Mining of Massive Datasets Project:

Implementing HITS and SimRank on Various Social Network Graph Datasets

Submitted by:
Sushovan Pan
ID: B2330054

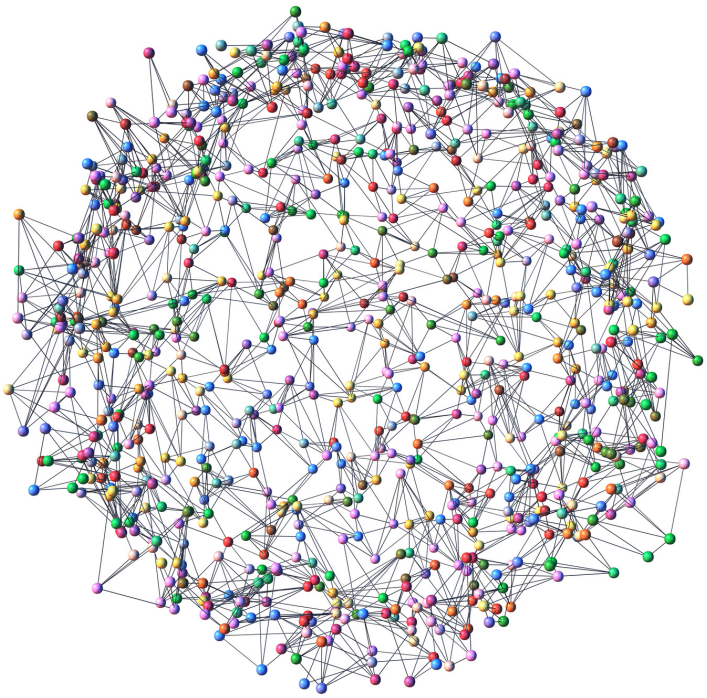

Figure 1: Sample Image

Department of Computer Science
Ramakrishna Mission Educational and Research Institute
Belur Math, Howrah 711202, West Bengal, India

February 15, 2025

Abstract

This project report presents the implementation of the HITS and SimRank algorithms on various social network graph datasets. The study aims to analyze the effectiveness of these algorithms in identifying influential nodes and measuring node similarity within large-scale networks. The report covers the theoretical foundations, implementation workflows, dataset selections, and potential applications of both algorithms.

Acknowledgments

I would like to thank Swami Shastravidyananda Maharaj of the Department of Computer Science for their guidance and support throughout this project.

Contents

1 Introduction to HITS Algorithm **4**
- 1.1 Mathematical Formulation . . . 4
- 1.2 Applications of HITS . . . 4

2 How HITS Works **5**

3 Dataset Selection **6**
- 3.1 Criteria for Selection . . . 6
- 3.2 Examples of Datasets Used . . . 6

4 Implementation Workflow **7**
- 4.1 Step-by-Step Process . . . 7

5 Use Cases **8**
- 5.1 Influential Pages on Social Media Graphs . . . 8
- 5.2 Detecting Malicious Users . . . 8

6 Introduction to SimRank Algorithm **9**
- 6.1 Key Features of SimRank . . . 9

7 How SimRank Works **10**
- 7.1 Key Steps in SimRank . . . 10
- 7.2 Benefits of SimRank . . . 10

8 Dataset Selection for SimRank **11**
- 8.1 Criteria for Selection . . . 11
- 8.2 Examples of Datasets Used . . . 11

9 Implementation Workflow for SimRank **12**
- 9.1 Step-by-Step Process . . . 12

10 Use Cases of SimRank **13**
- 10.1 Personalized Recommendations . . . 13
- 10.2 Link Prediction . . . 13

11 Sample Outputs **14**
- 11.1 HITS on Facebook Dataset . . . 14
- 11.2 HITS on Twitter Dataset . . . 21
- 11.3 SimRank on Twitter US Congress Dataset . . . 25
- 11.4 SimRank on Email Dataset . . . 27

12 Future Plans **32**
12.1 Optimizing Algorithms 32
12.2 Advanced Applications 32
12.3 Integration and Deployment 32

13 References **33**

Chapter 1

Introduction to HITS Algorithm

HITS (Hyperlink-Induced Topic Search) is a ranking algorithm that analyzes the structure of a directed graph, primarily used in web link analysis. It assigns two distinct scores to each node in the graph:

- **Authority Score:** Represents the value or credibility of the information that a node provides. Nodes with high authority scores are considered reliable sources of information.
- **Hub Score:** Reflects how well a node connects to other authoritative nodes, effectively acting as a guide to valuable sources.

1.1 Mathematical Formulation

The HITS algorithm operates using two equations:

$$h(i) = \sum_{j \in L(i)} a(j) \quad \text{(Hub score of node } i\text{)} \tag{1.1}$$

$$a(i) = \sum_{j \in L^{-1}(i)} h(j) \quad \text{(Authority score of node } i\text{)} \tag{1.2}$$

where $L(i)$ represents the set of nodes that node i points to, and $L^{-1}(i)$ represents the set of nodes that point to i. These scores are computed iteratively until convergence is reached.

1.2 Applications of HITS

HITS is widely used in various domains, including:

- **Social Network Analysis:** Identifying influential users or pages by analyzing connections between them.
- **Search Engines:** Improving the relevance of search results by ranking web pages based on their authority and hub scores.
- **Fraud Detection:** Identifying abnormal patterns or fake accounts that manipulate the network.

Chapter 2

How HITS Works

The HITS algorithm follows an iterative approach to calculate hub and authority scores for each node in a directed graph. The steps include:

1. **Initialization:** Assign an initial score (e.g., 1) to each node for both authority and hub values.
2. **Score Updates:**
   - **Hub Score:** Updated as the sum of the authority scores of all nodes it points to.
   - **Authority Score:** Updated as the sum of the hub scores of all nodes pointing to it.
3. **Normalization:** After each update, normalize the scores to prevent them from growing indefinitely.
4. **Convergence:** Repeat the score update process until the scores stabilize, indicating convergence.

The output consists of ranked lists of nodes based on their hub and authority scores, providing insights into their roles within the network.

Chapter 3

Dataset Selection

3.1 Criteria for Selection

Selecting the right dataset is crucial for the successful implementation of graph-based algorithms like HITS and SimRank. The datasets must fulfill specific criteria:

- **Graph-Structured Data:** The dataset should inherently represent a graph structure with nodes (representing entities) and edges (representing relationships between these entities).
- **Large-Scale Datasets:** Using large datasets helps in evaluating the scalability and performance of the algorithms under realistic conditions. The complexity and computation time of both HITS and SimRank depend heavily on the size of the graph.
- **Relevance to the Problem Domain:** The datasets should align with the use cases, such as social network analysis, recommendation systems, or fraud detection.

3.2 Examples of Datasets Used

- **Facebook Users Dataset:** Represents connections between users, capturing the friendship or interaction graph of a subset of Facebook users. It is often used to identify influential users and analyze social interactions.
- **Twitter Users Dataset:** Represents follower relationships among Twitter users. It is highly relevant for analyzing influence in social networks, detecting fake accounts, and understanding the spread of information.

