Precog Recruitment Task
Analyzing Citation Networks

Chetan Mahipal

IIIT Hyderabad - February 18, 2024

"A picture is worth a thousand words, but a graph is worth a thousand tables." -Edward R. Tufte.

1 Tasks Information

1. I have done Task 1 & Task 2 completely with all the analysis and codding portion.
2. I haven't done the bonus task completely. The first task involving node2vec is done but GNN is not complete as I am not able to figure out the feature learning and batch dimension for fitting the model.

2 Graph Exploration

2.1 Introduction

The first task of the graph theme is exploring and knowing the graph. We study the graph from the time of it's first node till the entry of latest node. The best way to understand any topic is to start from scratch. So, in this task we start our journey of exploring the given dataset from the year it came into existence. We observe the features of the graph and their evolution over the time. These features learning will help in knowing importance of different components and will help us to predict the future changes that we can expect in the graph.

2.2 Processing the Dataset

The first part of process of analysis is pre-processing the given dataset for a better study. We have two dataset files, first: this file contain information about the edges of the graph, and second: this file contain the time stamp of entry of each node in dataset. On carefully working with the two dataset, I found that there were roughly around 3000 nodes of whose entry date data do not exist. So, I propose two ways to process the data.

1. This method assumes that the nodes, whose entry time data do not exist, does not exist and we remove all the edges which have those nodes.
2. This method makes an assumption that those nodes were from the start the of the graph and existed always so, they are independent of date and are considered to be oldest.

For this task and upcoming ones, I have done analysis using the first method.
Also, as advised we can reduce the dataset by taking 25% to 75% according to the computation power of the machine.

2.3 Density of Graph

2.3.1 What is density in graph?

Graph density represents the ratio between the edges present in a graph and the maximum number of edges that the graph can contain. Conceptually, it provides an idea of how dense a graph is in terms of edge connectivity.

1. For a complete directed or undirected graph, the density is always **1**.
2. A graph with all isolated vertices has a density of **0**

2.3.2 How to calculate dennsity?

Density Formula for Simple Undirected Graphs:

$$DEN_U = \frac{2|E|}{|V| \times (|V| - 1)}$$

Density Formula for Directed Graphs:

$$DEN_D = \frac{|E|}{|V| \times (|V| - 1)}$$

Legend:

- DEN_U - Density of a simple undirected graph.
- DEN_D - Density of a directed graph.
- $|V|$ - Number of vertices (nodes) in the graph.
- $|E|$ - Number of edges in the graph.

2.3.3 Density for Citation Network

I performed an yearly analysis on density of citation network. The density was calculated for whole graph over a period of 11 years (from 1992-2002). the following results were obtained

Table 1: *Density Data for Citation Networks (1992-2002)*

Year	Density
1992	0.004587651122625216
1993	0.0012522519506892535
1994	0.0007614148035924327
1995	0.0006344934194734903
1996	0.0005721215654973133
1997	0.0005192150246573317
1998	0.00046169052234757265
1999	0.00043087620694911275
2000	0.00038751107400653625
2001	0.00037525267902149335
2002	0.0003732938613676198

Using the above data we generated line graph to have a better view over the decreasing nature of the graph. The plot is shown in Figure 1.
Note: Please note 0 of 1992 is due to lack of SCC in that year in Figure 2.

