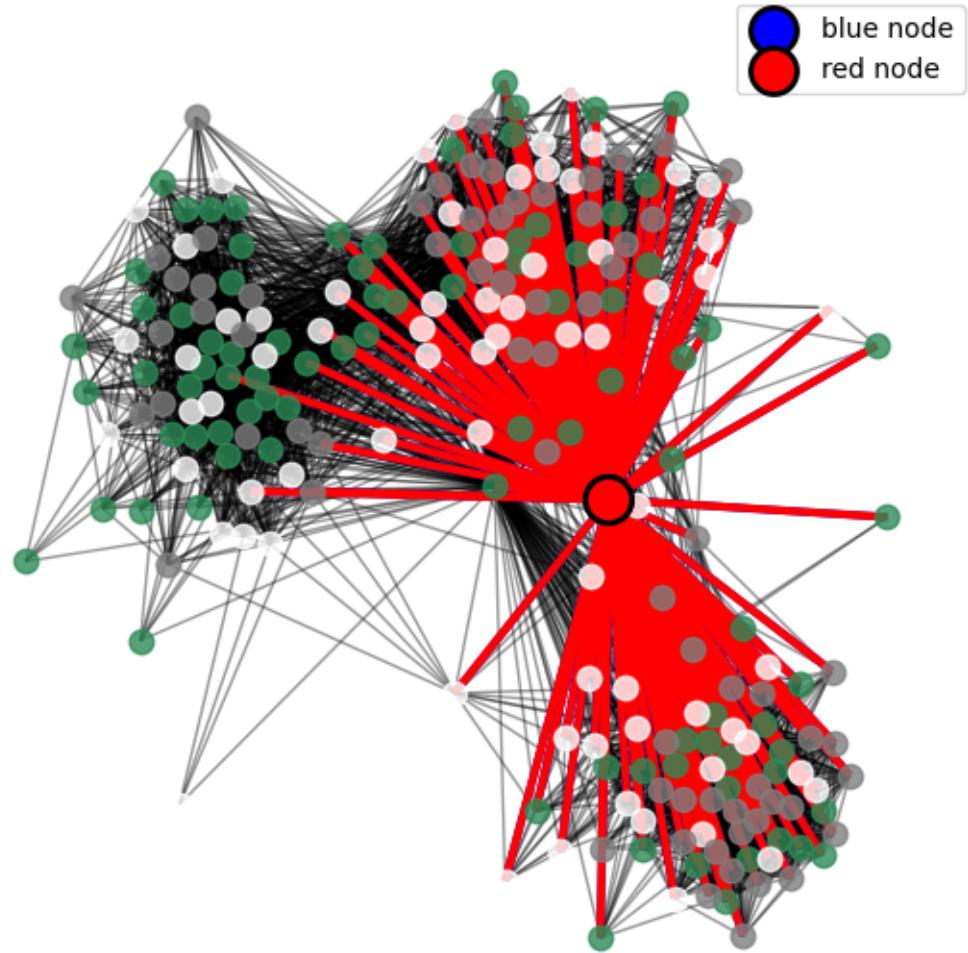
Core Node 349: Max Embeddedness Node = 377, Max Dispersion/Embeddedness Node = 377

Q14 - Core 349: Max Embed (Blue), Max Disp/Embed (Red)



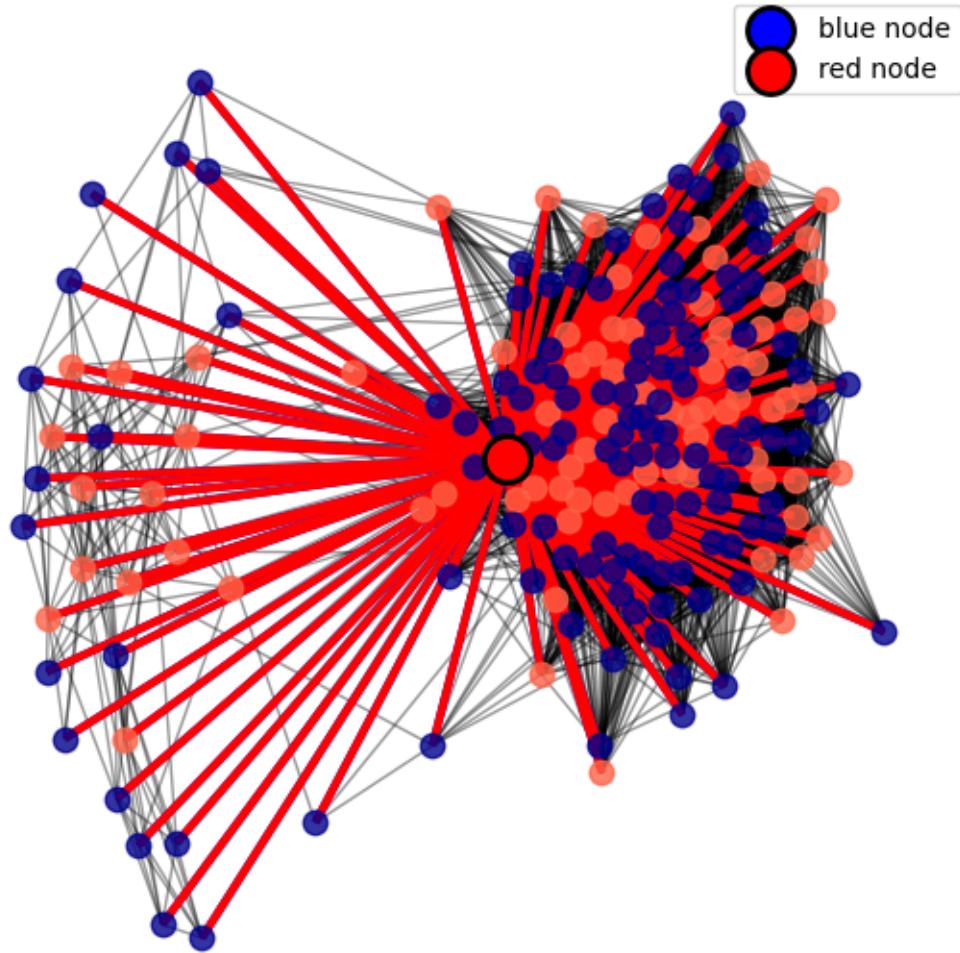
Core Node 484: Max Embeddedness Node = 108, Max Dispersion/Embeddedness Node = 108

Q14 - Core 484: Max Embed (Blue), Max Disp/Embed (Red)



Core Node 1087: Max Embeddedness Node = 108, Max Dispersion/Embeddedness Node = 108

Q14 - Core 1087: Max Embed (Blue), Max Disp/Embed (Red)



14 Question 15

Core Node	Maximum Dispersion	Maximum Embeddedness	Maximum Dispersion/Embeddedness
1	57	57	26
108	1889	1889	1889
349	377	377	377
484	108	108	108
1087	108	108	108

Dispersion quantifies the distance between a node's mutual friends and the core node after the node

and core node are removed. In our plots and results, nodes such as for Core 1 57 with maximum dispersion are considered structural bridges in the network.

