PRECOG TASK

2022101122
Karan Nijhawan

All the parts were Attempted

3. Analyzing citation networks

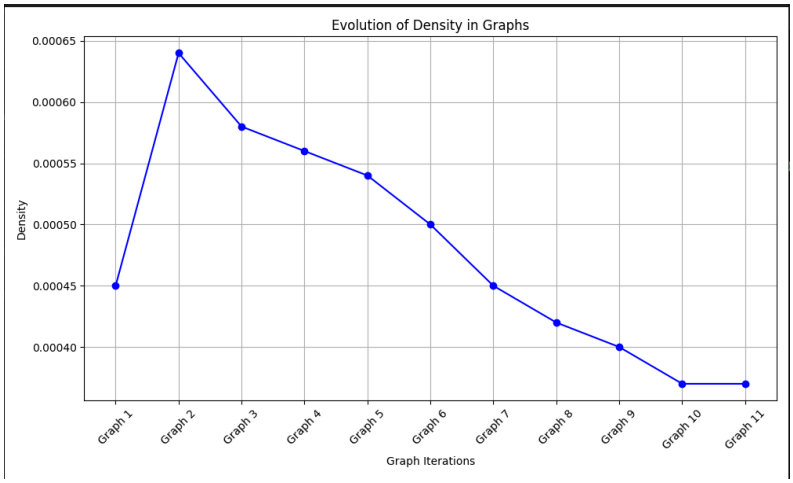
Task 1:

table.txt

Graph Iterations	Number of Nodes	Number of Edges
Graph 1	582	152
Graph 2	2117	2862
Graph 3	4421	11379
Graph 4	7276	29808
Graph 5	10481	58927
Graph 6	14013	98307
Graph 7	17736	142934
Graph 8	21739	201300
Graph 9	25775	263468
Graph 10	29878	334051
Graph 11	30558	347268

The graph evolved over the period of 11 years. These are the number of nodes and the number of edges, that got connected as the graph proceeded

1. **Density**



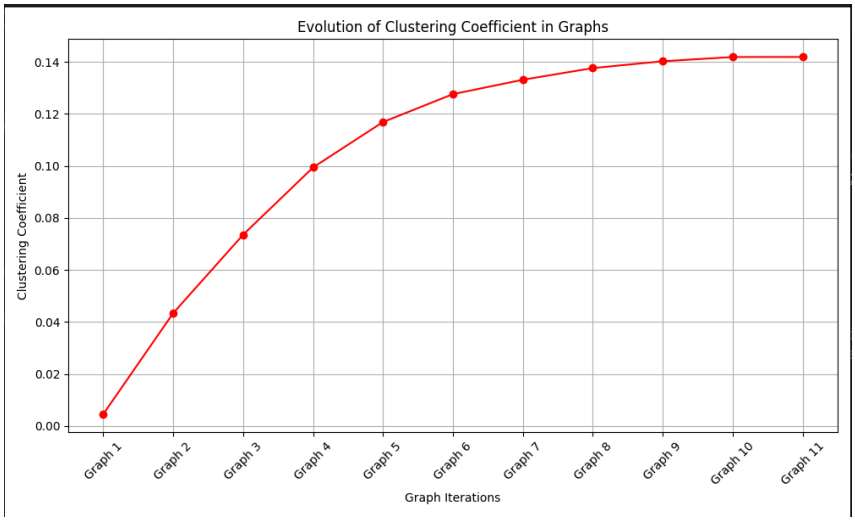
We saw a sudden rise in density from Graph 1 to Graph 2 , that was because just after the initial nodes were added , more and more paper that are based on those papers only got added , therefore an unexpected peak , after that it follows the general decreasing pattern

2. Clustering Coefficient

The clustering coefficient is a measure in graph theory that quantifies the degree to which nodes in a graph tend to cluster together. It provides insight into the local connectivity or cohesion within a network. The clustering coefficient is often used to assess the presence of clusters or communities in a graph.

$$C_i = \frac{2 \times \text{number of triangles centered on node } i}{\text{degree of node } i \times (\text{degree of node } i - 1)}$$

$$C_{\text{global}} = \frac{1}{N} \sum_{i=1}^{N} C_i$$



Graph 1:

- Low clustering coefficient suggests that, initially, research papers may not have formed tightly connected citation groups.

Graph 2:

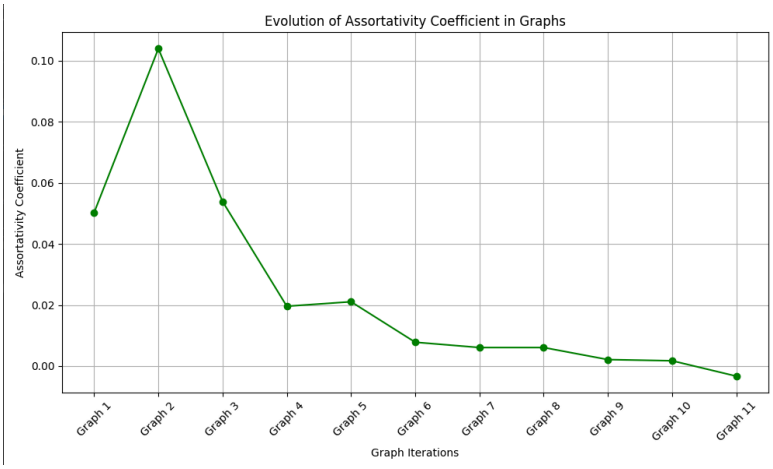
- A substantial increase in clustering coefficient indicates that, as the graph grows, papers may start forming more cohesive citation clusters or communities.

