**COLLEGE OF ENGINEERING, GUINDY**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**Project Report**

**CS6001: DATA MINING**

**Influential nodes in a Citation Network**

**Team Members**

**Niranjana S,2018103570**

**Sarikaa S, 2018103058**

**Objective Of This Project:**

Measuring the importance of the node in a citation network

**Problem Statement**

**Motivation**

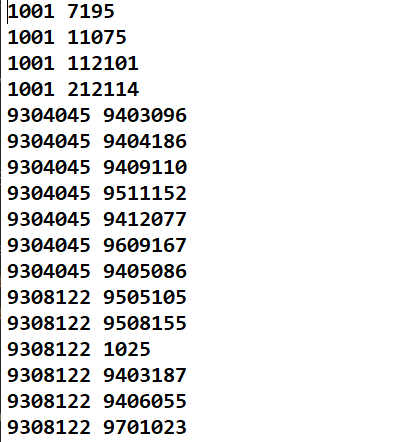
Finding the importance of node enable many interesting applications including ranking the items according to impact, examining the correlation with popular impact indicators, discovery of trends in the graph

**Problem Statement**

How to find the influential nodes in a citation network using a computationally cheap metric while taking the local neighborhood and topological characteristics’ into account?

**Dataset**

***Sample of the dataset***

******

This is the dataset for constructing a network to perform graph measures. The first column is taken as the beginning node and the second column as the target nodes and the edges are connected in this direction for directed graphs.

***Graph Mining***

It is a process to find patterns in graphs. By doing graph mining, the frequent subgraphs and the relationships in the graph can be found and this can be used to form communities in the graph and identify relationships between graphs.

***Networkx***

It is a python library used to construct and analyze graphs. This is used here to construct the graph and perform graph measures on it in order to find the most influential node.

***Graphia:***

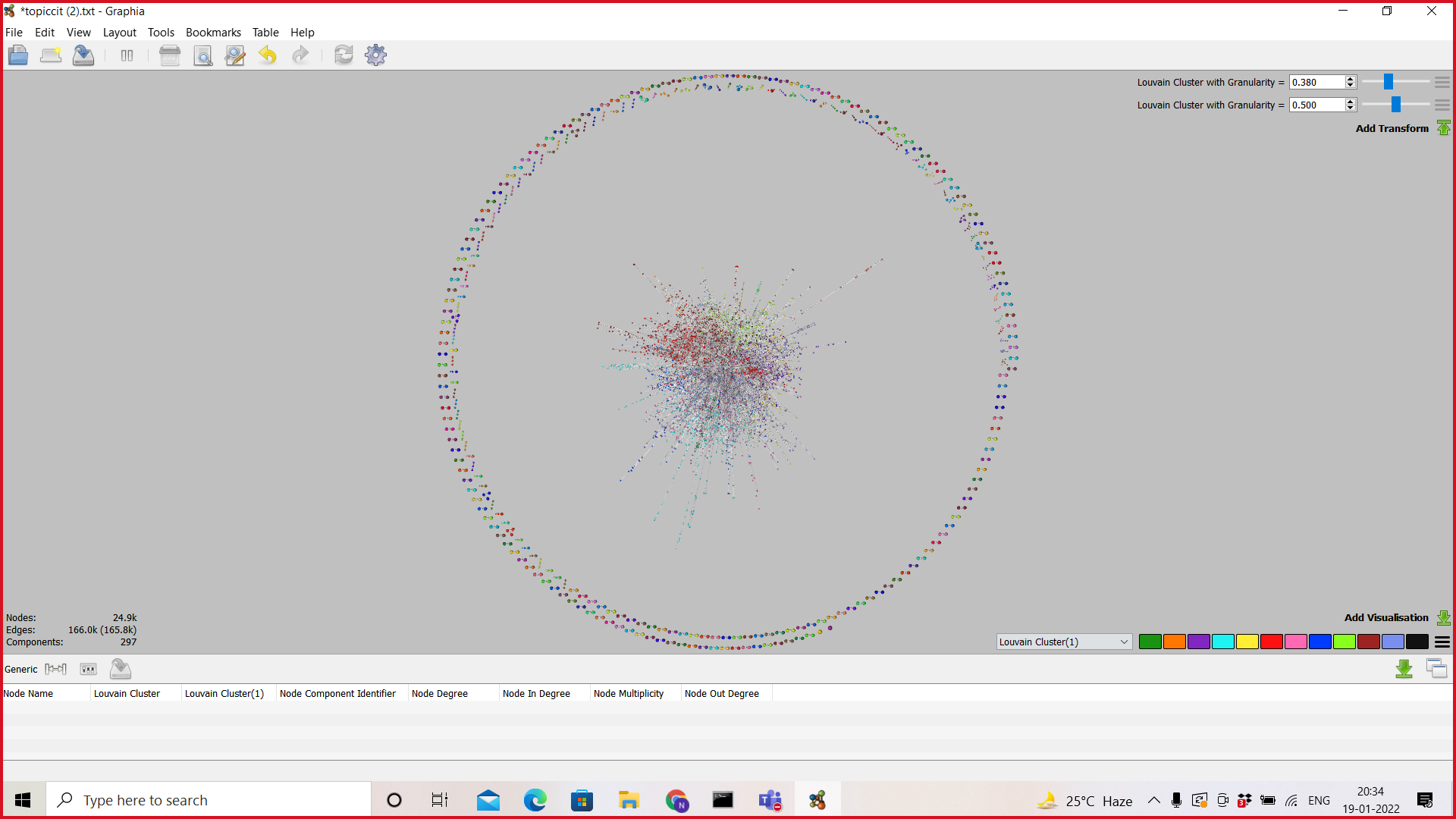
It is a visualization tool used to create graphs from given values and visualize them.

