Influence Maximization In Social Network By Using Centrality Measures

N.Sree Aravind Bhaskar | AP19110010196
Sai Vara Nitya Kotta | AP19110010239
Gowtham Nalluri | AP19110010254

Abstract:

The report is the long description of the work we have done on the problem statement of 'Finding Influential Nodes and its maximization'. We as a group have gone through a few research papers individually and mentioned the idea of the papers in a few lines in the section of related works. We implemented a paper with the title 'Identifying vital nodes from local and global perspectives in complex networks'. This report on the topic starts with Introduction where the problem statement and the terms involved are explained. Applications as the answer to where the problem statement is useful in the real world is written in one of the sections of the report. Different centrality measures are covered which is the basis for the topic. We ended the report with the Results from our implementation. For ease of understanding of the results we added the tables required and plotted graphs. Experimented through Local and Global Centrality (LGC), Closeness (CNC), PageRank (PRC), Eigenvector (EVC), Global and Local Structures (GLS), Global Structural Model (GSM) methods in real-world networks of various sizes. Our experiments have shown that LGC outperformed many of the compared methods.

Introduction:

Nodes are the building blocks that make up a network. They can represent anything – they could be people, computers, sensors, or anything else. In order to create a network, you need nodes and you should have some of them be influential because they serve as the center of a traffic pattern.
Influential nodes are the nodes of a network that possess high degrees and play an important role in the network's structure. Influential nodes are often represented by large circles on graph illustrations, if they can be determined from the data. In a social media or an information networking site, which typically have many users and connections between them, some users will stand out due to their activity level or number of connections to other highly active users. These are called influential nodes. Influential nodes can have many followers and friends, as well as people who mention them in posts or share their posts with others. They may also be popular because they post content frequently or post content that others find compelling such as funny videos and memes. In a social network, an influential node may be more active, have more connections or have more friends online than an ordinary member. Achieving many connections also makes it easy for some people to spread information by forwarding it to their connections. If people tend to forward messages without checking their origins, they may forward misleading information or miss important information that is true but not related to them.
Influential nodes would be more vulnerable than other nodes because they are essential for the smooth flow of information and resources to travel across the network. If that node is disrupted or the data is somehow corrupted at that point, then there will be problems for the whole network. This is demonstrated by major events in our technological history such as the Y2K bug, the Syrian Electronic Army attack on Twitter, and even the recent Sony hack. A node that is

influential is one that facilitates or enriches network operations. It can be used as a point of attack or defense by both beneficial and malicious agents.

Applications of Influential Nodes:

1)Identifying influential nodes in complex networks plays an important role in understanding the dynamics of infection in each outbreak of infectious disease. In the outbreak of most infectious diseases, the activity of some major nodes can cause the rapid transmission of the disease in the population. Identifying and rapidly isolating these influential nodes can effectively prevent the transmission of the disease. A recent example is COVOID-19.

2)Publishing Ads for global reach for many organizations and influencers on social media create a great impact on people by advertising.Not surprisingly, more and more customers have come to trust influencers, and this marketing method has become mainstream and widely used.

3)Disease Modeling by addressing the issue of efficiently discovering influential nodes in social networks according to the Vulnerability / Infection / Vulnerability model.

4)Viral Marketing is a business strategy that uses existing social networks to promote their own products which indeed refers to how consumers spread the information about a product with different people.

5)Helps us in identifying influential rumorists which is very important to prevent and control the spread of rumors. A rumor begun at a random node of the Twitter network reaches to 45.6 million on average of the total of 51.2 million members within only eight rounds of communication.

6)Opinion Monitoring.A good example of a company that intelligently monitors online mentions and comments is British Airways. We see that the company answering to customer requests on Facebook directing the customers to the right communication channels.

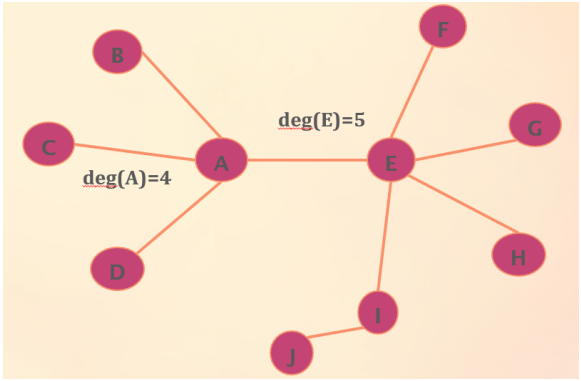
7)Finding Social Leaders by analyzing user interactions which can help determine the flow of impact within the network and provide important insights to leaders within the network.

8)There will be hikes in sales on special products. According to Newsweek, "The Kate Effect may be worth £1 billion to the UK fashion industry."(Kate is one of the people that has many friends on a social network).The Duchess of Cambridge has had an electric effect on fashion sales which is famously known as “The Duchess Effect”.

A centrality measure is a characteristic that measures how central an object is to other objects. This can either be a node with its neighbors in a graph, or the importance of nodes in a network. They are a vital tool for understanding networks. It is often used to help find the most important nodes in a network and why they are important. Centrality measures allow for comparison of different networks, such as graphs or complex networks, on equal footing. Measurements can be split into local (within network) and global (between networks) concepts of centrality. In graphs, a central node is a node with high degree. In social networks, centrality measures quantify the location of individuals within the network. Since nodes with high centrality have more links, information transmitted on those links will spread faster to other nodes, increasing the likelihood that it will reach others with high centrality. If a person has a high level of centrality in their social network then they will act as an information hub, helping to increase the diffusion rate even more.
Let's look at some social network analysis measures:

Degree Centrality:
Individual nodes are the centre of degree centrality, which simply counts the number of edges a node has. It is used for finding individuals who are likely to hold most individuals or information who can quickly connect with the wider network.

$$C_{Di} = \frac{\sum_{j=1}^{n} a_{ij}}{n-1}$$



Betweenness Centrality:
Betweenness is defined as the number of shortest paths between two nodes that pass through a given node. Betweenness centrality for a vertex v is defined as

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

For a given node v, the number of shortest paths between nodes i and j that pass through v is calculated, and divide by all shortest paths between nodes i and j. A node that is having high betweenness is vital to the communication of the network.

Closeness Centrality:

Closeness centrality is a parameter of closeness, used to produce a proximity measure. Closeness centrality measures the degree to which any one object is near another object in an undirected relationship. It does not take the direction of the relationship into account, but rather just how close an object is from its nearest neighbor in that relationship. A simple example of applications of closeness centrality as a measure of proximity is to ask whether it is better to be near the person who reported receiving the worst score on a test, or whether it is better to be at all points in between. In this example, the question can be divided into two parts: (1) which person was near the person with the bad score and (2) how near?
It can describe how far and close vertices are.

$$CC(v) = \frac{1}{\sum_{t \in V \setminus v} dG(v,t)} \times (n-1)$$

Clustering Coefficient Centrality:

Clustering coefficient centrality is also known as transitivity of a node defines the number of closed triplets in the neighborhood of the node over the total number of triplets in the neighborhood.

$$C_i = \frac{\text{number of triangles connected to node } i}{\text{number of triples centered around node } i},$$

Eigenvector Centrality:

Eigenvector looks at the quality of the links of the nodes to determine the popularity. Eigenvector centrality of a node is proportional to the sum of eigenvector centrality of all the nodes directly connected to it.

$$Ax = \lambda x$$

A node with high eigenvector centrality is connected to other nodes with high eigenvector centrality.

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t}\, x_t$$

PageRank Centrality:

PageRank is a tool for measuring the relative importance of pages on the World Wide Web. It is an open-source technology and was developed by Larry in 1996. It uses Internet popularity data, such as links from other web sites, to determine how important a given page is.

The link analysis algorithm used by Google assigns a numerical weighting factor (known as PageRank) to web pages that determines how they rank in search engine results pages.

1.Number of links - Pages with more in-links rank higher, than a page with fewer in-links.

2.Link Quality - A link from an important page is worth more than many links from relatively unknown sites.

3.Link Context - The text in and around links relates to the page they point at. Ranking boosts on text styles.

PageRank (*PR*) of page u is given by the summation of the PR of all pages in the set of all pages linking to page *u* ($v \in B_u$), divided by the number of *L(v)* of links from page *v.*

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

Katz Centrality:

Katz Centrality is a theory that helps explain why certain types of people seem to be more prominent in certain structures. It suggests that people will connect with others whose position they value, or that they admire. Individuals who are in these influential roles are likely to have an impact on the direction of their culture or the political structure and having them in these positions of power or importance may make those who view them positively more likely to form relationships. This is because those connected with this person may also gain some importance in the social structure, due to their relationship and connection with someone who holds a valued position.

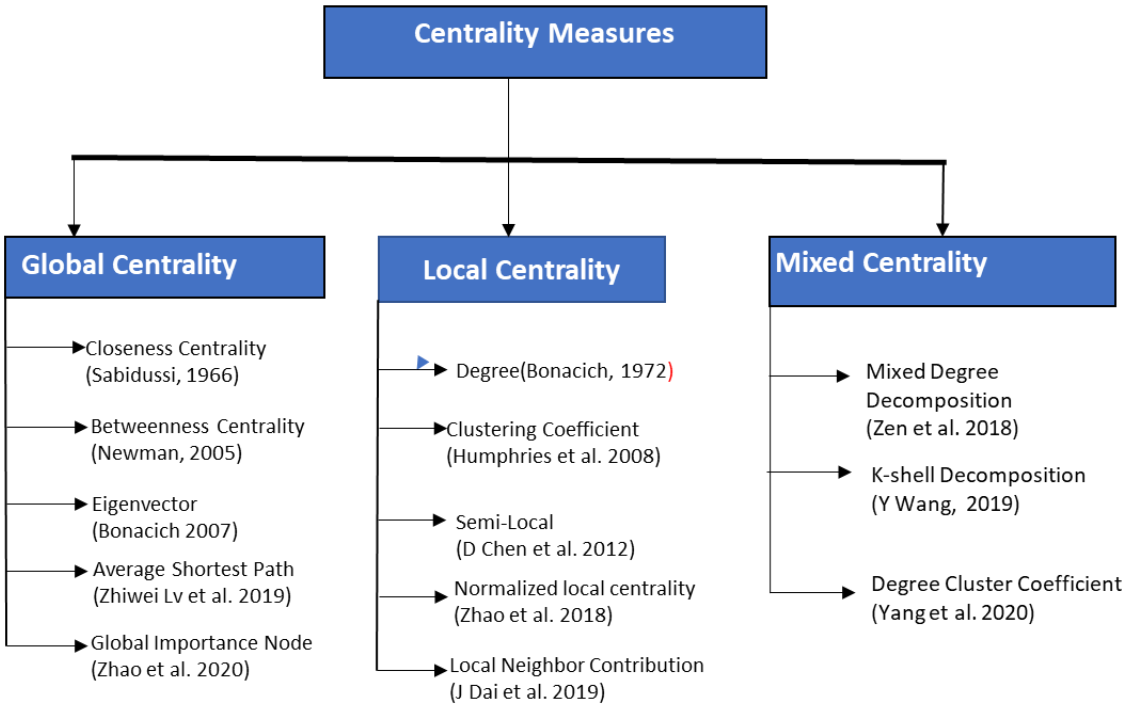
For example, hierarchical structures are prevalent in governments (monarchy), religions (religion), business environments (businesses) and even family relationships and marriages (husband/wife). These groups all develop their own culture based on what they need to survive and thrive.

Katz centrality for node *i* is:

$$x_i = \alpha \sum_j A_{ij} x_j + \beta,$$

where A is the adjacency matrix of the graph G with eigenvalues λ whereas the parameter β controls the initial centrality and $\alpha < 1/\lambda_{max}$.

Categories of Centrality Measures



Relative Average Shortest Path:

The average number of steps along the shortest paths for all possible pairs of network nodes is defined as average shortest-path length in network topology. It is a measure of the efficiency of information and mass transport on a network.

Relative change in ASP :-

$$AC[k] = \frac{|ASP[G'_k] - ASP[G]|}{ASP[G]}$$

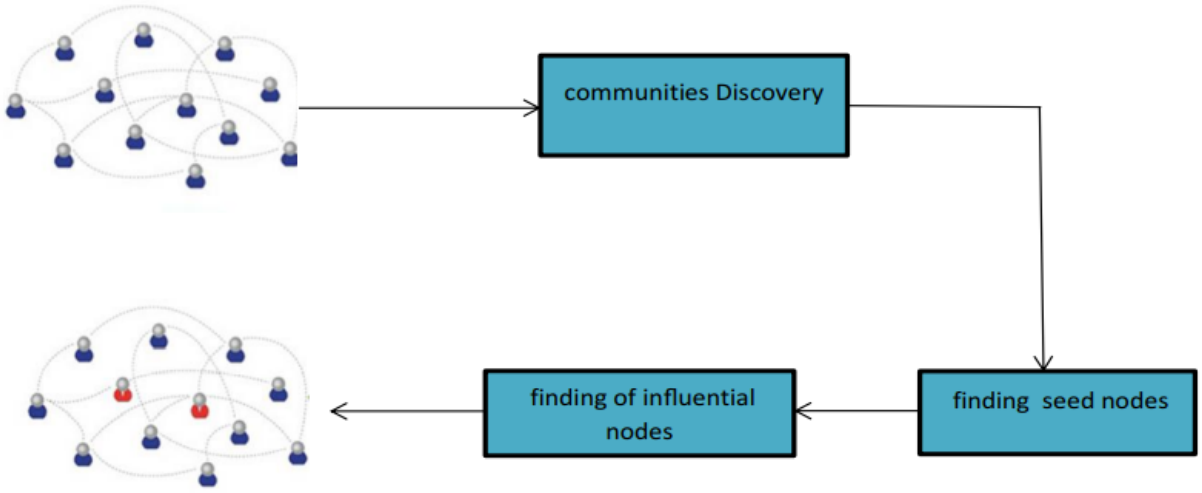
Where ASP[G] is the average shortest path and is a graph, the node k is removed from graph G.

$$ASP[G] = \frac{\sum_{i \neq j \in G} d_{ij}}{N(N-1)}$$

Methods:

Related Works:

1)Influence maximization is the process through which a person or group seeks to maximize the number of nodes that they can control in order to use those nodes for their own purposes. The paper will explore different algorithms for extracting influential nodes in a social network for community discovery. We then divide the social network into communities using a proposed algorithm called Community Discovery Based on Nodes Neighbor (CDBNN) algorithm. We also provide empirical evidence that these algorithms are effective in finding highly influential individuals, by comparing them with traditional metrics such as centrality measures.
The graph is the most significant data structure representing social networks in a form G(V, E) where (V) represents the vertices and (E) represents the edges. Influence is the most important thing in social networks as it is necessary for network analysis applications. Social influence is the relationship existing between two entities in a particular work. The first entity is called the influencer and the second entity is called the influence, so the first entity affects the second entity. The LT model and the IC model are the two basic models for disseminating an idea across the network.
The method proposed is influence maximization based on centrality measures and structural aspects(IMCS). It consists of three parts: community discovery, finding seed nodes and finding influential nodes. Firstly, the social network was divided into communities using CDBNN. Then in the second part the nodes that are playing the role of candidate nodes(seed) are extracted. Finally, from the active nodes the influential nodes are extracted.



2) The paper starts with the practical applications of finding influential nodes in a social network. Through this paper we get to know the effects of using a community finding approach besides the often proposed greedy approaches to find the best k nodes to activate, so that the diffusion of activation will spread to more nodes.
In the section of introduction, we get to know what kind of behavior an influenced node takes with the help of an example. We also get to know that nodes of similar types are often found

grouped together in localized areas of the network. And to identify such groups community finding algorithms are introduced.
In this paper, community finding is used to investigate how localized influence is and how we can use communities to find influential nodes.
Later, the paper mentions about the people or aspects where finding nodes that are highly influential is of interest.
In this paper the greedy approach by Kempe, et al is considered because it allows the number of nodes to be specified, which is important for comparisons.
The purpose of this paper is to study the process of influence with regard to the network communities which are defined by the structure of the links. The paper has 2 goals for the method implementation:
(i)To select the initial set of nodes such that the final set covers as many communities as possible.
(ii)To select the initial set of nodes to maximize the size of the final set.
The general method for finding community-based influential nodes consists of finding communities in the network and then selecting one node from each of the communities to activate.
Two algorithms of spectral clustering also known as normalized cut and implementation of the agglomerative method of Tang, et al. are used.
A method to apply an influence maximization algorithm to each one of the communities is used to decide how to select the nodes within the community after deciding on the community finding algorithms. Here, a faster way of selecting the node in each community that has the highest degree (the most number of links attached to it) is used. The complexity for the methods implemented is mentioned. The problem of influence maximization has been shown to be np hard. The greedy algorithm is tractable by still very slow. O(k.n.x.s).

3) A novel approach (Structure-based Identification Method, SIM) to perceive the influential nodes in industrial networks is proposed based on the community structure, which is going past using community metrics. The SIM approach extracts the weakly linked components, that are much more likely to live on after the critical nodes are attacked withinside the community. Evaluation outcomes show that the SIM approach obtains higher outcomes than today's strategies to identify influential nodes in real-global commercial networks and has a true prospect to be carried out in industrial application.Here in this paper they designed a network influential node identity approach, based on efficient extraction of vulnerable components. Compared to associated work, our approach (SIM) does now no longer depend upon rating nodes in step with their significance in the community. They evaluated the approach on 3 industrial networks with one-of-a-kind parameters. Their approach shows powerful overall performance on all the networks in phrases of massive components and identifying the influential nodes. Our approach can offer greater and accurate guidance for the design and operation of industrial networks, which include which influential nodes want to be allocated greater assisting resources

to ensure greater-solid operation of the network. SIM presents a trade-off among rapid identity method and effectiveness design, and thus, makes a vital contribution to future industrial network robustness estimation. They even gave the ideas of implementing in this arena as following:
(i) Intellectual identification of weak components as in the above method many components are getting repeated so whenever some sequences are identified with low probabilities in the final network they should be discarded in the early stages.
(ii) Adaptively divide k partitions. In the above SIM method, the number of influential nodes are divided into 2 parts but when the size of the network enlarges, then we should think of how to split k into a certain number of parts because when we are taking local and global networks into account the split should be considerable.

4) The paper proposes a two level approach, designed based on the Suspected-Infected (SI) epidemic model for maximizing the influence spread.
The paper starts with the mention of two categories of social networking mining - the study of structural characteristic and content analysis. It mentions a significant problem in the context of a social network, finding the most influential entities within the network. An application of this problem - viral marketing is highlighted. The type of problem i.e the problem being proved to be NP-hard is specified. The paper also proposes a multithreading approach for implementation of algorithms for the proposed SI model which adds in elevating the performance of the proposed approach in terms of influence spread per second. The paper further continues with the introduction in which viral marketing being an effective marketing strategy and an application of maximizing influence spread has been discussed. It is explained here that the companies started targeting key individuals called influencers, naming this indirect form of marketing as influencer marketing. Social media influencers are termed as the entities in the social network. The problem of influence spread is discussed in this section as well. The goal of the problem is to maximize the spread of information to a large population. The problem of influence maximization is redefined as the problem of forming an objective function for selecting appropriate target nodes in a social network such that it maximizes the influence spread. The target nodes will further propagate the influence to their connected nodes. And this is a helpful design marketing strategy. The major issue of how to improve the diffusion for the given seed selection is also mentioned in this section. The paper mentions related work. It includes the Kemp et al. proposed greedy approach for finding K influential nodes out of all the existing nodes. Cost-effective lazy forward selection approach is specified which helps in increasing the performance of greedy algorithms. In the next section, the problem is discussed further deeply.
The different models and frameworks defined by different researchers to obtain an optimal solution for the problem are discussed. There is also a mention about the classic progressive models - Independent Cascade model, Linear threshold model, Triggering model. Greedy algorithm proposed by Kempre et al. is discussed. And further, the proposed model is explained clearly. The approach is practically demonstrated by applying it on a dataset.

Working of the algorithm and the results are discussed in the last section.
Through the results, we get to know the SI model is based on the incremental approach where the spread is cumulative. The spanning tree enables the SI approach to find the best possible longest path which helps in increasing the influence speed. The comparison of performance gain for influence spread of different algorithms is depicted by applying all of them on a single data set.

Results:

There are many algorithms which use local and global structure of a network to find the influential nodes. The algorithm which we have implemented uses both topological aspects of the node network i.e both local as well as global influence named as Global Structure Model(GSM). To evaluate the influential nodes we have used various centrality measures in local structures like degree centrality, pagerank etc and global structures like betweenness centrality, closeness centrality etc. Though their performance is good, they have some limitations. For example, local structure based methods lose some of the global information And global structure models are too complex to find out the influential nodes, especially in huge networks. So, taking all these points into consideration we are implementing an algorithm which considers both local and global structure parameters to find the influential nodes. This algorithm is called LGC. LGC primarily follows three definitions which are:

(i)**Local Influence**: Local influence of node v_i represents the local information of graph G (degree of each node divided by total number of nodes in the graph). This is represented as follows:-

$LI_{v(i)} = d(v_i)/n$
where $d(v_i)$ represents the degree of the node i and n is the total number of nodes.

(ii)**Global Influence:** When we find an influential nod, it not only influences itself but also affects the neighboring nodes. The nodes with higher degree are likely to have greater influence but we cannot ignore shortest distance between the nodes as it is inversely proportional to the influence of the node. Therefore, we consider the sum of the degree of node j and shortest distance between nodes The global influence can be represented as follows:-

$GI_{v(i)} = \mathbf{\Sigma} (\sqrt{d(v_j)}+\alpha)/ d_{ij}$ (where $i \neq j$)
α is the tuning parameter which controls the influence of degree in a network which ranges between 0 and 1. We are taking the square root of it in order to normalize the influence of node j.

(iii)**Total Influence:** LGC is a combination of above two definitions which are local influence and global influence. LGC of a node i can be defined as follows:-

$LGC_{v(i)} = LI_{v(i)} \times GI_{v(i)}$

The proposed LGC can also be expressed as:

$LGC_{v(i)} = d(v_i)/n \times \mathbf{\Sigma} (\sqrt{d(v_j)}+\alpha)/ d_{ij}$

Computational Complexity of LGC:
The proposed LGC has two main components. In the initial step, the time complexity of the node's global influence is calculated and we used Dijkstra's algorithm to calculate the shortest path distance, and its complexity is $O(n^2)$. In the second step, the time complexity is $O(n)$. Hence, the total computational complexity of LGC is $O(n^2)$.

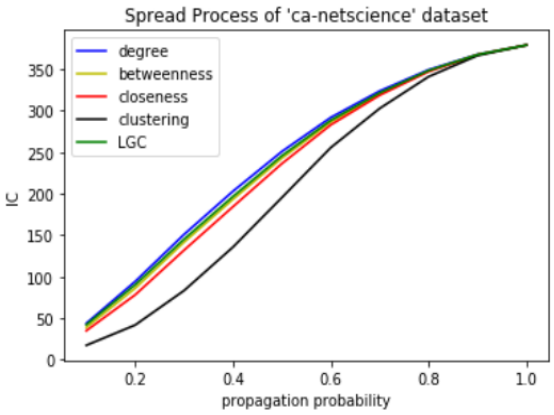
Now that we have calculated LGC and found the influential nodes, we now focus on maximizing the influence of the nodes. Influence Maximization (IM) is an area of network analysis with many applications, from viral marketing to disease models to public health interventions. IM is the task of finding a small subset of the nodes in your network and ensuring that the resulting "impact" propagated from that subset reaches the maximum number of nodes in your network. "Impact" means anything that can be propagated through connected peers in the network. B. Acceptance of information, behavior, illness or product. Here we have tried to implement the maximization using the Independent Cascade model. The IM algorithm solves a specific propagation or propagation process optimization problem. Therefore, you must first specify a function that simulates the propagation of a particular seed set across your network. The popular independent cascade model is used to simulate the distribution of effects, but there are many other models to choose from. The IC () function that describes the diffusion process is shown below. Calculate the expected variability of a particular seed set by averaging a large number (mc) Monte Carlo simulations. The loop outside the IC () function iterates over each of these simulations and stores each calculated spread in a spread list. The average of each of these entries is then a consistent and unbiased estimator of the expected spread of the seed set S and is returned as the function output. Within each Monte Carlo iteration, it simulates the propagation of its impact over the network over time. Check if a different "duration" occurs within each iteration of the while loop and a new node was activated in the previous time step. If the new node is not activated, the independent cascading process ends, saves the total spread, and then the function proceeds to the next simulation.This is the number of nodes that were finally activated.

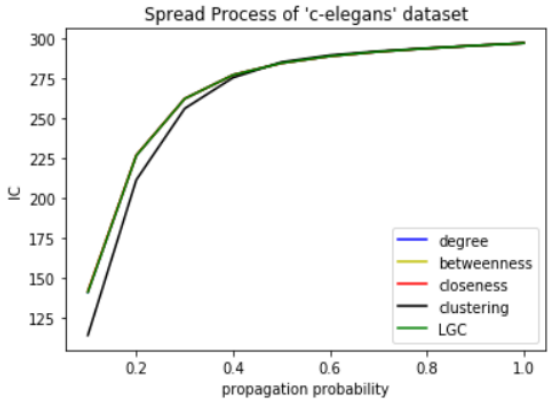
Explanation of LGC:

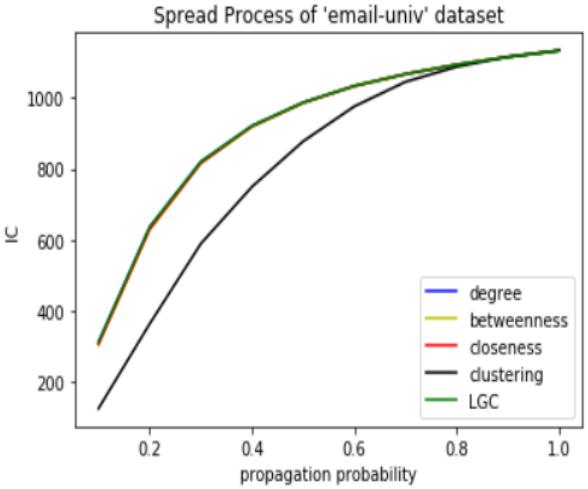
Sno	Dataset Name	Nodes	Edges	Max.Degree	Avg.Degree	Avg.Clustering	Diameter
1	ca-netscience	379	914	34	4.8	0.741	17
2	c-elegans	297	2148	83	15.2	0.292	5
3	email-univ	1133	5451	71	9	0.22	8
4	web-polblogs	643	2280	139	12.8	0.232	10
5	bio-c elegans	453	2025	237	8.94		