Graph 3 to Graph 11:

- The clustering coefficient continues to increase but at a diminishing rate. This might suggest that, as the network expands, new papers are more likely to connect with existing clusters rather than forming entirely new ones.
- It could also imply that the research topics covered by new research papers are diverse, leading to connections across different clusters, and preventing the formation of highly clustered groups.
- Alternatively, it could be influenced by the type of citation patterns in the field, where papers tend to cite works from multiple thematic areas.

3. Assortativity Coefficient

Assortativity measures the tendency of nodes with similar degrees to connect to each other in a network. In a graph, it can be positive, negative, or close to zero. Positive assortativity indicates that nodes with similar degrees are more likely to connect, while negative assortativity suggests that nodes with different degrees are more likely to connect.



Graph 1:

- A positive assortativity coefficient indicates that papers with similar numbers of references tend to cite each other. This could be due to common topics or themes that lead to interconnected citation patterns.

Graph 2:

- A substantial increase in assortativity suggests that as the graph grows, papers with similar degrees continue to connect. This might indicate the formation of citation clusters or communities with shared themes.

Graph 3 to Graph 11:

- The assortativity coefficient fluctuates but tends to decrease over time. This could suggest that, as the network expands, new papers are more likely to connect with papers of different degrees rather than primarily with papers of similar degrees.
- It might indicate that the research topics covered by new papers are diverse, leading to connections across different citation clusters, preventing a strong tendency for papers of similar degrees to connect.
- Alternatively, it could be influenced by the nature of citation patterns in the field, where papers tend to cite works from various thematic areas, and degree similarity becomes less pronounced.

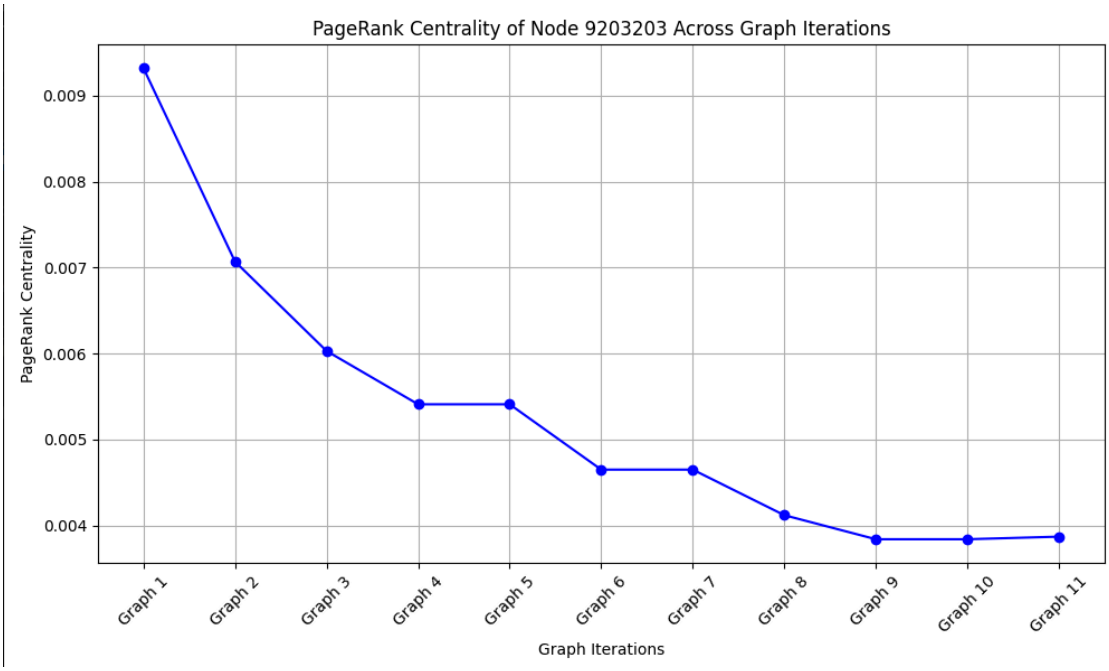
Interesting Insight:

This may suggest that original when research paper covered basic researches , they were tend to be connected similarly , but as time passed more advanced research evolved that was referred from basic papers causing an edge between basic paper(now which has more degree , because a lot of advanced papers are referring this paper now) to an advanced paper.

4. **PageRank Centrality:**

PageRank centrality is a measure used in network analysis to assess the importance or influence of nodes within a graph. Originally developed by Larry Page and Sergey Brin as part of the Google search engine algorithm, PageRank assigns a numerical weight to each node in a directed graph based on the structure of links between nodes. The basic idea is that a node's importance is influenced by the importance of nodes pointing to it.

PageRank Centrality of node 9203203:



Emerging Research Topics:

- Over time, new and innovative research topics emerge, shifting the focus of academic inquiry.
- Node 9203203, which initially held high centrality, experiences a decrease as attention turns to newer subjects.

