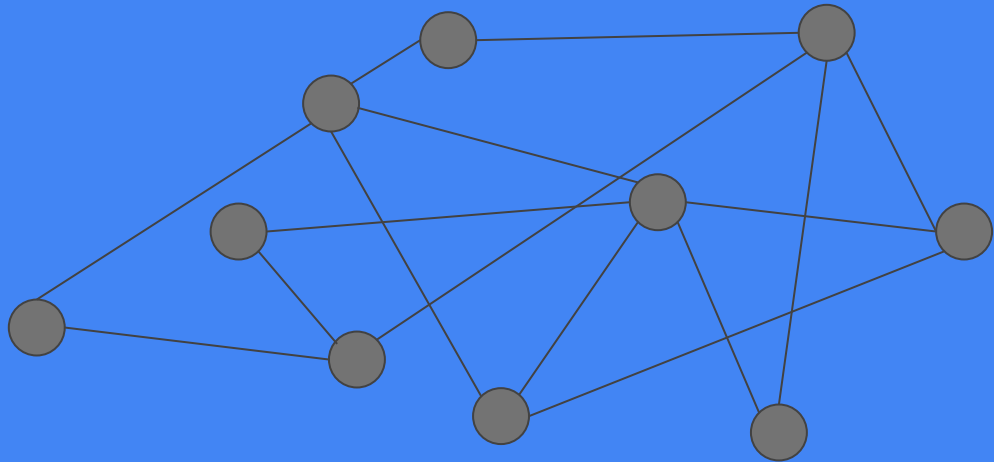


Title: Implementing HITS and SimRank on Various **Social Network** Datasets.

Presented by: Sushovan Pan.

Id : B2330054



topics

- **Introduction to HITS Algorithm**
- How HITS Works
- Dataset Selection
- Implementation Workflow
- Usage : Identifying the Most Influential Pages on Social Media Graphs
- Usage : Identifying Malicious Users (Practical Application)
- Sample output on facebook dataset
- Sample output on Twitter dataset
-
- **Introduction to SIMRANK Algorithm**
- How simrank work
- Dataset Selection
- Implementation Workflow
- Sample output on Twitter US congress dataset
- Sample output on email dataset
-
- Future plan
- References

Introduction to HITS Algorithm

HITS Overview: Hyperlink-Induced Topic Search (HITS) is an algorithm designed to analyze nodes in a directed graph, primarily for ranking web pages.

Key Concepts:

- **Authority:** Nodes that are sources of reliable information, often linked by other nodes.
- **Hub:** Nodes that link to valuable resources, acting as guides to authorities.

Applications: social network analysis, recommendation systems, search engines, and fraud detection.

How HITS Works

Algorithm Process:

- Starts with an initial score for each node.
- Iterative updates:
 - **Hub Score** = Sum of authority scores of linked nodes.
 - **Authority Score** = Sum of hub scores of linking nodes.
- **Convergence** achieved after several iterations.

Output: Two ranked lists of nodes (hubs and authorities).

Dataset Selection

- **Criteria for Choosing Datasets:**
 - Suitable for graph structure (nodes and edges).
 - Large datasets help demonstrate HITS scalability and performance.
- **Examples of Datasets Used:**
 - **Facebook users Dataset:** (Connections between users)
 - **Twitter users Dataset:** (Connections between users)

Implementation Workflow

Load the Graph Data:

- Convert raw edge data (CSV, MTX) to graph format.

Run the HITS Algorithm:

- Use NetworkX or a custom sparse implementation to calculate hub and authority scores.

Interpret Results:

- Identify top hubs and authorities, detect suspicious nodes.

Usage : Identifying the Most Influential Pages on Social Media Graphs

What Makes a Page Influential?

- **High Authority Score:** Indicates a page that others frequently refer to or recommend.
- **High Hub Score:** Shows a page that frequently points users to other important or popular pages.

Using HITS to Identify Influential Pages:

- **Authority Nodes (Influential Sources):** Pages with a high authority score are seen as reputable or valuable sources of information.
- **Hub Nodes (Influential Guides):** Pages with a high hub score are influential in directing users to high-authority pages.

Practical Application for Social Media:

- Identify key influencers and content aggregators in the network.
- Use insights to design targeted marketing, recommendations, or community-building strategies.

Example Output:

- Top 5 pages by authority and hub scores with potential insights on why they're influential.

Usage : Identifying Malicious Users (Practical Application)

Hub & Authority Score Thresholds: Nodes with very high scores in both categories may indicate suspicious behavior (e.g., fake accounts).


Detecting Suspicious Users:

- Set percentile-based thresholds to filter potentially malicious nodes.
- Examine nodes with unusually high hub or authority scores for fake account detection.

Sample output on facebook dataset

By Jure Leskovec

STANFORD UNIVERSITY



Social circles: Facebook

Dataset information

This dataset consists of 'circles' (or 'friends lists') from Facebook. Facebook data was collected from survey participants using this [Facebook app](#). The dataset includes node features (profiles), circles, and ego networks.

Facebook data has been anonymized by replacing the Facebook-internal ids for each user with a new value. Also, while feature vectors from this dataset have been provided, the interpretation of those features has been obscured. For instance, where the original dataset may have contained a feature "political=Democratic Party", the new data would simply contain "political-anonymized feature 1". Thus, using the anonymized data it is possible to determine whether two users have the same political affiliations, but not what their individual political affiliations represent.

Data is also available from [Google+](#) and [Twitter](#).

Dataset statistics

Nodes	4039
Edges	88234
Nodes in largest WCC	4039 (1.000)
Edges in largest WCC	88234 (1.000)
Nodes in largest SCC	4039 (1.000)
Edges in largest SCC	88234 (1.000)
Average clustering coefficient	0.6055
Number of triangles	1612010
Fraction of closed triangles	0.2647
Diameter (longest shortest path)	8
90-percentile effective diameter	4.7

Note that these statistics were compiled by combining the ego-networks, including the ego nodes themselves (along with an edge to each of their friends).

Source (citation)

- J. McAuley and J. Leskovec. [Learning to Discover Social Circles in Ego Networks](#). NIPS, 2012.

Files

File	Description
facebook.tar.gz	Facebook data (10 networks, anonymized)
facebook_combined.txt.gz	Edges from all egonets combined
readme-Ego.txt	Description of files

SNAP for C++

SNAP for Python

SNAP Datasets

BIOSNAP Datasets

What's new

People

Papers

Projects

Citing SNAP

Links

About

Contact us

Open positions

Open research positions in SNAP group are available at undergraduate, graduate and postdoctoral levels.

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:

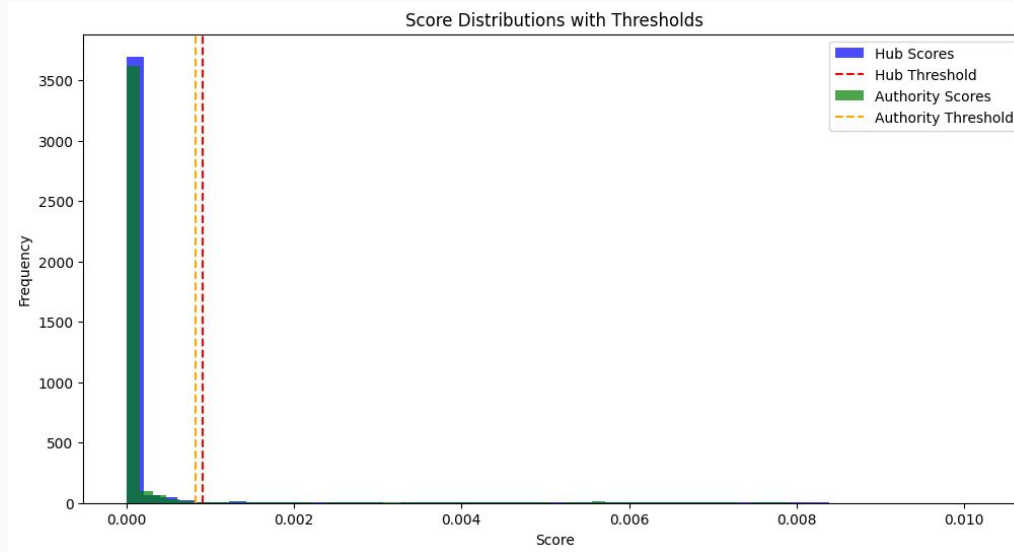
Node: 2604, Authority Score: 0.007932
Node: 2611, Authority Score: 0.007859
Node: 2590, Authority Score: 0.007836
Node: 2607, Authority Score: 0.007763
Node: 2601, Authority Score: 0.007698
Node: 2560, Authority Score: 0.007686
Node: 2624, Authority Score: 0.007661
Node: 2602, Authority Score: 0.007659
Node: 2625, Authority Score: 0.007645
Node: 2586, Authority Score: 0.007612

