Social Network Analysis

Final Project Report

National Institute of Technology, Tiruchirappalli

Student	Priyansh Kumar Paswan
Roll Number	205124071
Professor	Dr. S.R. Balasundaram
Department	Computer Applications
Institute	National Institute of Technology, Tiruchirappalli

1. Dataset Overview

Email-EU-core undirected graph. Structural summary and degree characteristics.

Observations: The network is sparse (low density) and, as in many communication graphs, the degree distribution typically exhibits a heavy tail. A small set of nodes act as hubs with substantially higher degree, while most nodes have moderate to low degree. Average clustering indicates the tendency of colleagues to form tightly knit triads. If the graph is not fully connected, insights should be interpreted within the giant component.

Interpretation: The subgraph snapshot provides intuition about the hub–periphery structure. The degree histogram helps motivate methods used later: centrality to identify important spreaders, community detection for group structure, and heuristics tailored to local neighborhoods for link prediction.

Metric	Value
Nodes	1005
Edges	16706
Density	0.0331
Avg clustering	0.3994
Connected	False

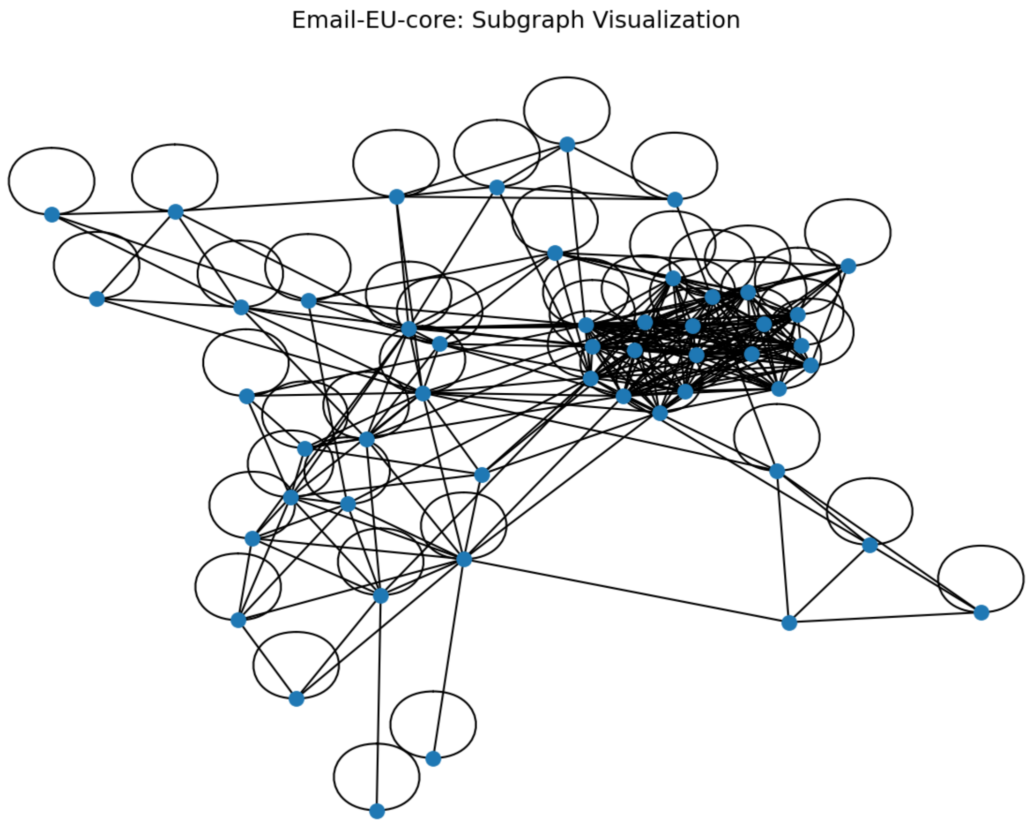

Figure 1: Subgraph visualization

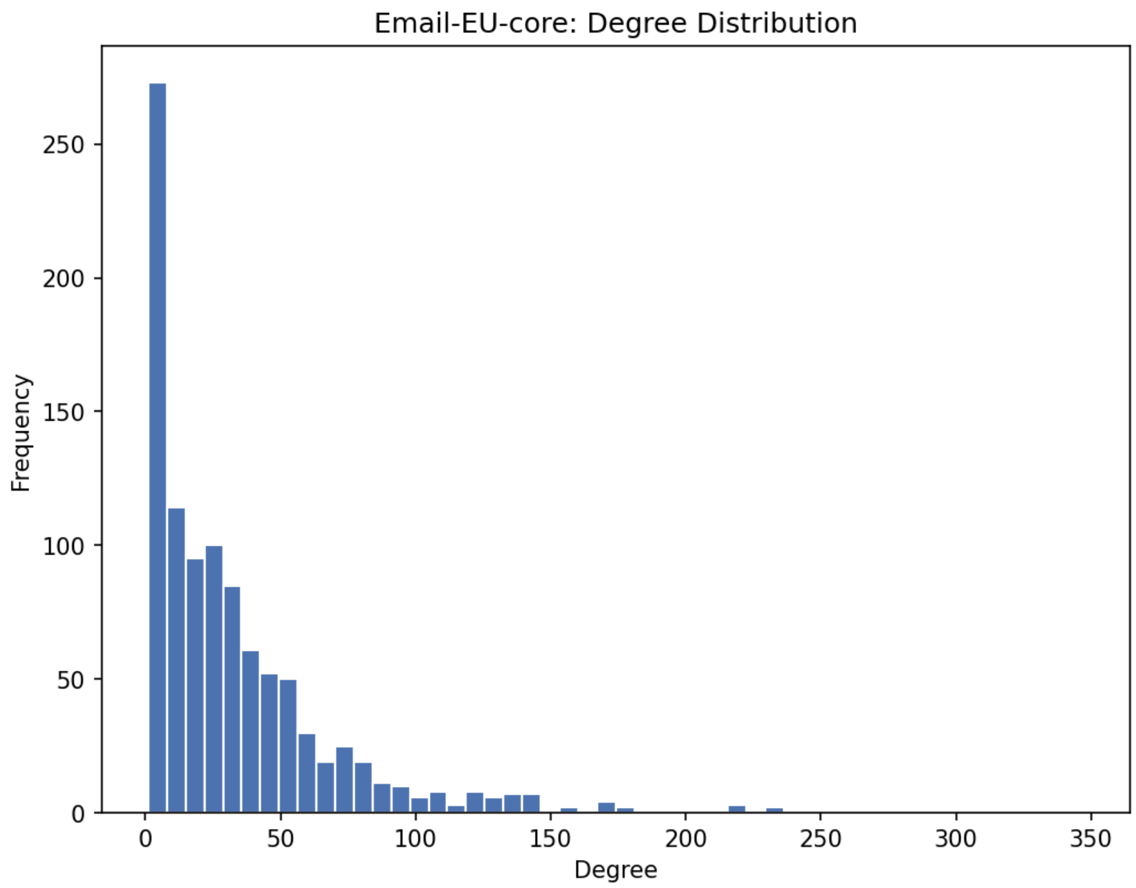

Figure 2: Degree distribution histogram