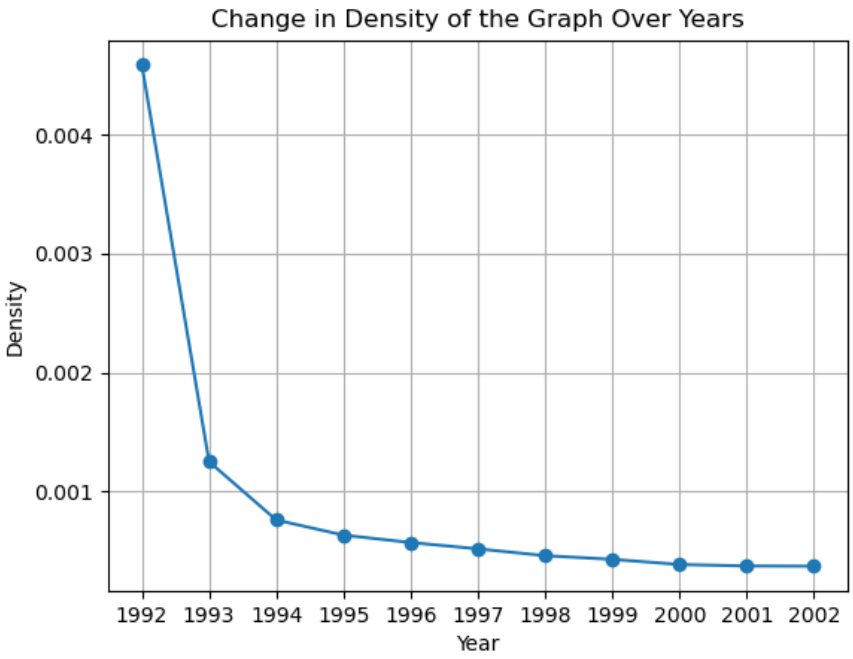

Figure 1: *Plot for Density Data*

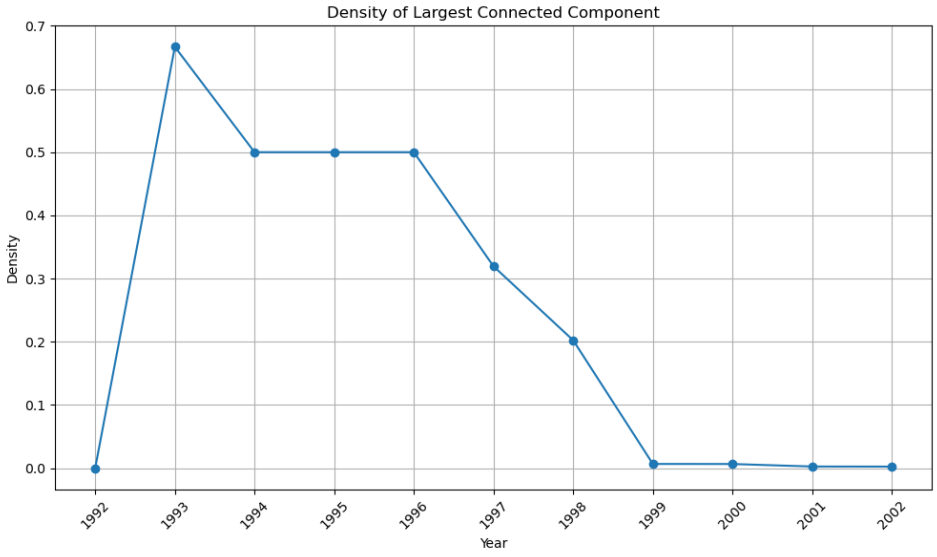

Figure 2: *Plot for Density Data of Largest Strongly Connected Component*

2.3.4 Analysis

The graph in the initial years had a lot of fluctuations that is evident from the sharp decrease in the value of density. The early 1990s was the time of discovery and exploration of new fields in physics. A lot of new topics were introduced in the field, which were then introduced as stand-alone topics, thus the number of isolated vertices or sub-graphs increased over time which explain the sharp decline in density. Over the years , saw collaboration between different topics, thus introduction of new edges between vertices and sub-graphs, this leads to decrease in the declining rate, thus marking that the graph is approaching stability. As the world of physics grew, lot of topics found their inter-dependence on one another thus we can say rating of network formation effected addition of isolated vertices or sub-graphs. The following inferences can be made from the plot:

1. The period from 1992-1995 experienced boom of new research fields, this also hints at technological advancements. This suggests increase in scholarly communities of physics.
   Example: The discovery of high-temperature superconductors, such as Yttrium-Barium-Copper-Oxide (YBCO) and Bismuth-Strontium-Calcium-Copper-Oxide (BSCCO), in the late 1980s and early 1990s sparked intense research activity.

2. The period from 1995-2002 experienced interconnection of different research areas, collaboration between different university or scientific communities.
   Example: The field of quantum information science emerged as a promising interdisciplinary area that brought together researchers from condensed matter physics, quantum mechanics, and computer science.

2.4 Average Out-Degree Nodes

2.4.1 Introduction

The out-degree of a node refers to the number of outgoing edges from that node, representing the extent of its connectivity or influence within the network. This metric is crucial for understanding how nodes interact and disseminate information within the network.

2.4.2 Result

1. **Increasing Trend**: The average out-degree tends to increase over the years, consistent with the densification law as shown in Figure 3. This suggests a growing number of connections per node in the network.

2. **Variability**: While the general trend is upward, there may be fluctuations or periods of stability in the average out-degree due to various factors such as network growth, structural changes, or external influences.

The results obtained from the average node degree analysis suggest a trend indicative of densifying graphs over time. In the next section, we delve deeper into the factors contributing to this phenomenon and explore the densification law in more detail.

2.5 Edges v/s Nodes

2.5.1 Densification Law and Calculation

The densification law refers to the phenomenon observed in networks, such as citation networks, where the number of edges grows faster than the number of nodes over time. Mathematically, it can be expressed as:

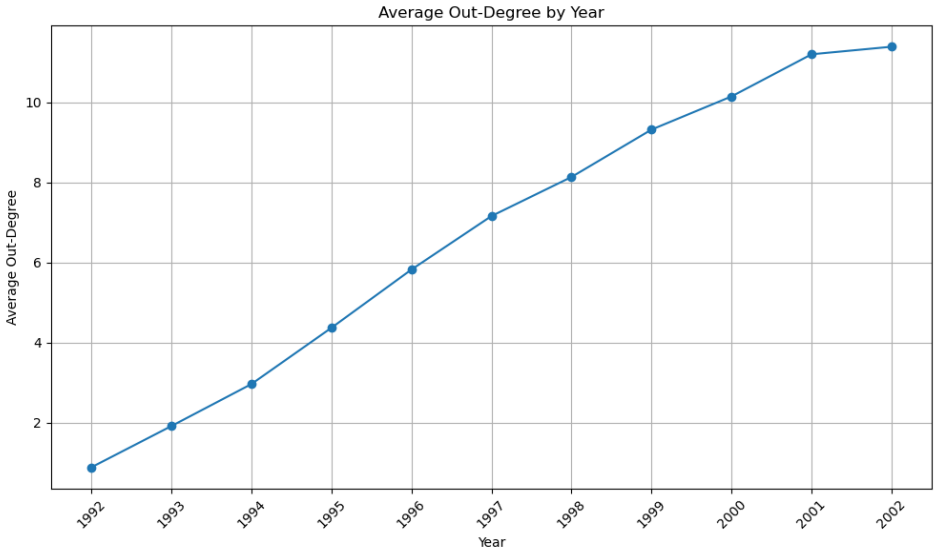

Figure 3: *Rate of Average Out-Degree Nodes (log-log scale)*

$$E \propto V^a$$

where E is the number of edges, V is the number of vertices (nodes), and a is the densification exponent.

To calculate the densification exponent a, one typically performs a regression analysis on a log-log plot of the number of edges versus the number of nodes over time. By fitting a line to the data points and determining the slope of the line, the densification exponent a can be estimated.