Embeddedness captures the amount of mutual friends relative to the core node, identifying nodes that are closely integrated in dense communities. Nodes like 1889 for Core 108 which had the highest embeddedness, displayed great local connectedness.

The ratio of dispersion to embeddedness singles out nodes with few mutual friends who are scattered across various communities. Nodes with high ratios like Node 26 for Core 1 are believed to connect distraught clusters with weak direct ties.

As a whole, dispersion picks out bridging nodes while embeddedness locates nodes deeply integrated in a cluster. The ratio shows nodes that connect communities.

15 Question 16

```
[108]: # Code for Q16

# Read Facebook combined list
G = nx.read_edgelist('facebook_combined.txt', create_using=nx.Graph(), ↴
                     nodetype=int)

# Create the personalized network
core_node = 415 - 1
G_415 = nx.ego_graph(G, node, radius=1)

# Find all nodes in G_415 with degree exactly 24
Nr = [node for node, degree in G_415.degree() if degree == 24]

# Print result
print(f"The length of the list Nr: {len(Nr)}")
```

The length of the list Nr: 11

16 Question 17

```
[112]: # Code for Q17

# Common Neighbors score
def common_neighbors_score(G, i, j):
    return len(list(nx.common_neighbors(G, i, j)))

# Jaccard score
def jaccard_score(G, i, j):
    try:
        neighbors_i = set(G.neighbors(i))
        neighbors_j = set(G.neighbors(j))
        return len(neighbors_i & neighbors_j) / len(neighbors_i | neighbors_j)
```

```

except ZeroDivisionError:
    return 0

# Adamic-Adar score
def adamic_adar_score(G, i, j):
    neighbors_i = set(G.neighbors(i))
    neighbors_j = set(G.neighbors(j))
    common = neighbors_i & neighbors_j
    score = 0
    for k in common:
        deg_k = len(list(G.neighbors(k)))
        if deg_k > 1:
            score += 1 / math.log(deg_k)
    return score

```

```

[120]: # Run one iteration of the recommendation algorithm
def run_one_iteration(G_personal, user, similarity_measure):
    neighbors = list(G_personal.neighbors(user))

    # Randomly delete neighbors with 25% probability
    Ri = [friend for friend in neighbors if random.random() < 0.25]

    if len(Ri) == 0:
        return None # Skip this iteration if no edges are removed

    G_temp = G_personal.copy()
    for friend in Ri:
        if G_temp.has_edge(user, friend):
            G_temp.remove_edge(user, friend)

    # Recommend |Ri| new friends
    candidates = set(G_temp.nodes()) - set(G_temp.neighbors(user)) - {user}
    if len(candidates) == 0:
        return None # No candidates to recommend

    scores = {}
    for candidate in candidates:
        score = similarity_measure(G_temp, user, candidate)
        scores[candidate] = score

    sorted_candidates = sorted(scores.items(), key=lambda x: x[1], reverse=True)
    top_candidates = [cand for cand, _ in sorted_candidates[:len(Ri)]]

    # Calculate accuracy
    intersection = set(top_candidates) & set(Ri)
    accuracy = len(intersection) / len(Ri)

```

```

    return accuracy

[132]: # evaluate accuracy across multiple iterations
def evaluate_accuracy(G_personal, Nr, similarity_measure, num_iterations=10):
    user_accuracies = []

    for user in Nr:
        accuracies = []
        for _ in range(num_iterations):
            acc = run_one_iteration(G_personal, user, similarity_measure)
            if acc is not None:
                accuracies.append(acc)
        if accuracies:
            user_accuracies.append(sum(accuracies) / len(accuracies))

    if user_accuracies:
        return sum(user_accuracies) / len(user_accuracies)
    else:
        return 0

acc_common = evaluate_accuracy(G_415, Nr, common_neighbors_score)
print(f"Average accuracy (Common Neighbors): {acc_common:.4f}")

acc_jaccard = evaluate_accuracy(G_415, Nr, jaccard_score)
print(f"Average accuracy (Jaccard): {acc_jaccard:.4f}")

acc_adamic = evaluate_accuracy(G_415, Nr, adamic_adar_score)
print(f"Average accuracy (Adamic-Adar): {acc_adamic:.4f}")

```

Average accuracy (Common Neighbors): 0.8540
Average accuracy (Jaccard): 0.7955
Average accuracy (Adamic-Adar): 0.8397

17 2: Extraction & Exploration of Google+ network Data

```

[ ]: from google.colab import drive
drive.mount('/content/drive')

Mounted at /content/drive

[ ]: # Dataset extraction
import os
import tarfile

drive_path = '/content/drive/My Drive/gplus.tar.gz'
extract_dir = 'gplus'

```

```

# Check if file exists
if os.path.exists(drive_path):
    print("File found, extracting...")
    os.makedirs(extract_dir, exist_ok=True)
    with tarfile.open(drive_path, "r:gz") as tar:
        tar.extractall(extract_dir)
    print("Extraction complete.")
else:
    print("File not found. Check the path.")

```

File found, extracting...

Extraction complete.

```

[ ]: # Print content of gplus directory
for root, dirs, files in os.walk("gplus"):
    print(f"Directory: {root}")
    for file in files[:5]: # Print a few files
        print("      ", file)

```

Directory: gplus

Directory: gplus/gplus

```

101133961721621664586.feat
107040353898400532534.featnames
101626577406833098387.feat
110241952466097562819.circles
103537112468125883734.followers

```

```

[ ]: # Print content of first few .circles files
for file in os.listdir('gplus/gplus'):
    if file.endswith(".circles"):
        print(f"\n--- {file} ---")
        with open(os.path.join('gplus/gplus', file), "r") as f:
            for line in f:
                print(line.strip())
        break # Print only the first few

```

```

--- 110241952466097562819.circles ---
9mENop2BfTg      103650208770613583188
3RSyt9yB8p2      103650208770613583188

```

```

[ ]: # Print number of users with >2 circles
valid_users = []

for file in os.listdir('gplus/gplus'):
    if file.endswith(".circles"):
        path = os.path.join('gplus/gplus', file)

        with open(path, "r") as f:

```

```

# Count lines as number of circles
circle_lines = [line.strip() for line in f if line.strip()]
num_circles = len(circle_lines)

if num_circles > 2:
    user_id = file.split(".")[0]
    valid_users.append(user_id)

print(f"Users with >2 circles: {len(valid_users)}")

```

Users with >2 circles: 57

```

[ ]: from itertools import islice

# Get all the .circles files
circle_files = [file for file in os.listdir('gplus/gplus') if file.endswith(".circles")]

# Print number of circles for the first 5 users
for file in islice(circle_files, 5):
    path = os.path.join('gplus/gplus', file)
    with open(path, "r") as f:
        num_lines = sum(1 for line in f if line.strip())
        print(f"{file}: {num_lines} circles")

```

110241952466097562819.circles: 2 circles
 104105354262797387583.circles: 10 circles
 116807883656585676940.circles: 3 circles
 108883879052307976051.circles: 3 circles
 114104634069486127920.circles: 2 circles