2. Link Analysis (PageRank & Eigenvector)

PageRank (importance via random walks) and Eigenvector centrality (importance via influential neighbors).

Observations: PageRank elevates nodes that attract many paths, directly or via iterative reinforcement; Eigenvector centrality favors nodes connected to other well-connected nodes. Overlap between the two measures typically signals strong hubs embedded in a highly connected core. Disagreements often reveal local elites or structurally peripheral nodes endorsed by a single dominant neighbor.

Interpretation: In the network view, node size tracks PageRank while color follows an eigenvector gradient. Warm colors indicate high eigenvector centrality; large, warm nodes are the core influencers. The top PageRank bar chart complements the map by listing specific IDs you'd prioritize for information diffusion or monitoring.

node	pagerank	eigenvector	degree
160.0	0.0090709484950265	0.1658461052562613	347.0
121.0	0.0060687854256318	0.1484213057175369	234.0
82.0	0.0060307053368408	0.145251809177294	233.0
107.0	0.0058380960432493	0.139876476623585	221.0
86.0	0.0057215196123223	0.1122173025767131	218.0
62.0	0.0054316159798508	0.1314982021050847	216.0
5.0	0.0049141637345112	0.0794635644309267	171.0
13.0	0.004589938693459	0.0856933490779659	180.0
166.0	0.0045516651162006	0.1103349118134628	177.0
434.0	0.0045327987830447	0.1253049110776443	185.0