2.5.2 Analysis

The plot obtained on plotting these two in normal scale is shown in Figure 4

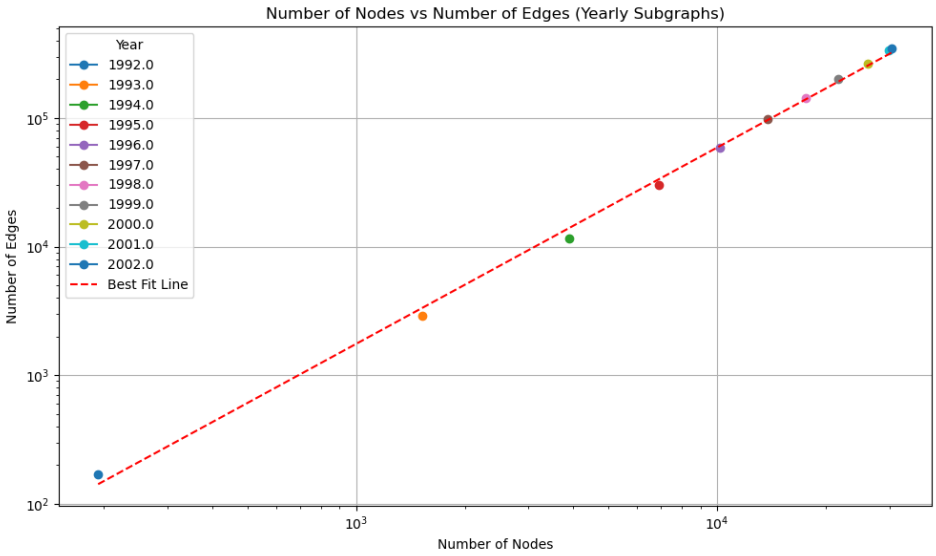

Figure 4: *Plot for Number of Edges v/s Number of Nodes (log-log scale)*

The graph nodes were trained to fit linear regressing which produced value of $a = \mathbf{1.52}$. This implies that our dataset follows below listed properties according to Densification Law:

1. **Property**: A slope greater than 1 (a $>$ 1) suggests that the average degree of the graph increases over time.
   Inference: This suggests a trend towards more interconnected and collaborative research within the field and an average increase in bibliographies of paper.

2. **Property**: A DPL (Degree-Edge Power Law) plot suggests a non-linear growth between number of edges and number of nodes.
   Inference: Information flows more readily between researchers, leading to the dissemination of new findings, methodologies, and theoretical frameworks in simplified terms, which in turn leads to more collaboration between different scientific communities.

3. **Property**: A similar analysis is conducted on a citation graph from the HEP–PH section of the arXiv, exhibiting a densification exponent ('a') of 1.56.
   Inference: By comparing the results with another dataset (the HEP–PH section), the analysis validates the observed behavior and highlights common patterns across different subfields of high-energy physics research.

2.6 Shrinking Diameters

2.6.1 Effective Diameter of Graph and it's Calculation

The diameter δ is an index measuring the topological length or extent of a graph by counting the number of edges in the shortest path between the most distant vertices. It is defined as:

$$\delta = \max_{i,j}\{s(i,j)\}$$

where $s(i,j)$ is the number of edges in the shortest path from vertex i to vertex j. With this formula, first, all the shortest paths between all the vertices are searched; then, the longest path is chosen. This measure therefore describes the longest shortest path between two random vertices of a graph.

Given an undirected network, let $g(d)$ represent the fraction of connected node pairs whose shortest connecting path has length at most d. The hop-plot is the set of pairs $(d, g(d))$, providing the cumulative distribution of distances between connected node pairs. We extend this plot to interpolate linearly between consecutive points $(d, g(d))$ and $(d+1, g(d+1))$ for all d. The effective diameter is defined as the value d at which this interpolated function reaches 0.9.

The plot for effective diameter is shown in Figure 5 which only for 6 years due to computation limitations.

2.6.2 Analysis

1. **Community Formation**: The limiting value of the effective diameter may indicate the formation of stable and distinct communities within the graph. As the diameter stabilizes, it suggests that researchers within each community are closely interconnected, facilitating efficient communication and collaboration within their respective domains.

2. **Scalability and Efficiency**: As the graph stabilizes, it indicates that the network can efficiently handle the growth of new nodes and edges while maintaining consistent levels of connectivity and communication. This scalability ensures that the network remains effective in facilitating collaboration and information exchange, even as it expands over time.

3. **Multiple Ambassadors**: By connecting with multiple ambassadors, newly arriving nodes establish links between previously distant parts of the graph more quickly. This accelerated

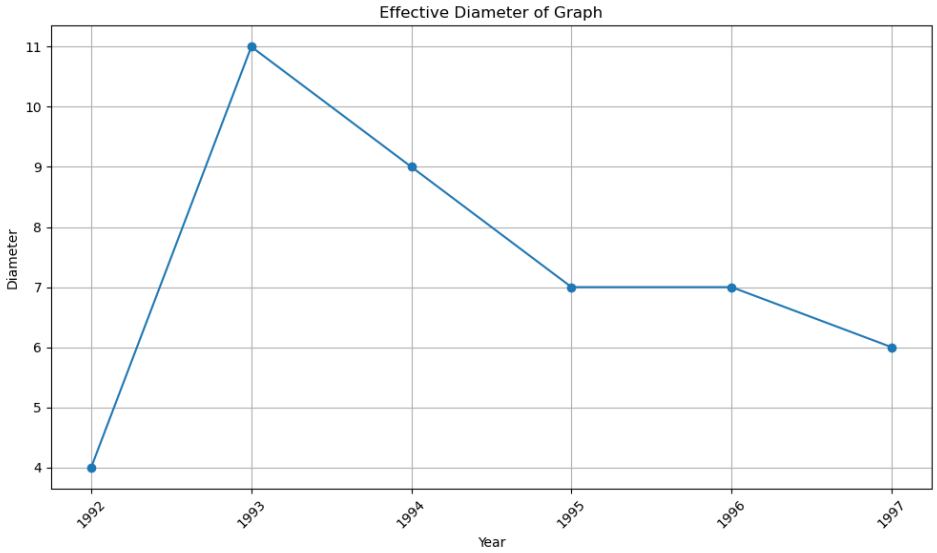

Figure 5: *Plot for Effective Diameter*

integration facilitates efficient information flow and collaboration, fostering a tightly interconnected network where knowledge and ideas can easily propagate.