Chapter 4

Implementation Workflow

4.1 Step-by-Step Process

The implementation of HITS and SimRank involves the following detailed steps:

1. **Load Graph Data:**
   - Prepare raw data in formats like CSV or MTX, where edges represent relationships between nodes.
   - Use libraries such as NetworkX to convert the raw data into a graph representation. Each node in the graph represents an entity, and edges define their relationships.
   - Clean and preprocess the data to remove redundant edges, self-loops, or missing values.
2. **Run HITS Algorithm:**
   - Initialize the hub and authority scores for all nodes to 1.
   - Iteratively update the hub and authority scores as follows:
     - Update hub scores based on the sum of authority scores of linked nodes.
     - Update authority scores based on the sum of hub scores of linking nodes.
   - Normalize the scores after each iteration to prevent overflow and ensure convergence.
   - Continue the iterative process until the scores stabilize or a predefined number of iterations is reached.
3. **Analyze Results:**
   - Identify the nodes with the highest hub and authority scores. High authority scores indicate nodes that are reliable sources of information, while high hub scores identify nodes that effectively direct users to authoritative sources.
   - Visualize the results using tools such as heatmaps, graphs, or network diagrams to interpret the findings better.
   - Detect anomalies or trends, such as clusters of influential nodes or outliers indicating suspicious activity.

Chapter 5

Use Cases

5.1 Influential Pages on Social Media Graphs

In social media networks, certain pages or users hold significant influence due to their centrality in the network. The HITS algorithm can identify such influential entities by analyzing their hub and authority scores:

- **High Authority Scores:** Pages or users with high authority scores are considered reputable and are frequently referred to by other nodes in the network. For instance, a well-established news outlet or a popular public figure often has a high authority score.
- **High Hub Scores:** Nodes with high hub scores effectively link to authoritative sources, acting as information aggregators or influencers. For example, a page that curates content from various trusted sources can have a high hub score.

These insights can be utilized for:

- Designing targeted marketing campaigns.
- Identifying key influencers for promotional activities.
- Enhancing recommendation systems by focusing on authoritative and hub-like nodes.

5.2 Detecting Malicious Users

Malicious users or entities in a network often exhibit unusual patterns that can be detected through their hub and authority scores:

- **High Scores in Both Categories:** A node with exceptionally high hub and authority scores may indicate suspicious activity, such as bots creating fake links to manipulate rankings or influence metrics.
- **Threshold-Based Filtering:** By setting percentile-based thresholds for hub and authority scores, nodes with abnormal behavior can be flagged for further investigation.

These insights can help:

- Detect and mitigate fake accounts or coordinated bot activity in social networks.
- Enhance the robustness of recommendation systems by filtering out manipulated content.
- Improve network security by identifying and isolating malicious nodes.

Chapter 6

Introduction to SimRank Algorithm

SimRank is a graph-based similarity measure that quantifies how similar two nodes are based on their neighbors. The core idea is that "*two nodes are similar if their neighbors are similar.*"

Mathematically, SimRank is defined as:

$$S(i,j) = \begin{cases} 1 & \text{if } i = j \\ \frac{c}{|N(i)||N(j)|} \sum_{u \in N(i)} \sum_{v \in N(j)} S(u,v) & \text{if } i \neq j \end{cases}$$

where:

- $S(i,j)$ is the similarity score between nodes i and j.
- c is the decay factor ($0 < c < 1$) controlling the influence of neighbors.
- $N(i)$ and $N(j)$ are the sets of neighbors of nodes i and j, respectively.

The algorithm initializes all nodes with $S(i,j) = 0$ for $i \neq j$ and $S(i,i) = 1$. It then iteratively updates scores using the formula until the values converge or a maximum number of iterations is reached.

6.1 Key Features of SimRank

- **Recursive Definition:** The similarity of two nodes is influenced by the similarity of their respective neighbors.
- **Similarity Scale:** Scores range from 0 (no similarity) to 1 (identical nodes).
- **Decay Factor:** A parameter (c) controls the influence of neighbors and ensures convergence during computation.

Chapter 7

How SimRank Works

The SimRank algorithm is a similarity measure designed to quantify the similarity between two nodes in a graph. The underlying principle of SimRank is that "two nodes are similar if their neighbors are similar." This recursive definition makes SimRank a powerful tool for various graph-based applications.

7.1 Key Steps in SimRank

- **Similarity Initialization:** Each node in the graph is assigned a similarity score of 1 with itself ($S(i,i) = 1$), and a score of 0 with all other nodes ($S(i,j) = 0$ for $i \neq j$).
- **Similarity Propagation:** The similarity score between two nodes i and j is updated iteratively using the formula:

$$S(i,j) = \frac{c}{|N(i)||N(j)|} \sum_{u \in N(i)} \sum_{v \in N(j)} S(u,v) \tag{7.1}$$

  where:
  - $N(i)$ and $N(j)$ represent the sets of neighbors of nodes i and j.
  - c is the decay factor (a value between 0 and 1) that controls the influence of neighbors.
- **Convergence:** The algorithm iteratively updates the similarity scores until they stabilize (i.e., the changes between iterations become negligible) or a predefined number of iterations is reached.

7.2 Benefits of SimRank

- Captures structural similarity between nodes based on their connectivity patterns.
- Suitable for various applications, such as recommendation systems, link prediction, and clustering.
- Scalable with sparse matrix optimizations for large graphs.

Chapter 8

Dataset Selection for SimRank

8.1 Criteria for Selection

Datasets for SimRank must meet the following criteria:

- **Graph Structure:** The dataset should represent relationships as nodes and edges, such as social or communication networks.
- **Scalability:** Large datasets are preferred to test the computational efficiency and scalability of the SimRank algorithm.

8.2 Examples of Datasets Used

- **Twitter US Congress Dataset:** Captures interactions between members of Congress on Twitter, including mentions, retweets, and replies.
- **Email Dataset:** Represents communication links within an organization or institution, ideal for analyzing relationships and connectivity patterns.

Chapter 9

Implementation Workflow for SimRank

9.1 Step-by-Step Process

The implementation of SimRank involves the following steps:

1. **Load Graph Data:**
   - Prepare raw data (e.g., edge lists, adjacency matrices) and convert them into a graph representation.
   - Use graph processing libraries such as NetworkX or custom parsers to handle large datasets.
2. **Run SimRank Algorithm:**
   - Initialize the similarity matrix with $S(i, i) = 1$ and $S(i, j) = 0$ for $i \neq j$.
   - Iteratively compute similarity scores using the propagation formula, leveraging sparse matrix optimizations for large graphs.
3. **Interpret Results:**
   - Analyze the resulting similarity matrix to identify patterns, clusters, or anomalies.
   - Visualize similarity scores using heatmaps, similarity matrices, or graph layouts.

Chapter 10

Use Cases of SimRank

10.1 Personalized Recommendations

SimRank can be used to provide personalized recommendations by identifying nodes with high similarity scores:

- **Product Recommendations:** Suggest items that are structurally similar to those already purchased or liked by a user.
- **User Recommendations:** Recommend users with similar behavior or connections in a social network.
- **Content Discovery:** Highlight content that aligns with a user's preferences based on graph similarity.

10.2 Link Prediction

SimRank helps predict potential connections by identifying structurally similar nodes:

- **Network Growth:** Suggest new links in a social or professional network, fostering connectivity.
- **Relationship Discovery:** Identify potential collaborations or interactions based on structural similarity in research or organizational networks.
- **Fraud Detection:** Highlight suspicious or unexpected connections that may indicate fraudulent behavior.