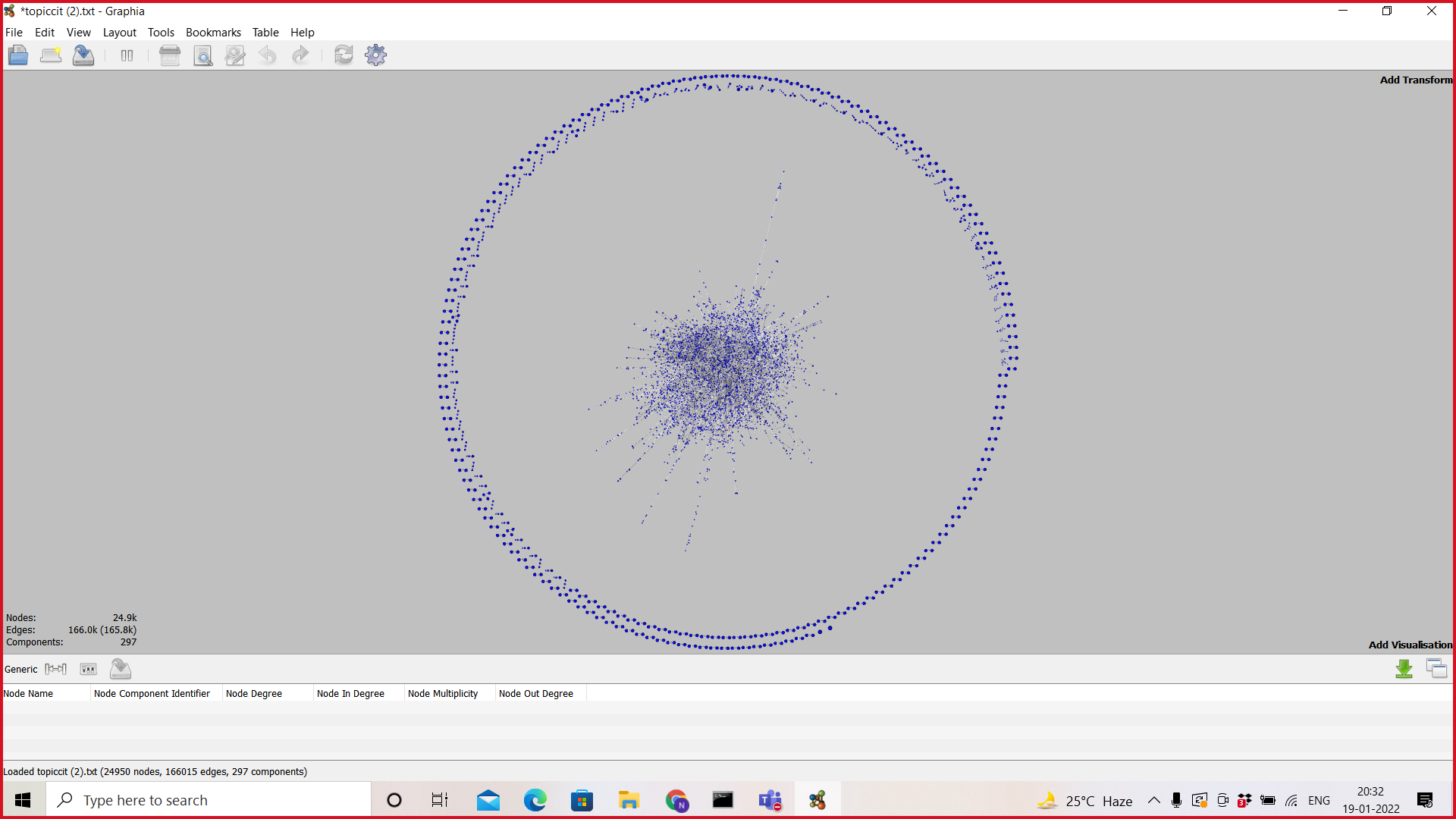