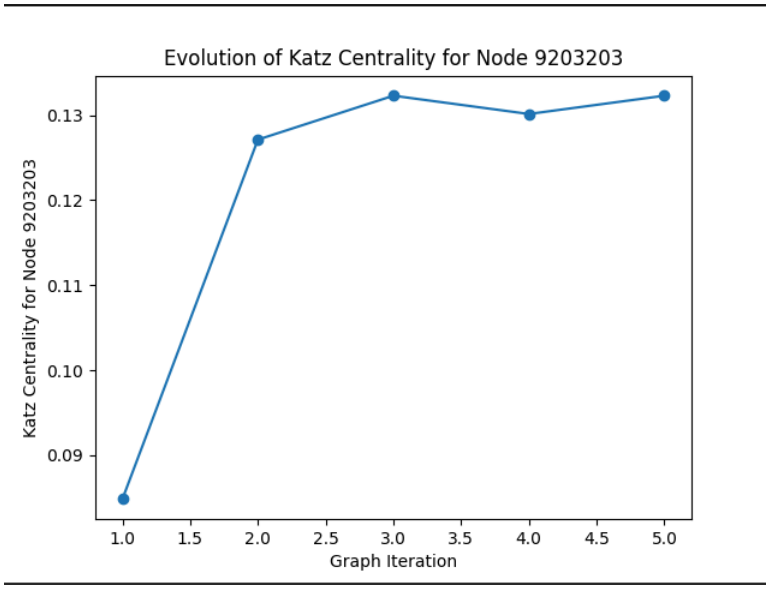
Here is the list of top 5 Page Rank Nodes :

	Graph Iteration	Top Node 1	Top Node 2	Top Node 3	Top Node 4	Top Node 5
0	1	9203206	9203203	9203220	9206242	9204220
1	2	9203203	9206203	9206242	9203206	9204220
2	3	9203203	9206203	9303255	9310316	9206242
3	4	9303255	9206203	9303255	9310316	9206242
4	5	9303255	9206203	9303255	9310316	9206242
5	6	9303255	9206203	9206242	9310316	9204220
6	7	9303255	9206203	9206242	9310316	9204220
7	8	9303255	9206203	9206242	9310316	9204220
8	9	9303255	9206203	9206242	9310316	9204220
9	10	9303255	9206203	9206242	9310316	9204220
10	11	9303255	9206203	9206242	9310316	9204220

5. **Katz Centrality**

   Katz centrality is a measure of the relative influence or importance of a node within a network. It considers both direct connections and indirect paths through other nodes. Nodes with higher Katz centrality scores are not only well-connected but also have connections to nodes that, in turn, have connections to many others.

Katz Centrality of node 9203203:



Insights:

Initially, when the node was introduced, it held a moderate Katz centrality value. However, the sudden surge in centrality indicates a rapid influx of advanced research papers. Initially, when the node was introduced, it held a moderate Katz centrality value. However, the sudden surge in centrality indicates a rapid influx of advanced research papers referencing this particular node. As the field matured and subsequent papers delved into related topics, the centrality values stabilized, reflecting a consistent influence.

However, with the saturation of research on the initial theme, node 9203203 faced challenges in maintaining its position within the top 5. The competitive landscape intensified, and the node gradually faded from prominence in later iterations as new research topics emerged, demonstrating the dynamic nature of academic networks.ncing this particular node. As the field matured and subsequent papers delved into related topics, the centrality values stabilized, reflecting a consistent influence.

However, with the saturation of research on the initial theme, node 9203203 faced challenges in maintaining its position within the top 5. The competitive landscape intensified, and the node gradually faded from prominence in later iterations as new research topics emerged, demonstrating the dynamic nature of academic networks.

A comparison on Node 9203203

Katz Centrality:

- *Adaptive Nature:* Katz centrality considers both direct and indirect connections. The sudden peak indicates a quick adaptation to a surge in citations, while stabilization suggests adaptability to ongoing research trends.
- *Saturation in Research:* The subsequent decrease in Katz centrality can be attributed to saturation in research on the initial theme. As newer topics emerge, the node faces competition and gradually fades.

PageRank:

- *Algorithm Emphasis:* PageRank emphasizes the importance of nodes with high-quality inbound links. Node 9203203, with consistently high PageRank centrality, is consistently cited by influential papers.
- *Resilience:* The node's continued presence in the top 5 PageRank centralities suggests resilience to changing research dynamics. Its enduring influence contributes to its high PageRank centrality.

TASK 2:

1. Implement any two algorithms/ ML methods for community detection on the graph at any time T
    a. Louvain Method

The Louvain method is a community detection algorithm designed to find modular structures in complex networks. It is based on the optimization of modularity, a metric that quantifies the quality of a partition of a graph into communities. The algorithm is iterative and consists of two phases: the "greedy" phase and the "aggregation" phase.

I. Greedy Phase:

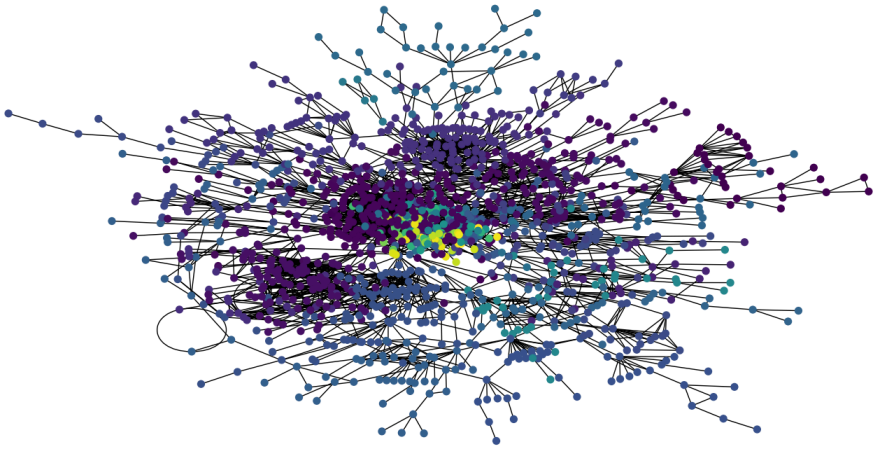
- Initialization: Each node starts as its own community.
- Iteration:
    - For each node, it evaluates the change in modularity if the node were to move to the community of one of its neighbors.
    - If the modularity increases, the node is moved to the community with the maximum modularity gain.
    - This process is repeated until no further improvement in modularity is possible.