Chapter 11

Sample Outputs

11.1 HITS on Facebook Dataset

Outputs show the top-ranked hubs and authorities with a brief analysis of their significance.

```
Original graph: 4039 nodes, 88234 edges.
Reduced graph: 4039 nodes, 88234 edges.

Top Hub Nodes:
Node: 1912, Hub Score: 0.010229
Node: 1993, Hub Score: 0.008594
Node: 1985, Hub Score: 0.008440
Node: 1917, Hub Score: 0.008364
Node: 1983, Hub Score: 0.008334
Node: 1938, Hub Score: 0.008301
Node: 1943, Hub Score: 0.008301
Node: 2078, Hub Score: 0.008275
Node: 1962, Hub Score: 0.008273
Node: 2059, Hub Score: 0.008257

Top Authority Nodes:
Node: 2604, Authority Score: 0.007932
Node: 2611, Authority Score: 0.007859
Node: 2590, Authority Score: 0.007836
Node: 2607, Authority Score: 0.007763
Node: 2601, Authority Score: 0.007698
Node: 2560, Authority Score: 0.007686
Node: 2624, Authority Score: 0.007661
Node: 2602, Authority Score: 0.007659
```

Figure 11.1: Top Hub Nodes and Top Authority Nodes.

```
Suspicious Hub Nodes (Potential Malicious Users):
Node: 1912, Hub Score: 0.010229
Node: 2007, Hub Score: 0.001706
Node: 2189, Hub Score: 0.001050
Node: 2543, Hub Score: 0.002174
Node: 1941, Hub Score: 0.005639
Node: 2266, Hub Score: 0.006909
Node: 2347, Hub Score: 0.005091
Node: 2542, Hub Score: 0.002520
Node: 2026, Hub Score: 0.001109
Node: 2468, Hub Score: 0.002764
```

Figure 11.2: Sushpicious Hub Nodes.

```
Suspicious Authority Nodes (Potential Malicious Users):
Node: 2543, Authority Score: 0.006835
Node: 2266, Authority Score: 0.005464
Node: 2347, Authority Score: 0.006185
Node: 2542, Authority Score: 0.007212
Node: 2468, Authority Score: 0.005556
Node: 1983, Authority Score: 0.000912
Node: 1984, Authority Score: 0.000937
Node: 1985, Authority Score: 0.001105
Node: 1993, Authority Score: 0.001305
Node: 1997, Authority Score: 0.000888
Node: 2005, Authority Score: 0.001221
Node: 2020, Authority Score: 0.001245
```

Figure 11.3: Sushpicious Authority Nodes.

```
Common nodes in Top Hub Nodes and Suspicious Hub Nodes: {'1912', '1983', '1943', '2059', '1917', '1993', '1985', '1938', '2078', '1962'}
Common nodes in Top Authority Nodes and Suspicious Authority Nodes: {'2560', '2607', '2601', '2604', '2586', '2624', '2602', '2590', '2625', '2611'}
```

Figure 11.4: common nodes in top Hub and suspicious Hub nodes and common nodes in top Authority and suspicious Authority nodes.

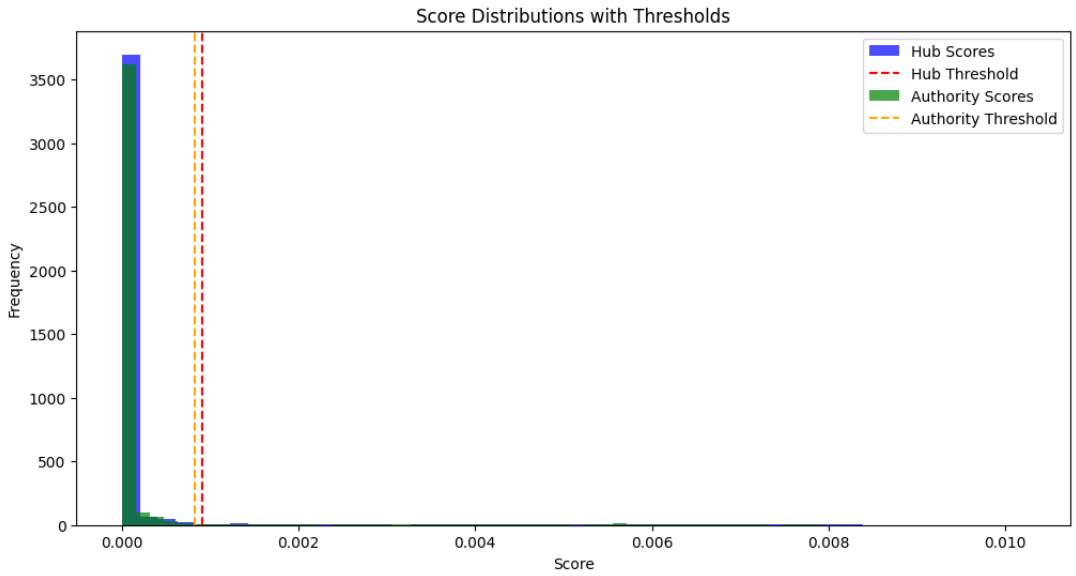

Figure 11.5: Hub and Authority Score distribution of the nodes,along with thereshold.

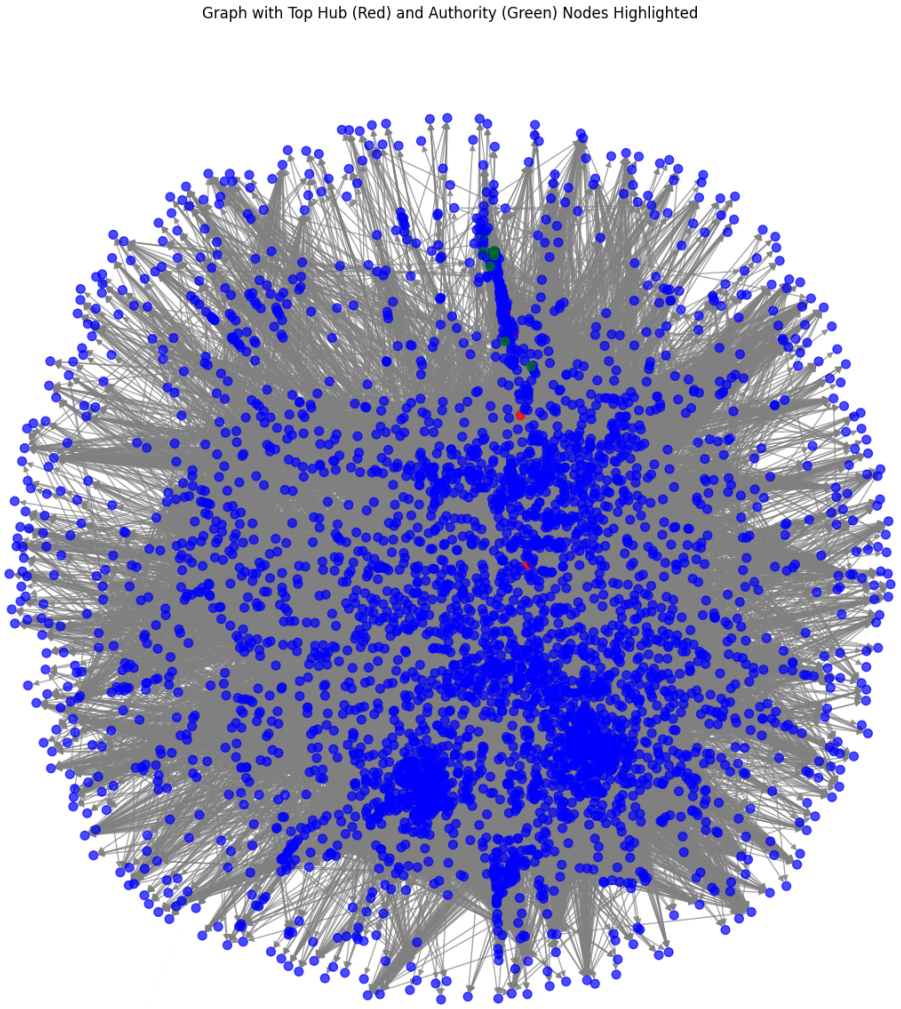

Figure 11.6: top Hub and Athority nodes highlighted in red and green.