3 Centrality

Centrality is a crucial concept in graph analytics that deals with distinguishing important nodes in a graph. Simply put, it recognizes nodes that are important or central among the whole list of other nodes in a graph. With so many angles that can define the importance of every node, various metrics are taken into consideration to study each node from a different perspective. The different perspectives of a particular node are studied under different indices, which are collectively known as centrality measures. They help in gaining a better grasp of the network and cutting through the chaos while extracting information from a network.

3.1 Degree Centrality

3.1.1 Definition and Calculation

Degree is a simple centrality measure that counts how many neighbors a node has. If the network is directed, we have two versions of the measure: in-degree is the number of in-coming links, or the number of predecessor nodes; out-degree is the number of out-going links, or the number of successor nodes. Typically, we are interested in in-degree, since in-links are given by other nodes in the network, while out-links are determined by the node itself.

Let $A = (a_{i,j})$ be the adjacency matrix of a directed graph. The in-degree centrality x_i of node i is given by:

$$x_i = \sum_k a_{k,i}$$

or in matrix form (1 is a vector with all components equal to unity):

$$x = 1A$$

The out-degree centrality y_i of node i is given by:

$$y_i = \sum_k a_{i,k}$$

or in matrix form:

$$y = A1$$

3.1.2 Results on Citation Network

We perform two types of analysis on the citation network. The first aims to find the *NodeId* which has maximum centrality for a particular year as shown in Figure 6.The second aims at find the change in value of maximum centrality over the years as shown in Figure 7.

Max Centrality Node and Value over Years

Year	Node with Max Centrality	Max Centrality Value
1992.0	9203203.0	0.057291666666666664
1993.0	9203203.0	0.035386631716906945
1994.0	9209232.0	0.023925906869050682
1995.0	9209232.0	0.021514755051606337
1996.0	9209232.0	0.019071962249311836
1997.0	9306320.0	0.01886381774649931
1998.0	9306320.0	0.018167366867264674
1999.0	9407339.0	0.01715924332824569
2000.0	9606399.0	0.01726574735474999
2001.0	9803315.0	0.020778872578591057
2002.0	9803315.0	0.021639344262295083

Figure 6: *Table of NodeId with Max Degree Centrality*

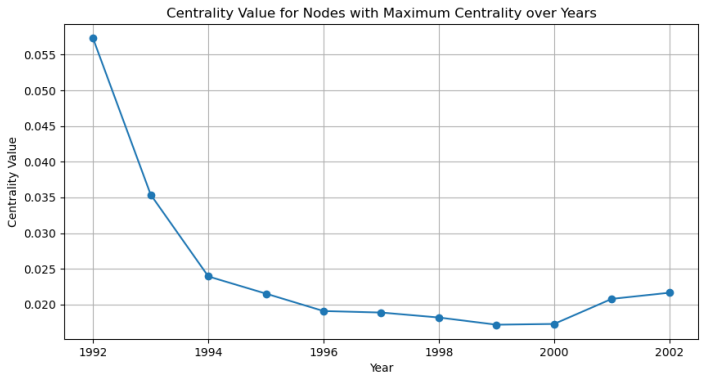

Figure 7: *Plot for Degree Centrality*

3.1.3 Analysis-Max Centrality *NodeId*

1. The NodeId will lead to that research paper that has been cited most, which will help in most identifying most worked domain in that year. The domain of that research will have most collaborations and research papers as the degree centrality is in terms of neighbours.

2. The NodeId also helps to identifying the breakthrough research that leads to numerous more studies and also acted as prerequisites in solving other problems. This helps to order the papers in accordance to their importance in network.

3.1.4 Analysis-Degree Centrality Plot

The max centrality over the years decreases as shown in Figure 7. This hints at formation isolated communities with dependence on one research paper decreases. This is expected as per when one paper gets introduced into bifurcates into many new fields and this process continues thus centrality of one node decreases as new fields may start their own isolated graph and they continue in it's own. Thus, the importance of one node gets divided into multiple components.

3.2 Betweenness Centrality

3.2.1 Definition and Calculation

Betweenness centrality defines the importance of any node based on the number of times it occurs in the shortest paths between other nodes. It measures the percent of the shortest path in a network and where a particular node lies in it.

A node with high betweenness centrality is considered the most influential one over other nodes in the network. This is because this measure can provide insights into the most critical path as disrupting them will disrupt the network.

Directed graph $G = (V, E)$
$\sigma(s, t)$: number of shortest paths between nodes s and t
$\sigma(s, t|v)$: number of shortest paths between nodes s and t that pass through v
$CB(v)$, the betweenness centrality of v:

$$CB(v) = \sum_{s,t \in V} \frac{\sigma(s, t|v)}{\sigma(s, t)}$$

If $s = t$, then $\sigma(s, t) = 1$
If $v \in \{s, t\}$, then $\sigma(s, t|v) = 0$

3.2.2 Results on Citation Network

We perform two types of analysis on the citation network. The first aims to find the *NodeId* which has maximum centrality for a particular year as shown in Figure 8.The second aims at find the change in value of maximum centrality over the years as shown in Figure 9.