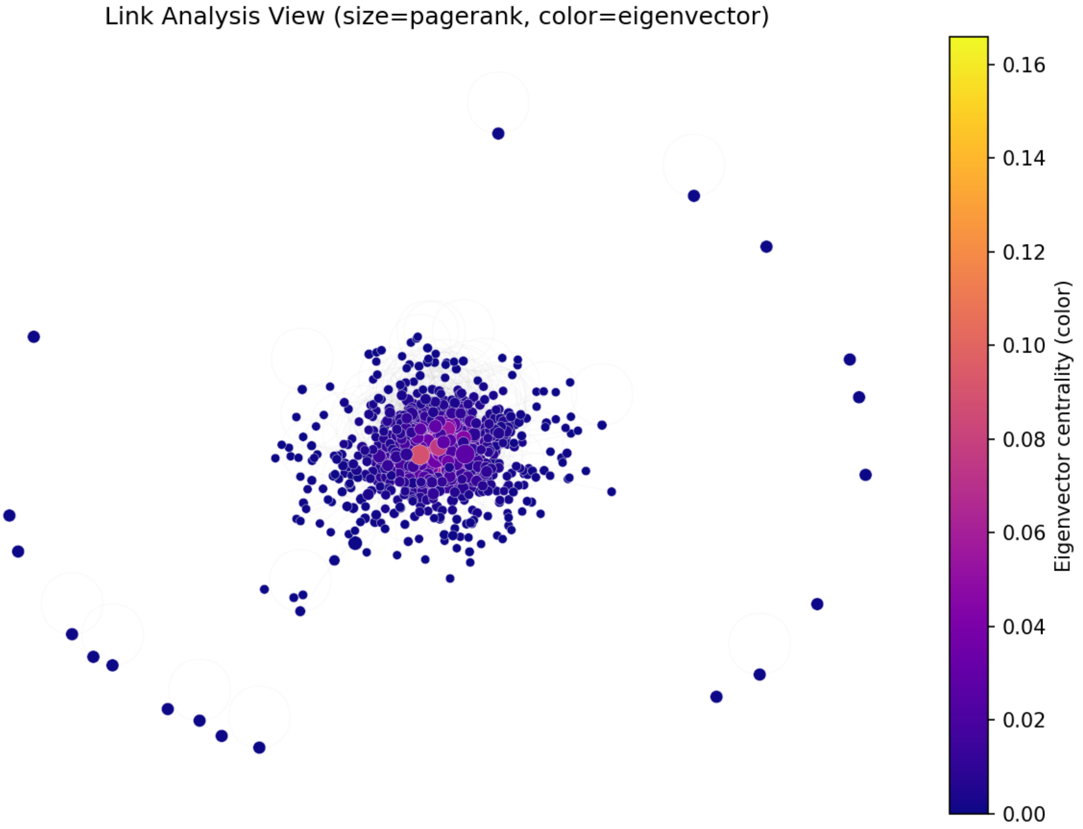

Figure 3: Link Analysis view (size=pagerank, color=eigenvector)

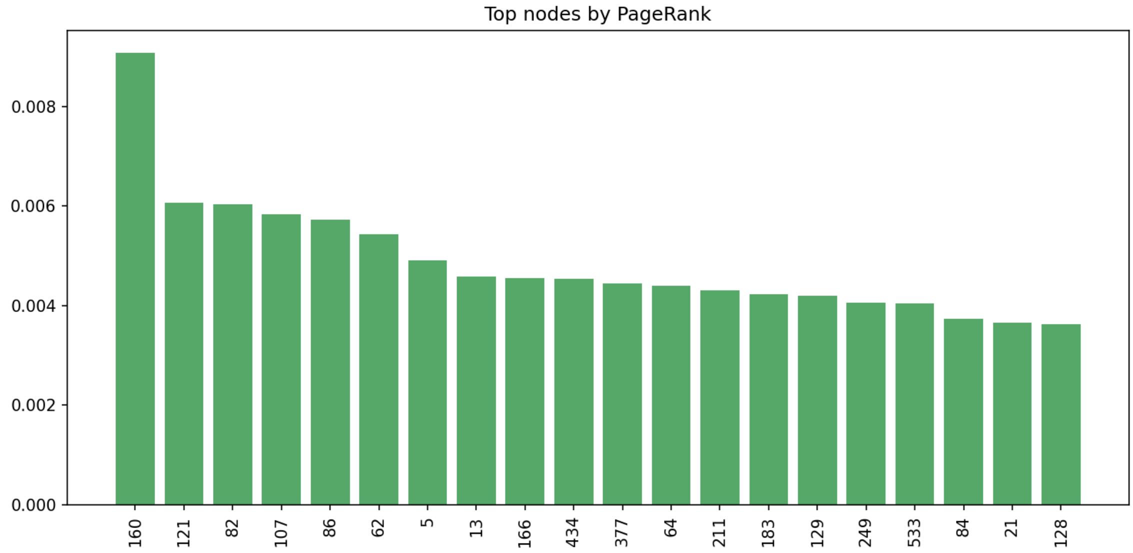

Figure 4: Top nodes by PageRank (bar chart)

3. Node Classification (Label Propagation)

Asynchronous label propagation communities.