Suspicious Hub Nodes (Potential Malicious Users):

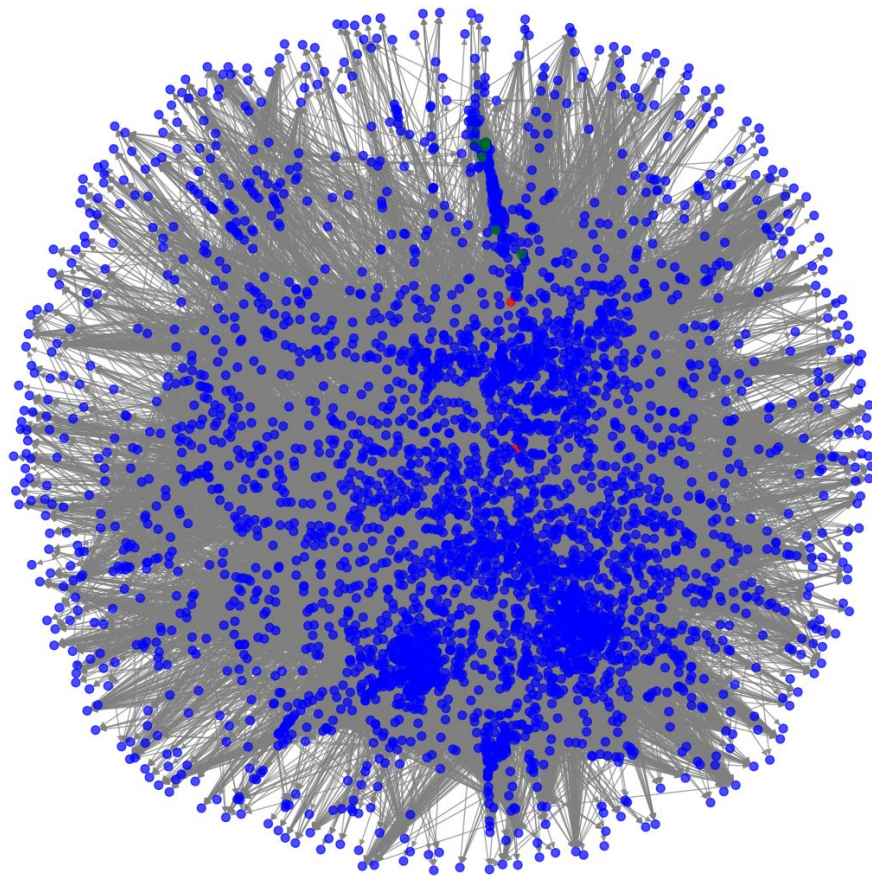
Node: 1912, Hub Score: 0.010229
Node: 2007, Hub Score: 0.001706
Node: 2189, Hub Score: 0.001050
Node: 2543, Hub Score: 0.002174
Node: 1941, Hub Score: 0.005639
Node: 2266, Hub Score: 0.006909
Node: 2347, Hub Score: 0.005091
Node: 2542, Hub Score: 0.002520
Node: 2026, Hub Score: 0.001109
Node: 2158, Hub Score: 0.003754

Suspicious Authority Nodes (Potential Malicious Users):

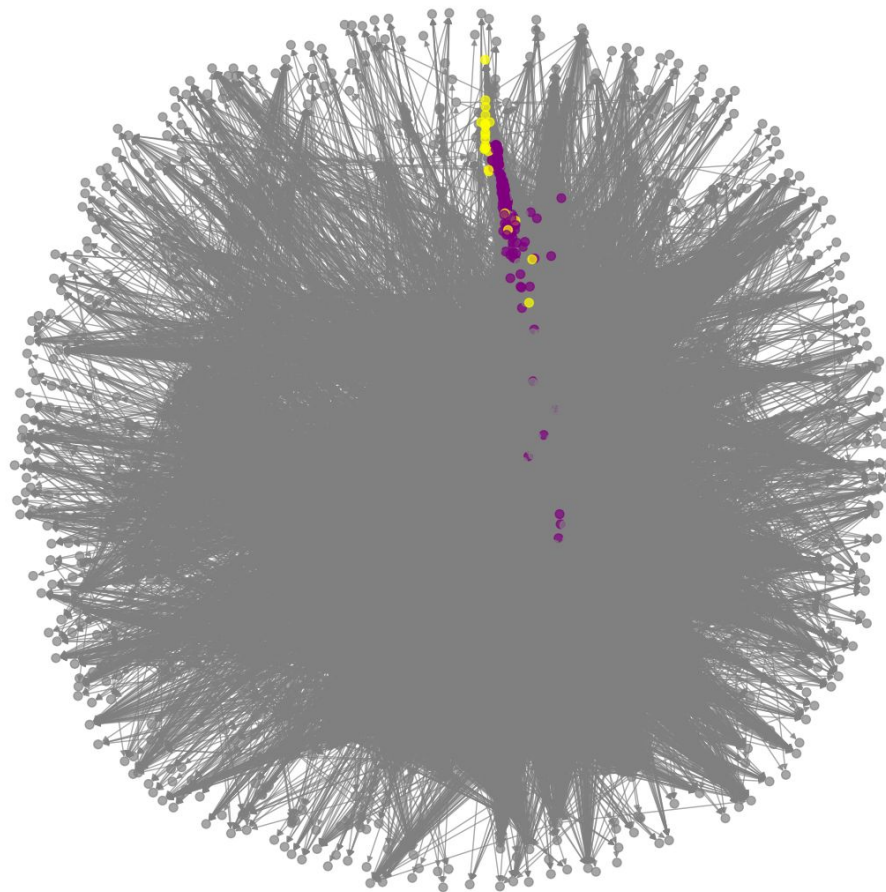
Node: 2543, Authority Score: 0.006835
Node: 2266, Authority Score: 0.005464
Node: 2347, Authority Score: 0.006185
Node: 2542, Authority Score: 0.007212
Node: 2468, Authority Score: 0.005556
Node: 1983, Authority Score: 0.000912
Node: 1984, Authority Score: 0.000937
Node: 1985, Authority Score: 0.001105
Node: 1993, Authority Score: 0.001305
Node: 1997, Authority Score: 0.000888



Graph with Top Hub (Red) and Authority (Green) Nodes Highlighted




Graph with Suspicious Hub (Purple) and Authority (Yellow) Nodes Highlighted



Sample output on Twitter dataset

By Jure Leskovec

STANFORD UNIVERSITY



Social circles: Twitter

Dataset information

This dataset consists of 'circles' (or 'lists') from Twitter. Twitter data was crawled from public sources. The dataset includes node features (profiles), circles, and ego networks.

Data is also available from [Facebook](#) and [Google+](#).