18 Question 18

```

[ ]: import os
import networkx as nx

data_dir = "gplus/gplus"
personal_networks = {} # Dictionary to store user_id → DiGraph

circle_files = [f for f in os.listdir(data_dir) if f.endswith(".circles")]

for file in circle_files:
    user_id = file.split(".")[0]
    circle_path = os.path.join(data_dir, file)

    with open(circle_path, "r") as f:
        circles = [line.strip() for line in f if line.strip()]

```

```

if len(circles) <= 2:
    continue # Skip users with 2 or fewer circles

edge_path = os.path.join(data_dir, f"{user_id}.edges")
if not os.path.exists(edge_path):
    continue # Skip users without .edges file

# Build the directed graph from the .edges file
G = nx.DiGraph()
with open(edge_path, "r") as f:
    for line in f:
        u, v = line.strip().split()
        G.add_edge(u, v)

# Store the graph in a dictionary
personal_networks[user_id] = G

print(f"Total directed personal networks created: {len(personal_networks)}")

```

Total directed personal networks created: 57

Total Number of personal networks = 57

19 Question 19

```

[ ]: import networkx as nx
import matplotlib.pyplot as plt

user_ids = [
    "109327480479767108490",
    "115625564993990145546",
    "101373961279443806744"
]

def plot_degrees(user_id, data_dir):
    edge_path = os.path.join(data_dir, f"{user_id}.edges")

    if not os.path.exists(edge_path):
        print(f"Edges file not found for user {user_id}")
        return

    G = nx.DiGraph()

    with open(edge_path, "r") as f:
        for line in f:
            u, v = line.strip().split()
            G.add_edge(u, v)

```

```

in_degrees = [d for n, d in G.in_degree()]
out_degrees = [d for n, d in G.out_degree()]

plt.figure(figsize=(12, 4))

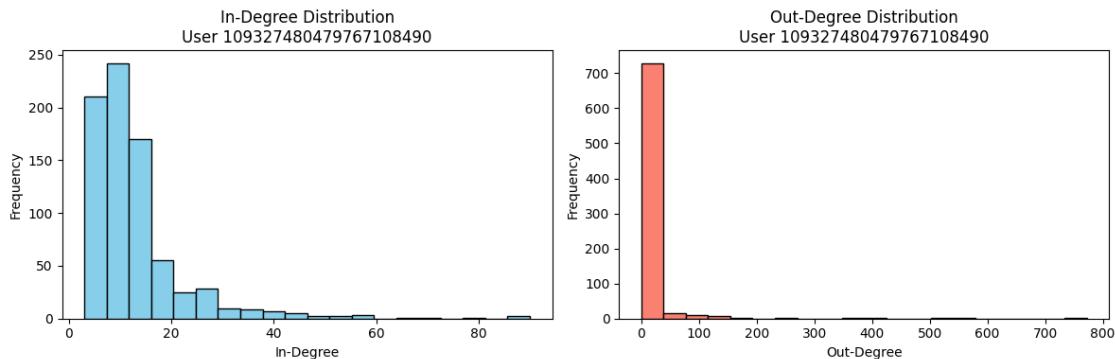
plt.subplot(1, 2, 1)
plt.hist(in_degrees, bins=20, color="skyblue", edgecolor="black")
plt.title(f"In-Degree Distribution\nUser {user_id}")
plt.xlabel("In-Degree")
plt.ylabel("Frequency")

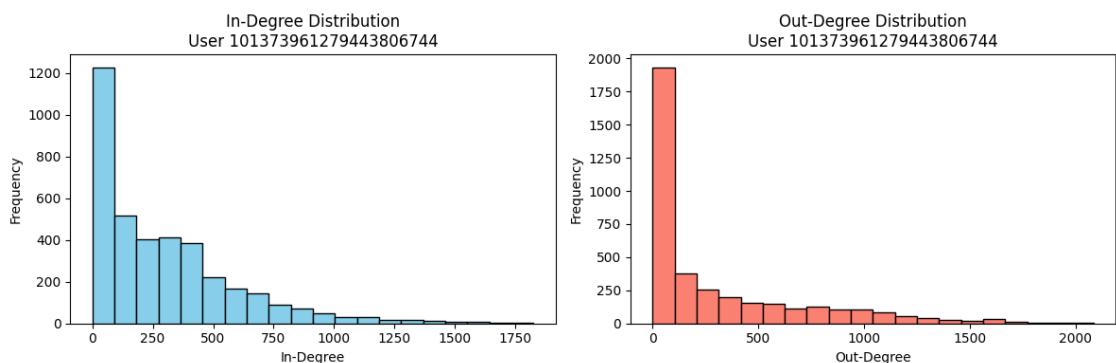
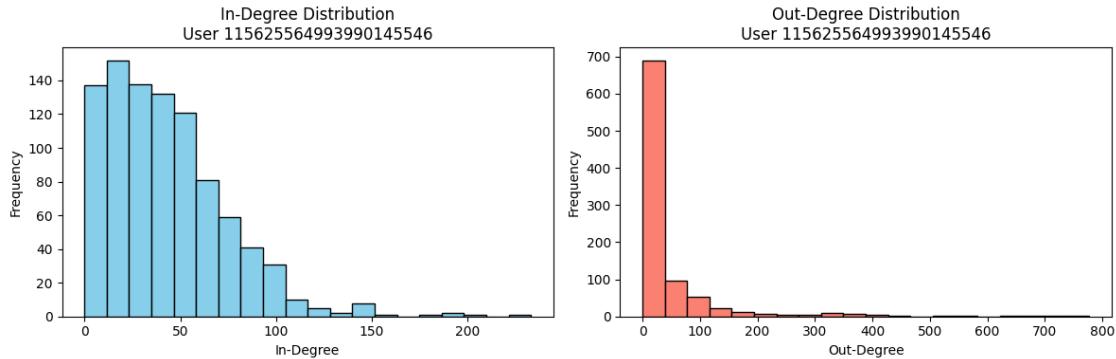
plt.subplot(1, 2, 2)
plt.hist(out_degrees, bins=20, color="salmon", edgecolor="black")
plt.title(f"Out-Degree Distribution\nUser {user_id}")
plt.xlabel("Out-Degree")
plt.ylabel("Frequency")

plt.tight_layout()
plt.show()

# Plot for each of the 3 users
for uid in user_ids:
    plot_degrees(uid, data_dir)