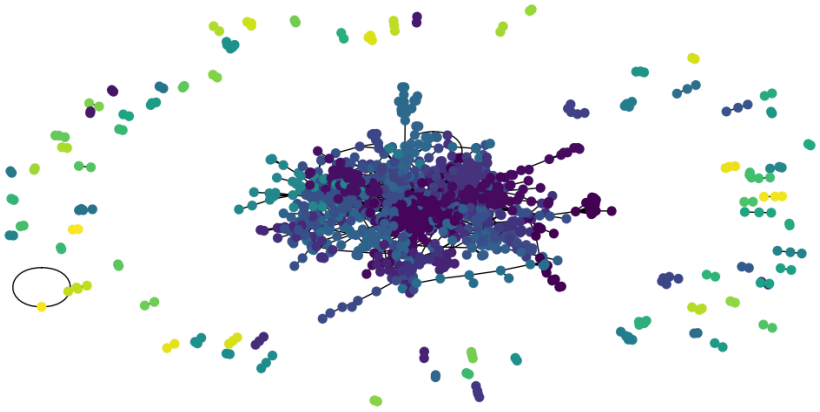
Ii. Aggregation Phase:

- The communities found in the first phase are considered as nodes in a new network, and edges between these nodes are weighted by the sum of the edge weights between nodes in the original network.
- The algorithm then goes through the greedy phase again on this new network, iteratively refining the partitioning.

This graph was made with the help of data at time 1994-01-01.

This graph contains communities that were cut using Louvain Method



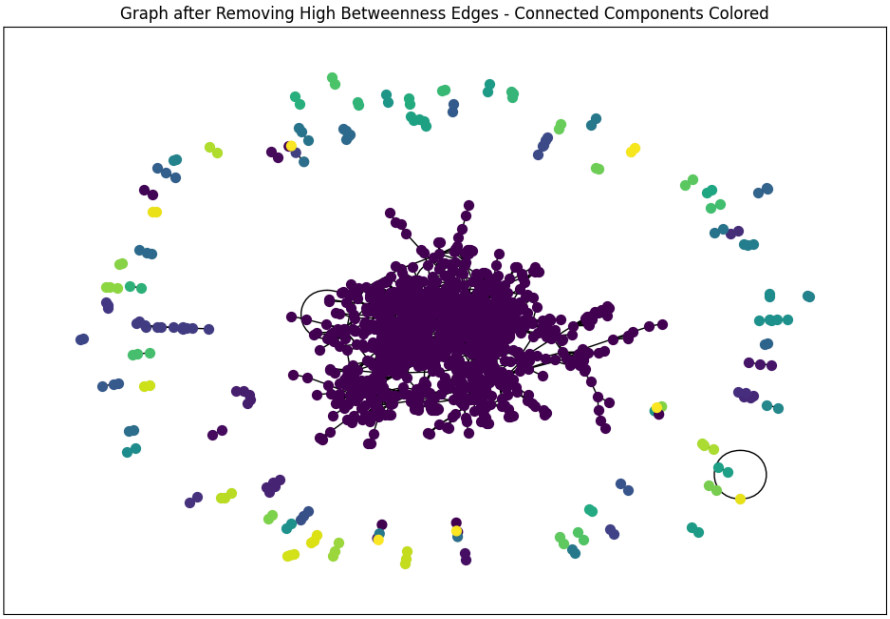


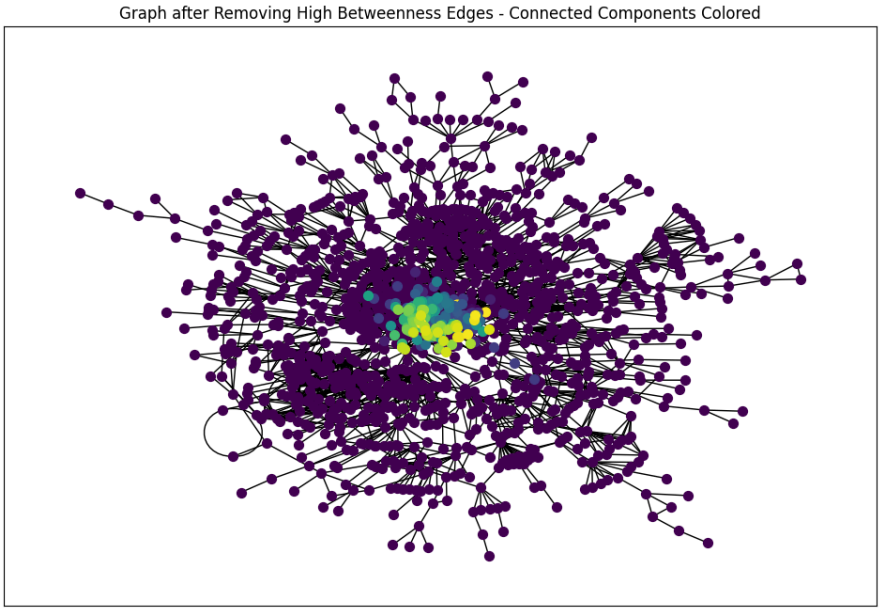
b. Using Betweenness Centrality:

Steps include:

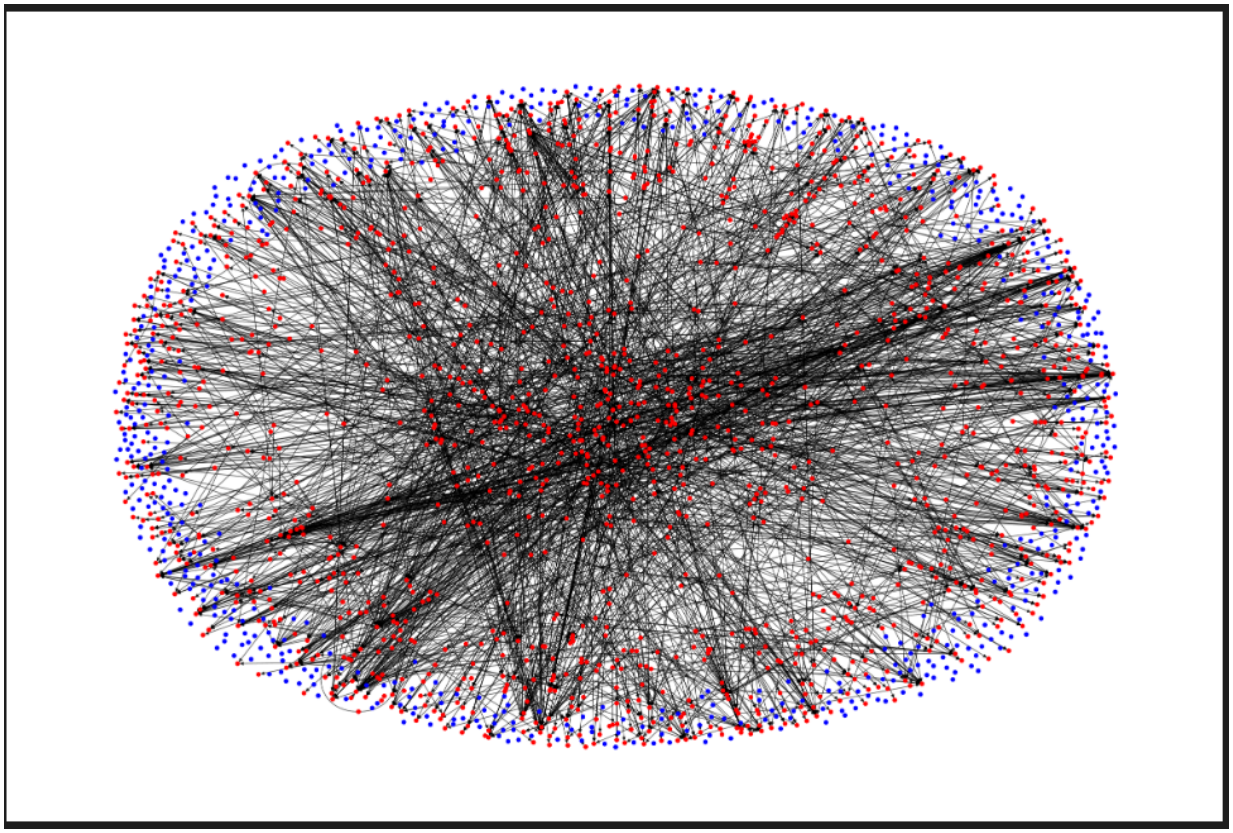
- **Setting Up the Network**:
  - Begin with a graph that illustrates relationships or citations.
- **Node Selection Based on Date**:
  - Create a subgraph by filtering nodes according to a specific date.
- **Computing Betweenness Centrality**:
  - Evaluate the betweenness centrality for each node in the subgraph.
- **Identifying Nodes with High Betweenness**:
  - Nodes surpassing a set threshold in betweenness centrality are deemed influential.
- **Eliminating Edges Connected to High-Betweenness Nodes**:
  - Disconnect nodes associated with high-betweenness by removing their edges.
- **Visualizing the Altered Network**:
  - Observe the network's modified form post-edge removal to uncover potential clusters.
- **Improving Cluster Separation**:
  - Enhance the distinction between clusters for better interpretability.
- **Simplifying the Network and Reducing Noise**:
  - Streamline the network by minimizing noise introduced by influential nodes.
- **Exploring Community Detection Possibilities**:
  - Disconnected subgraphs may indicate the presence of distinct communities.
- **Gaining Insights into Network Dynamics**:
  - Understand alterations in connectivity and potential vulnerabilities within the network.
- **Fine-Tuning and Iterating**:
  - Optionally, adjust parameters and repeat the process for further refinement.

The clustering was done on the same data as above ie 2 years





2.Analyzing the communities



This is the same graph without community detection,

Community detection algorithms, such as the Louvain Method, aim to identify groups or communities of nodes within a graph based on patterns of connectivity. The results of community detection may provide insights into the underlying structure or organization of the network. Analyzing why a community formed as it did can involve considering several factors:

A. Density of Connections: Communities often form because nodes within a community have a higher density of connections among themselves compared to connections outside the community. This suggests a more cohesive and interconnected subgroup.
B. Modularity Score: The Louvain Method optimizes modularity, a measure of the quality of community structure within a network. Modularity is calculated based on the difference between the observed and expected connections

within communities. A higher modularity score indicates a better-defined community structure.
C. Node Similarity: Nodes within a community may share similar attributes, such as node degree, centrality, or other topological features. Analyzing the characteristics of nodes within a community can help understand why they are grouped together.
4. Edge Betweenness: The Louvain Method often considers the edge betweenness, which measures the number of shortest paths passing through an edge. Communities may form around edges with lower betweenness, indicating that these edges act as "bridges" between communities.
5. Graph Topology: The overall structure of the graph, including its size, connectivity, and the distribution of nodes, can influence community formation. For example, densely connected subgraphs may naturally form distinct communities.
6. Dynamic Processes: Consider any dynamic processes that might have influenced community formation, such as the evolution of the network over time or the influence of external factors.
7. Domain-Specific Knowledge: Depending on the nature of the graph, domain-specific knowledge can provide additional insights into the reasons behind community formation.