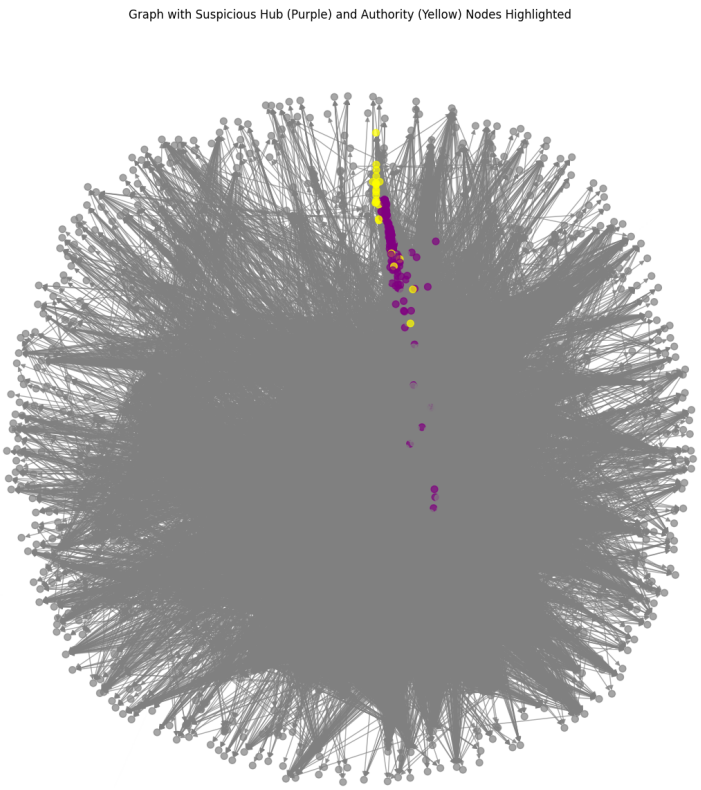

Figure 11.7: suspicious Hub and Athority nodes highlighted in purple and yellow.

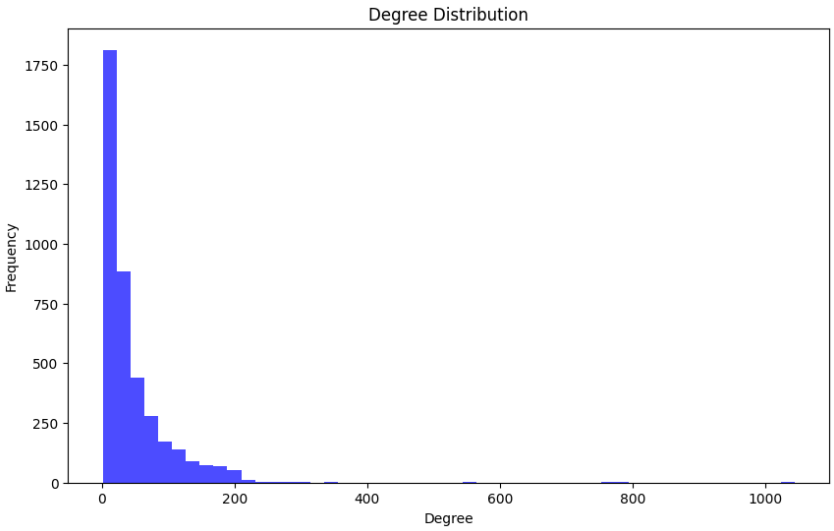

Figure 11.8: Degree distribution of the social media graph.

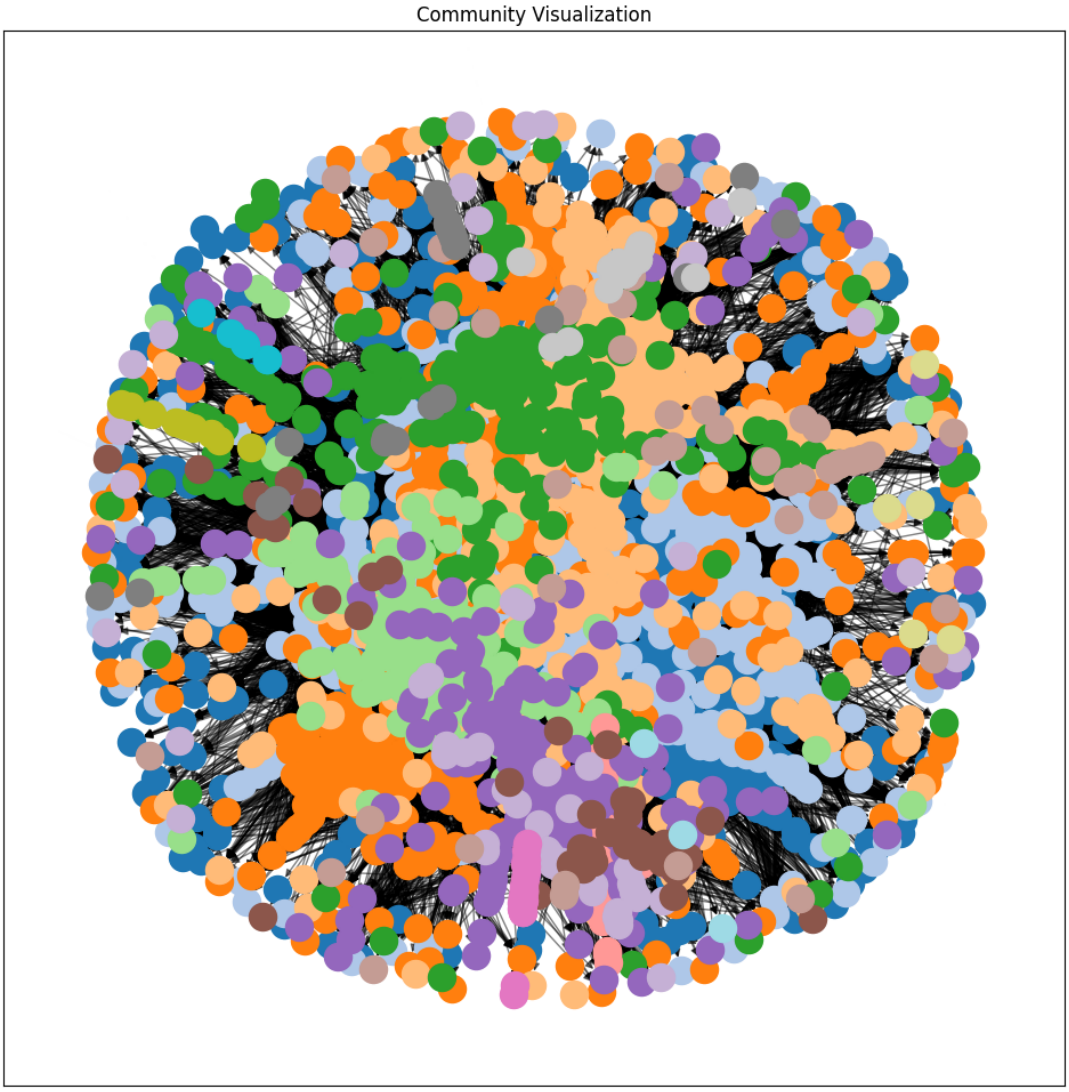

Figure 11.9: community visualization of the Facebook graph.

```
Top 5 Degree Centrality Nodes:
Node: 107, Score: 0.2588
Node: 1684, Score: 0.1961
Node: 1912, Score: 0.1870
Node: 3437, Score: 0.1355
Node: 0, Score: 0.0859

Top 5 Betweenness Centrality Nodes:
Node: 1684, Score: 0.0330
Node: 1912, Score: 0.0271
Node: 1718, Score: 0.0266
Node: 563, Score: 0.0130
Node: 1405, Score: 0.0101

Top 5 Closeness Centrality Nodes:
Node: 2642, Score: 0.1180
Node: 2649, Score: 0.1179
Node: 2629, Score: 0.1163
Node: 2643, Score: 0.1159
Node: 2543, Score: 0.1159
```

Figure 11.10: centrality scores.

```
Average Clustering Coefficient: 0.6055
```

Figure 11.11: average clustering coefficient of the social media graph.

```
Graph Diameter: 8
```

Figure 11.12: diameter of the social graph.