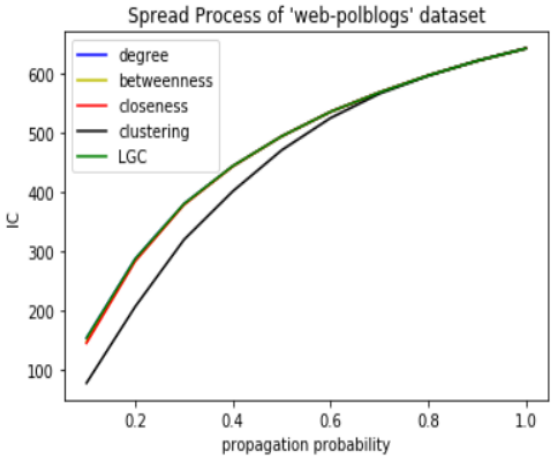
Datasets	**Degree Centrality**	**Betweenness**	**Closeness**	**Clustering**	**LGC**
ca-netscience	250.887	242.855	236.255	195.373	245.731
c-elegans	284.517	284.517	284.517	285.175	284.517
email-univ	986.932	986.13	986.145	877.962	986.932
web-polblogs	494.355	494.235	494.255	470.841	494.355

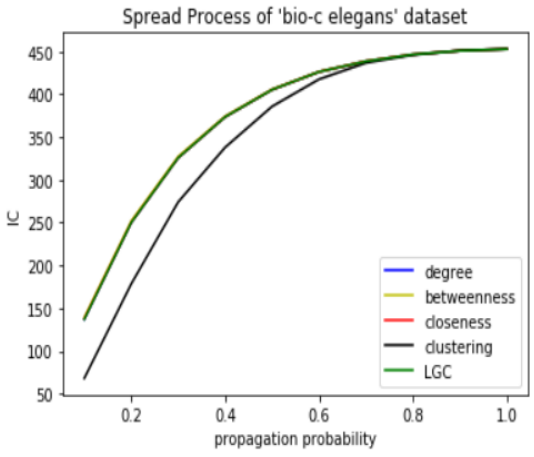
bio-c elegans	405.455	405.837	405.455	385.868	405.455











Conclusion:

We explored the issue of identifying critical nodes in a complex network from a local and global perspective. In this reclassification, we suggested a new LGC algorithm that uses local and global information of the node on the network. LGC is practically applied to real networks of various sizes using SIR and Kendall measurements for evaluation. Experimental results are based on these real networks. We have shown that the proposed LGC is superior to existing methods.

However, LGC's current design still has some operational limitations that need to be effectively addressed. For example, the design proposed by LGC is limited to unweighted undirected networks. We will continue to get used to overcoming these challenges. Networks such as weighted directed, weighted undirected, bipartite graph, etc.

References:

1) Jerry Scripps1 1Grand Valley State University, Allendale, MI 49401, USA scrippsj@gvsu.edu
2) Wang, T.; Zeng, P.; Zhao, J.; Liu, X.; Zhang, B. Identification of Influential Nodes in Industrial Networks Based on Structure Analysis. Symmetry 2022, 14, 211. https://doi.org/10.3390/sym14020211
3) www.elsevier.com/locate/eswa
4) Zainab Naseem Attuah1 *, Firas Sabar Miften1 , Evan Abdulkareem Huzan1 1 University of Thi-Qar, College of Education for Pure Science, Iraq.
5) School of Computer and Information Science, Southwest University, Chongqing, China2 Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong, SAR, China. https://doi.org/10.3389/fphy.2021.766615
6) https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8259501
7) https://doi.org/10.1155/2021/8403738
8) https://rdcu.be/cM6oo
9) https://appliednetsci.springeropen.com/articles/10.1007/s41109-017-0047-y
10) https://www.hindawi.com/journals/ddns/2019/8938195/