In Summary,

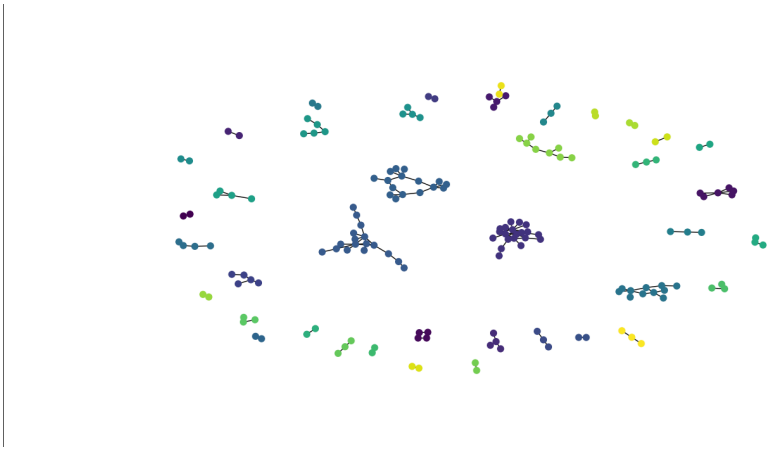
Using community detection in this graph can help us separate domains of research, As nodes which are similar or more tend to be closely connected are separated out in a community. The Louvain Method could not cut a very discrete line between nodes , but help nodes in the same cluster separate out. The Edge-Centrality Method on the other hand could not separate out nodes that are somewhat similar but can help create absolute lines between domains of the research.

Another approach could be applying the centrality algorithm and then applying the Louvain Method of Clustering. This could help in separating out absolute domains and also help us see the minor separations between subtopics of the same domain.

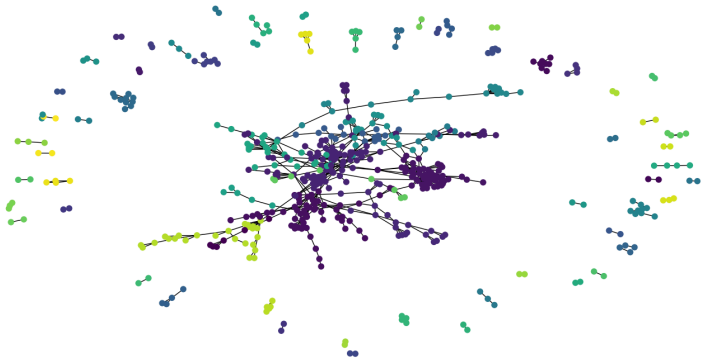
In conclusion, the combination of community detection and centrality algorithms offers a nuanced exploration of the graph, providing insights into both overarching domains and finer distinctions within interconnected research topics.

3. Temporal Community Detection

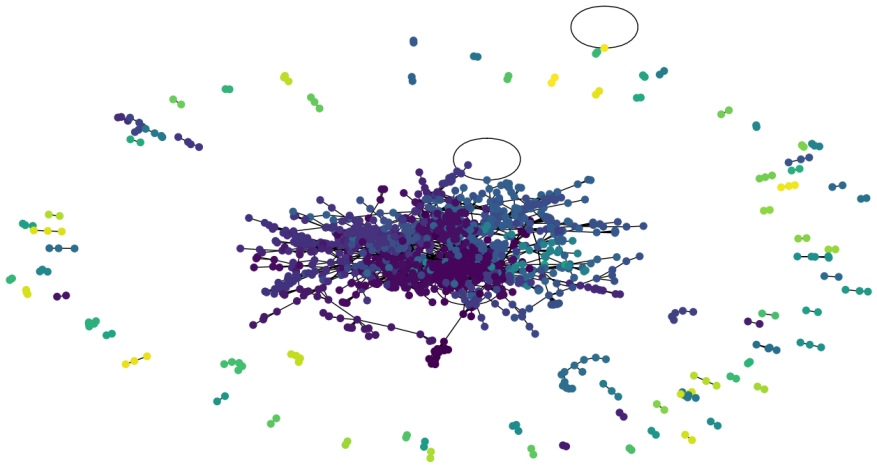
Below are the phases the graph evolved through and how communities evolved with time.



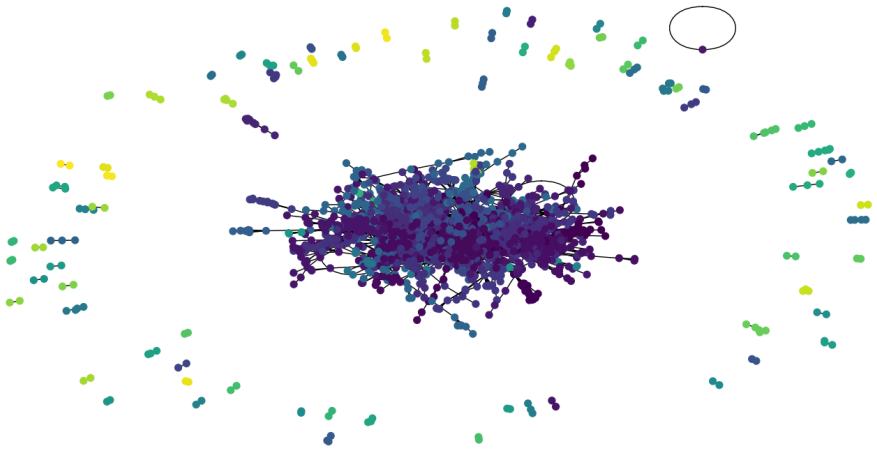
Phase 1



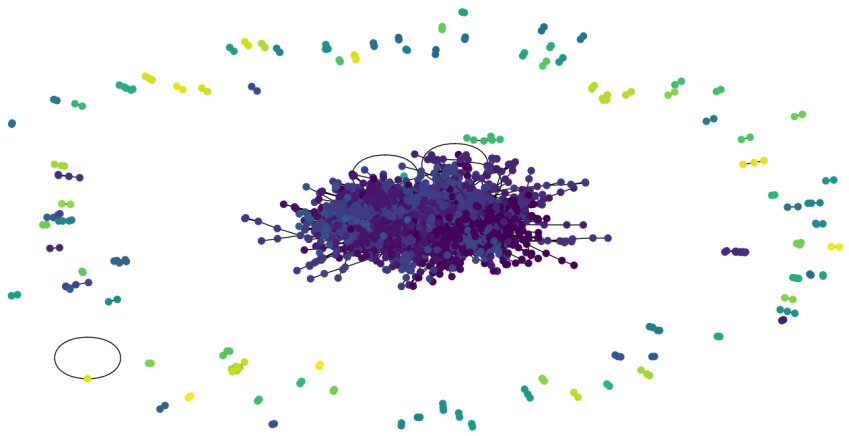
Phase 2



Phase 3



Phase 4



Phase 5

These are the evolving Phases of graph 2.5 yrs, each phase took 6 months to develop.

Interesting Insights

1. Phase 1 shows how initially different domains boomed in research independently.
2. Phase 2 shows how these different domains started linking through research that was a part of both, Though the topology of the graph was sufficient enough to differentiate these domains.
3. Phase 3 shows how these links between some domains increased so much in number that they were soon considered the parts of the same community.
4. While some nodes that were originally clustered into the same community grew branches that they got separated into different communities.
5. This community analysis would be beneficial enough to understand that things like “Wave Nature” and “Particle Nature” that were

initially studied as different topics are now considered under the same domain of “Quantum Physics”
The graph temporal Community Detection Algorithms will help us to find such more topics and study how physics evolved through time.

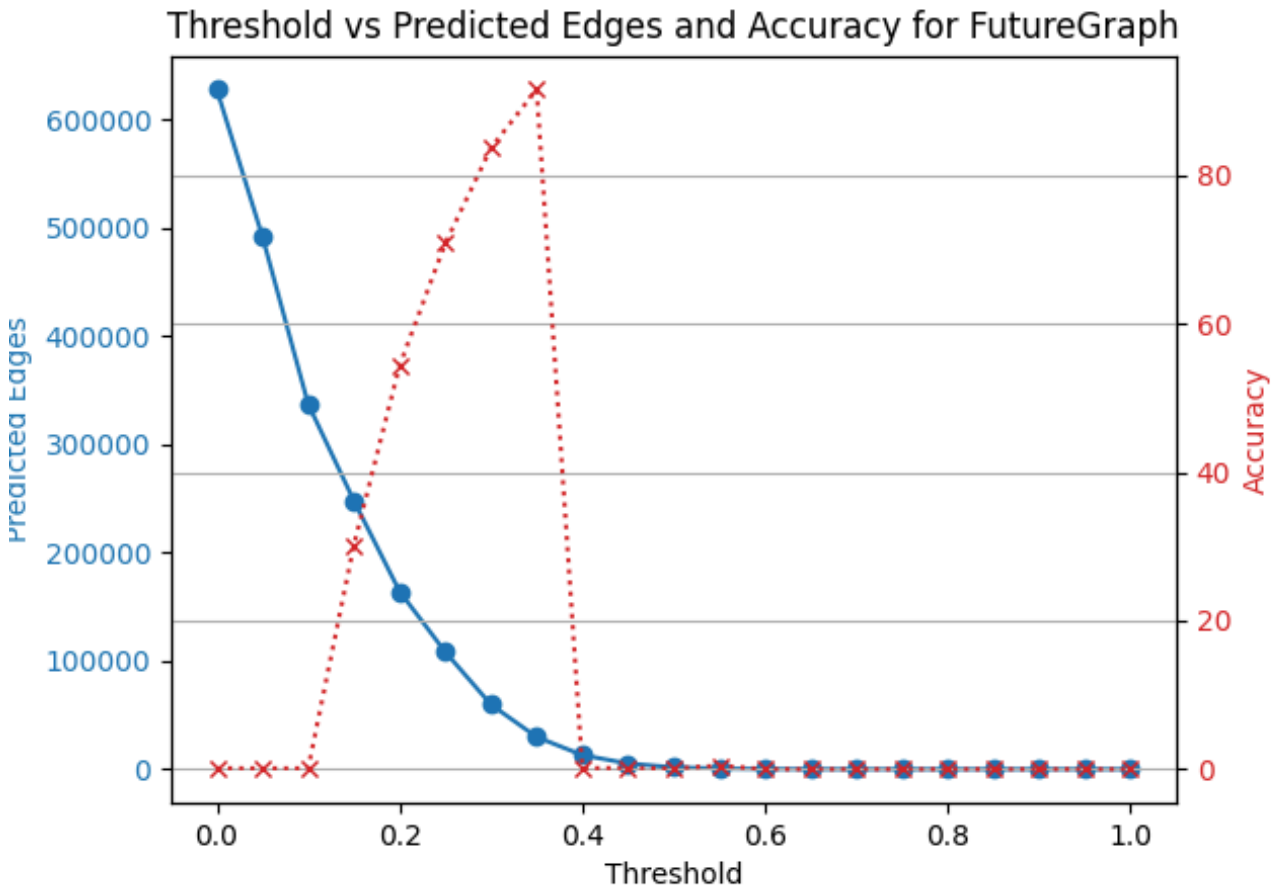
6. Another such analysis that can be drawn from such algorithms is how topics that were originally researched under the same domain separated out with the course of time.
   Such examples include, that the “study of particles at atomic level” was considered a domain that people researched in. Later this thing got separated out and it broke into two communities “Nuclear Studies” and “Nano Technology” that have nothing in common now other than the original roots.
   This is an example of how a community broke into 2 different communities with time and new and new research progressed on the particular sides of the node.so vast in number