```





20 Question 20

```
[ ]: !pip install igraph
```

Collecting igraph
 Downloading
 igraph-0.11.8-cp39-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
 (3.8 kB)
 Collecting texttable>=1.6.2 (from igraph)
 Downloading texttable-1.7.0-py2.py3-none-any.whl.metadata (9.8 kB)
 Downloading
 igraph-0.11.8-cp39-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (3.1 MB)
 3.1/3.1 MB
 26.4 MB/s eta 0:00:00
 Downloading texttable-1.7.0-py2.py3-none-any.whl (10 kB)
 Installing collected packages: texttable, igraph
 Successfully installed igraph-0.11.8 texttable-1.7.0

```
[ ]: import igraph as ig
import matplotlib.pyplot as plt
```

```

def plot_walktrap_communities(nx_graph, user_id, max_edges=2000):
    # Convert to undirected for Walktrap
    G_undirected = nx_graph.to_undirected()

    # Mapping for igraph conversion
    mapping = {node: idx for idx, node in enumerate(G_undirected.nodes())}
    reverse_mapping = {idx: node for node, idx in mapping.items()}
    edges = [(mapping[u], mapping[v]) for u, v in G_undirected.edges()]

    ig_graph = ig.Graph(edges=edges)
    ig_graph.vs["name"] = list(mapping.keys())

    # Walktrap community detection
    clusters = ig_graph.community_walktrap().as_clustering()
    modularity = clusters.modularity
    membership = clusters.membership
    print(f"User {user_id} - Walktrap Modularity: {modularity:.4f}")

    layout = ig_graph.layout("fr")
    coords = list(zip(*layout.coords))

    plt.figure(figsize=(8, 8))
    plt.scatter(coords[0], coords[1], c=membership, cmap='tab20', s=30)

    # draw only a few edges
    if ig_graph.ecount() <= max_edges:
        for edge in ig_graph.es:
            source, target = edge.tuple
            x0, y0 = layout[source]
            x1, y1 = layout[target]
            plt.plot([x0, x1], [y0, y1], color='gray', linewidth=0.3, alpha=0.4)
    else:
        print(f"Skipping edge drawing for user {user_id} (too many edges:{ig_graph.ecount()})")

    plt.title(f"Community Structure (Walktrap) - User {user_id}")
    plt.axis('off')
    plt.tight_layout()
    plt.show()
    plt.close()

# List of users to process
user_ids = [
    "109327480479767108490",
    "115625564993990145546",
    "101373961279443806744"
]

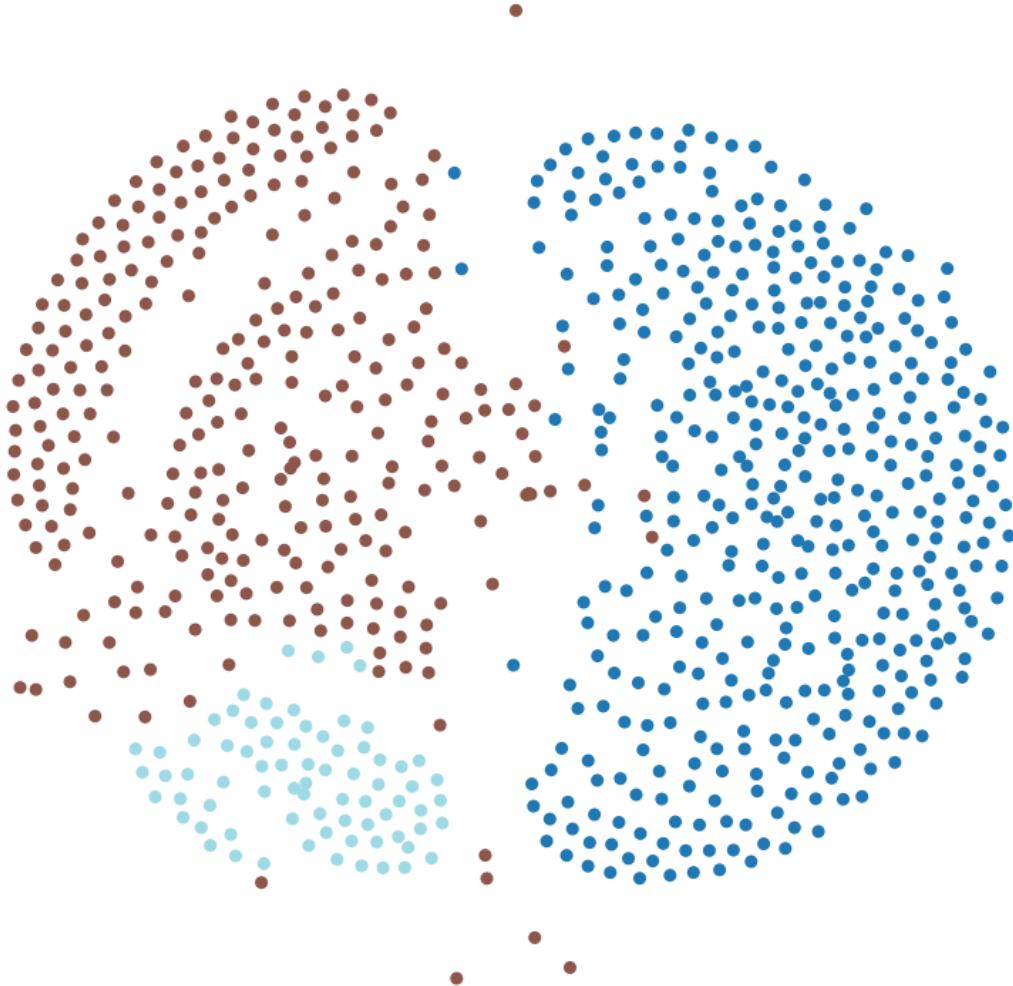
```

```
]
```

```
# Run with RAM-safe plotting
for uid in user_ids:
    plot_walktrap_communities(personal_networks[uid], uid)
```

User 109327480479767108490 - Walktrap Modularity: 0.2577
Skipping edge drawing for user 109327480479767108490 (too many edges: 9070)

Community Structure (Walktrap) — User 109327480479767108490



User 115625564993990145546 - Walktrap Modularity: 0.3087
Skipping edge drawing for user 115625564993990145546 (too many edges: 34022)

Community Structure (Walktrap) — User 115625564993990145546



User 101373961279443806744 - Walktrap Modularity: 0.1734

Skipping edge drawing for user 101373961279443806744 (too many edges: 954581)

Community Structure (Walktrap) — User 101373961279443806744



The modularity scores obtained from the algorithm are:
* User 109327480479767108490: 0.2577
* User 115625564993990145546: 0.3087
* User 101373961279443806744: 0.1734

The modularity scores are not similar. User 115625564993990145546 exhibits the highest modularity, indicating a more clearly defined community structure. User 101373961279443806744 has the lowest modularity, suggesting a less distinct division into communities.

While modularity scores indicate how well communities are separated in terms of edge density, they do not always align with the visual clarity of communities in 2D plots. In this case, the network with a lower modularity score (0.2577) appears more visually clustered due to its size and layout, while the network with higher modularity (0.3087) shows more overlap despite stronger mathematical separation.

21 Question 21

Homogeneity and completeness are two entropy-based measures used to evaluate the relationship between user-defined circles and algorithmically detected communities in a social network.

Homogeneity measures the extent to which each community contains members from a single circle. It is defined as:

$$h = 1 - (H(C|K)/H(C))$$

where $H(C)$ is the entropy of the circle distribution and $H(C|K)$ is the conditional entropy of circles given communities. A homogeneity score close to 1 indicates that the communities are “pure,” meaning most of the nodes in a community belong to the same circle. A lower value indicates that communities contain a mix of nodes from different circles.