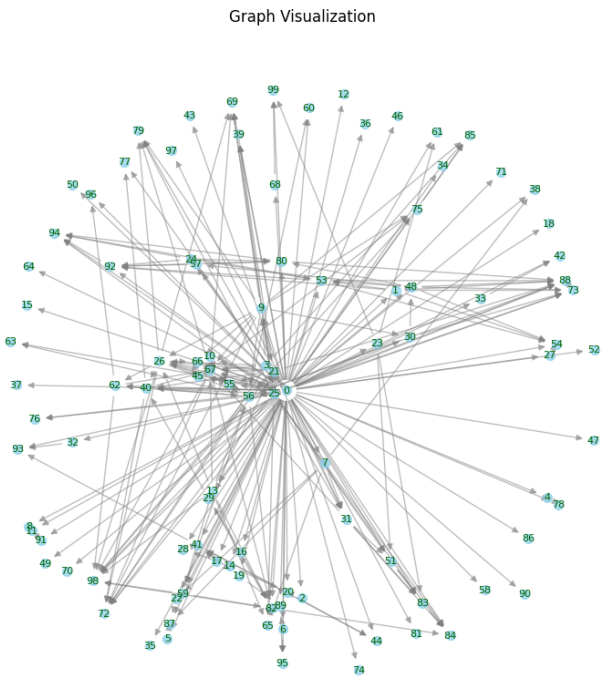

Figure 11.13: first 100 nodes visulaization of the graph.

11.2 HITS on Twitter Dataset

Highlights influential accounts, with insights into their role in the network.

```
Original graph: 81306 nodes, 1768149 edges.
Reduced graph: 68413 nodes, 1685163 edges.
Hub Threshold: 4.579473933737e-06, Suspicious Hubs: 6842
Authority Threshold: 4.545274087216759e-06, Suspicious Authorities: 6842
```

Figure 11.14: numeber of suspicious hub nodes and authority nodes, and their threshold score.

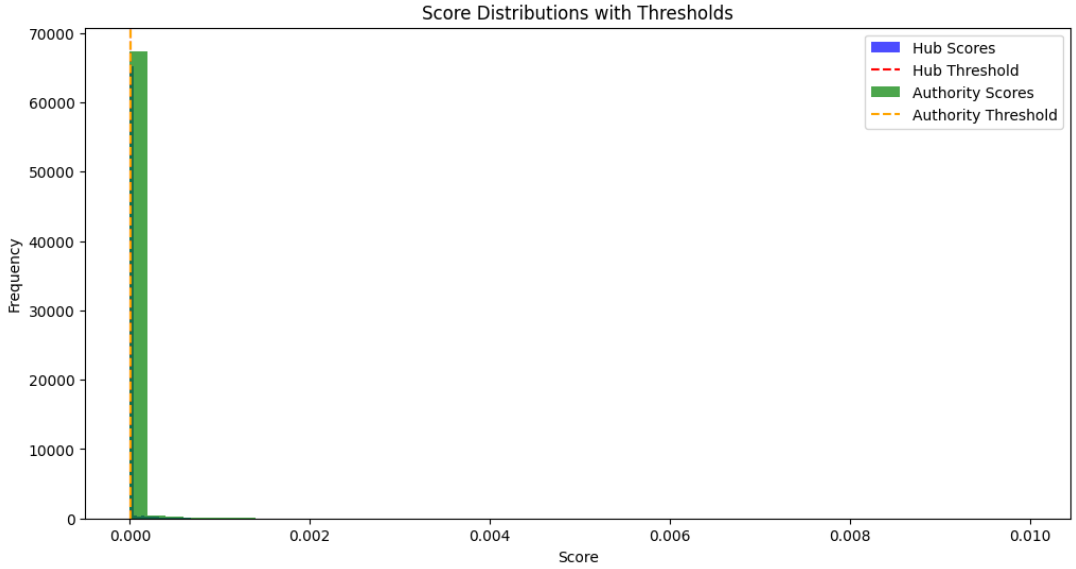

Figure 11.15: common nodes in top Hub and suspicious Hub nodes, and common nodes in top Authority and suspicious Authority nodes.

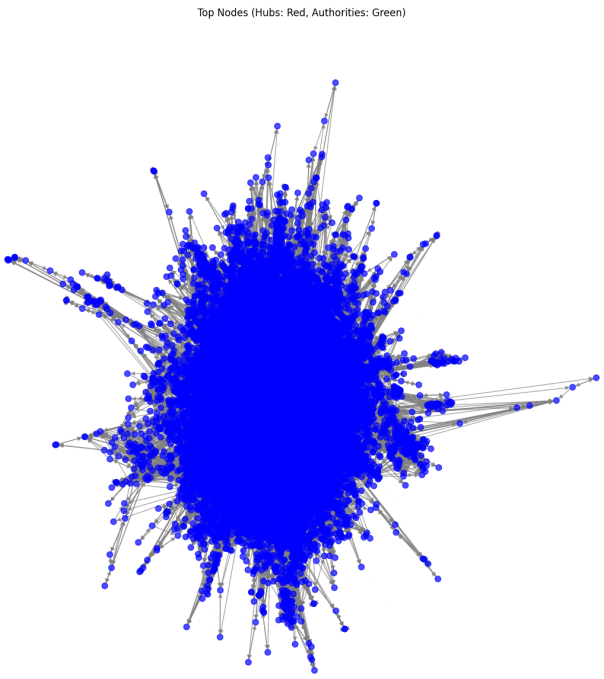

Figure 11.16: top Hub and Athority nodes highlighted in red and green.

```
Top Hub Nodes:
Node: 208132323, Hub Score: 0.002000
Node: 440963134, Hub Score: 0.001738
Node: 153226312, Hub Score: 0.001567
Node: 197504076, Hub Score: 0.001498
Node: 100581193, Hub Score: 0.001454
Node: 279787626, Hub Score: 0.001442
Node: 221829166, Hub Score: 0.001427
Node: 184097849, Hub Score: 0.001391
Node: 274153775, Hub Score: 0.001391
Node: 463952369, Hub Score: 0.001366

Top Authority Nodes:
Node: 40981798, Authority Score: 0.009715
Node: 43003845, Authority Score: 0.009363
Node: 22462180, Authority Score: 0.009183
Node: 34428380, Authority Score: 0.009161
Node: 31331740, Authority Score: 0.005457
Node: 27633075, Authority Score: 0.005448
Node: 18996905, Authority Score: 0.004948
Node: 83943787, Authority Score: 0.004809
Node: 117674417, Authority Score: 0.004786
Node: 238260874, Authority Score: 0.004449
```

Figure 11.17: Top Hub Nodes and Top Authority Nodes.

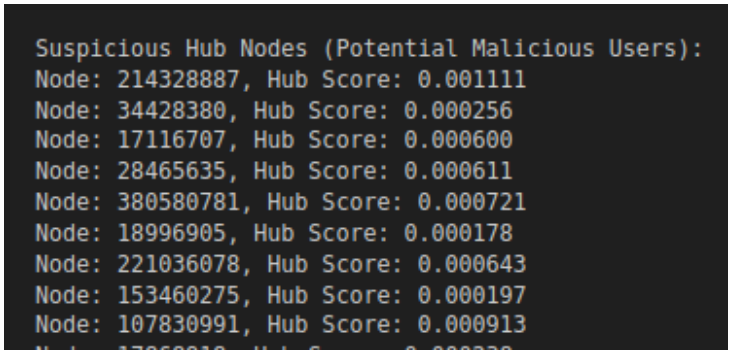

Figure 11.18: Sushpicious Hub Nodes.