Bonus Task:

Link Prediction is a task in graph and network analysis where the goal is to predict missing or future connections between nodes in a network. As before, multiple algorithms exist for this task. However, you are required to implement both a graph neural network, and a classic algorithm like DeepWalk or Node2Vec. For training, you can use the citation network at any time interval [0-T] and for testing/validation, use nodes that appear after time T. ,. Compare the results of the two approaches, and analyze whether the GNN performs better, and if so, why. You will be evaluated on how well the model can predict these edges, as well as your understanding of the link prediction task in graphs

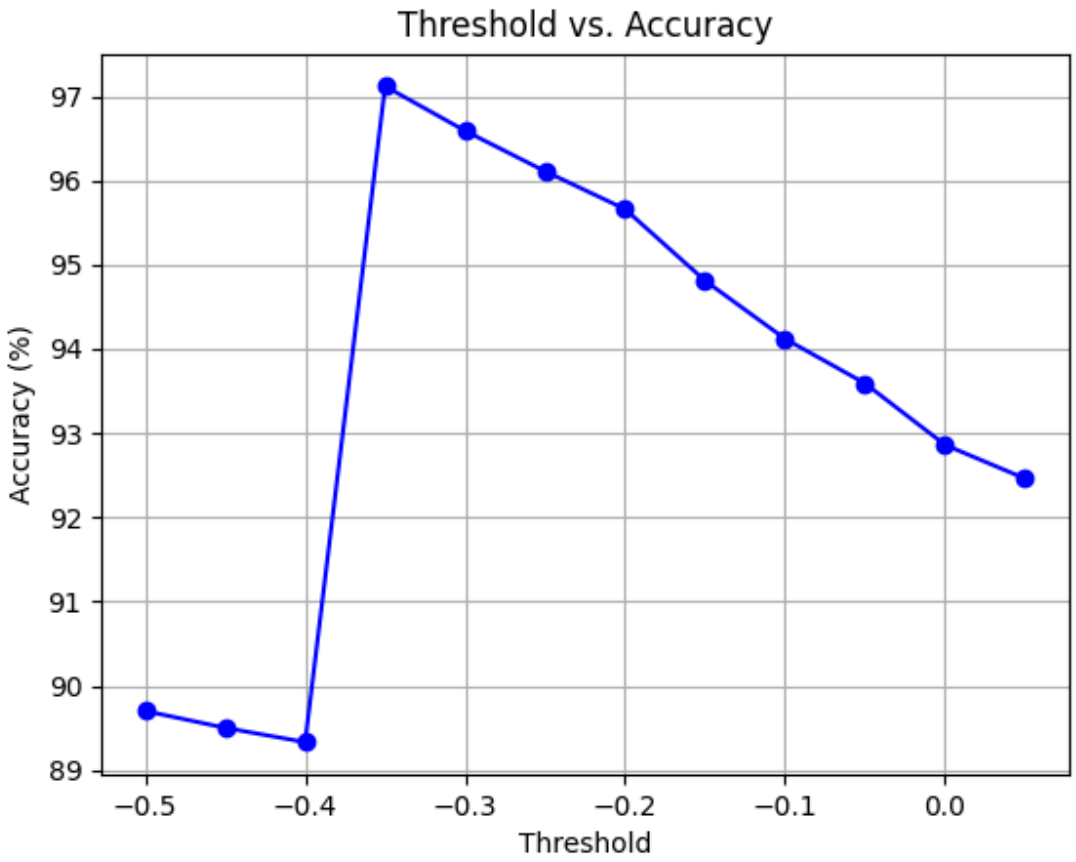
For link prediction graph at time 1993-12-01 was taken and the two models NODE2VEC and GNN were trained and testing was done on the graph at time 1994-03-01

A threshold was set for Node2Vec , and accuracy was compared for different values of threshold ….



This graph accuracy decreases with increase in no of nodes and suddenly drops to 0 when predicted nodes goes to minimal constant set.
Therefore actual accuracy is
87.4 %

The GNN model:
Similar graph- linkage was predicted using neural networks and the plot was plotted



The GNN provides the predicted values from -1 to 1 and therefore such values of threshold. Also note with increase in threshold no of predicted edges decreases , and no decreases to such minimal value that after 0.05 , accuracy either shoots up or shoots down randomly every time you run the script.

The Accuracy being around 97 %

Clearly the accuracy of GNN is better.

Possible Reasons:

- Node2Vec is effective for capturing local neighborhood structures and can generate embeddings that preserve the structural similarity between nodes. Though, Node2Vec may not explicitly consider global graph structure, potentially missing important higher-level relationships.
- GNNs are designed to capture both local and global graph structures, making them potentially more powerful for capturing complex relationships.
  They can adapt to different types of graph-structured data and learn hierarchical representations.
- Since the citation network has a global complex structure and the new edge don't just depend on the local structure but the global topography of the graph GNN performs better.
- GNNs are more expressive in terms of modeling intricate relationships within the graph, potentially making them more suitable for tasks requiring a deeper understanding of the overall structure.