Completeness measures the extent to which all members of a circle are assigned to the same community. It is defined as:

$$c = 1 - (H(K|C)/H(K))$$

where $H(K)$ is the entropy of the community distribution and $H(K|C)$ is the conditional entropy of communities given circles. A completeness score close to 1 indicates that members of a circle are grouped together in the same community. A lower value means the nodes in a circle are spread across multiple communities.

In summary, homogeneity reflects how well communities align internally with single circles, while completeness reflects how well circles are preserved as cohesive groups within the detected communities. Both measures range from 0 to 1, with higher values indicating stronger alignment between circles and communities.

22 Question 22

```
[ ]: import os
import math
import networkx as nx
import igraph as ig
from collections import defaultdict, Counter

# Helper to load circles
def load_circle_labels(user_id, data_dir):
    circles_path = os.path.join(data_dir, f"{user_id}.circles")
    circles = []
    with open(circles_path, "r") as f:
        for line in f:
            parts = line.strip().split()
            if len(parts) > 1:
                circles.append(parts[1:])
    return circles

# Compute homogeneity and completeness
```

```

def compute_homogeneity_completeness(circles, communities):
    # Build sets
    C = [set(c) for c in circles]
    K = [set(comm) for comm in communities]

    # Nodes with circle labels
    all_labeled_nodes = set(node for circle in C for node in circle)
    N = len(all_labeled_nodes)

    if N == 0:
        return 0.0, 0.0

    # ai: size of circle i
    ai = [len(circle & all_labeled_nodes) for circle in C]

    # bi: number of labeled nodes in community
    bi = [len(comm & all_labeled_nodes) for comm in K]

    # Aji: number of nodes in community j and circle i
    A = [[len(comm & circle) for circle in C] for comm in K]

    # Entropy H(C)
    H_C = -sum((a / N) * math.log(a / N) for a in ai if a > 0)

    # Entropy H(K)
    H_K = -sum((b / N) * math.log(b / N) for b in bi if b > 0)

    # Conditional entropy H(C/K)
    H_C_given_K = 0
    for j in range(len(K)):
        for i in range(len(C)):
            Aji = A[j][i]
            if Aji > 0:
                H_C_given_K -= (Aji / N) * math.log(Aji / bi[j])

    # Conditional entropy H(K/C)
    H_K_given_C = 0
    for i in range(len(C)):
        for j in range(len(K)):
            Aji = A[j][i]
            if Aji > 0:
                H_K_given_C -= (Aji / N) * math.log(Aji / ai[i])

    # Final scores
    h = 1 - H_C_given_K / H_C if H_C > 0 else 1.0
    c = 1 - H_K_given_C / H_K if H_K > 0 else 1.0
    return h, c

```

```

# Convert NetworkX to igraph undirected
def nx_to_igraph_undirected(nx_graph):
    G_undirected = nx_graph.to_undirected()
    mapping = {node: idx for idx, node in enumerate(G_undirected.nodes())}
    reverse_mapping = {idx: node for node, idx in mapping.items()}
    edges = [(mapping[u], mapping[v]) for u, v in G_undirected.edges()]
    ig_graph = ig.Graph(edges=edges)
    ig_graph.vs[["name"]] = list(mapping.keys())
    return ig_graph

# Users and data_dir
user_ids = [
    "109327480479767108490",
    "115625564993990145546",
    "101373961279443806744"
]

data_dir = "gplus/gplus"

# Final computation loop
for uid in user_ids:
    # Load network and convert
    G_nx = personal_networks[uid]
    G_ig = nx_to_igraph_undirected(G_nx)

    # Run Walktrap
    clusters = G_ig.community_walktrap().as_clustering()
    community_sets = []
    for cluster in clusters:
        members = [G_ig.vs[node]["name"] for node in cluster]
        community_sets.append(set(members))

    # Load circle labels
    circles = load_circle_labels(uid, data_dir)

    # Compute h and c
    h, c = compute_homogeneity_completeness(circles, community_sets)
    print(f"User {uid}: Homogeneity = {h:.4f}, Completeness = {c:.4f}")

```

User 109327480479767108490: Homogeneity = 0.8881, Completeness = 0.3955
User 115625564993990145546: Homogeneity = 0.4647, Completeness = -2.9843
User 101373961279443806744: Homogeneity = 0.0034, Completeness = -1.5098

User ID	Homogeneity (h)	Completeness (c)
109327480479767108490	0.8881	0.3955
115625564993990145546	0.4647	-2.9843

User ID	Homogeneity (h)	Completeness (c)
101373961279443806744	0.0034	-1.5098

The homogeneity values are valid and show varying degrees of purity of communities with respect to circle membership. User 109327480479767108490 has high homogeneity, indicating that most communities contain nodes from a single circle. The other users show poor alignment.

The completeness values for two users are negative, which is theoretically invalid. This occurred due to numerical instability when the entropy of community distribution $H(K)$ was close to zero, causing the ratio $H(K|C)/H(K)$ to become large and result in a negative value for completeness. This suggests weak correspondence between circles and communities and potential mismatches or sparsity in the circle information.

23 3. Cora Dataset

```
[ ]: ! pip install torch_geometric

Requirement already satisfied: torch_geometric in
/usr/local/lib/python3.11/dist-packages (2.6.1)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-
packages (from torch_geometric) (3.11.15)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages
(from torch_geometric) (2025.3.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages
(from torch_geometric) (3.1.6)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages
(from torch_geometric) (2.0.2)
Requirement already satisfied: psutil>=5.8.0 in /usr/local/lib/python3.11/dist-
packages (from torch_geometric) (5.9.5)
Requirement already satisfied: pyparsing in /usr/local/lib/python3.11/dist-
packages (from torch_geometric) (3.2.3)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-
packages (from torch_geometric) (2.32.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages
(from torch_geometric) (4.67.1)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->torch_geometric) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->torch_geometric) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-
packages (from aiohttp->torch_geometric) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->torch_geometric) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->torch_geometric) (6.4.3)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->torch_geometric) (0.3.1)
```

```

Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->torch_geometric) (1.19.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2->torch_geometric) (3.0.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests->torch_geometric) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests->torch_geometric) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests->torch_geometric) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests->torch_geometric)
(2025.1.31)

```

```

[ ]: from torch_geometric.datasets import Planetoid
      import torch

      # Load Cora
      dataset = Planetoid(root='data/Cora', name='Cora')

      # Access graph
      data = dataset[0]

      # Access components
      X = data.x
      edge_index = data.edge_index
      y = data.y
      train_mask = data.train_mask
      val_mask = data.val_mask
      test_mask = data.test_mask

      print(data)

```

```

Data(x=[2708, 1433], edge_index=[2, 10556], y=[2708], train_mask=[2708],
val_mask=[2708], test_mask=[2708])