Max Centrality Node and Value over Years

Year	Node with Max Centrality	Max Centrality Value
1992.0	9203210.0	0.00010907504363001745
1993.0	9212305.0	3.308768235824936e-05
1994.0	9302210.0	5.497505155397219e-05
1995.0	9503358.0	0.00031585166459378465
1996.0	9603310.0	0.0009408465733847559
1997.0	9606399.0	0.0024683027055698 82
1998.0	9606399.0	0.0032745428979581367
1999.0	9901214.0	0.038739205613360646
2000.0	9901214.0	0.028107585394979234
2001.0	9806471.0	0.04197462867594767
2002.0	9806471.0	0.040449799922551516

Figure 8: *Table of NodeId with Max Betweenness Centrality*

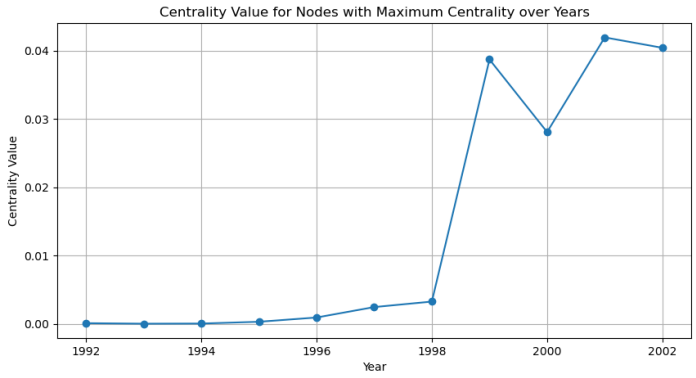

Figure 9: *Plot of Betweenness Centrality*

3.2.3 Analysis-Max Centrality *NodeId*

1. The NodeId will lead to that research paper that has been incoming in the shortest path most of the time. This marks the importance of node as this node leads to variety of different research groups. This implies, person with knowledge of that paper can change fields easily and transition would not be time consuming.

2. Nodes with high betweenness centrality play a critical role in controlling the flow of information within the network, as highlighted by Brandes (2008, p. 137). Investigators possessing high betweenness centrality are better equipped to facilitate or restrict the exchange of information between different research groups, as noted by Freeman (1977) and Abbasi et al. (2012). Their strategic position allows them to effectively relay or withhold information, granting them an advantage in information search processes.

3.2.4 Analysis-Betweenness Centrality Plot

The max centrality over the years increases as shown in Figure 9. An increase in the maximum betweenness centrality implies that certain individuals or entities are gaining more control over the dissemination of knowledge and ideas within the network. These key players may hold significant power in directing research agendas, influencing decision-making processes, or shaping scholarly discourse within their respective domains. The researcher in this position, by controlling flow of information change the network's work direction.

3.3 PageRank Centrality

3.3.1 Definition and Calculation

PageRank Centrality is defined as the measure of directional influence of nodes and, thus, is most suited for directed graphs.PageRankit accounts for link direction. Each node in a network is assigned a score based on its number of incoming links (its 'indegree'). These links are also weighted depending on the relative score of its originating node.

Let $A = (a_{i,j})$ be the adjacency matrix of a directed graph. The PageRank centrality x_i of node i is given by:

$$x_i = \alpha \sum_k \frac{a_{k,i}}{d_k} x_k + \beta$$

where α and β are constants and d_k is the out-degree of node k if such degree is positive, or $d_k = 1$ if the out-degree of k is null. In matrix form we have:

$$x = \alpha x D^{-1} A + \beta$$

where β is now a vector whose elements are all equal a given positive constant and D^{-1} is a diagonal matrix with i-th diagonal element equal to $1/d_i$. Notice that, as seen for Katz centrality, PageRank is determined by an endogenous component that takes into consideration the network topology and by an exogenous component that is independent of the network structure. It follows that x can be computed as:

$$x = \beta(I - \alpha D^{-1} A)^{-1}$$

The damping factor α and the personalization vector β have the same role seen for Katz centrality. In particular, α should be chosen between 0 and $1/\rho(D^{-1}A)$. Also, when the network is large, it is preferable to use the power method for the computation of PageRank.

3.3.2 Results on Citation Network

We perform two types of analysis on the citation network. The first aims to find the *NodeId* which has maximum centrality for a particular year as shown in Figure 10.The second aims at find the change in value of maximum centrality over the years as shown in Figure 11.

Max Centrality Node and Value over Years

Year	Node with Max Centrality	Max Centrality Value
1992.0	9203206.0	0.022990113377578 19
1993.0	9203203.0	0.008589678625435235
1994.0	9203203.0	0.006405414425564468
1995.0	9303255.0	0.005113187895704255
1996.0	9303255.0	0.0054554798415574444
1997.0	9303255.0	0.005180611298073127
1998.0	9303255.0	0.004745856582918326
1999.0	9303255.0	0.00443471123562891
2000.0	9303255.0	0.004097306871113333
2001.0	9303255.0	0.003873631026992555
2002.0	9303255.0	0.003847035800912828

Figure 10: *Table of NodeId with Max PageRank Centrality*

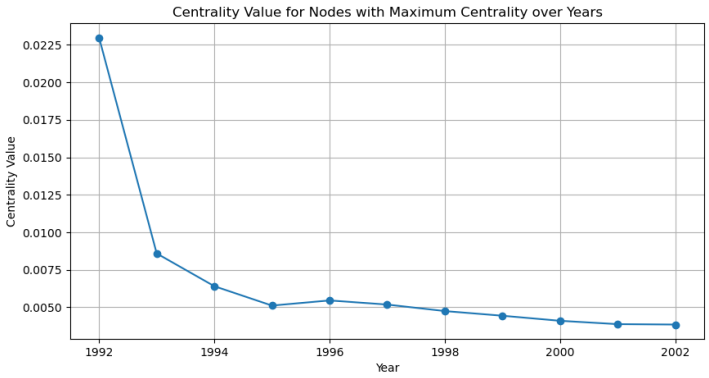

Figure 11: *Plot for PageRank Centrality*

3.3.3 Analysis

1. PageRank serves as an unbiased metric reflecting a paper's citation importance, aligning closely with researchers' subjective perceptions of significance. This alignment makes PageRank a valuable tool for prioritizing the influence of research papers that might have received fewer citations, thereby uncovering the latent impact of scientific publications.