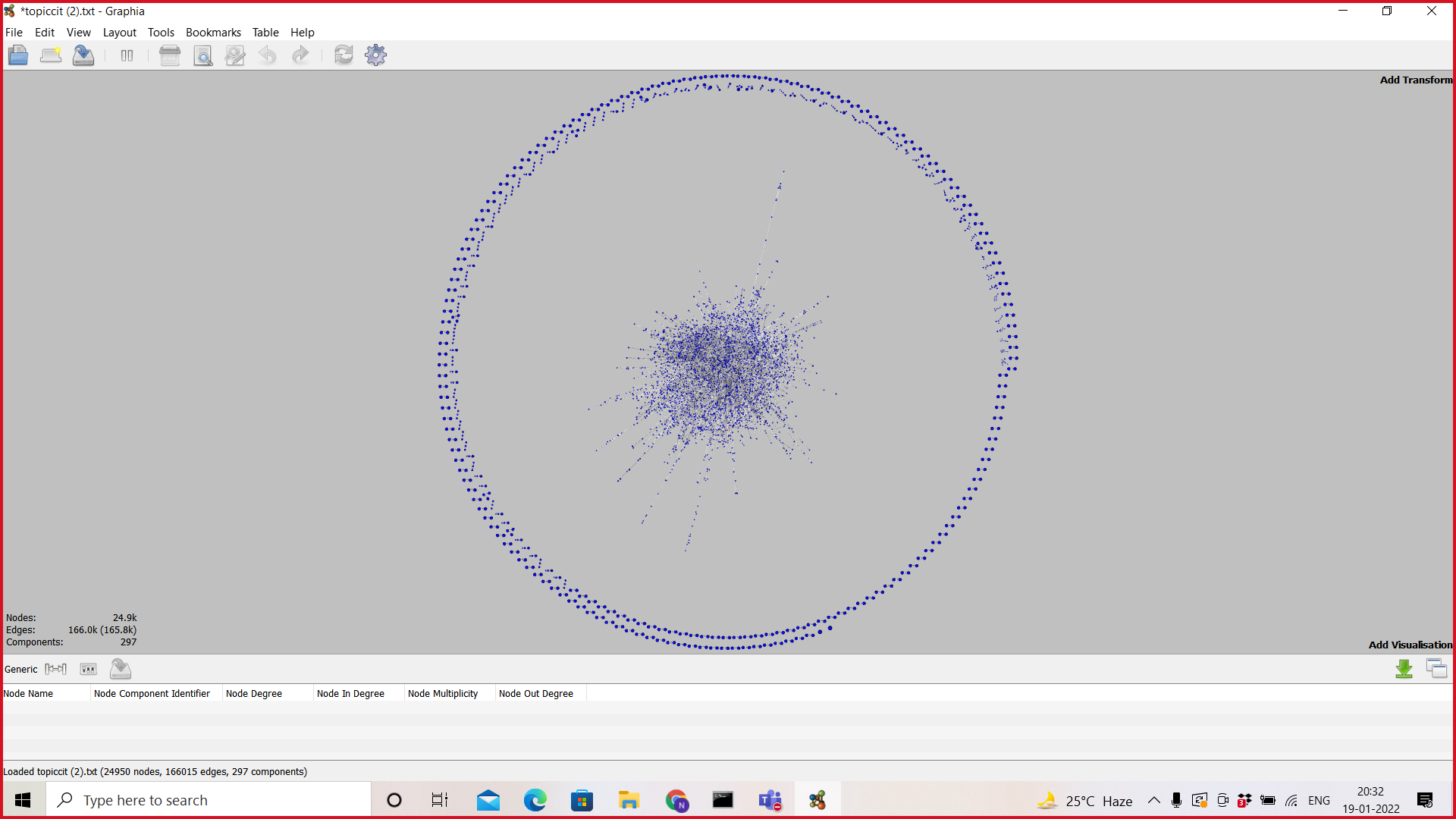
***Graph Visualization***