```

Note: In the pdf there are 5429 edges mentioned but in the above output shows 10556 edges, this is expected as PyG doubles edges to support undirected GCN operations (each undirected edge become two directed edge)

24 Question 23

```

[ ]: import torch
      import torch.nn.functional as F
      from torch_geometric.nn import GCNConv

      # 2 - layered GCN
      class GCN(torch.nn.Module):

```

```

def __init__(self, in_channels, hidden_channels, out_channels, dropout):
    super(GCN, self).__init__()
    self.conv1 = GCMConv(in_channels, hidden_channels)
    self.conv2 = GCMConv(hidden_channels, out_channels)
    self.dropout = dropout

def forward(self, data):
    x, edge_index = data.x, data.edge_index
    x = self.conv1(x, edge_index)
    x = F.relu(x)
    x = F.dropout(x, p=self.dropout, training=self.training)
    x = self.conv2(x, edge_index)
    return F.log_softmax(x, dim=1)

```

```

[ ]: def train(model, data, optimizer):
    model.train()
    optimizer.zero_grad()
    out = model(data)
    loss = F.nll_loss(out[data.train_mask], data.y[data.train_mask])
    loss.backward()
    optimizer.step()
    return loss.item()

@torch.no_grad()
def test(model, data):
    model.eval()
    out = model(data)
    pred = out.argmax(dim=1)
    accs = []
    for mask in [data.train_mask, data.val_mask, data.test_mask]:
        correct = pred[mask].eq(data.y[mask]).sum().item()
        acc = correct / mask.sum().item()
        accs.append(acc)
    return accs # [train_acc, val_acc, test_acc]

```

```

[ ]: import itertools

# Hyperparameter grid
hidden_sizes = [16, 32]
dropouts = [0.3, 0.5]
weight_decays = [5e-4, 1e-3]
lrs = [0.01, 0.005]

# Grid search
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
best_val_acc = 0
best_config = {}

```

```

best_test_acc = 0

for hidden, dropout, wd, lr in itertools.product(hidden_sizes, dropouts, weight_decays, lrs):
    model = GCN(dataset.num_node_features, hidden, dataset.num_classes, dropout).to(device)
    data = data.to(device)
    optimizer = torch.optim.Adam(model.parameters(), lr=lr, weight_decay=wd)

    for epoch in range(1, 201):
        train(model, data, optimizer)

    train_acc, val_acc, test_acc = test(model, data)

    print(f"[H={hidden}, D={dropout}, WD={wd}, LR={lr}] Val: {val_acc:.4f} | Test: {test_acc:.4f}")

    if val_acc > best_val_acc:
        best_val_acc = val_acc
        best_test_acc = test_acc
        best_config = {
            'hidden_size': hidden,
            'dropout': dropout,
            'weight_decay': wd,
            'learning_rate': lr
        }

# Final best config
print("\nBest Configuration:")
for k, v in best_config.items():
    print(f"{k}: {v}")
print(f"Validation Accuracy: {best_val_acc:.4f}")
print(f"Test Accuracy: {best_test_acc:.4f}")

```

[H=16, D=0.3, WD=0.0005, LR=0.01] Val: 0.7600 | Test: 0.8040
[H=16, D=0.3, WD=0.0005, LR=0.005] Val: 0.7720 | Test: 0.8030
[H=16, D=0.3, WD=0.001, LR=0.01] Val: 0.7700 | Test: 0.8170
[H=16, D=0.3, WD=0.001, LR=0.005] Val: 0.7740 | Test: 0.7990
[H=16, D=0.5, WD=0.0005, LR=0.01] Val: 0.7720 | Test: 0.8110
[H=16, D=0.5, WD=0.0005, LR=0.005] Val: 0.7720 | Test: 0.7980
[H=16, D=0.5, WD=0.001, LR=0.01] Val: 0.7700 | Test: 0.8110
[H=16, D=0.5, WD=0.001, LR=0.005] Val: 0.7780 | Test: 0.8020
[H=32, D=0.3, WD=0.0005, LR=0.01] Val: 0.7720 | Test: 0.8050
[H=32, D=0.3, WD=0.0005, LR=0.005] Val: 0.7780 | Test: 0.8000
[H=32, D=0.3, WD=0.001, LR=0.01] Val: 0.7820 | Test: 0.8100
[H=32, D=0.3, WD=0.001, LR=0.005] Val: 0.7840 | Test: 0.8150
[H=32, D=0.5, WD=0.0005, LR=0.01] Val: 0.7700 | Test: 0.8100

```
[H=32, D=0.5, WD=0.0005, LR=0.005] Val: 0.7760 | Test: 0.8050  
[H=32, D=0.5, WD=0.001, LR=0.01] Val: 0.7660 | Test: 0.7970  
[H=32, D=0.5, WD=0.001, LR=0.005] Val: 0.7720 | Test: 0.8030
```

Best Configuration:
hidden_size: 32
dropout: 0.3
weight_decay: 0.001
learning_rate: 0.005
Validation Accuracy: 0.7840
Test Accuracy: 0.8150

```
[ ]: import torch  
  
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')  
model = GCN(in_channels=1433, hidden_channels=32, out_channels=7, dropout=0.3).  
        to(device)  
data = data.to(device)  
  
optimizer = torch.optim.Adam(model.parameters(), lr=0.005, weight_decay=0.001)  
  
# Train the 2 - layer GCN model  
for epoch in range(1, 201):  
    loss = train(model, data, optimizer)  
    train_acc, val_acc, test_acc = test(model, data)  
    if epoch % 20 == 0:  
        print(f'Epoch: {epoch:03d}, Loss: {loss:.4f}, Train: {train_acc:.4f},  
              Val: {val_acc:.4f}, Test: {test_acc:.4f}')
```

```
Epoch: 020, Loss: 0.3030, Train: 0.9857, Val: 0.8040, Test: 0.8140  
Epoch: 040, Loss: 0.0547, Train: 1.0000, Val: 0.7760, Test: 0.8010  
Epoch: 060, Loss: 0.0358, Train: 1.0000, Val: 0.7740, Test: 0.8000  
Epoch: 080, Loss: 0.0357, Train: 1.0000, Val: 0.7700, Test: 0.8050  
Epoch: 100, Loss: 0.0330, Train: 1.0000, Val: 0.7700, Test: 0.8080  
Epoch: 120, Loss: 0.0355, Train: 1.0000, Val: 0.7680, Test: 0.8080  
Epoch: 140, Loss: 0.0302, Train: 1.0000, Val: 0.7720, Test: 0.8050  
Epoch: 160, Loss: 0.0230, Train: 1.0000, Val: 0.7760, Test: 0.8030  
Epoch: 180, Loss: 0.0230, Train: 1.0000, Val: 0.7760, Test: 0.8100  
Epoch: 200, Loss: 0.0235, Train: 1.0000, Val: 0.7740, Test: 0.8130
```

```
[ ]: # 3-layer GCN  
class GCN3(torch.nn.Module):  
    def __init__(self, in_channels, hidden_channels, out_channels, dropout):  
        super().__init__()  
        self.conv1 = GCNConv(in_channels, hidden_channels)  
        self.conv2 = GCNConv(hidden_channels, hidden_channels)  
        self.conv3 = GCNConv(hidden_channels, out_channels)  
        self.dropout = dropout
```

```

def forward(self, data):
    x, edge_index = data.x, data.edge_index
    x = F.relu(self.conv1(x, edge_index))
    x = F.dropout(x, p=self.dropout, training=self.training)
    x = F.relu(self.conv2(x, edge_index))
    x = F.dropout(x, p=self.dropout, training=self.training)
    x = self.conv3(x, edge_index)
    return F.log_softmax(x, dim=1)