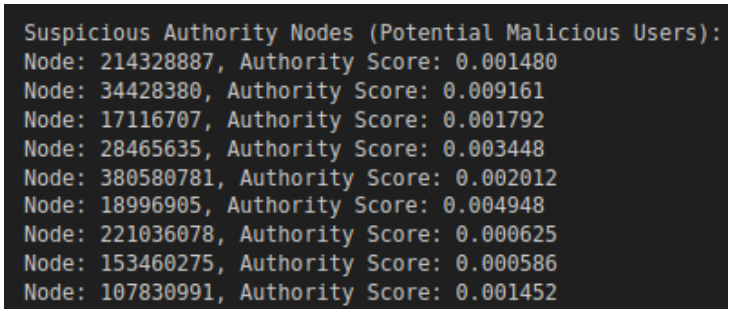

Figure 11.19: Sushpicious Authority Nodes.

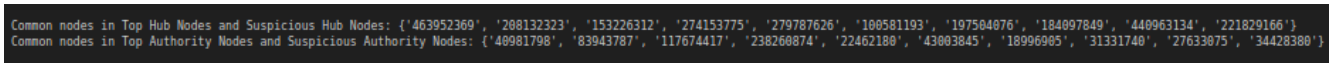

Figure 11.20: common nodes in top Hub and suspicious Hub nodes and common nodes in top Authority and suspicious Authority nodes.

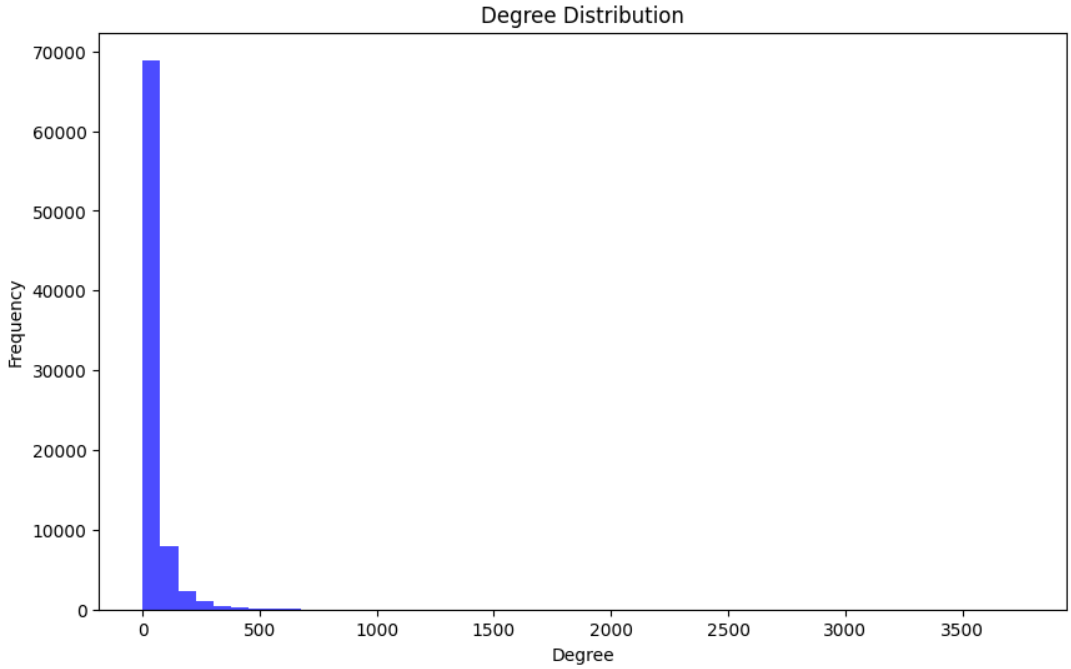

Figure 11.21: Degree distribution of the graph.

```
Detected 240 communities.
```

Figure 11.22: total communities of the Tweeter Graph.

11.3 SimRank on Twitter US Congress Dataset

Finiding the most similar graph pair of nodes from the Twitter US Congress Dataset.

SimRank Scores (Partial):

	0	4	12	18	25	30	46	55	58	59	...
0	1.000000	0.022752	0.021619	0.022940	0.022034	0.017925	0.020992	0.020598	0.016835	0.024535	...
4	0.022752	1.000000	0.018124	0.027585	0.024085	0.019527	0.019741	0.022202	0.019606	0.023637	...
12	0.021619	0.018124	1.000000	0.023134	0.018363	0.018759	0.018562	0.016986	0.016254	0.019562	...
18	0.022940	0.027585	0.023134	1.000000	0.023399	0.020691	0.020863	0.019786	0.018919	0.022959	...
25	0.022034	0.024085	0.018363	0.023399	1.000000	0.020788	0.020363	0.018978	0.016629	0.021845	...

5 rows × 475 columns

Figure 11.23: Sample simrank score of the nodes pair of the dataset garph.

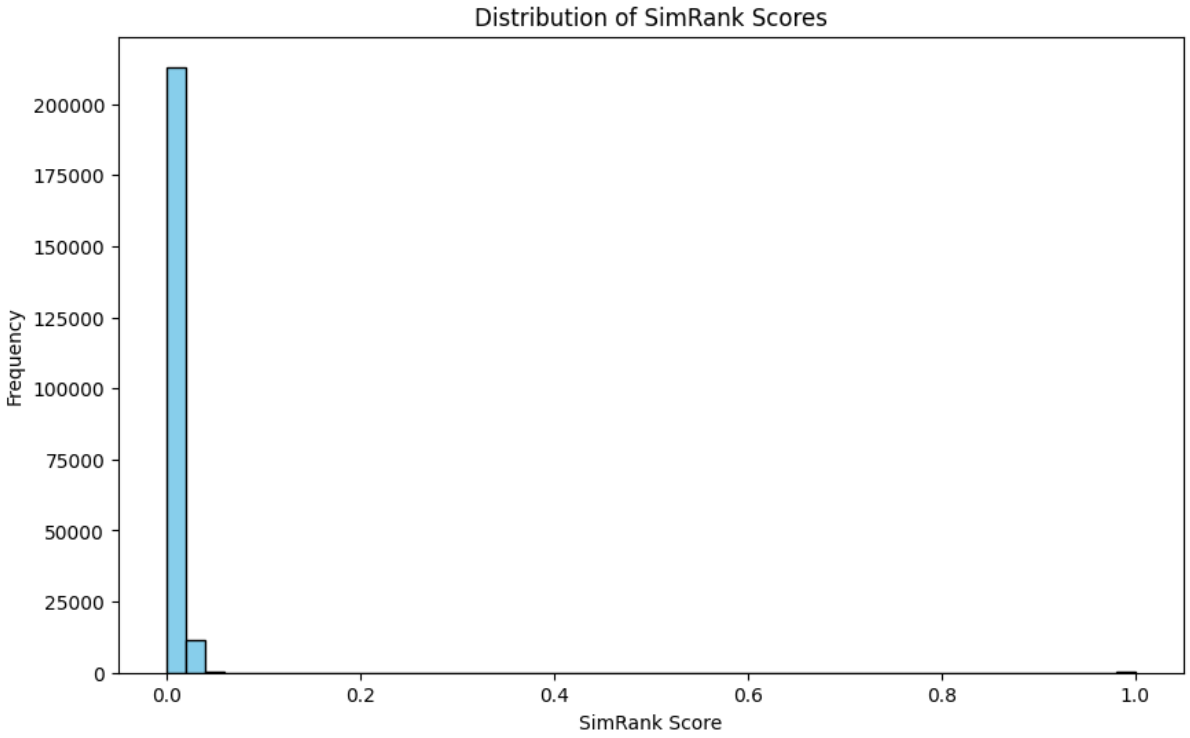

Figure 11.24: Simrank score distribution of the dataset graph.