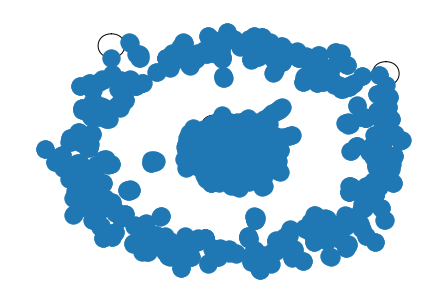
The given file “topiccit.txt” is first visualized. Graphia tool is used for visualization of graphs here.

Graph visualization is done by Graphia software and networkx package (nx.draw)

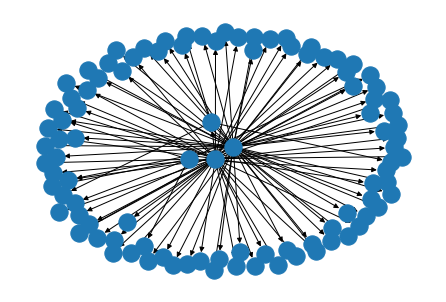
***Louvian clustering***

******

******



For 100 samples



**Background**

Common measures for the influence of nodes is evaluated from global structure information, such as betweenness centrality , closeness centrality and Katz centrality. These methods show good performance in node sorting but they are computationally expensive. The influence of nodes is quantified by local information, such as degree centrality, semilocal centrality , hybrid degree centrality, average shortest path centrality, and h index . Local measures are less efficient because they only consider local neighborhood information. TResearch based on random walk evaluates the influence of nodes through multiple iterative operations with high-computational complexity such as feature vector centrality , PageRank, LeaderRank, VoteRank and HITS

***Graph properties***

The properties of the graph obtained from the above graph:

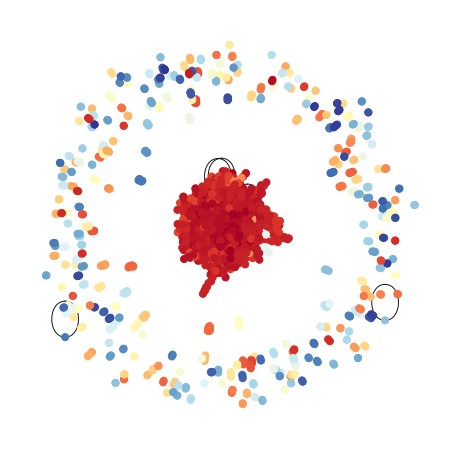
* Number of nodes : **24950**
* Number of edges : **166015**
* Cycles : **141116**
* Self loops : **31**

**Louvain Clustering:**

This method is used to detect communities in graphs. It finds communities by maximizing the modularity for each community. It simply checks how densely connected within the community is than with other communities.