```

```

[ ]: device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
model = GCN3(in_channels=1433, hidden_channels=32, out_channels=7, dropout=0.3).
        to(device)
data = data.to(device)

optimizer = torch.optim.Adam(model.parameters(), lr=0.005, weight_decay=0.001)

# Train the 3 - layer GCN model
for epoch in range(1, 201):
    loss = train(model, data, optimizer)
    train_acc, val_acc, test_acc = test(model, data)
    if epoch % 20 == 0:
        print(f'Epoch: {epoch:03d}, Loss: {loss:.4f}, Train: {train_acc:.4f}, Val: {val_acc:.4f}, Test: {test_acc:.4f}')

```

Epoch: 020, Loss: 0.2976, Train: 0.9929, Val: 0.7760, Test: 0.8030
 Epoch: 040, Loss: 0.0206, Train: 1.0000, Val: 0.7580, Test: 0.7850
 Epoch: 060, Loss: 0.0144, Train: 1.0000, Val: 0.7720, Test: 0.7980
 Epoch: 080, Loss: 0.0145, Train: 1.0000, Val: 0.7800, Test: 0.8070
 Epoch: 100, Loss: 0.0267, Train: 1.0000, Val: 0.7800, Test: 0.8050
 Epoch: 120, Loss: 0.0185, Train: 1.0000, Val: 0.7800, Test: 0.8060
 Epoch: 140, Loss: 0.0139, Train: 1.0000, Val: 0.7680, Test: 0.8050
 Epoch: 160, Loss: 0.0146, Train: 1.0000, Val: 0.7680, Test: 0.8020
 Epoch: 180, Loss: 0.0128, Train: 1.0000, Val: 0.7840, Test: 0.8140
 Epoch: 200, Loss: 0.0122, Train: 1.0000, Val: 0.7640, Test: 0.7970

A Graph Convolutional Network (GCN) was implemented for semi-supervised node classification on the Cora dataset, which consists of 2,708 nodes, 1,433-dimensional feature vectors, and 7 classes. Only 140 nodes (20 per class) are labeled during training.

Hyperparameter tuning was performed using grid search over the following values:

- Hidden size: [16, 32]
- Dropout: [0.3, 0.5]
- Learning rate: [0.01, 0.005]
- Weight decay: [5e-4, 1e-3]
- Epochs: 200 (fixed)

The best hyperparameter configuration based on validation accuracy was:

- Hidden size: 32
- Dropout: 0.3

- Learning rate: 0.005
- Weight decay: 0.001

Performance of 2-layer CNN with best configuration: - Training accuracy: 100%

- Validation accuracy: 77.4%
- Test accuracy: 81.3%

A 3-layer GCN was also evaluated using the same hyperparameters. It achieved a peak validation accuracy of 78.4% and test accuracy of 81.4%, but exhibited higher variance and lower stability compared to the 2-layer model. Due to oversmoothing and inconsistent performance, the 2-layer GCN was selected as optimal.

25 Question 24

```
[ ]: # Install PyTorch (CUDA 11.8) + torch-geometric + required extensions
!pip install -q torch==2.0.0+cu118 torchvision==0.15.1+cu118 torchaudio==2.0.
    ↵1+cu118 --index-url https://download.pytorch.org/whl/cu118
!pip install -q pyg-lib torch-scatter torch-sparse torch-cluster
    ↵torch-spline-conv torch-geometric -f https://data.pyg.org/whl/torch-2.0.
    ↵0+cu118.html
```

	2.3/2.3 GB
437.0 kB/s eta 0:00:00	6.1/6.1 MB
99.0 MB/s eta 0:00:00	4.4/4.4 MB
82.5 MB/s eta 0:00:00	63.3/63.3 MB
11.5 MB/s eta 0:00:00	132.3/132.3 kB
12.3 MB/s eta 0:00:00	
	Preparing metadata (setup.py) ... done
	Building wheel for lit (setup.py) ... done
	2.6/2.6 MB
29.9 MB/s eta 0:00:00	10.2/10.2 MB
125.3 MB/s eta 0:00:00	4.9/4.9 MB
120.5 MB/s eta 0:00:00	3.3/3.3 MB
83.2 MB/s eta 0:00:00	886.5/886.5 kB
58.9 MB/s eta 0:00:00	

```
[ ]: from torch_geometric.nn import Node2Vec

device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
```

```

node2vec = Node2Vec(
    data.edge_index,
    embedding_dim=128,
    walk_length=20,
    context_size=10,
    walks_per_node=10,
    num_negative_samples=1,
    sparse=True
).to(device)

loader = node2vec.loader(batch_size=128, shuffle=True)
optimizer = torch.optim.SparseAdam(list(node2vec.parameters()), lr=0.01)

def train_node2vec():
    node2vec.train()
    total_loss = 0
    for pos_rw, neg_rw in loader:
        optimizer.zero_grad()
        loss = node2vec.loss(pos_rw.to(device), neg_rw.to(device))
        loss.backward()
        optimizer.step()
        total_loss += loss.item()
    return total_loss / len(loader)

for epoch in range(1, 101):
    loss = train_node2vec()
    if epoch % 20 == 0:
        print(f"Epoch {epoch}, Loss: {loss:.4f}")

```

Epoch 20, Loss: 0.9742
 Epoch 40, Loss: 0.8433
 Epoch 60, Loss: 0.8297
 Epoch 80, Loss: 0.8259
 Epoch 100, Loss: 0.8238

```

[ ]: node2vec.eval()
      with torch.no_grad():
          embeddings = node2vec.embedding.weight.detach().cpu().tolist()

      import numpy as np
      embeddings = np.array(embeddings)
      print("Embeddings shape:", embeddings.shape)

```

Embeddings shape: (2708, 128)

```

[ ]: import numpy as np
      from sklearn.neural_network import MLPClassifier

```

```

from sklearn.metrics import accuracy_score

# Convert text features and labels safely
text_features = np.array(data.x.cpu().tolist())      # shape: [2708, 1433]
labels = np.array(data.y.cpu().tolist())              # shape: [2708]

# Train/test split using masks
train_idx = data.train_mask.cpu().tolist()
test_idx = data.test_mask.cpu().tolist()

# Filter features and labels by index
X_n2v_train = embeddings[train_idx]
X_n2v_test = embeddings[test_idx]

X_text_train = text_features[train_idx]
X_text_test = text_features[test_idx]

# Combine features
X_combined_train = np.concatenate([X_n2v_train, X_text_train], axis=1)
X_combined_test = np.concatenate([X_n2v_test, X_text_test], axis=1)

y_train = labels[train_idx]
y_test = labels[test_idx]