Observations: Label propagation uncovers cohesive groups that likely correspond to organizational units, projects, or frequent correspondents. A few large communities typically account for most nodes, with several smaller groups at the periphery.

Interpretation: The distribution of community sizes suggests modular structure. Visual clusters in the community map align with these sizes; boundaries are porous where bridging nodes connect two modules—these nodes often reappear with high betweenness in the influence analysis.

community	size
0	969
19	1
21	1
22	1
23	1
24	1
25	1
26	1
27	1
28	1
29	1
30	1
31	1
32	1
33	1

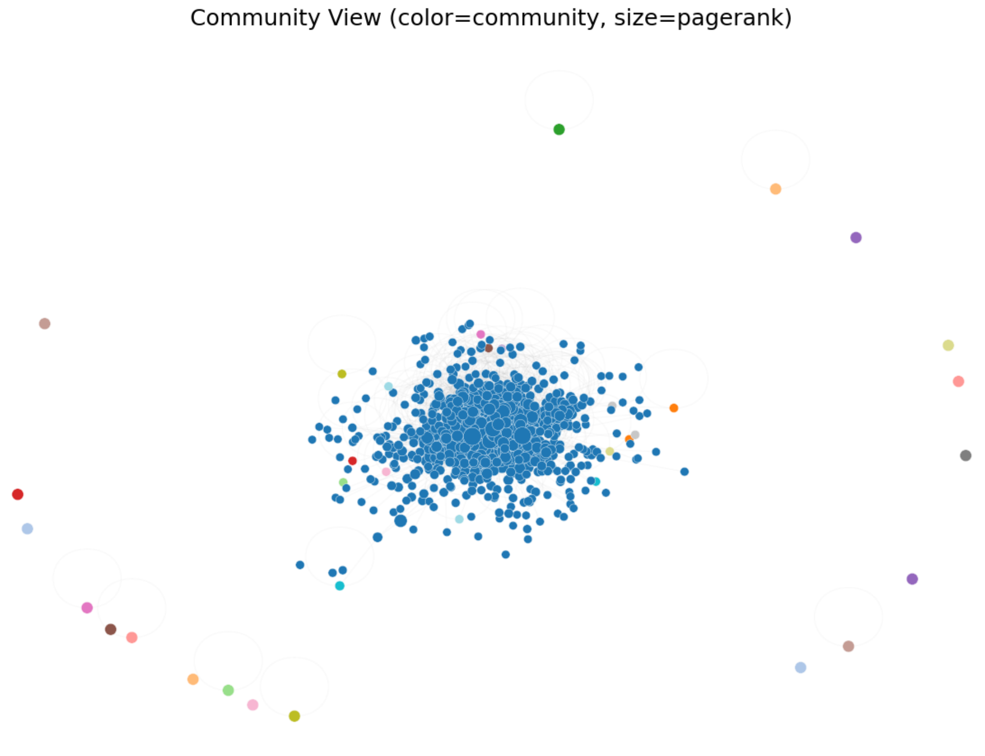

Figure 5: Communities (Python-generated equivalent)

4. Influence Analysis (PageRank & Betweenness)

PageRank for global influence; Betweenness for brokerage across communities.

Observations: High PageRank nodes generally sit in the dense core, while top betweenness nodes are often boundary spanners connecting modules. When a node ranks highly on both, it's a critical actor for both diffusion and bridging.