Louvain Modularity Partition: **24950**

Girvan connected components: **297**



***Graph Measures***

**Total degree measures**

Node with highest degree

1193

9711200.0

Node with highest Outdegree

1177

9711200.0

Node with highest indegree

118

9611190.0

Nodes with zero outdegree (sink nodes): 5618

Nodes with zero indegree : (source nodes) 3963

Nodes with both zero indegree and outdegree : 0 (means there are no isolated nodes)

There are two nodes with only self loop

8 nodes with outdegree 1. Does that mean they are sink nodes? Yes.

1 node with indegree 1. Does that mean it is a source node? Yes.

There are 2 selfloop containing nodes with degree (2). That means there are 2 isolated nodes. Will isolated nodes affect the communities? They are not counted as separate communities. They are outliers.

***Clustering coefficient***

The highest coefficient is for a node with degree=2. The lowest coefficient is for an isolated node with only a self loop. There are many nodes with coefficient 0. The highest degree node and influential node in the overall graph has lesser value of coefficient. The higher value coefficient belongs to the community with the number of nodes 7332.

***Community detection***

A community where a group of nodes is more associated with each other than other groups. We find communities to know how the graph is partitioned or hierarchically organized. The quality of the communities is measured using “modularity”.

Here, communities are formed based on the source nodes.There are 3964 source nodes. The community with the largest number of nodes (13217). The self loops are removed before forming communities. There are a total 3964 communities based on source nodes.

Communities are constructed using depth first search based on the source node.

If we consider a community based on the connected components then there are 297 connected components, which are found using Girvan Newman method and also evident from the graph visualization using Graphia.

All communities have only one source node and each source node is a part of exactly one community, although the communities are not connected components.

**Influential Node Identification**

Various measures are implemented to find the influential node in each community. Computationally expensive measures are implemented for a lesser number of communities.

**Centrality measures:**

**Degree centrality:**

**Definition**

It is the total number of edges of a node. The nodes with the maximum degree are the influential nodes.

**Formula**

where A is the adjacency matrix of G

**Drawback**

Degree Centrality doesn't tell which node is the center in the graph, it just tells the number of connections. One graph can have the same degree but far off the edge of the network and the other node has the same degree but be central.

**Values: 9711200.0**

**Keys: 0.04781754779750692**

**Closeness centrality:**

If a node can reach more other nodes both directly and by way of shortest indirect paths than another node; then that node may be more important. Closeness centrality is a measure of the average shortest distance from each vertex to each other vertex.

**Formula:**



where d(u, v) is the geodesic distance between u and v.

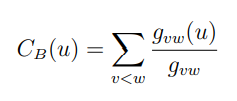
**Drawback:**

It considers only the shortest path and excludes the contributions of longest and other paths.

**Betweenness centrality:**

It looks for the shortest paths through a particular node. When a node connects two or more nodes of the network together, it can be a key node by controlling the flow between them. If two nodes are selected in a connected graph, at least one shortest path will exist between them.

**Formula:**

****

where  is the number of shortest paths between v and w, and  is the number of shortest paths between v and w containing u.

**Drawback:**

The main drawback is high computational cost, especially for large networks. It considers only the shortest path and excludes the contributions of longest and other paths.

**Feedback Centralities**

Feedback centralities form an especially appealing class of centrality measures. These measures assess the importance of a node recursively by looking at the importance of its neigh- bors or, in directed graphs, direct predecessors. Eigen, Katz, Pagerank are the feedback centralities.

**Eigen centrality:**

Eigenvector centrality is useful for understanding which nodes can get information to many other nodes quickly.it looks at a combination of a node’s edges and the edges of that node’s neighbors.

**Formula:**





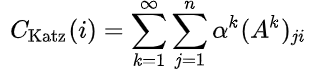
**Drawback:**

The central node passes its centrality along to its neighbors, and then, in the next step, that centrality gets reflected back to the central node. This repeated “reflection” can result in a heavy, and perhaps unjustifiably large, accumulation of eigenvector centrality near high-degree nodes in a network.

**Katz Centrality**

Katz centrality computes the relative influence of a node within a network by measuring the number of the immediate neighbors (first degree nodes) and also all other nodes in the network that connect to the node under consideration through these immediate neighbors. Unlike other centrality measures which consider only the shortest path between a pair of nodes, Katz centrality measures influence by taking into account the total number of walks between a pair of nodes.

Formula:



Where the element at location (i,j) of reflects the total number of k numbers of connections between i and j.

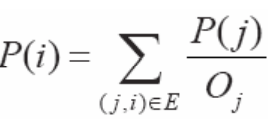
**Drawback:**

The drawback is that when a node becomes very central in a network, it passes its centrality to all of its outgoing links, making all those nodes very popular.

**Page rank**

A page with a high prestige score is more important than a page with a low prestige score. It relies on the link structure to tell the quality of the page. The importance of a page is determined by the sum of all PageRank scores of pages pointing to it. The prestige score of a page should be shared pages that it points to.

**Formula:**



Where  is the out degree of a page j.

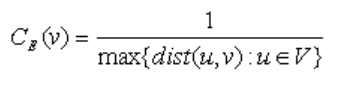
**Drawback:**

The major drawback of the PageRank algorithm of Google is that ranking is based not only on the page ranks produced but also on the number of hits to the Web page. This paved the way for illegitimate means of boosting page ranks. As a result, Web pages whose page rank is zero are also ranked in top-order.

**Reciprocal of Eccentricity**

The eccentricity of a node v in a network is the maximum distance from v to any other node. It calculates the degree to which a network is dominated by a particular node.

Formula:



Drawback:

This centrality measure is more meaningful if only its value is high. The nodes with low reciprocal of eccentricity value have less role to play.

**Hub score and authority score**

A high hub node points to many good authorities and a high authority node receives from many good hubs. The authority score of a vertex is therefore proportional to the sum of the hub scores of the vertices on the in-coming ties and the hub score is proportional to the authority scores of the vertices on the out-going ties.

**Formula**:

*Each node’s Hub score = (Authority score of each node it points to).*

*Each node’s Authority score = (Hub score of each node pointing to it).*

**Drawback:**

* Topic drift (Topic drift can happen when a page with high effective Page Rank and low delta links to a page with a high topic rank).
* Poor performance due to poor selection of k (convergence factor)

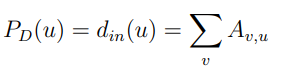
***Prestige measures:***

It tells the importance of nodes more than centrality. It is applied for directed graphs.

**Degree prestige:**

Degree prestige is defined as the in-degree of each node.All nodes have indegree 1 except source nodes in every community, Thus degree prestige is same for all nodes except source node which is 0.

Formula:

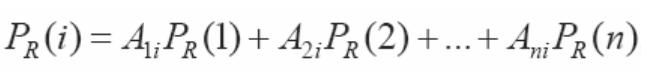


**Drawback:**

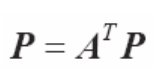
As a measure of prestige, the influence domain of a vertex does not distinguish between direct and indirect choices, which is not completely desirable. Usually, we consider direct choices more prestigious than indirect choices.

**Rank prestige:**

An node’s prestige depends on the prestige of those nodes that it is connected to



This equation could be written in a matrix form



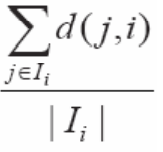
**Drawback:**

It disregards the concern of diversity, or handles it with non-optimized heuristics.

**Proximity prestige:**

Proximity prestige of selected nodes by dividing the influence domain of a node by the average distance from all nodes in the influence domain. A larger influence domain and a smaller distance yield a higher proximity prestige score.

Formula:



Where  is the size of 

**K core algorithm**

A k-core is a maximal subnetwork in which each vertex has at least degree k within the subnetwork. A k-core is not necessarily a cohesive group itself. The *k*-cores define a hierarchy of *k*-shells, each of which consists of all the nodes in the *k*-core but not the (*k* + 1)–core. A decomposition in terms of the *k*-cores highlights a network’s core-periphery structure and is typically obtained by iteratively removing the *k*-shells , starting with peripheral low-degree nodes in the outer shells and working toward embedded high-degree nodes in the inner cores.

**Drawback:**

Must be careful that removing a vertex reduces the degree of all the vertices adjacent to it, hence the degree of adjacent vertices can also drop below-‘K’.

**VoteRank:**

VoteRank computes a ranking of the nodes in a graph G based on a voting scheme. With VoteRank, all nodes vote for each of its in-neighbours and the node with the highest votes is elected iteratively. The voting ability of out-neighbors of elected nodes is decreased in subsequent turns.

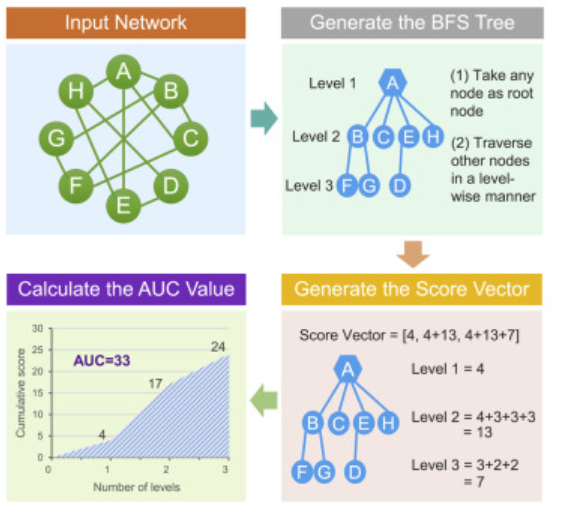
**Drawback:**

It does not distinguish the voting ability of each node.

**TARank:**

TARank method that integrates the information collected from the breadth-first search tree to identify influential nodes. It is a graph-traversal approach to identify influential nodes in a network.

**Algorithm:**



**Cocited papers:**

The pair of papers with common successors (papers which cite both) are called cocited papers. We have found the cocited papers for a particular node and the common successors between each pair.

**Hybrid degree centrality:**

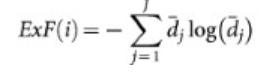
A hybrid degree centrality is proposed to evaluate nodes’ spreading ability. It can take the advantages of degree centrality or local centrality depending on the given spreading probability. A hybrid centrality based on the linear combination of degree centrality and cohesion centrality.

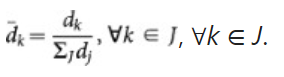
**Drawback:**

It has service coverage problems.

**Expected force of infection:**

The expected force is a node property derived from local network topology, independent of the rest of the network or any specific spreading process. The expected force of infection can be approximated by the entropy of the *dj* after normalisation,

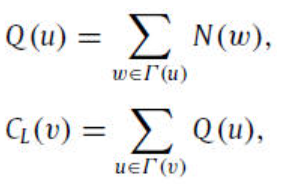


where *i* refers to the seed node and 

**Drawback:**

Expected forces are consistently strong, only greater with mean correlations and small variance.

**Semilocal centrality**

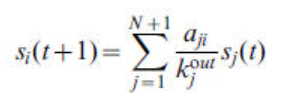


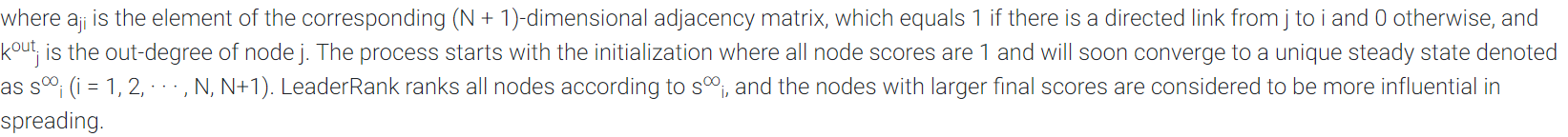
where Γ(u) is the set of the nearest neighbors of node u and N(w) is the number of the nearest and the next nearest neighbors of node w.Local centrality measure is likely to be more effective to identify influential nodes than degree centrality measure as it utilizes more information, while it has much lower computational complexity than the betweenness and closeness centralities.