```

```

[ ]: # To train and evaluate MLP
def train_evaluate(X_train, y_train, X_test, y_test):
    clf = MLPClassifier(hidden_layer_sizes=(64,), max_iter=500, random_state=42)
    clf.fit(X_train, y_train)
    y_pred = clf.predict(X_test)
    return accuracy_score(y_test, y_pred)

# Evaluate on all 3 feature setups
acc_node2vec = train_evaluate(X_n2v_train, y_train, X_n2v_test, y_test)
acc_text = train_evaluate(X_text_train, y_train, X_text_test, y_test)
acc_combined = train_evaluate(X_combined_train, y_train, X_combined_test, y_test)

# Print results
print(f"Accuracy using Node2Vec only: {acc_node2vec:.4f}")
print(f"Accuracy using Text features only: {acc_text:.4f}")
print(f"Accuracy using Combined features: {acc_combined:.4f}")

```

Accuracy using Node2Vec only: 0.6920
 Accuracy using Text features only: 0.5060
 Accuracy using Combined features: 0.7230

To evaluate the impact of different node features on classification, Node2Vec was used to extract

structure-based embeddings from the Cora network. A multi-layer perceptron (MLPClassifier) was chosen to evaluate classification accuracy across three feature setups:

- Node2Vec embeddings only (128-dimensional)
- Text features only (original 1433-dimensional vectors)
- Combined Node2Vec + text features (1561-dimensional)

All models were trained on the labeled nodes defined by train_mask and evaluated using test_mask.

Feature Type	Test Accuracy
Node2Vec only	69.2%
Text features only	50.6%
Combined	72.3%

The classifier performed best when using combined features (Accuracy of 72.3%), indicating that graph structure and textual content provide complementary information. The text-only accuracy was relatively low, likely due to the sparsity of features and the limited size of the training set. Node2Vec alone performed significantly better, highlighting the predictive value of structural relationships in the citation network.

26 Question 25

```
[ ]: import torch
import numpy as np
from collections import defaultdict
from sklearn.metrics import accuracy_score, f1_score

# Move data to CPU
edge_index = data.edge_index.cpu()
features = data.x.cpu().numpy()
labels = data.y.cpu().numpy()
train_mask = data.train_mask.cpu().numpy()
test_mask = data.test_mask.cpu().numpy()

# Build adjacency list
adj_list = defaultdict(list)
for u, v in edge_index.numpy().T:
    adj_list[u].append(v)
    adj_list[v].append(u) # undirected

# Build seed nodes per class (20 per class)
num_classes = labels.max() + 1
seeds_per_class = {}
for c in range(num_classes):
    class_indices = np.where((labels == c) & train_mask)[0]
    seeds_per_class[c] = class_indices[:20] # take first 20
```

```
[ ]: from sklearn.metrics.pairwise import cosine_similarity

def transition_probs(node, neighbors, features):
    if len(neighbors) == 0:
        return [], []

    node_feat = features[node].reshape(1, -1)
    neighbor_feats = features[neighbors]
    sims = cosine_similarity(node_feat, neighbor_feats).flatten()
    exp_sims = np.exp(sims)
    probs = exp_sims / exp_sims.sum()
    return neighbors, probs

[ ]: def run_random_walks(adj_list, features, seeds_per_class, alpha=0.1, walk_length=10, num_walks=1000):
    visit_counts = np.zeros((features.shape[0], num_classes))

    for c in range(num_classes):
        seeds = seeds_per_class[c]
        for seed in seeds:
            for _ in range(num_walks):
                curr = seed
                for _ in range(walk_length):
                    if np.random.rand() < alpha:
                        curr = np.random.choice(seeds) # teleport
                    else:
                        neighbors = adj_list[curr]
                        neighbors, probs = transition_probs(curr, neighbors, features)
                        if len(neighbors) == 0:
                            break
                        curr = np.random.choice(neighbors, p=probs)
                visit_counts[curr][c] += 1 # visited by class c

    return visit_counts

[ ]: visit_counts = run_random_walks(adj_list, features, seeds_per_class, alpha=0)

# Predict class for test nodes
test_indices = np.where(test_mask)[0]
y_true = labels[test_indices]
y_pred = visit_counts[test_indices].argmax(axis=1)

# Evaluate
acc = accuracy_score(y_true, y_pred)
f1 = f1_score(y_true, y_pred, average='macro')
```

```
print(f"Accuracy when alpha=0: {acc:.4f}")
print(f"Macro F1-score when alpha=0: {f1:.4f}")
```

Accuracy when alpha=0: 0.6880
Macro F1-score when alpha=0: 0.6990

```
[ ]: visit_counts = run_random_walks(adj_list, features, seeds_per_class, alpha=0.1)

# Predict class for test nodes
test_indices = np.where(test_mask)[0]
y_true = labels[test_indices]
y_pred = visit_counts[test_indices].argmax(axis=1)

# Evaluate
acc = accuracy_score(y_true, y_pred)
f1 = f1_score(y_true, y_pred, average='macro')

print(f"Accuracy when alpha=0.1: {acc:.4f}")
print(f"Macro F1-score when alpha=0.1: {f1:.4f}")
```

Accuracy when alpha=0.1: 0.6860
Macro F1-score when alpha=0.1: 0.6979

```
[ ]: visit_counts = run_random_walks(adj_list, features, seeds_per_class, alpha=0.2)

# Predict class for test nodes
test_indices = np.where(test_mask)[0]
y_true = labels[test_indices]
y_pred = visit_counts[test_indices].argmax(axis=1)

# Evaluate
acc = accuracy_score(y_true, y_pred)
f1 = f1_score(y_true, y_pred, average='macro')

print(f"Accuracy when alpha=0.2: {acc:.4f}")
print(f"Macro F1-score when alpha=0.2: {f1:.4f}")
```

Accuracy when alpha=0.2: 0.6810
Macro F1-score when alpha=0.2: 0.6897

A class-wise personalized PageRank-style classifier was implemented using biased random walks on the Cora dataset. For each class, 20 labeled seed nodes were selected, and 1000 random walks were performed from each seed. The transition probabilities were based on cosine similarity between node features, and teleportation (jumping back to a seed) was introduced with probability $\{0.0, 0.1, 0.2\}$.

Each unlabeled node was assigned the class that visited it most frequently across all walks. The results are reported using test accuracy and macro-averaged F1-score.

Teleport	Accuracy	Macro F1
0.0	0.6880	0.6990
0.1	0.6860	0.6979
0.2	0.6810	0.6897

The best performance was achieved when $\gamma = 0$, indicating that purely local walks based on graph structure and feature similarity are most effective on Cora. Increasing teleportation slightly degraded performance, likely due to reduced graph exploration and overemphasis on seed nodes. This behavior aligns with the strong structural homophily in Cora, where nodes of the same class tend to form tightly connected communities.