Top 5 Most Similar Node Pairs:

	Node1	Node2	SimRank Score
0	67	34	0.213189
1	19	34	0.142825
2	258	337	0.127791
3	458	34	0.121901
4	402	258	0.111217

Figure 11.25: Top 5 most similar nodes pairs.

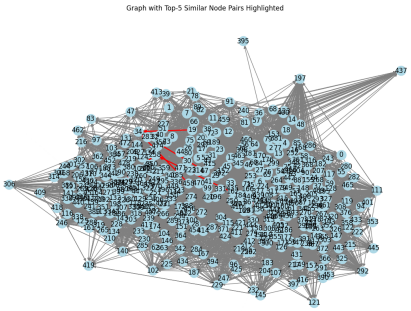

Figure 11.26: Graph with the Top similar Nodes pair highlighted.

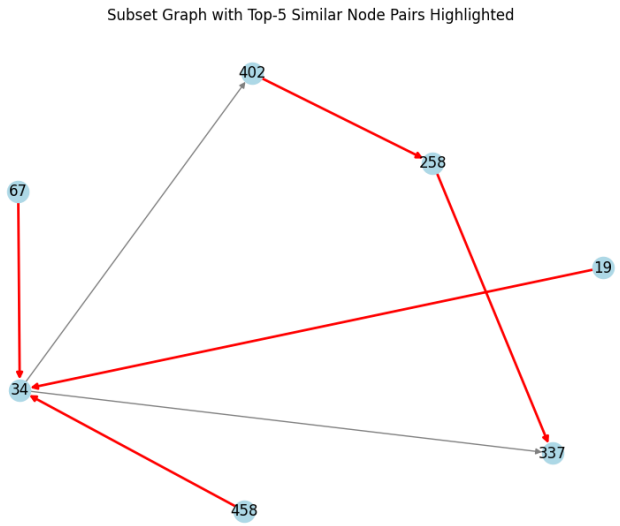

Figure 11.27: Top 5 most similar nodes pairs highlighted in the subgarph.

11.4 SimRank on Email Dataset

Demonstrates similarity scores for specific nodes, revealing communication patterns.

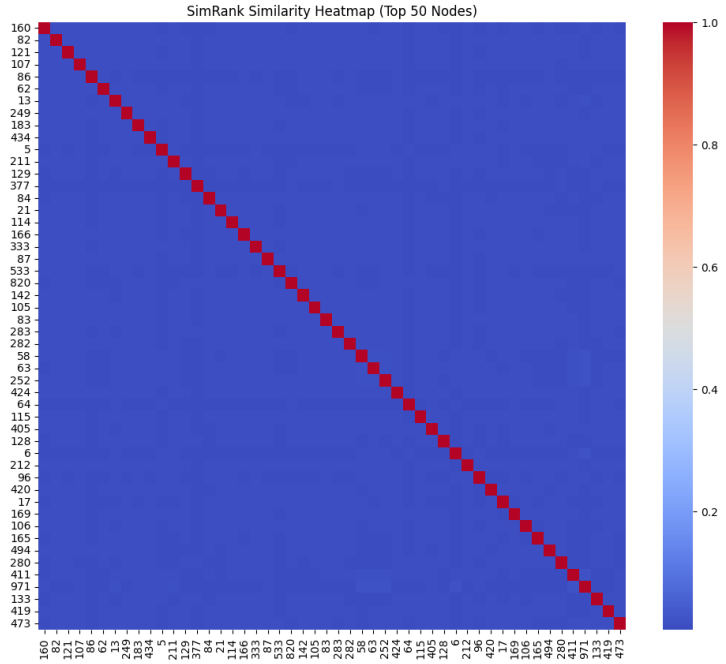

Figure 11.28: Simaranks simialrity heatmap of the top 50 nodes.

department	0	1	2	3	4	5	6	7	8	9	...
0	0.010867	0.010738	0.008005	0.008163	0.008381	0.009884	0.009819	0.006959	0.007576	0.006248	...
1	0.031071	0.028253	0.007771	0.007875	0.007879	0.009008	0.008102	0.006854	0.006988	0.005944	...
2	0.008238	0.008788	0.007975	0.008161	0.008163	0.009062	0.008739	0.008158	0.008969	0.007212	...
3	0.010203	0.010738	0.009454	0.009501	0.010163	0.010253	0.008874	0.008099	0.009027	0.006423	...
4	0.010080	0.010133	0.008212	0.008527	0.008807	0.009254	0.009231	0.007515	0.007860	0.006543	...

5 rows × 1005 columns

Figure 11.29: Sample simrank score of the Graph.

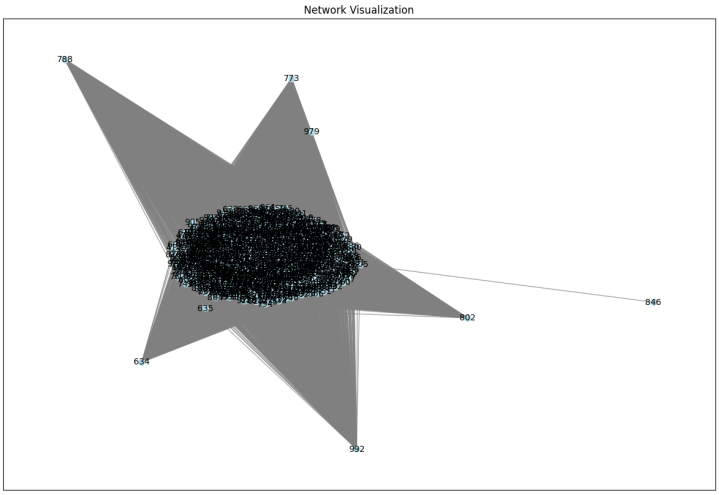

Figure 11.30: Visualization of the whole Graph.

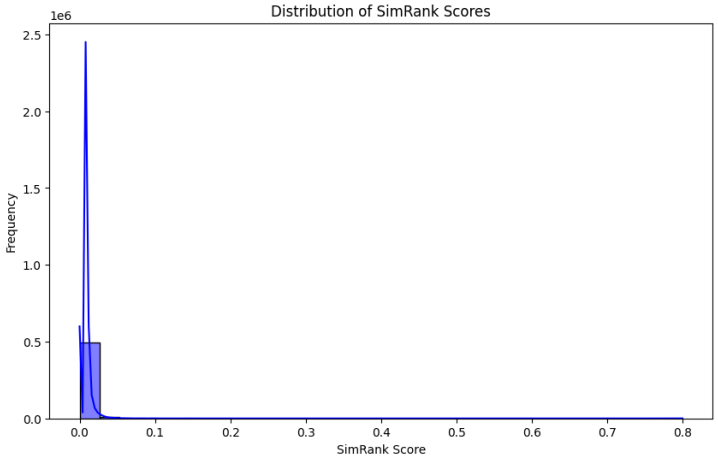

Figure 11.31: SimRanks Score Distribution, of the nodes of the graph.