Interpretation: Use the top-influencers chart to identify specific candidates for targeted messaging. In the influence map, larger nodes (betweenness) with strong color (PageRank) warrant special attention as both gatekeepers and amplifiers.

node	pagerank	betweenness	degree
160.0	0.0090709484950265	0.0874147349363879	347.0
121.0	0.0060687854256318	0.0278415388258006	234.0
82.0	0.0060307053368408	0.0278807411351142	233.0
107.0	0.0058380960432493	0.0243403121826939	221.0
86.0	0.0057215196123223	0.0377885326911519	218.0
62.0	0.0054316159798508	0.0225098451925391	216.0
5.0	0.0049141637345112	0.0309946865452777	171.0
13.0	0.004589938693459	0.0235649895706901	180.0
166.0	0.0045516651162006	0.0176393735896517	177.0
434.0	0.0045327987830447	0.0154127096447369	185.0

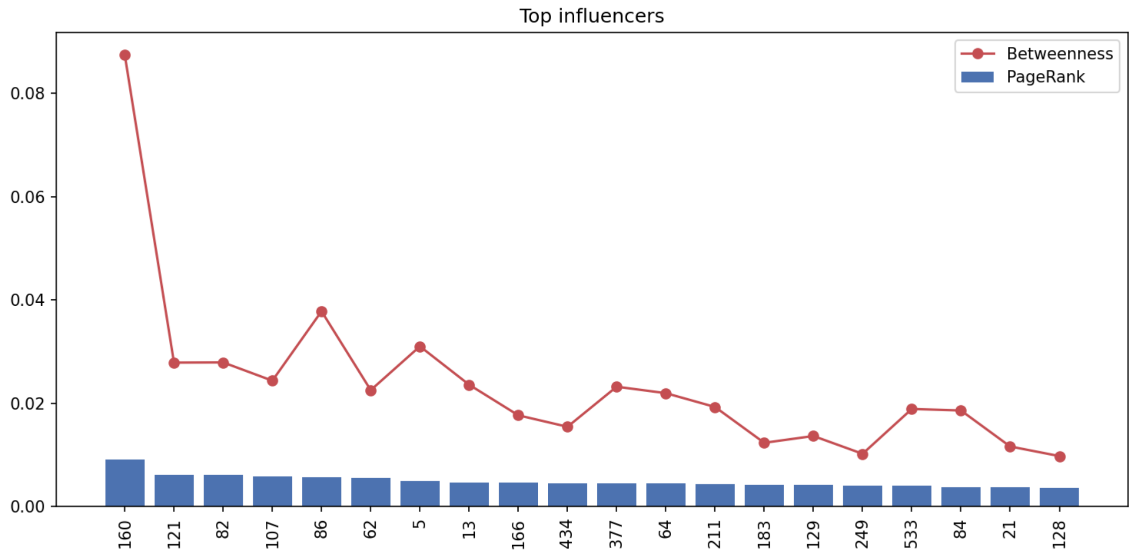

Figure 6: Top influencers (bar/line)

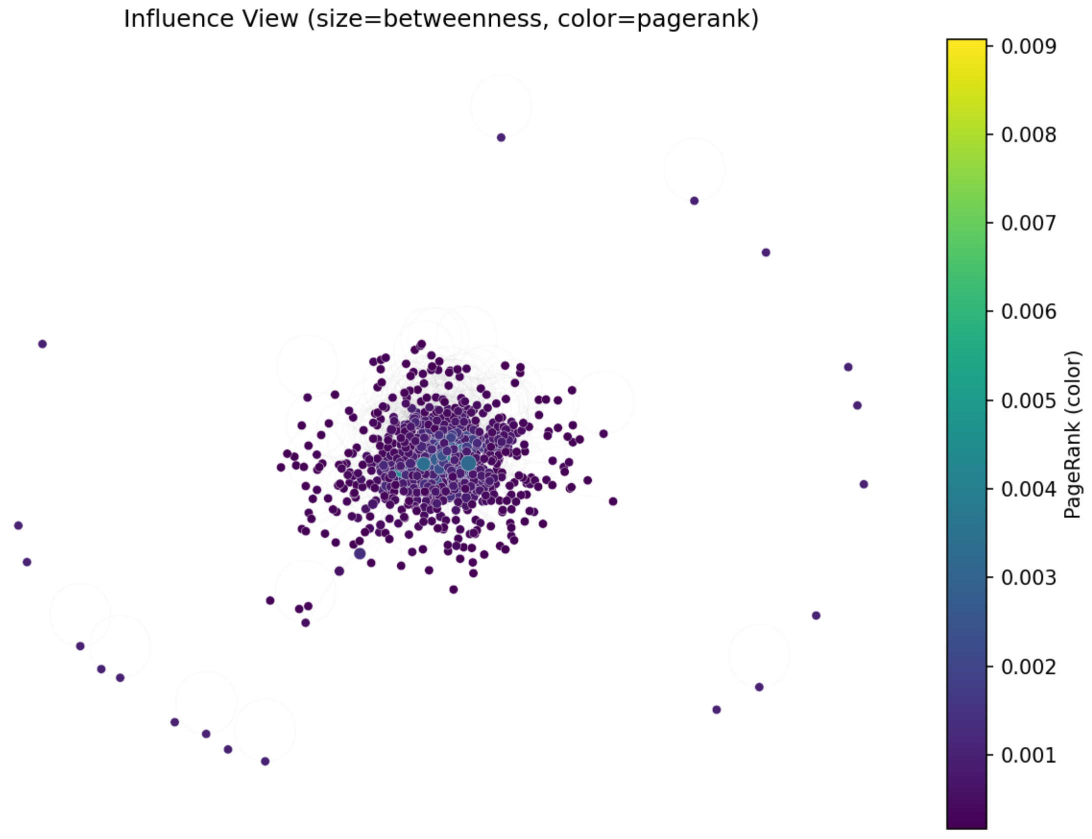

Figure 7: Influence view (size=betweenness, color=pagerank)

5. Link Prediction (Adamic-Adar, Jaccard, Preferential Attachment)

Hold-out 10% edges; compute heuristic scores on train graph; evaluate ROC/AUC.

Observations: Local-neighborhood metrics like Adamic–Adar often perform strongly in social graphs where shared neighbors are informative. Jaccard rewards exclusive overlap, while Preferential Attachment favors globally high-degree pairs—useful when growth is driven by popularity.

Interpretation: The ROC curves and AUC table summarize ranking quality across thresholds. The best curve bows furthest toward the top-left. Differences between metrics suggest whether closure (common neighbors) or popularity (degree) drives new links in this network.

metric	auc
adamic_adar	0.9468674701384132
jaccard	0.935174248697092
pref_attach	0.8597603122394544