Dataset statistics

Nodes	81306
Edges	1768149
Nodes in largest WCC	81306 (1.000)
Edges in largest WCC	1768149 (1.000)
Nodes in largest SCC	68413 (0.841)
Edges in largest SCC	1685163 (0.953)
Average clustering coefficient	0.5653
Number of triangles	13082506
Fraction of closed triangles	0.06415
Diameter (longest shortest path)	7
90-percentile effective diameter	4.5

Source (citation)

- J. McAuley and J. Leskovec. [Learning to Discover Social Circles in Ego Networks](#). NIPS, 2012.

Files

File	Description
twitter.tar.gz	Twitter data (973 networks)
twitter_combined.bt.gz	Edges from all egonets combined
readme-Ego.txt	Description of files

Open positions

Open research positions in **SNAP** group are available at [undergraduate](#), [graduate](#) and [postdoctoral](#) levels.

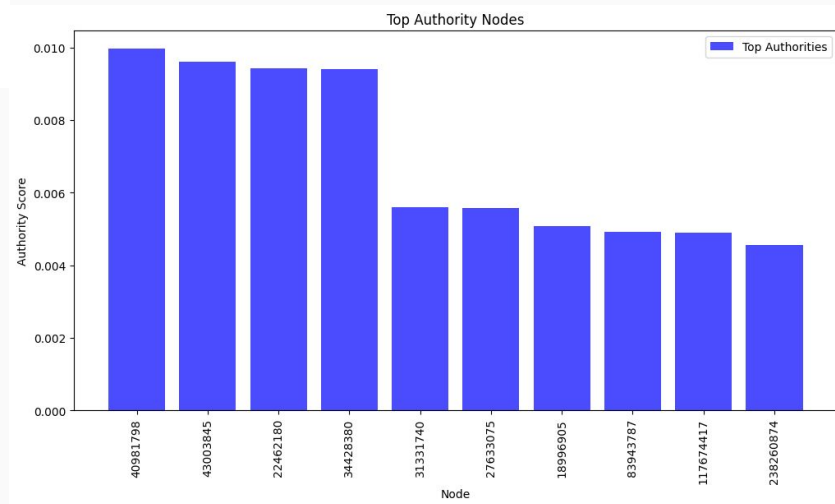
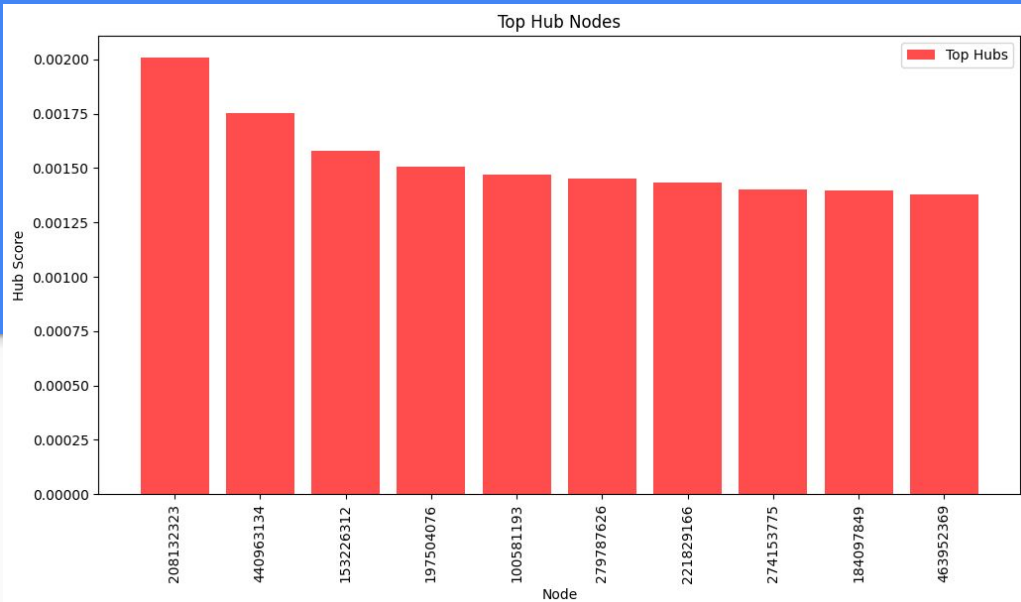
- SNAP for C++
- SNAP for Python
- SNAP Datasets
- BIO-SNAP Datasets
- What's new
- People
- Papers
- Projects
- Citing SNAP
- Links
- About
- Contact us

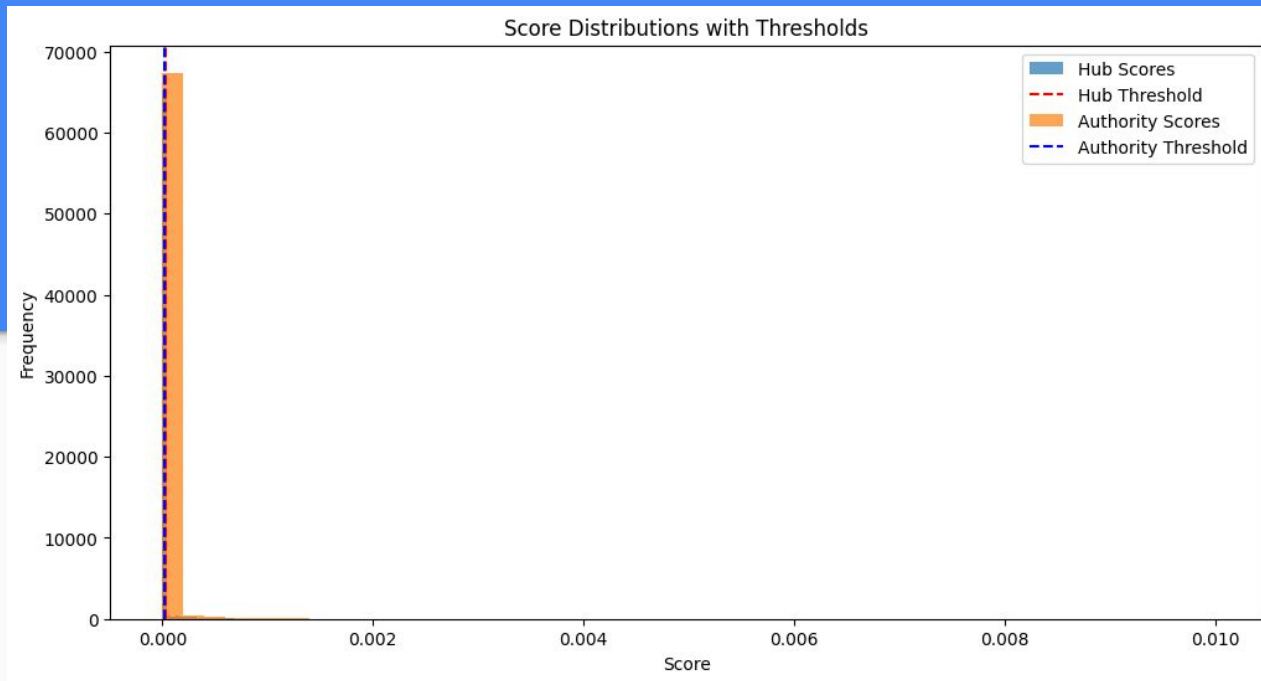
Top Hub Nodes:

```
Node: 208132323, Hub Score: 0.002007
Node: 440963134, Hub Score: 0.001751
Node: 153226312, Hub Score: 0.001580
Node: 197504076, Hub Score: 0.001505
Node: 100581193, Hub Score: 0.001468
Node: 279787626, Hub Score: 0.001453
Node: 221829166, Hub Score: 0.001435
Node: 274153775, Hub Score: 0.001403
Node: 184097849, Hub Score: 0.001399
Node: 463952369, Hub Score: 0.001381
```