**LeaderRank:**

the LeaderRank algorithm to **identify influential spreaders in directed networks**, which is a simple variant of PageRank, namely a so-called ground node connected with every other node by a bidirectional link is introduced into the original network, and then the standard random walk process is applied to dig out influential spreaders.

**Formula:**





**Drawbacks:**

* the ability to find out more influential spreaders;
* the higher tolerance to noisy data;
* the higher robustness to intentional attacks.

**Discussion and Inference**

* Overall Influential node in the graph : **9711200.0. Cocitation strength:** 561, 2681 cocited papers.
* Global metrics and metrics involving paths of the node have higher time complexity, while metrics focusing on local neighbors require low computational resources.
* Paper with more cocited papers tends to have higher degree centrality and tends to be the popular node. This is explained by the density of neighbors and co-citation emphasizes the connectivity between the papers. Papers with more co-cited papers have larger co-citation strength with its pair also.

**References**

* Liu et al., A graph-traversal approach to identify influential nodes in a network, Patterns (2021), <https://doi.org/10.1016/j.patter.2021.100321>
* [Hubs and Authorities (stanford.edu)](https://nlp.stanford.edu/IR-book/html/htmledition/hubs-and-authorities-1.html)
* [Ranking pages: Hubs and Authorities - Safecont](https://safecont.com/en/ranking-urls-hubs-authorities/)
* [Microsoft PowerPoint - Link\_Analysis.pptx (ncsu.edu)](https://www.csc2.ncsu.edu/faculty/nfsamato/practical-graph-mining-with-R/slides/pdf/Link_Analysis.pdf)
* [CentiServer - Centrality - Eccentricity Centrality](https://www.centiserver.org/centrality/Eccentricity_Centrality/#:~:text=Once%20this%20path%20with%20length%20dist%20%28v%2CK%29%20is,means%20that%20all%20other%20nodes%20are%20in%20proximity.)
* [Katz Centrality (Centrality Measure) - Tutorialspoint.dev](https://tutorialspoint.dev/data-structure/graph-data-structure/katz-centrality-centrality-measure#:~:text=Katz%20Centrality%20%28Centrality%20Measure%29%20In%20graph%20theory%2C%20the,an%20actor%20%28or%20node%29%20within%20a%20social%20network.)
* [A graph-traversal approach to identify influential nodes in a network: Patterns (cell.com)](https://www.cell.com/patterns/fulltext/S2666-3899(21)00169-0?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2666389921001690%3Fshowall%3Dtrue)
* [Identifying influential nodes in social networks: A voting approach - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0960077921006639)
* [The H-index of a network node and its relation to degree and coreness | Nature Communications](https://www.nature.com/articles/ncomms10168#Sec2)
* [Identifying and ranking influential spreaders in complex networks with consideration of spreading probability - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0378437116305581#:~:text=A%20hybrid%20degree%20centrality%20is%20proposed%20to%20evaluate,integrates%20degree%20and%20local%20centrality%20with%20spreading%20probability.)
* [Node-weighted centrality: a new way of centrality hybridization | Computational Social Networks | Full Text (springeropen.com)](https://computationalsocialnetworks.springeropen.com/articles/10.1186/s40649-020-00081-w)
* [Understanding the influence of all nodes in a network | Scientific Reports (nature.com)](https://www.nature.com/articles/srep08665/)
* [CentiServer - Centrality - Semi Local Centrality](https://www.centiserver.org/centrality/Semi_Local_Centrality/#:~:text=Semi%20Local%20Centrality%20Definition%20The%20local%20centrality%20C,and%20the%20next%20nearest%20neighbors%20of%20node%20w.)
* [CentiServer - Centrality - LeaderRank](https://www.centiserver.org/?q1=centrality&q2=LeaderRank)
* [Twitterrank: Finding topic-sensitive influential Twitterers (smu.edu.sg)](https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1503&context=sis_research)