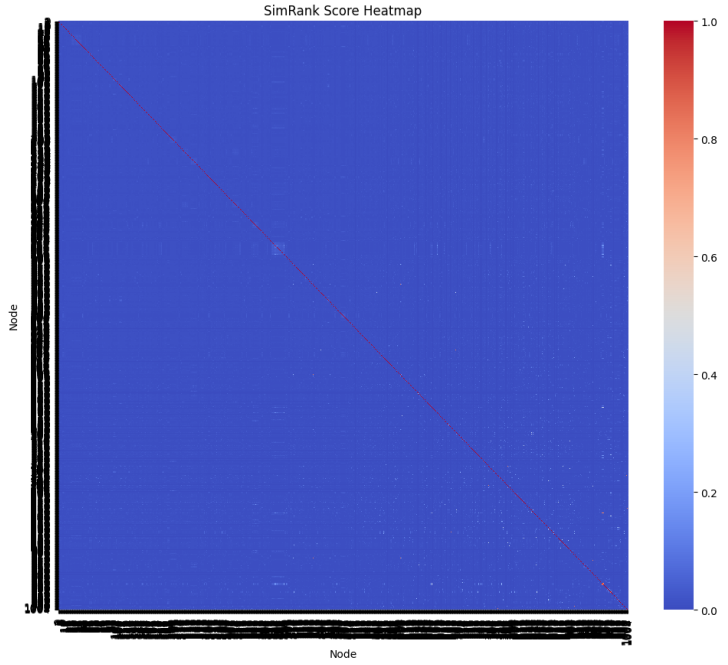

Figure 11.32: heatmap of the whole graph, Simrank Score.

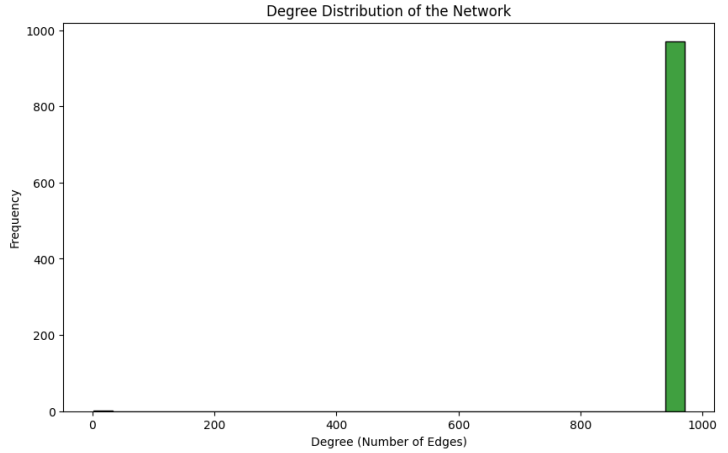

Figure 11.33: Degree DIstribution of the whole Graph.

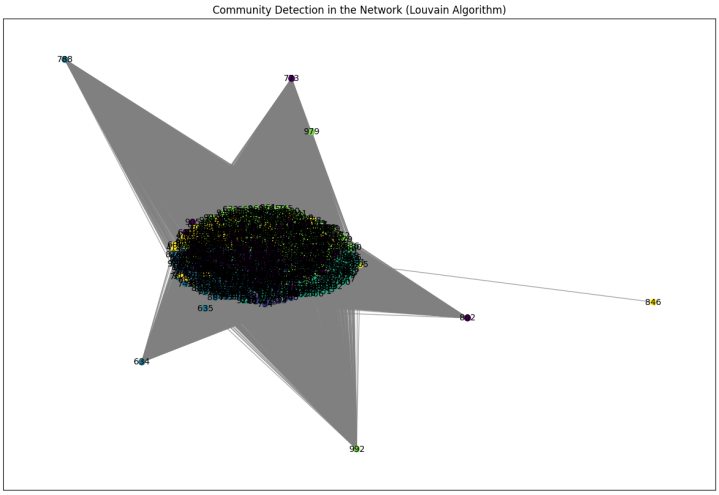

Figure 11.34: Community Detection In the Social Network Graph.

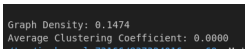

Figure 11.35: Graph density and average clustering coefficient.

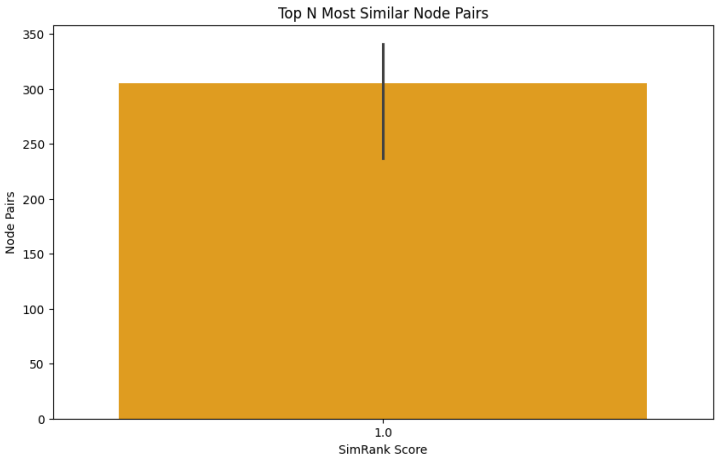

Figure 11.36: Top 10 most simlar nodes pair.

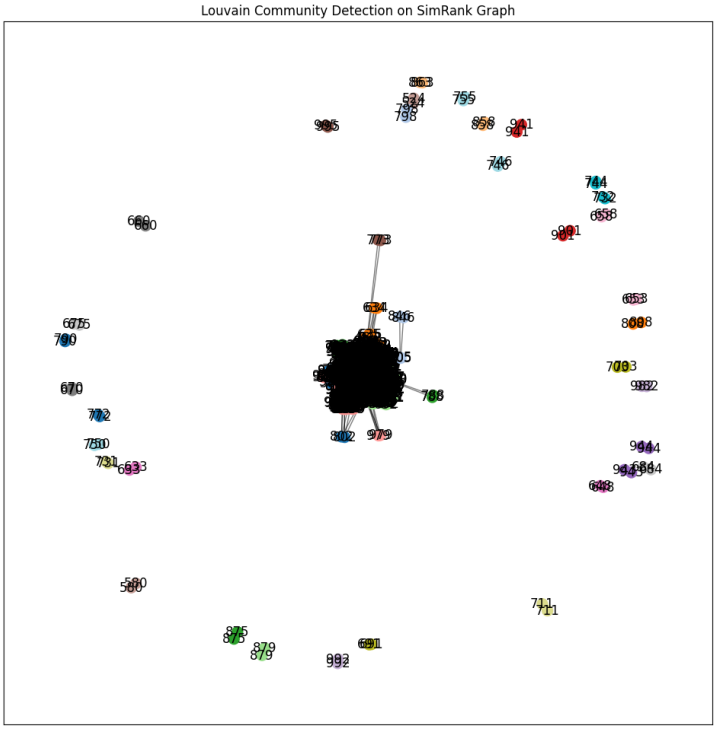

Figure 11.37: communities in the whole Graph.

Chapter 12

Future Plans

12.1 Optimizing Algorithms

- Use approximate SimRank and sparse techniques for efficient computation.
- Implement distributed frameworks (e.g., Apache Spark) for scalability.

12.2 Advanced Applications

- Apply HITS for influencer detection and SimRank for personalized recommendations.
- Extend analysis to temporal and multi-modal graphs for dynamic insights.

12.3 Integration and Deployment

- Deploy solutions on cloud platforms with tools like Neo4j for real-world use.
- Integrate scores as features in machine learning models.

Chapter 13

References

Datasets:

- Ego-Facebook Dataset
- Ego-Twitter Dataset
- Email-Eu-core Dataset
- Congress-Twitter Dataset

Code:

- SimRank Similarity Measure - GeeksforGeeks
- HITS Algorithm - GeeksforGeeks
- ChatGPT
- The code files for this project are hosted in this repository.