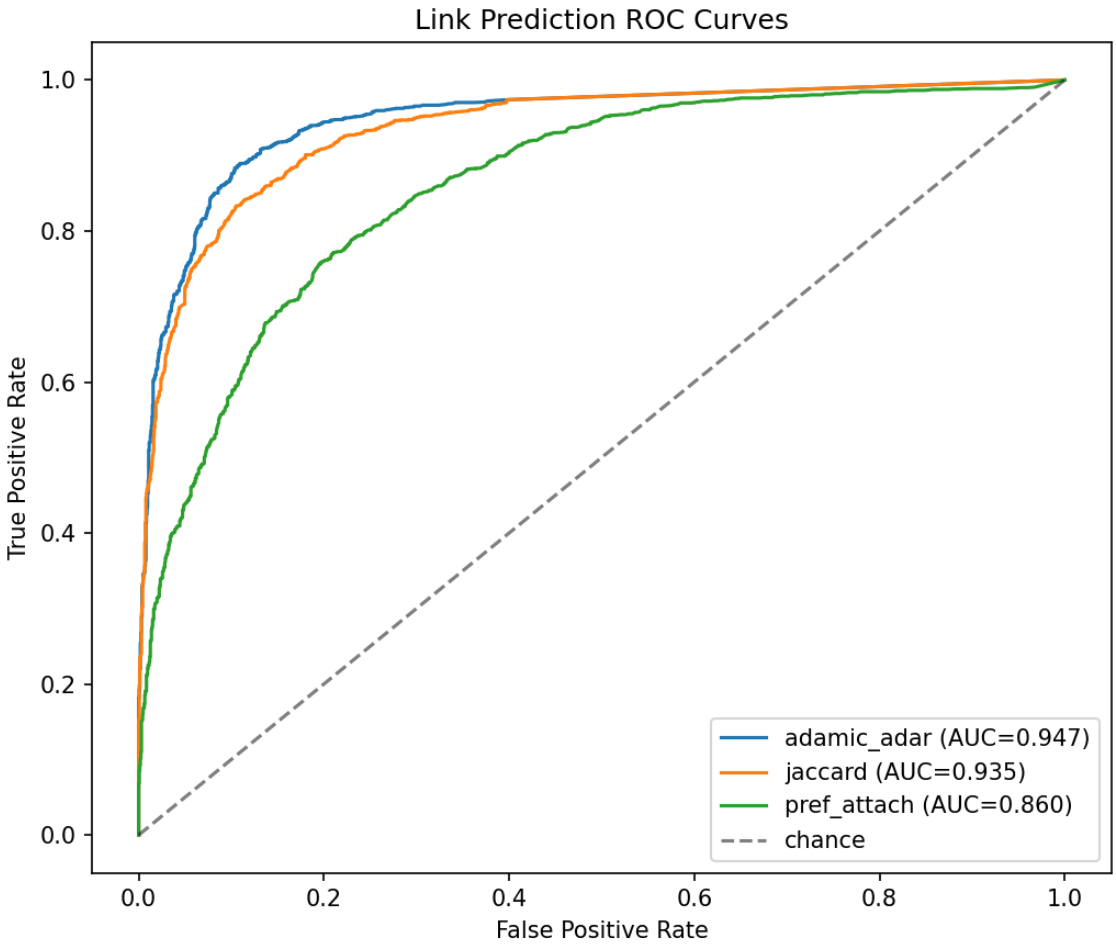

Figure 8: ROC curves for link prediction

6. Anomaly Detection (IsolationForest on egonet features)

Features: degree, clustering, average neighbor degree, egonet edges; Model: IsolationForest.

Observations: The model flags structurally unusual nodes—e.g., high degree but low clustering (broadcast hubs), or low degree with unexpectedly high clustering (insular ties). These outliers can reflect unique roles, errors, or atypical communication patterns.

Interpretation: Treat anomalies as hypotheses for follow-up, not conclusions. Cross-reference with domain knowledge (department, role, time) to determine whether they are benign (e.g., mailing lists) or risk-relevant (e.g., chokepoints or isolated actors).

node	degree	clustering	avg_neighbor_degree	ego_edges	decision_function	
160.0	347.0	0.0935119649477586	56.73198847262248	6177.0	1.0	-0.1259716934781769
121.0	234.0	0.1728989401403194	70.74358974358974	5056.0	1.0	-0.0815577059253059
82.0	233.0	0.1660831921701487	69.78111587982832	4828.0	1.0	-0.0814038916550311
107.0	221.0	0.1700389594068116	70.77828054298642	4462.0	1.0	-0.0651324610742785
62.0	216.0	0.1520336975121758	68.41203703703704	3856.0	1.0	-0.0565494375542041
86.0	218.0	0.1205857019810508	61.821100917431195	3194.0	1.0	-0.052512605438006
434.0	185.0	0.2011049060229388	76.41621621621621	3685.0	1.0	-0.0213113652596759
13.0	180.0	0.110899511204215	58.111111111111114	2082.0	1.0	-0.0128058498699874
5.0	171.0	0.1070019723865877	57.046783625730995	1840.0	1.0	-0.0072600496357142
882.0	2.0	1.0	262.0	5.0	1.0	-0.0025955469404698