2. Analyzing the trend can reveal insights into the evolving dynamics of the citation network. For instance, consistent increases in centrality values of certain nodes may indicate growing importance or relevance of specific research topics or authors over time. This helps in ranking the authors of the papers.

3. In the context of a citation network, the node with maximum centrality represents a research paper or author that has received a substantial number of citations from other important papers or authors. This indicates the widespread recognition and impact of the research associated with this node.

3.3.4 Analysis-PageRank Centrality Plot

1. Changes in publication patterns, such as increased volume or diversity of research output, may dilute the centrality of individual papers or authors over time. This could be due to factors like interdisciplinary collaboration, increased competition, or changes in citation practices.

2. A decrease in maximum centrality could also signal fragmentation within the citation network. As the network expands or diversifies, the concentration of influence among a few key nodes may decrease, resulting in a more decentralized structure.

3. The decrease in maximum centrality may reflect the rise of new influential nodes within the network. As the field evolves, new research topics, authors, or papers may gain prominence, leading to a redistribution of centrality across the network.

4 Community Detection

Community detection in a network identifies and groups the more densely interconnected nodes in a given graph. This graph can take the form of a social network graph, a biological network, or a representation of a local network of computers, for example. Clusters of related nodes can be grouped using various algorithms. In this article, eight such algorithms are tested against three datasets, and their performance is evaluated.

4.1 Girvan-Newman Algorithm

4.1.1 Methodology

For the detection and analysis of community structures, the Girvan-Newman algorithm relies on the iterative elimination of edges that have the highest number of shortest paths between nodes passing through them. By removing edges from the graph one-by-one, the network breaks down into smaller pieces, so-called communities.

The idea is to find which edges in a network occur most frequently between other pairs of nodes by finding edge betweenness centrailities. The edges joining communities are then expected to have a high edge betweenness. The underlying community structure of the network will be much more fine-grained once the edges with the highest betweenness are eliminated which means that communities will be much easier to spot.

4.1.2 Algorithm

The Girvan-Newman algorithm can be divided into four main steps:

1. For every edge in a graph, calculate the edge betweenness centrality.
2. Remove the edge with the highest betweenness centrality.
3. Calculate the betweenness centrality for every remaining edge.
4. Repeat steps 2–4 until there are no more edges left.
5. Check if number of weakly components increases, algorithm stops.

4.2 Result and Analysis

The communities obtained on running Girvan-Newman algorithm is shown in Figure 12.

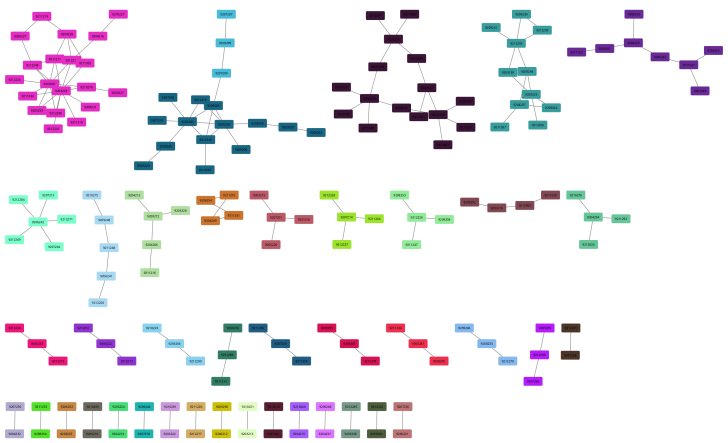

Figure 12: *Graph with coloured Communities*

Let's do analysis for getting the communities:

1. **Disjoint Components**: The communities majorly formed on the basis of the disconnected or isolated sub-graphs. As all the components do not have edges connecting them, we treat each component as different graphs to work on.

2. **Existence of single shortest path**: In most of the the components there exist only single shortest path or the shortest path length is same, thus those edges will have low Edge Betweenness Centrality.

3. **Edge Removal**: On first run, due to graph being directed, the edge with highest edge centrality is present in 2^{nd} as it have the most number of shortest path. On removing this number of connected components increases,algorithm stops.

4.3 Label Propogation Algorithm

4.3.1 Methodology

The Label Propagation algorithm (LPA) is a fast algorithm for finding communities in a graph. It detects these communities using network structure alone as its guide, and doesn't require a predefined objective function or prior information about the communities.

LPA works by propagating labels throughout the network and forming communities based on this process of label propagation.

The intuition behind the algorithm is that a single label can quickly become dominant in a densely connected group of nodes, but will have trouble crossing a sparsely connected region. Labels will get trapped inside a densely connected group of nodes, and those nodes that end up with the same label when the algorithms finish can be considered part of the same community.