Top Authority Nodes:

```
Node: 40981798, Authority Score: 0.009957
Node: 43003845, Authority Score: 0.009598
Node: 22462180, Authority Score: 0.009415
Node: 34428380, Authority Score: 0.009393
Node: 31331740, Authority Score: 0.005594
Node: 27633075, Authority Score: 0.005586
Node: 18996905, Authority Score: 0.005075
Node: 83943787, Authority Score: 0.004932
Node: 117674417, Authority Score: 0.004907
Node: 238260874, Authority Score: 0.004558
```





Suspicious Hub Nodes (Potential Malicious Users):

Node: 214328887, Hub Score: 0.001123
Node: 34428380, Hub Score: 0.000258
Node: 17116707, Hub Score: 0.000606
Node: 28465635, Hub Score: 0.000617
Node: 380580781, Hub Score: 0.000730

Suspicious Authority Nodes (Potential Malicious Users):

Node: 214328887, Authority Score: 0.001519
Node: 34428380, Authority Score: 0.009393
Node: 17116707, Authority Score: 0.001840
Node: 28465635, Authority Score: 0.003533
Node: 380580781, Authority Score: 0.002066

Introduction to SIMRANK Algorithm

SimRank is a graph-based similarity measure that quantifies the similarity between two nodes in a network. It is based on the principle that "two nodes are similar if their neighbors are similar." This recursive definition makes SimRank a powerful tool for various applications, including recommendation systems, link prediction, and network analysis.

Key Features of SimRank:

1. **Recursive Nature:** SimRank defines the similarity between two nodes A and B based on the average similarity of their respective neighbors.
2. **Similarity Scale:** The similarity score ranges between 0 and 1, where 1 indicates identical nodes and 0 indicates no similarity.
3. **Parameter c :** A decay factor c (typically $0 < c < 1$) controls the influence of neighbors on the similarity score, ensuring convergence during computation.

How simrank work

How SimRank Works:

1. **Initialization:** All nodes are considered maximally similar to themselves ($S(i,i)=1$) and dissimilar to others ($S(i,j)=0$ for $i \neq j$).
2. **Propagation:** The similarity between two nodes i and j is updated iteratively by averaging the similarity of their neighbors.
3. **Convergence:** The process repeats until the similarity scores stabilize or the maximum number of iterations is reached.

Mathematical Formula:

For two nodes i and j :

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

Where:

- $N(i)$ and $N(j)$ are the sets of neighbors of nodes i and j .
- c is the decay factor.
- $S(u, v)$ is the similarity of nodes u and v .

Dataset Selection

- **Criteria for Choosing Datasets:**
 - Suitable for graph structure (nodes and edges).
 - Large datasets help demonstrate SimRank's scalability and performance.
- **Examples of Datasets Used:**
 - **Twitter US congress Dataset:** (Twitter connection between US congress member)
 - **Email Dataset:** (email communication between member of a research institution)

Implementation Workflow

Load the Graph Data:

- Convert raw edge data (CSV, MTX) to graph format.

Run the SimRank Algorithm:

- Use a custom sparse implementation to calculate simrank.

Interpret Results:

- Identify top similar nodes.

By Jure Leskovec
STANFORD UNIVERSITY

- SNAP for C++
- SNAP for Python
- SNAP Datasets
- BIO2SNAP Datasets
- What's new
- People
- Papers
- Projects
- Citing SNAP
- Links
- About
- Contact us

Twitter Interaction Network for the US Congress

Dataset information

This network represents the Twitter interaction network for the 117th United States Congress, both House of Representatives and Senate. The base data was collected via the Twitter's API, then the empirical transmission probabilities were quantified according to the fraction of times one member retweeted, quote tweeted, replied to, or mentioned another member's tweet. See the publication for more details.

Open positions

Open research positions in **SNAP** group are available at [undergraduate](#), [graduate](#) and [postdoctoral](#) levels.

Dataset statistics

Directed	Yes
Node features	No
Edge features	Yes
Nodes	475
Edges	13,289

Source (citation)

C.G. Fink, K. Fullin, G. Gutierrez, N. Omodt, S. Zinnecker, G. Sprint, and S. McCulloch: A centrality measure for quantifying spread on weighted, directed networks. *Physica A*, 2023.

```
@article{fink2023centrality,
  title={A centrality measure for quantifying spread on weighted, directed networks},
  author={Fink, Christian G and Fullin, Kelly and Gutierrez, Guillermo and Omodt, Nathan and McCulloch, Sydney},
  journal={Physica A},
  year={2023}}

```

Dataset Description

C.G. Fink, N. Omodt, S. Zinnecker, and G. Sprint: A Congressional Twitter network dataset quantifying pairwise probability of influence. *Data in Brief*, 2023.

```
@article{fink2023twitter,
  title={A Congressional Twitter network dataset quantifying pairwise probability of influence},
  author={Fink, Christian G and Omodt, Nathan and Zinnecker, Sydney and Sprint, Gina},
  journal={Data in Brief},
  year={2023}}

```

Files

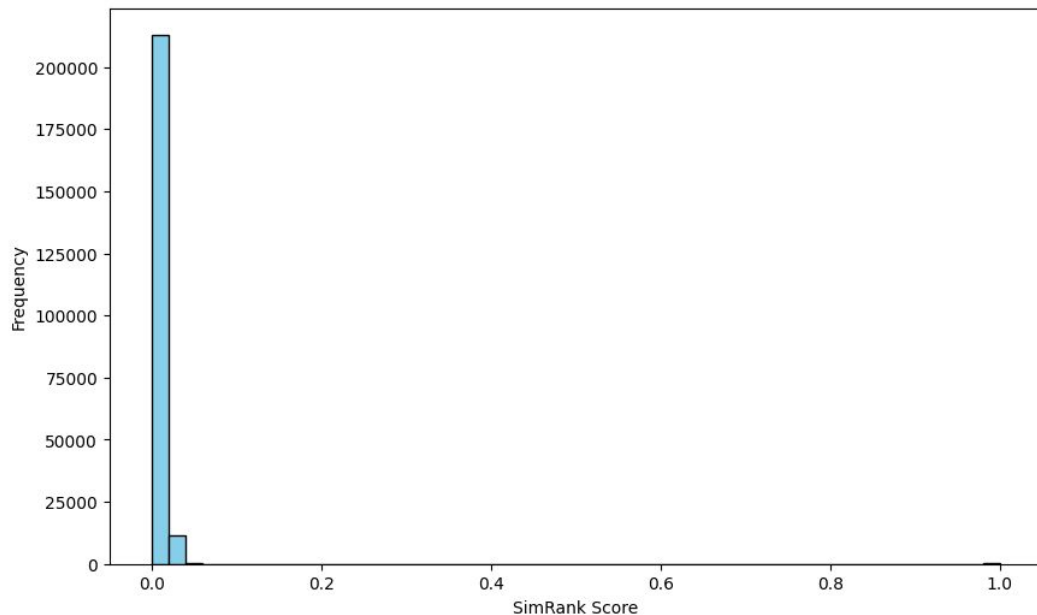
File	Description
congress_network.zip	Twitter Interaction Network for the US Congress

SimRank Scores (Partial):

	0	4	12	18	25	30	46	55	58	59	...	34	404	31	158	395	227	240	356	434	456
0	1.000000	0.022752	0.021619	0.022940	0.022034	0.017925	0.020992	0.020598	0.016835	0.024535	...	0.010624	0.008309	0.0	0.006182	0.009135	0.0	0.0	0.0	0.0	0.0
4	0.022752	1.000000	0.018124	0.027585	0.024085	0.019527	0.019741	0.022202	0.019606	0.023637	...	0.009794	0.008566	0.0	0.006154	0.008728	0.0	0.0	0.0	0.0	0.0
12	0.021619	0.018124	1.000000	0.023134	0.018363	0.018759	0.018562	0.016986	0.016254	0.019562	...	0.010124	0.008706	0.0	0.007155	0.008935	0.0	0.0	0.0	0.0	0.0
18	0.022940	0.027585	0.023134	1.000000	0.023399	0.020691	0.020863	0.019786	0.018919	0.022959	...	0.012084	0.007969	0.0	0.006210	0.008532	0.0	0.0	0.0	0.0	0.0
25	0.022034	0.024085	0.018363	0.023399	1.000000	0.020788	0.020363	0.018978	0.016629	0.021845	...	0.009638	0.008394	0.0	0.006169	0.009381	0.0	0.0	0.0	0.0	0.0

5 rows × 475 columns

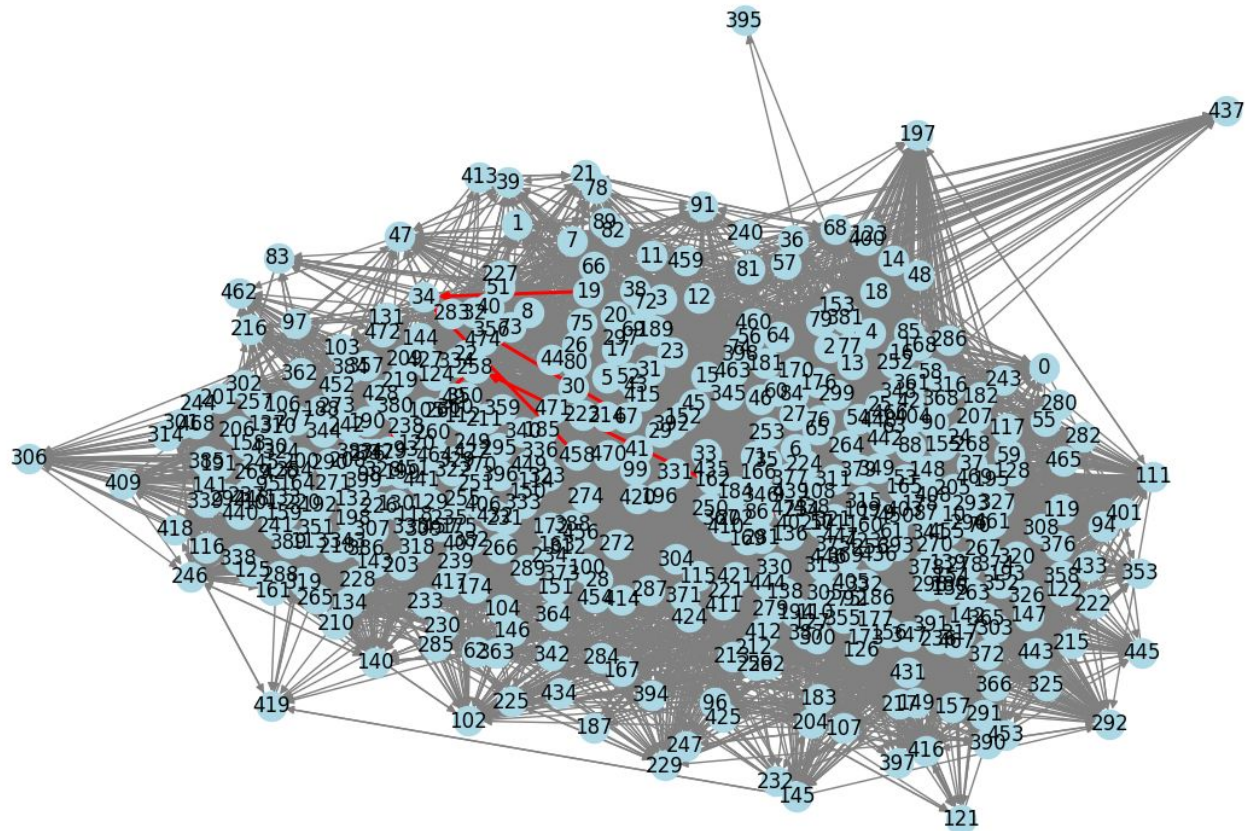
Distribution of SimRank Scores



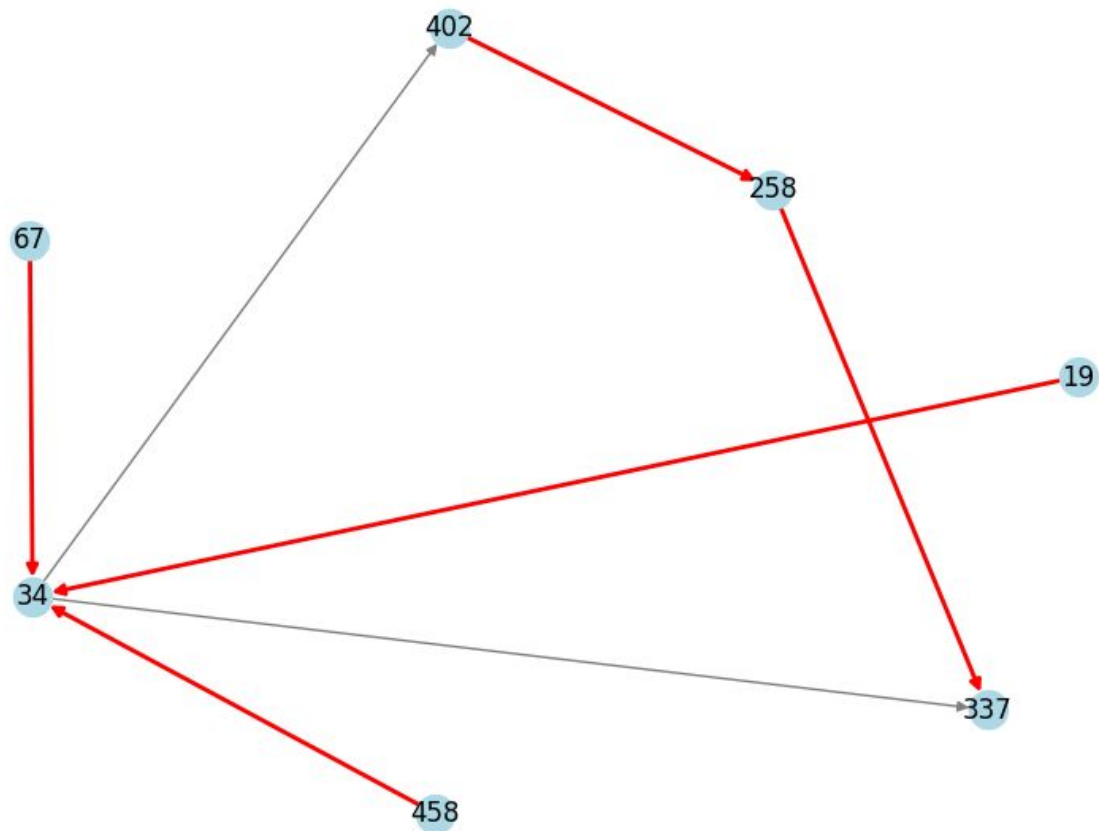
Top 5 Most Similar Node Pairs:

	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

Graph with Top-5 Similar Node Pairs Highlighted




Subset Graph with Top-5 Similar Node Pairs Highlighted



By Jure Leskovec

STANFORD UNIVERSITY



SNAP for C++

SNAP for Python

SNAP Datasets

BIDSnap Datasets

What's new

People

Papers

Projects

Citing SNAP

Links

About

Contact us

email-Eu-core network

Dataset information

The network was generated using email data from a large European research institution. We have anonymized information about all incoming and outgoing email between members of the research institution. There is an edge (u, v) in the network if person u sent person v at least one email. The e-mails only represent communication between institution members (the core), and the dataset does not contain incoming messages from or outgoing messages to the rest of the world.

The dataset also contains "ground-truth" community memberships of the nodes. Each individual belongs to exactly one of 42 departments at the research institute.

This network represents the "core" of the email-EuAll network, which also contains links between members of the institution and people outside of the institution (although the node IDs are not the same).

Dataset statistics

Nodes	1005
Edges	25571
Nodes in largest WCC	986 (0.981)
Edges in largest WCC	25552 (0.999)
Nodes in largest SCC	803 (0.799)
Edges in largest SCC	24729 (0.967)
Average clustering coefficient	0.3994
Number of triangles	105461
Fraction of closed triangles	0.1085
Diameter (longest shortest path)	7
90-percentile effective diameter	2.9

Source (citation)

- Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich. "Local Higher-order Graph Clustering." In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017.
- J. Leskovec, J. Kleinberg and C. Faloutsos. [Graph Evolution: Densification and Shrinking Diameters](#). ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.

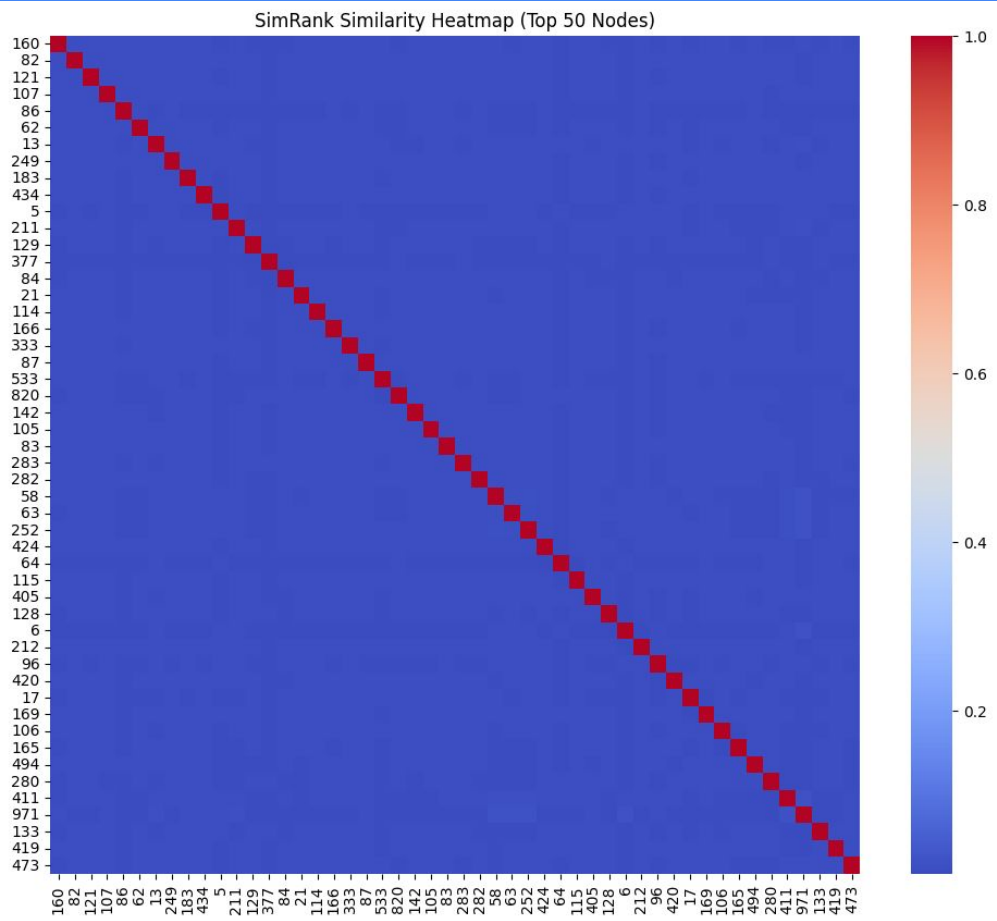
Files

File	Description
email-Eu-core.txt.gz	Email communication links between members of the institution
email-Eu-core-department-labels.txt.gz	Department membership labels

Data format for community membership

NODEID DEPARTMENT

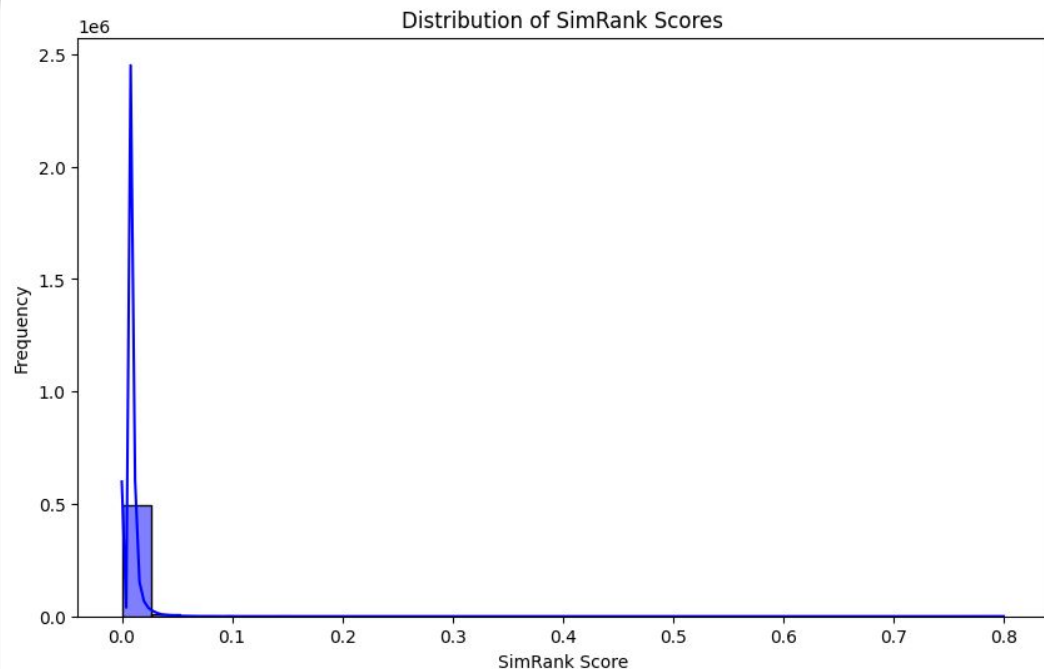
- NODEID: id of the node (a member of the institute)
- DEPARTMENT: id of the member's department (number in 0, 1, ..., 41)



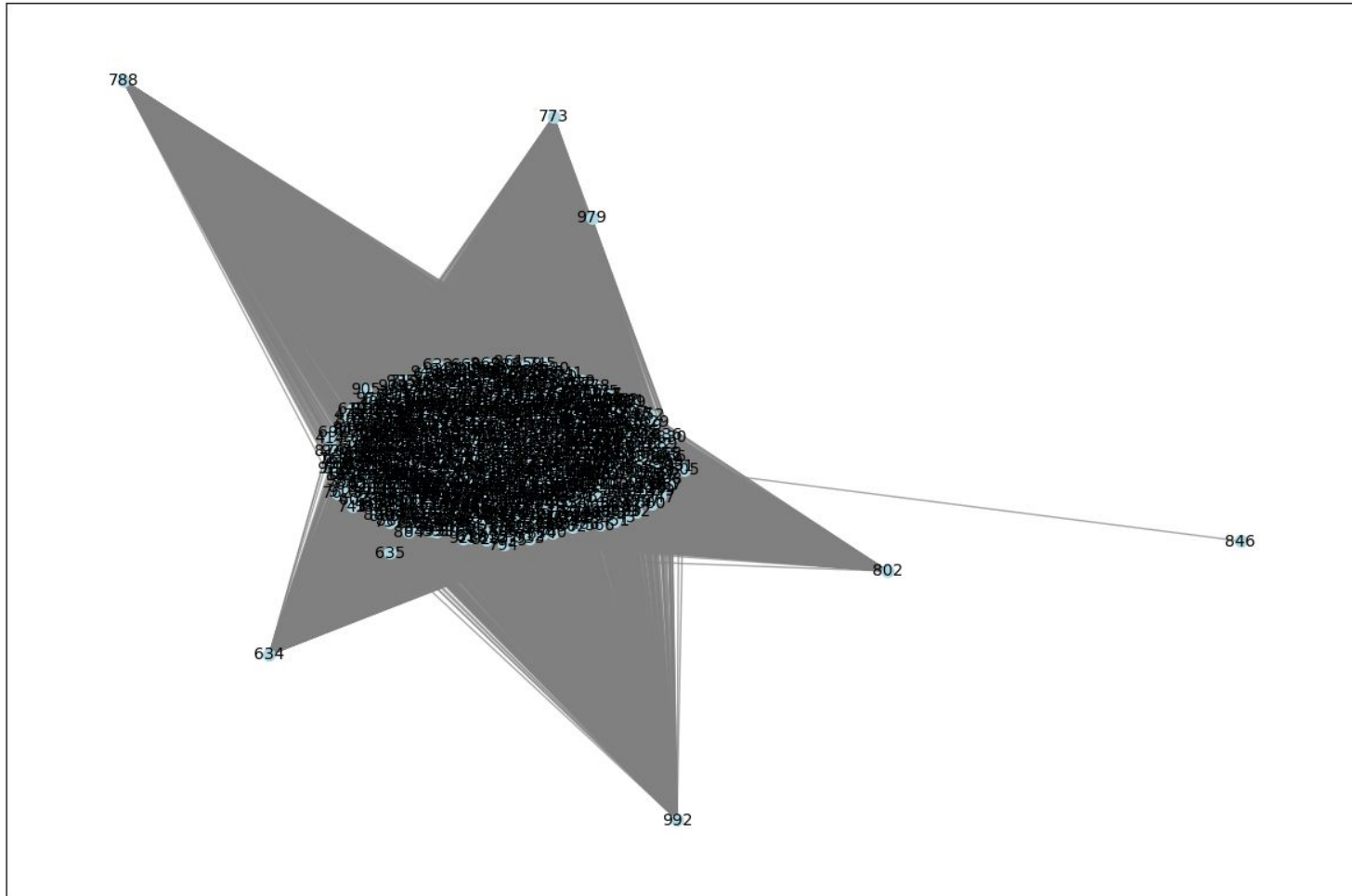
Simrank value of some sample nodes

department	0	1	2	3	4	5	6	7	8	9	...	995	996	997	998	999	1000	1001	1002	1003	1004
0	0.010867	0.010738	0.008005	0.008163	0.008381	0.009884	0.009819	0.006959	0.007576	0.006248	...	0.0	0.007924	0.006666	0.012906	0.008674	0.013459	0.008699	0.007293	0.009941	0.007173
1	0.031071	0.028253	0.007771	0.007875	0.007879	0.009008	0.008102	0.006854	0.006988	0.005944	...	0.0	0.006725	0.006886	0.009418	0.008226	0.009486	0.007924	0.042429	0.008117	0.006886
2	0.008238	0.008788	0.007975	0.008161	0.008163	0.009062	0.008739	0.008158	0.008969	0.007212	...	0.0	0.007799	0.006942	0.008209	0.008262	0.007782	0.007985	0.007756	0.007669	0.007425
3	0.010203	0.010738	0.009454	0.009501	0.010163	0.010253	0.008874	0.008099	0.009027	0.006423	...	0.0	0.007884	0.007998	0.013391	0.009782	0.010165	0.008115	0.008085	0.009073	0.007643
4	0.010080	0.010133	0.008212	0.008527	0.008807	0.009254	0.009231	0.007515	0.007860	0.006543	...	0.0	0.007536	0.007706	0.033119	0.008914	0.022556	0.007946	0.007956	0.008867	0.007315

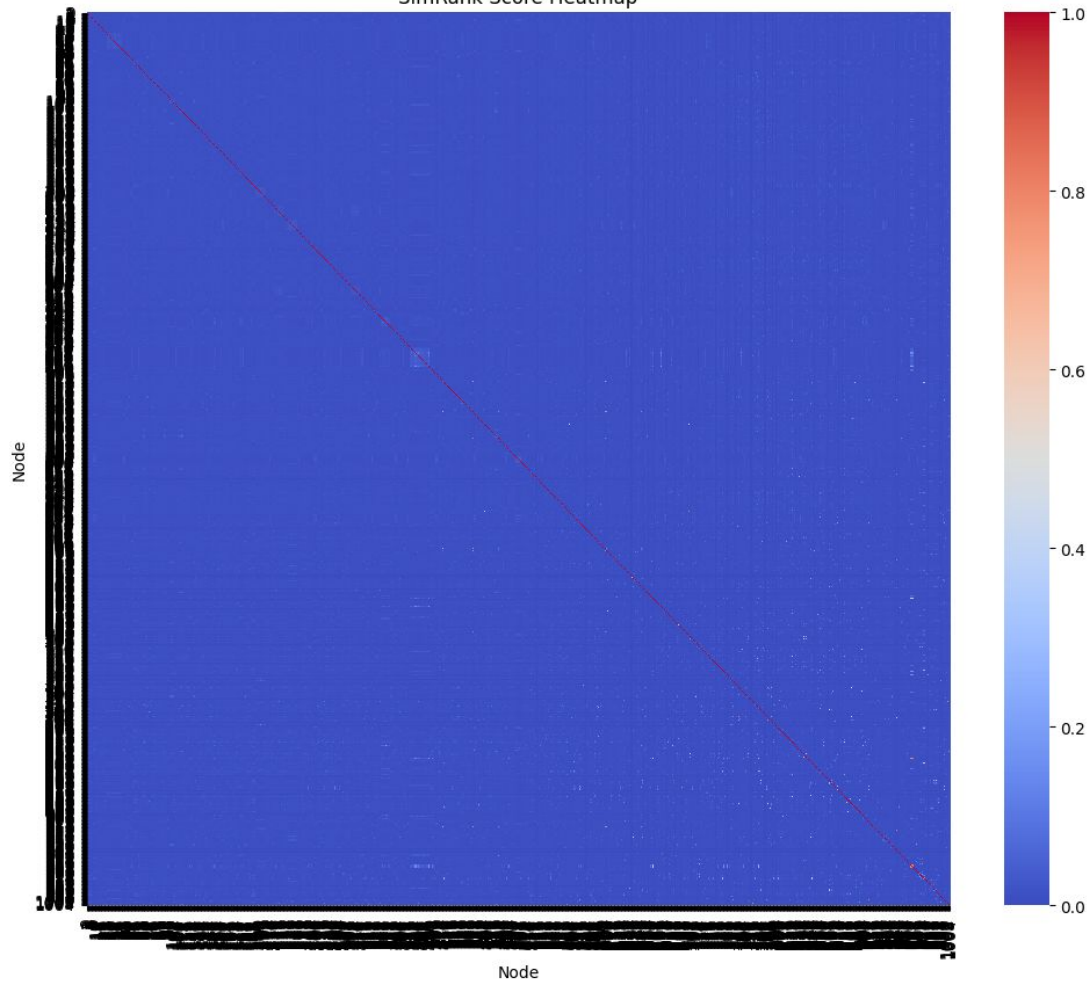
5 rows × 1005 columns



Network Visualization

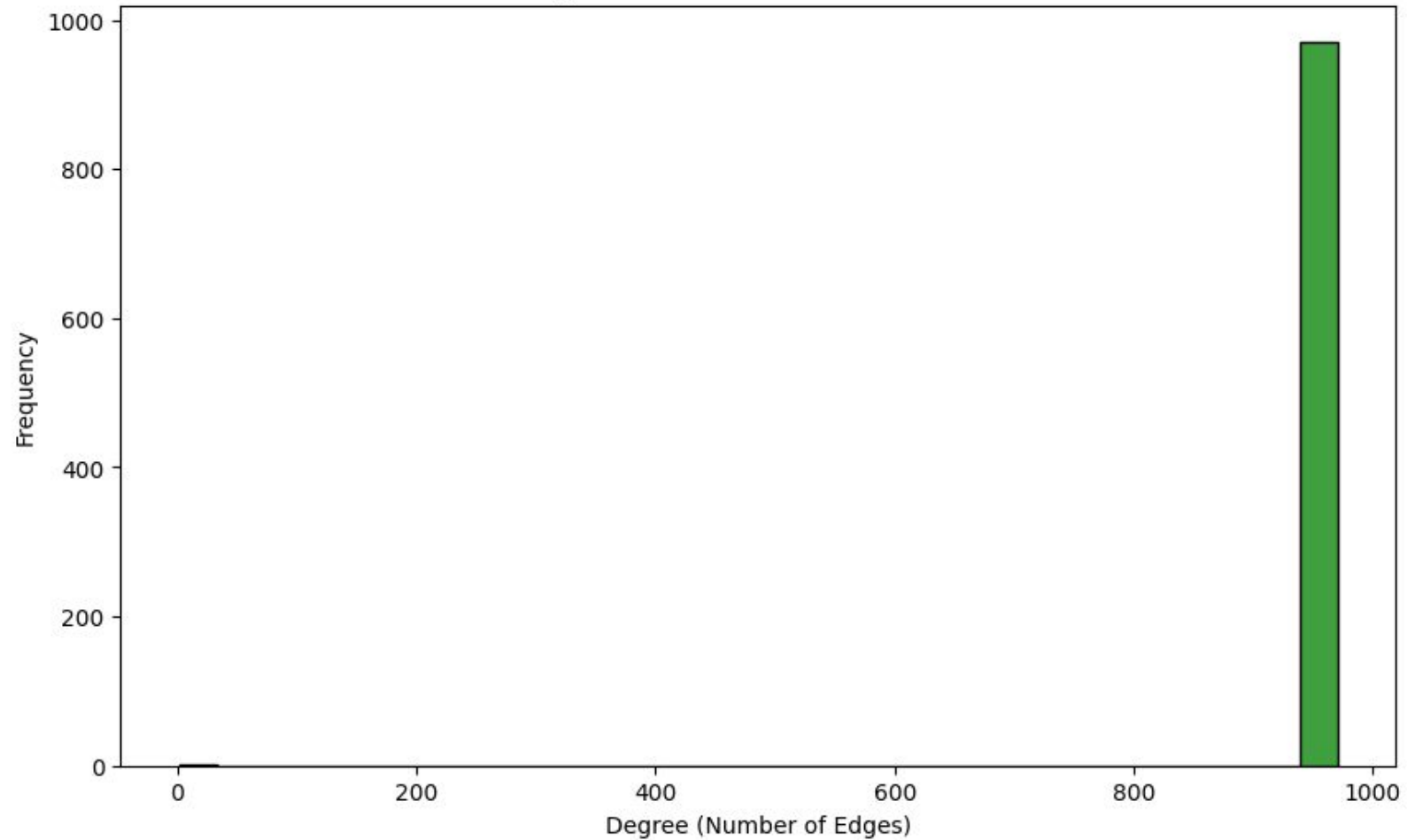


SimRank Score Heatmap

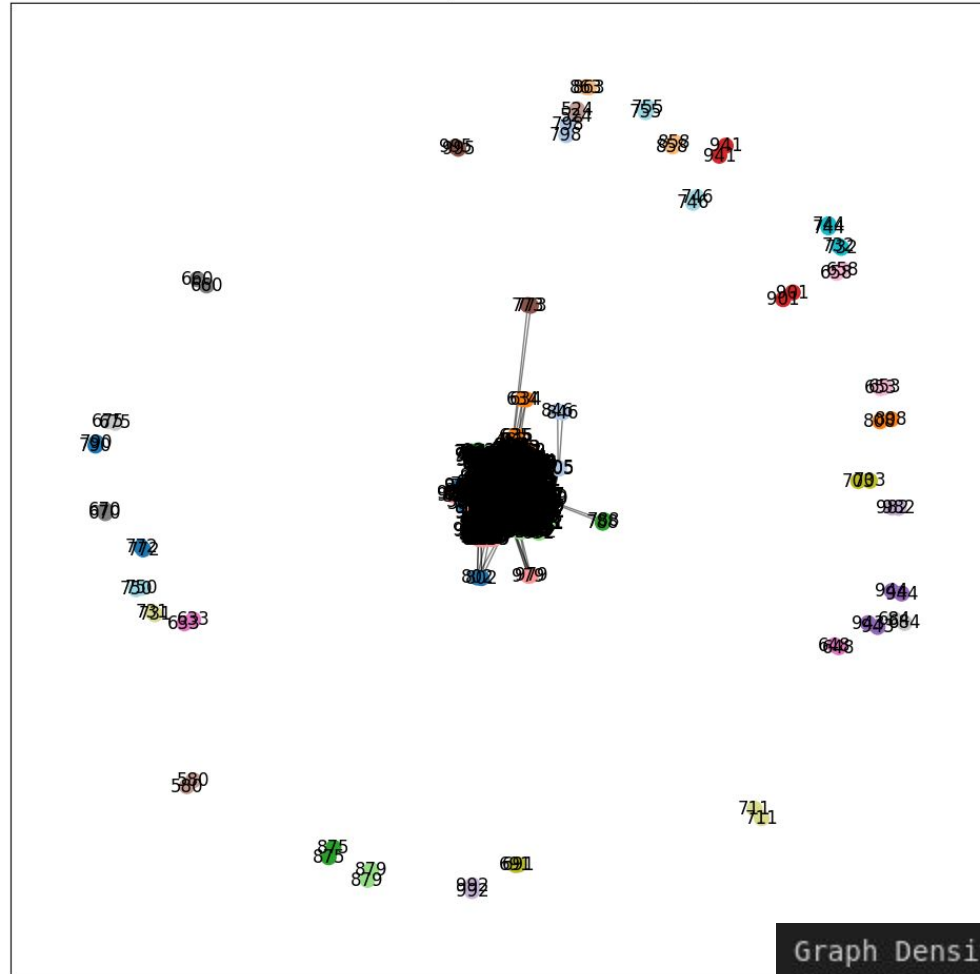


Heatmap for all nodes

Degree Distribution of the Network



Louvain Community Detection on SimRank Graph



Graph Density: 0.1474

Average Clustering Coefficient: 0.0000

Future plan

Algorithm Optimization and Scalability

- Optimize memory and computation using approximate SimRank and sparse matrix techniques.
- Implement parallel or distributed processing with frameworks like Apache Spark to handle large-scale datasets.

Advanced Applications and Insights

- Use HITS for influencer identification and SimRank for personalized content recommendations.
- Extend the analysis to temporal graphs and multi-modal networks, tracking changes and relationships over time.

Integration and Deployment

- Incorporate HITS and SimRank scores as features in machine learning tasks like node classification or link prediction.
- Deploy scalable solutions using cloud platforms and graph databases like Neo4j for real-world social media analysis.

References

- Dataset 1 - <https://snap.stanford.edu/data/ego-Facebook.html>
- Dataset 2 - <https://snap.stanford.edu/data/ego-Twitter.html>
- dataset 3 - <https://snap.stanford.edu/data/email-Eu-core.html>
- Dataset 4 - <https://snap.stanford.edu/data/congress-twitter.html>

code

- <https://www.geeksforgeeks.org/simrank-similarity-measure-in-graph-based-text-mining/h-hits-algorithm-using-networkx-module-python/>
- chatgpt

Thank You.