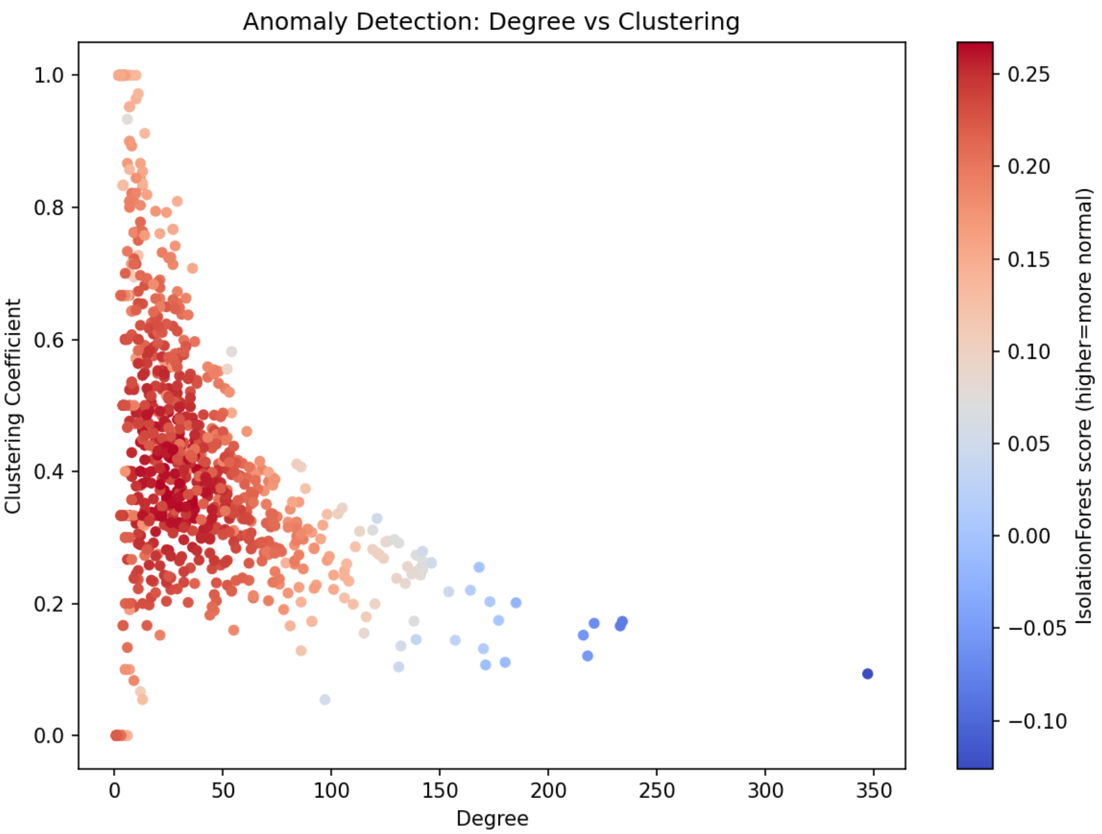

Figure 9: Anomaly scatter (Degree vs Clustering)

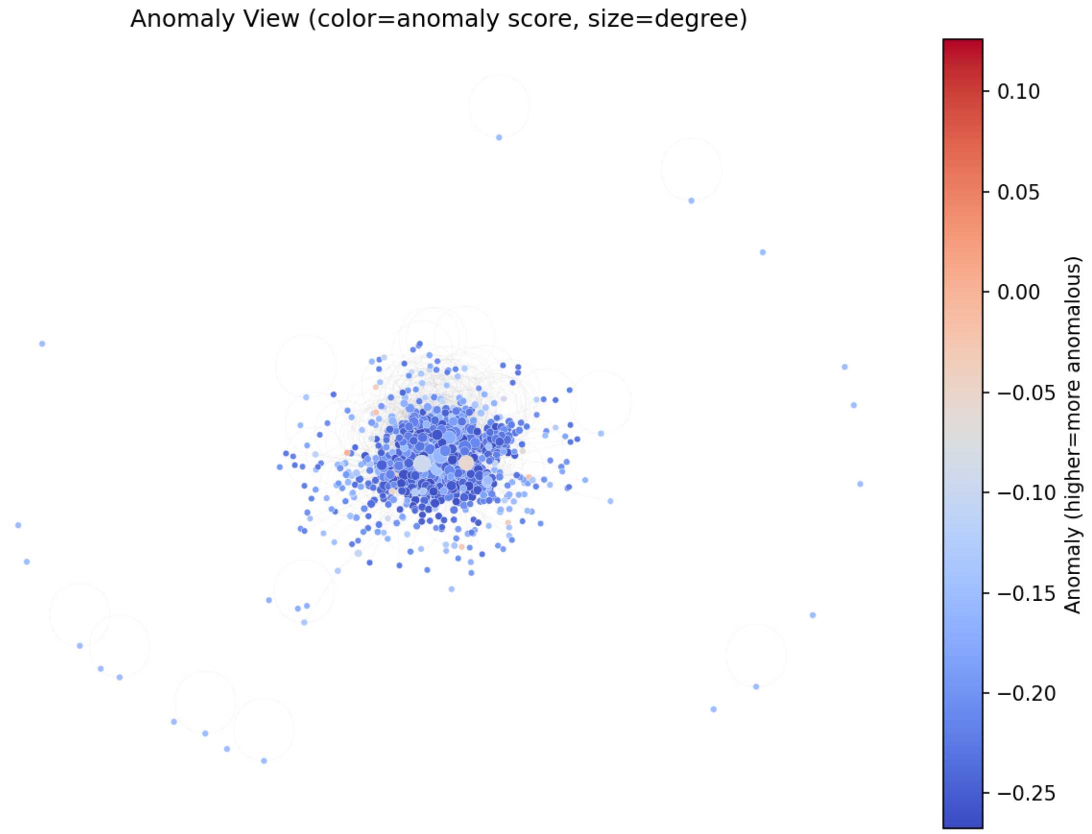

Figure 10: Anomaly view (color=anomaly score, size=degree)

***