4.3.2 Algorithm

The algorithm follows as:

Algorithm 1 Label Propagation Algorithm (LPA)

Input: Graph $G = (V, E)$
Output: Community labels for each node
Initialization: Each node $v \in V$ is initialized with a unique community label.
while not converged or maximum iterations not reached **do**
    for each node $v \in V$ **do**
        Let $N(v)$ be the neighbors of node v.
        Let $L(v)$ be the set of labels of neighbors of node v.
        Let $l_{\max}$ be the label with the maximum frequency in $L(v)$.
        Set the label of node v to $l_{\max}$.
    end for
end while

4.3.3 Result and Analysis

The communities obtained on running label Propagation algorithm is shown in Figure 13.

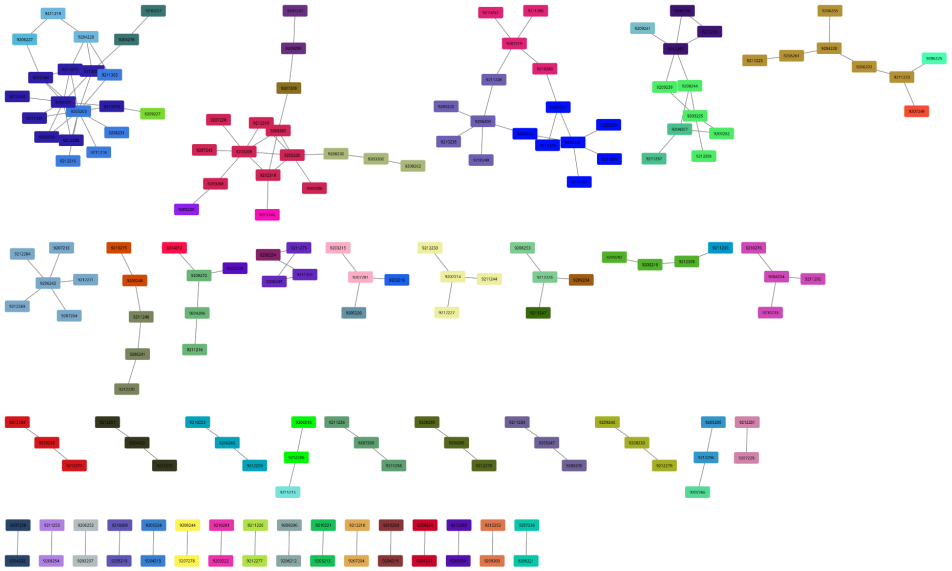

Figure 13: *Graph with coloured Communities*

The community detection in this algorithm is based on the neighbour nodes in the graph. The outcome of the algorithm is due to the different scientific research communities. Each community has it's own set of research papers, it may happen that a few papers of that community is cited by other but majority papers are connected among themselves, this leads to labeling on those communities. The semi supervised nature of the algorithm helps to detect these communities. Th number of communities in Label Propagation is more as compared to Girvan-Newman as it study's the features of the graph (taking neighbours into consideration) in terms of nodes not edges as neighbouring nodes are more important in graph like citation network. Thus, through these communities we can bifurcate different research fields and can also do independent analysis on these, for example, ranking the authors or papers according to it's importance etc.

5 Temporal Community Detection

5.1 Introduction

Temporal community detection refers to the process of identifying and analyzing communities or clusters within a dynamic network that evolves over time. Unlike traditional community detection, which assumes static network structures, temporal community detection considers the temporal aspect of network data, where nodes and edges may appear, disappear, or change over time.

The goal of temporal community detection is to uncover recurring patterns of interactions, collaborations, or relationships within the evolving network across different time intervals. This involves identifying groups of nodes that exhibit dense connections or interactions over time, indicating shared characteristics, functions, or roles within the network.

5.2 Processing Dataset

Please refer to the Section 2.2. The analysis is don for two year, with a period of 3 months, implies a total 8 scenarios have been taken into consideration of which we will discuss only 7, leaving the first one due to it's size which deviates the starting results.

5.3 Results

We performed temporal analysis on mainly three attributes and these are:

1. Modularity
2. Size of Clusters
3. Number of Communities

We also did visual analysis on the graph using the nodes and will be providing with analysis only. For the rest of three the results are shown in Figure 14, 15, 16.

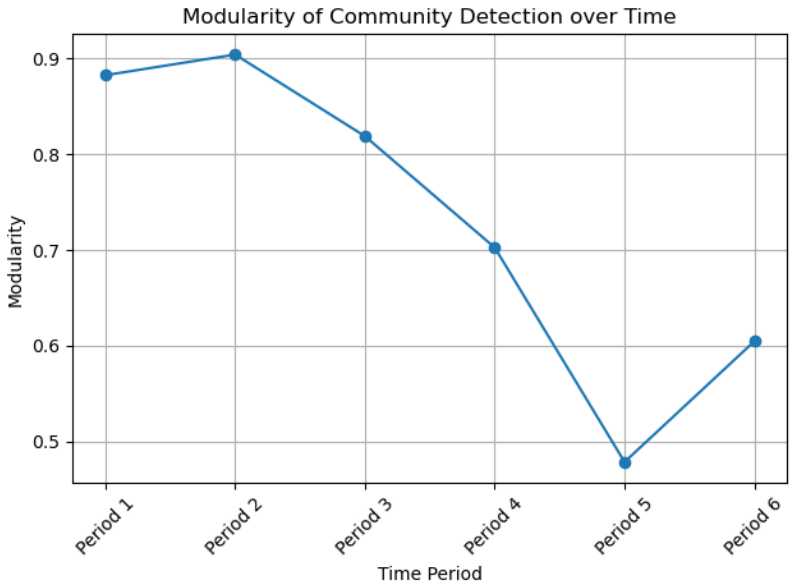

Figure 14: *Graph with coloured Communities*

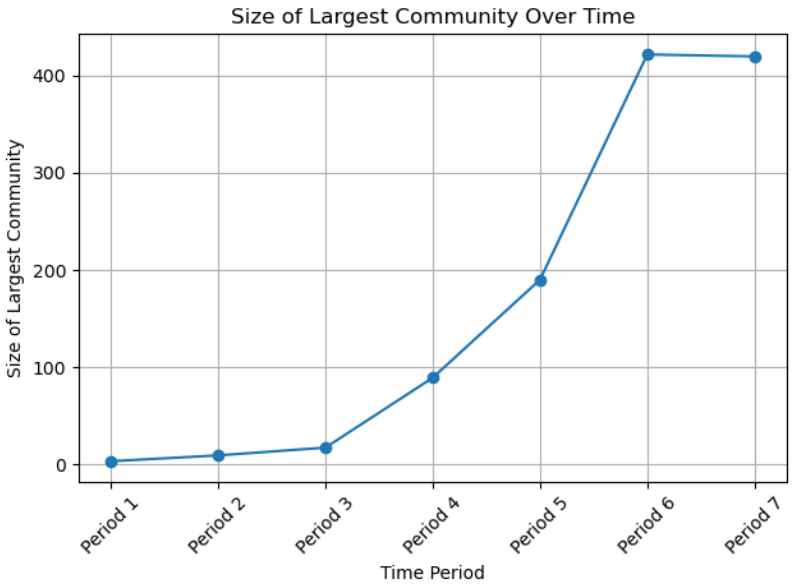

Figure 15: *Graph with coloured Communities*

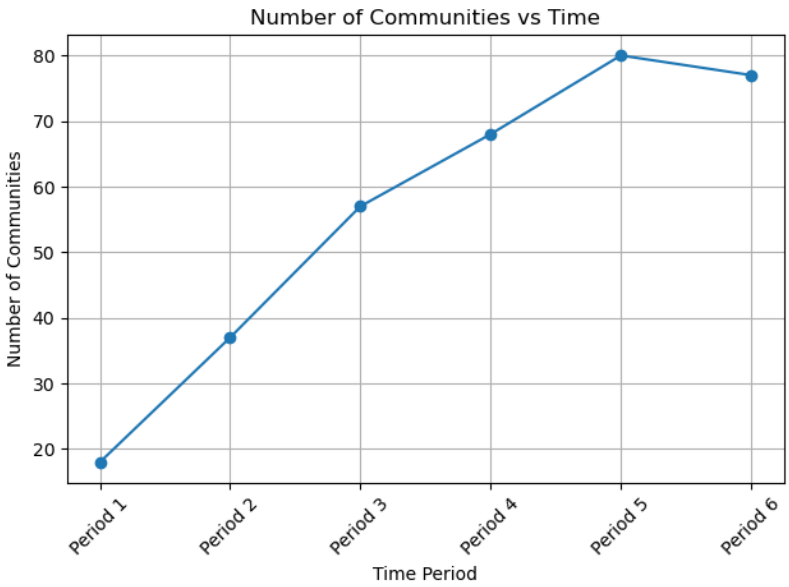

Figure 16: *Graph with coloured Communities*

5.4 Analysis

5.4.1 Modularity

Modularity is a measure of the structure of networks or graphs which measures the strength of division of a network into modules (also called groups, clusters or communities). Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules.

Modularity is the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random. It is positive if the number of edges within groups exceeds the number expected on the basis of chance. For a given division of the network's vertices into some modules, modularity reflects the concentration of edges within modules compared with random distribution of links between all nodes regardless of modules.

The formula for modularity (Q) is given by:

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \frac{s_v s_w + 1}{2} \tag{1}$$

where:

Q : Modularity
m : Total number of edges in the network
vw : Pairs of nodes in the network
A_{vw} : Actual number of edges between nodes v and w
k_v, k_w : Degrees of nodes v and w
s_v, s_w : Community indicators for nodes v and w

Following analysis can be drawn with decrease in modularity over years:

1. A decrease in modularity may also create communication barriers between specialized research communities. As the network becomes more interconnected, researchers may face challenges in effectively communicating with each other due to differences in terminology, methodologies, and views.

2. Decreased modularity may lead to the breakdown of networks into smaller, more isolated clusters or sub-communities. This fragmentation can result in the formation of echo chambers, where ideas are reinforced within homogeneous groups without exposure to diverse perspectives.

3. At the same time, overcoming the challenges open ways to new venture, joint conferences and strategical discussion for spreading and diffusing ideas.

5.4.2 Size and Number of Community

Following analysis can be drawn with increase in size and number of community over years:

1. **Network Growth**: As expected the researches open doors to new topics, which leads to more papers citing the previous one , so overtime the network should grow and community size increases and becomes dense.

2. **Bridges**: It may also happen that overtime due to knowledge diffusion sub-graphs becomes bridging points for two communities, forming larger communities thus establishing deep relationship in the work of those sub-graphs.

3. **Newer Fields**: Research leads to new fields which not only forms new communities but can leads to growth of other communities as well. The number of communities indicate the progress in the field and can be used to predict the future of making of new research communities.

One thing to notice is that the size of community graph is convex graph while number of communities graph is concave, this means with time the communities are merging to form larger communities indicating inter dependency. The rate of increase in size is more than rate on increase in number and it's effect can be seen in form of multiple collaborations.

5.4.3 Visual Interpretation

By carefully analyzing each years graph, we see that new communities tends to grow around the older communities, this may help in back-tracking and studying the origin of any kind of research or thesis. We can also see little fragmentation as the community becomes more interconnected, it kind of detaches from parent one, basically flow information tends to block from the parent community.

References

[1] Aditya Grover & Jure Leskovec(2016). node2vec: Scalable Feature Learning for Networks

[2] J. Leskovec, J. Kleinberg and C. Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations.