

IDENTIFYING INFLUENTIAL SPREADERS IN COMPLEX NETWORKS

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE ENGINEERING

Submitted By:

PULKIT SHARMA

(2K18/CSE/14)

Under the supervision of

MR. SANJAY KUMAR
(ASST. PROFESSOR CSE)

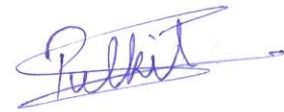


DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

JUNE, 2020

CANDIDATE'S DECLARATION

I, Pulkit Sharma (2K18/CSE/12) student of M.Tech Computer Science Engineering, hereby declare that the project Dissertation titled “Identifying Influential Spreaders in Complex Networks” which is submitted by me to the Department of Information Technology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any degree, Diploma Associateship, Fellowship or other similar title or recognition.



Place: Delhi

Pulkit Sharma

Date: 8/12/2020

CERTIFICATE

I hereby certify that the Project Dissertation titled “Identifying Influential Spreaders in Complex Networks” which is submitted by Pulkit Sharma, Roll No 2K18/CSE/12 Computer Science Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 09/12/2020



MR. SANJAY KUMAR

SUPERVISOR

ACKNOWLEDGEMENT

I express my gratitude to my major project guide MR. SANJAY KUMAR, Cse Dept., Delhi Technological University, for the unrelenting support and guidance he provided in making this major project. It is my pleasure to record my sincere thanks to my respected guide for his invaluable insight and constructive criticism without which the project would not have shaped as it has.

I humbly extend my words of gratitude to other faculty members of this department for providing their valuable help and time whenever it was required.



PULKIT SHARMA

Roll No. 2K18/CSE/12

M.Tech (Computer Science Engineering)

E-mail: - s.pulkit295@gmail.com

ABSTRACT

One thing that we can't disregard in this day and age is incorporation of social media in everyday life. Because of expansion in reach of information to people in developing nations, consistently huge number of individuals access social network, a lot of them do that for the first time. Impact of social media is so much that it can affect one's point of view towards an issue. Since each social media brags about its social reach, client is the main element of the social network. A lot of research is being done on the problem of user classification in any social network. Identifying the most influential nodes is a significant issue in controlling the spreading cycle in complex networks. Centrality measures are utilized to rank the network nodes relying on different properties captured by that centrality. Researchers have been pursuing for quite a long time to outline a widespread strategy for user classification in a social network. In this project, we have attempted to devise a new technique, Improved Gravity centrality to group nodes in a complex social network utilizing the mix of network structure of the diagram and Gravity centrality, using H-index as the mass of the network node. We have thought about the aftereffect of our proposed strategy with different existing models in social organization writing on various constant datasets with the assistance of SIR epidemic model (Suspected-Infected-Recovered). When contrasted with existing techniques for node ranking, our outcomes provide quite an improvement.

Keywords: Social Network. Influential Nodes. Influence Maximization. Node Centrality. Gravity Centrality. H-index. SIR model.

CONTENTS

<u>CANDIDATE’S DECLARATION</u>	ii
<u>CERTIFICATE</u>	iii
<u>ACKNOWLEDGEMENT</u>	iv
<u>ABSTRACT</u>	v
<u>CONTENTS</u>	vi
<u>List of Figures</u>	viii
<u>List of Tables</u>	ix
<u>List of Formulas</u>	x
<u>CHAPTER 1 INTRODUCTION</u>	1
1.1 <u>DIFFERENT CENTRALITY MEASURES</u>	2
1.2 <u>PRESENCE OF COMMUNITIES IN NETWORKS</u>	5
1.3 <u>INFORMATION DIFFUSION MODELS</u>	7
<u>CHAPTER 2 RELATED WORK</u>	8
<u>CHAPTER 3 METHODOLOGY</u>	10
3.1 <u>H-INDEX</u>	10
3.2 <u>GRAVITY CENTRALITY</u>	10
3.3 <u>SELECTION OF SPREADERS</u>	11
3.4 <u>SIR MODEL</u>	13
<u>CHAPTER 4 DATASETS AND PERFORMANCE METRICS</u>	14
4.1 <u>DATASETS</u>	14
4.2 <u>PERFORMACE METRICS</u>	14
4.2.1 <u>INFECTED SCALE</u>	14
4.2.2 <u>FINAL INFECTED SCALE</u>	15

4.2.3	SHORTEST PATH LENGTH	15
4.2.4	KENDALL TAU	15
	<u>CHAPTER 5 RESULTS</u>	17
	<u>CHAPTER 6 CONCLUSION AND FUTURE WORK</u>	28
	<u>References</u>	Error! Bookmark not defined.

List of Figures

<u>Figure 1.1</u>	K-Shell Algorithm	4
<u>Figure 3.2</u>	Different Information Diffusion Models	6
<u>Figure 3.3</u>	Flowchart for the steps in our methodology	12
<u>Figure 5.4</u>	Infected Scale ($F(t)$) vs Time for Facebook	17
<u>Figure 5.5</u>	Infected Scale ($F(t)$) vs Time for Gnutella	19
<u>Figure 5.6</u>	Infected Scale ($F(t)$) vs Time for PGP	20
<u>Figure 5.7</u>	Infected Scale ($F(t)$) vs Time for Jazz	22
<u>Figure 5.8</u>	Final Infected Scale ($F(t_c)$) vs Spreaders for Facebook	24
<u>Figure 5.9</u>	Final Infected Scale ($F(t_c)$) vs Spreaders for Gnutella	24
<u>Figure 5.10</u>	Final Infected Scale ($F(t_c)$) vs Spreaders for PGP	25
<u>Figure 5.11</u>	Final Infected Scale ($F(t_c)$) vs Spreaders for Jazz	26
<u>Figure 5.12</u>	Shortest Path (L_s) vs Spreaders for different datasets	26

List of Tables

<u>Table 1.1</u>	Explanations of various Symbols used in the paper	1
<u>Table 4.2:</u>	Overview of the used Network Datasets	12

List of Formulas

<u>Formula 1.1:</u> Eigen Vector Centrality	2
<u>Formula 1.2:</u> Betweenness Centrality	3
<u>Formula 1.3:</u> Closeness	3
<u>Formula 1.4:</u> Clustering Coefficient	Error! Bookmark not defined.
<u>Formula 1.5:</u> Coreness	4
<u>Formula 1.6:</u> Extended Coreness	4
<u>Formula 1.7:</u> Page Rank	5
<u>Formula 3.8:</u> H-Index	9
<u>Formula 3.9:</u> Extended H-Index	9
<u>Formula 3.10:</u> Gravity Centrality	10
<u>Formula 3.11:</u> Extended Gravity Centrality	10
<u>Formula 3.12:</u> Improved Gravity Centrality	10
<u>Formula 3.13:</u> Extended Improved Gravity Centrality	10
<u>Formula 4.14:</u> Infected Scale	13
<u>Formula 4.15:</u> Final Infected Scale	13
<u>Formula 4.16:</u> Shortest Path Length	14
<u>Formula 4.17:</u> Kendall Tau	27

1 INTRODUCTION

The expression "social networks" was coined by Barnes in the year 1954. It is believed that social networks started from email services. With the development of social networks, various new social media platforms are arising nowadays, for example Facebook (2003), YouTube (2005), Twitter (2006), Instagram (2010). There has been a great change in the way individuals get data. Rather than individuals being uninvolved recipients, like in the past, they are now becoming active propagators. Digital revolution everywhere in the world has paved way for an ever increasing number of users to join these social networks making a successful ground for digital marketing, community targeting and information dispersal. Social Networks are represented as graphs $G(V, E)$ where G is the graph, V and E are the sets of its vertices and edges. The users of the social network are represented by vertices and the type of connection between the users is represented by the edges. An asymmetrical sort of relationship will prompt a directed graph, for instance, twitter follower-followed network. Undirected graphs are utilized for those social networks that have even connection between users like on Facebook. Although each node is vital piece of the network, a few nodes have exceptionally serious level of impact in their area and in the network as whole. Determining the node's impact has been of late, the territory of exploration in social networks as top-K nodes have tremendous authority over the spread of data in the network. In numerical terms, discover most notable nodes in the graph that have higher significance than others. This significance relies on the data dissemination ability of that node. In this way we need to discover top nodes that have higher data dissemination capacity in the social network. Ranking of the nodes is done depending upon the various values obtained for them through centrality measures. Then we may grade and sort the nodes by utilizing different methods like number of neighbors (degree), coreness of the node, significance of neighbors of the node and so on.

S.No.	Variable Abbreviation	Description of the variable
1	G	The graph representing the social network
2	V	The set of vertices in a graph, i.e. users in a network
3	E	The set of edges in a graph, i.e. connections in a network
4	$e_{i,j}$ or i,j	Denotes the edge connecting vertices 'i' and 'j'
5	D_i	Degree of vertex 'i'
6	$P_{j,k}$	Shortest path between nodes 'j' and 'k'
7	$eigen_i$	The computed Eigen vector centrality of node 'i'
8	CC_i	Clustering Coefficient of node 'i'
9	C_b	Betweenness centrality of node 'i'
10	C_i	Closeness centrality of node 'i'

11	PR_i	Page Rank centrality of node 'i'
12	ks_i	K-Shell centrality of node 'i'
13	Cr_i	Coreness centrality of node 'i'
14	ECr_i	Extended Coreness of node 'i'
15	h_i	h-index of node 'i'
16	Gr_i	Gravity Centrality
17	EGr_i	Extended Gravity Centrality

TABLE 1: Explanations of various Symbols used in the paper

1.1 DIFFERENT CENTRALITY MEASURES

Degree: Degree of a node (vertex) in a graph tells us regarding the neighbors of that node. In regard to an undirected graph, there is just one kind of degree centrality that is defined, which is given by the number of nodes that are adjacent to this node. Typically we normalize this value to catch the overall picture in the graph. We can say generally that, on the off chance that a node has a high degree centrality, it is highly probable that the node is a well known node. For directed graphs, there are two kinds of degree centralities for each of the nodes, outdegree-centrality (denoting edges beginning at this node) and indegree-centrality (denoting edges finishing at this node). For a straightforward twitter type of network, the number of followers can be considered as indegree and number of individuals followed as outdegree. Degree centrality overlooks locale of the node in the graph, which might not be genuinely mirroring the concerned nodes' impact.

Eigen Vector: Eigenvector centrality (additionally called Eigen centrality) is a proportion of the impact of a node on a network. Eigen vector centrality catches the fundamental idea that was overlooked by degree centrality. Assume an individual on twitter has relatively few followers yet he/she is followed by quite a powerful individual. At that point that individual is probably going to be famous as his/her tweets will reach and influence the masses without any problem. In this manner the number of the connections isn't the only figure that tells us about the significance of the node, but the nature of those connections is also important. Each node is assigned a relative score, keeping in mind that association with a node having high score increases the score of the node more than the associations to the nodes with low score. A high value of eigen vector centrality implies that the node is associated with a node that has a high score and along these lines its significance in the network is more. Mathematically, it can be calculated as:

$$eigen_i = \frac{1}{\lambda} \sum_{e_{i,j} \in E} x_j \quad (1)$$

Here, λ is a constant and is chosen in a way that all the calculated values in the eigenvector are non negative.

Betweenness Centrality: Betweenness of a node is higher if a lot of shortest paths between other nodes are passing through it. For a graph $G(V, E)$ with n vertices, the betweenness centrality $C_b(i, g)$ for vertex 'i' is computed as follows:

For each pair of nodes (k, j):

1. Compute every shortest path between those nodes.
2. Compute the proportion of shortest paths with the current node (node i) as transit.
3. Sum this fraction for each pair of nodes in the graph.

The betweenness centrality can mathematically be represented as:

$$C_b(i, g) = \frac{2}{(n-1)(n-2)} \sum_{k \neq j, i \notin \{j, k\}} \frac{P_i(kj)}{P(kj)} \quad (2)$$

Where, $\frac{P_i(kj)}{P(kj)}$ is the probability that a randomly selected path between the nodes k and j passes through node i.

Closeness Centrality: The value of closeness centrality of a node is high if it is proximal to other nodes in the network, i.e. if it has short distances to other vertices. Closeness centrality is generally positively related to other centrality measures like degree as it assigns higher values to more central nodes, i.e. the ones having shortest-path length.

The closeness centrality is:

$$C_c(i, g) = \frac{(n-1)}{\sum_{i \neq j} d(i, j; g)} \quad (3)$$

Where, $d(i, j; g)$ is the distance between i and j.

Clustering Coefficient: The extent to which cluster formation of nodes is present in the graph is known as clustering coefficient. Alternatively, it can also be defined as measure of how close the graph is to completeness. If the number of neighbors of node 'i' is m, the clustering coefficient can be calculated as:

$$CC_i = \frac{2 * (|e_{j,k \in E, v_j, v_k \in Neighbors_i}|)}{m * (m-1)} \quad (4)$$

K-Shell: Recent studies have shown that a node that has a large value of degree centrality may not automatically be the most predominant node. A node with high value that is positioned at the outskirts of the network is less impactful than a node with same value present in the centre of the network. Hence, this algorithm assigns a higher value to a more central node. To compute the k-shell values in a network we remove all the vertices with remaining degree less than or equal to k, starting with k=1. When all such vertices are removed, we repeat the process by incrementing k by 1. We do this until no vertex is remaining in the graph. The k-shell value of a vertex is equal

to the corresponding value of k upon its removal. The computational complexity of this algorithm is quite low and thus is scalable. However, one shortcoming of this method is that it assigns same value to a large number of vertices which distorts the ranking order.

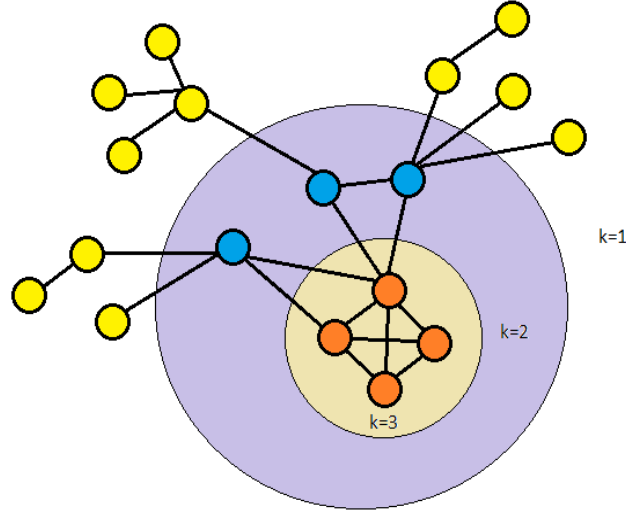


Figure 1: K-Shell algorithm

Coreness: Coreness centrality is an extension of the k -shell centrality. Here we assume that a node with neighbours having high k -shell values is also influential. In comparison with k -shell centrality, it does a better job of classifying a node into a greater range of labels. Therefore we observe a higher precision. This centrality approach considers both the arrangement as well as the degree of the node while counting the k -shell values of its neighbour nodes.

Mathematically, Coreness of a node can be expressed as:

$$Cr_i = \sum_{j \in N_i} ks_j \quad (5)$$

Where, N_i represents the set of neighbour nodes of 'i'.

Extended Coreness: It is a more updated version of Coreness centrality that fuses the node neighbourhood factors with its coreness. It is computed as the summation of coreness values of its neighbourhood nodes, given by the following equation:

$$ECr_i = \sum_{j \in N_i} Cr_j \quad (6)$$

PageRank: It is safe to assume that page rank is perhaps the most famous centrality measure in the area of social media research. Google makes use of this centrality measure for its search

result optimization and ranking. In addition, this algorithm is also used for classifying users in a social network quite predominantly. PageRank value of a node is calculated by taking into consideration both number and quality of links to that node. The algorithm requires heavy computations and resources but gives great results. It is represented by following equation:

$$PR_i = \alpha \sum_{e_{i,j} \in E} \frac{PR_j}{D_j} + D_i \quad (7)$$

H-Index: The degree of a node has an impact on most of the other centrality measures in some manner. A new centrality measure with a fairly good efficiency, called H-index was put forth by Liu which distinguishes the most predominant disseminators by computing the h-indices of the nodes in the network. H-index provides a thought regarding the outstanding propagation ability of a node depending on its neighbors. H-Index of a node i is characterized as the largest value of h where in any event the node has a minimum of h adjacent nodes which have a minimum degree of h . Additionally, the Extended H-index of a node was characterized as the summation of the h-indices of that node's neighbors.

Gravity Centrality: While coming up with the k-shell algorithm, kitsak made a hypothesis that nodes belonging to the same shell will have similar influence and those belonging to a higher value shell will have higher influence. However, this hypothesis has proven to not hold its ground in all cases as indicated by recent researches. It is not uncommon for nodes of a shell to exhibit different impact and the approach may entirely crash in some networks that do not have a kernel structure.

Taking inspiration from the gravity equation, Ma et al. put forth gravity centrality where they considered the k-shell estimation of every node to be its mass and the shortest path connecting two nodes their distance to locate the high impact diffusers in networks. They recommend that the communication impact between two nodes is in directly correlation to the result of their k-shell values and is in inverse correlation to the square of the distance between them.

1.2 PRESENCE OF COMMUNITIES IN NETWORKS

Connections between constituents of complex frameworks can be thought of as networks. The elements of the framework are depicted as nodes and connections between them are depicted as edges. At a casual glance, this informal community may look like having complex nature, yet they show some type of comprehensible structure at moderate level. Mathematically it is feasible to isolate the collection of nodes that have strong connections with one another and weak connections with the remainder of the network. Such interconnected node groups are regularly described as communities and can be seen in various network frameworks. A better way to understand a relationship of this type is to think of it as your Facebook friends circle, where there

is a large overlap between the interests, dislikes, thoughts and perspectives of you and your friends. Subsequently there is normally a higher dependability on each other amongst you and your friends, so their thoughts on a subject can reshape your viewpoint as well. This idea is utilized by numerous organizations in community choice during promotions and marketing. They typically target most significant node in a community (like community leader/celebrated figure). Subsequently community identification has developed into a major and exceptionally pertinent issue in the field of social network research. Likewise, it surmises unique connections between nodes that may not be effectively open from direct observational tests. Given their significance, numerous community detection algorithms have been created through tools and techniques from different fields of study, for example, sociology, statistical physics, applied mathematics, computer science, biology, and public health. They generally cannot be characterized objectively, rather they have multiple ways by which they can be broken down and analyzed. Accordingly the pertinence of any community detection algorithm is also in correlation with graph structure in some sense.

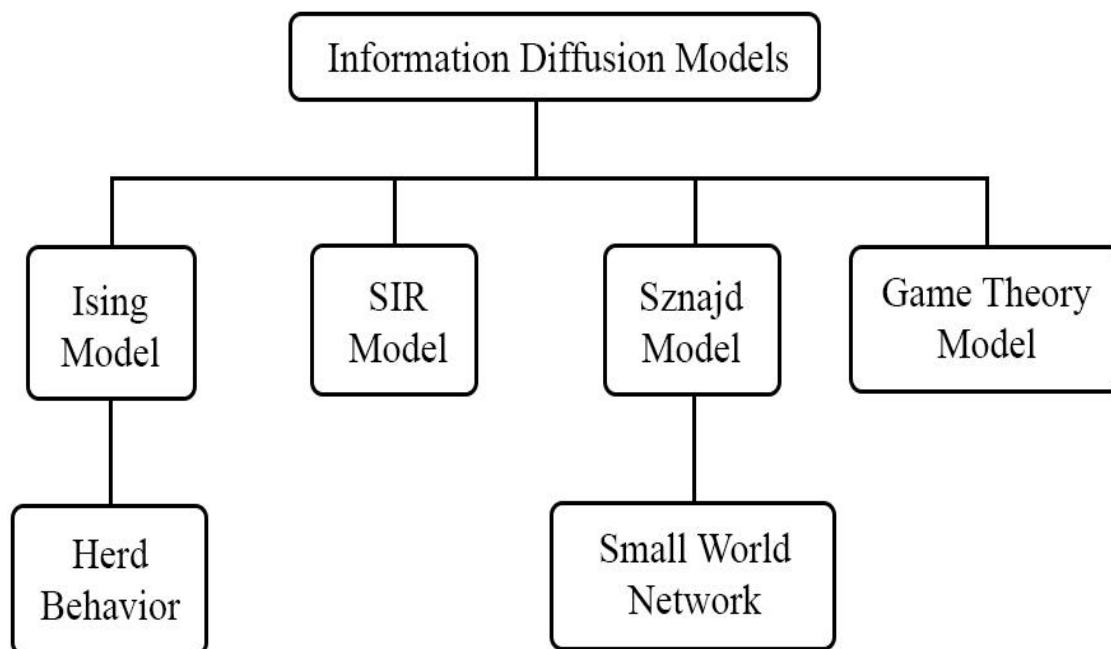


Figure 2: Different Information Diffusion Models

1.3 INFORMATION DIFFUSION MODELS

The study of impact of information dissemination has been a key area of research for a very long time. The systems and frameworks designed to assess this impact are called information diffusion models. Once the spreader nodes have been chosen through centrality measures, we use the information diffusion models to test the impact of our selection. There have been various models proposed and implemented by researchers all around the world like Ising model, SIR model, Sznajd Model, Independent Cascade Model etc. The dissemination of information of online platforms and social media networks resembles the spread of an infectious disease, therefore we have chosen SIR epidemic model (Susceptible-Infected-Recovered) for our analysis.

2. RELATED WORK

There have been numerous endeavors in the past to anticipate the progression of data in the mind boggling networks. Numerous creators have attempted different models and their blends to clarify the data dispersion in a network all the more successfully, however numerous reviews have demonstrated that they alluded to just one or other issue in informal communities.

Guille dissects the discernment of subjects, the demonstrating of the dispersal of data and the techniques for recognizing persuasive spreaders. Dong partitions flow models into hypothetical dispersion models and falling models of data scattering dependent on exploration information, some from a genuine interpersonal organization and others not. We will talk about the boundaries that influence the dispersal of data. These boundaries incorporate the strength of the connection, homophily, networks, assessment, client jobs and topics. Kumaran thinks about strategies, calculations and spreader detection procedures. His vision of the examination is identified with the dispersion of the impact. Dey analyzes related references in theme examination, data spread and the properties of social connections with regards to online informal communities. In spite of the fact that the scientists' perspectives contrast from the examination of the writing, they all speak to various features of the scattering of data research.

Chenxu proposed a strategy to demonstrate and quantify the impact of assessment pioneers on microblogs dependent on the transmission of data. This strategy depends exclusively on the structure of the network. Bo proposed a strategy to explore assessment pioneers dependent on the conduct and associations of one client with another. As per the administration abilities model, informal community clients can be partitioned into four classifications, in particular customary clients, dynamic clients, assessment pioneers present and network executives. Assessment pioneers can be discovered utilizing predominant and verifiable elements. Reference is made to explicit client practices, the examination of data and the association connections between clients. At long last, the assessment heads of the network can be acquired through three layers of separating. The impact of the scattering of data will be boosted through the extraction of feeling pioneers. Jiaxin proposed a technique to quantify social impact by anticipating a client's capacity to disperse data. The impact rating depends on the retweet account. Xianhu proposed a calculation dependent on the arrangement by theme. They consolidated the thickness of the node (client movement, connection between two nodes and subject), fringe thickness, client association property and substance characteristic to extricate assessment pioneers from a particular theme. Saito has embraced an EM calculation to anticipate the likelihood of spread by this model. It isn't appropriate for the use of a lot of information on interpersonal organizations because of its utilization of time. Wang and Jung center around augmenting impact dependent on the IC model. They accept that the versatility of the calculation is the way to expanding impact

research, to adjust to huge scope informal communities. Arora has proposed an ASIM calculation that consolidates runtime and memory utilization to expand impact research, which makes it reasonable for the investigation of genuine informal communities.

Different endeavors have been made in client arrangement in informal organization. We can't just overlook the commitment of the client in an informal community. Accomplishment of each informal organization relies on the quantity of clients dynamic on it. A portion of these clients may go about as influencer in the network. There have been broad investigations to recognize the predominant nodes in the network. Liu and Tang put forth a neighbor oriented centrality technique that took into consideration the neighbor's impact in choosing the centrality proportion for the node. Ling Ma and Feng Zhang associated the centrality gauge to the current standards of material science. They formulated the gravity centrality and utilized the exemplary attraction equation formalized by Newton. Kitsak exhibited the presence of a structure comparable to that of a shell in graph and contrived k-shell centrality gauge for client categorization. He accepted that the nodes in a similar shell will have same effect however this articulation was just halfway obvious. Zeng additionally enhanced the k-shell calculation by putting forth a mixed degree distribution strategy by joining the lingering and the depleted degree. Chen planned a semi-nearby file by considering the following closest I neighbors. Lin introduced an improved positioning technique by considering the most brief way between distance between an objective node and the node set with most noteworthy k-guiding principle. Bae characterized a fresh measure, coreness centrality list, which is evaluated by adding all neighbor's k-shell worth. Sara Hajjam and Hasan Badir proposed a half and half position calculation by consolidating the impacts of both improved coreness centrality and Eigen centrality.

Qingcheng Hao proposed a way to deal with incorporate the community identification strategy in client categorization. He incorporated the commitments of both neighborhood and inner properties of the node.

3. METHODOLOGY

Determination of predominant nodes through improved gravity centrality focuses at recognizing the nodes that are situated in the center of the network, and are all around associated i.e., not just to one level but rather up to specific levels. The essential thought of our hypothesized strategy is: computing improved gravity centrality of each of the nodes existing in the network, ranking them as per the obtained values of the centrality measure and then we choose top-k powerful nodes from the list. Here, k is a hyperparameter to the calculation whose value is derived by performing repeated simulations with varying k and noticing its value that amplifies the dissemination impact. Improved gravity calculation considers two centralities, i.e., H-index centrality and Gravity centrality.

3.1 H-INDEX

The idea of H-index (Hirsch index) has been inspired from the citation networks. A similar idea has been repeated to social networks where H-index of a node i is characterized as the largest value, h with the end goal that there occur a minimum of h of its neighbors with degree at most h . It is very clear, in H-index centrality a node is deemed significant on the basis that it has significant connections, i.e., persuasive and predominant neighboring nodes up to two hops. H-index, accordingly underlines on the vicinity of a node to decide its ability of being a predominant disseminator in the network. For a node i , H-index can be characterized as:

$$h_i = \max(\min(D_j) \geq h); \forall e_{i,j} \in E \quad (8)$$

An updated version of H-index is also available to us which is known as the Extended H-Index. This centrality measure also takes into account the h-indices of all the neighbors of the node we are investigating. The formula for extended h-index is:

$$EH_i = h_i + \sum_{j \in N_i} h_j \quad (9)$$

Where, N_i is the set of neighbor nodes of node i .

3.2 GRAVITY CENTRALITY

We know that a node has higher influence if its neighbors have high k-shell value, also the effect of interactions between two diminishes with increasing distance. Keeping this in mind, gravity centrality is defined using the gravitational formula where the mass of the node is taken to be its

k-shell value and the distance is taken to be the shortest path between two nodes. It can mathematically be represented as:

$$Gr_i = \sum_{j \in \varphi_i} \frac{ks_i * ks_j}{d_{ij}^2} \quad (10)$$

Where, φ_i is the set of neighboring nodes up to a distance of 3.

While computing the gravity centrality of a node, we generally take into account the effect of nodes that are present at a distance of at most 3. This is done so due to the following reasons:

1. Effect of a neighbour node is inversely proportional to square of the distance between them. Hence, for distances greater than 3 the denominator becomes very large, thereby reducing the effect of the node drastically.
2. Restricting the distance to 3 increases the efficiency of the algorithm, making it easily scalable in case we have large networks.

Building on this concept, another index called Extended Gravity index is developed, which can mathematically be expressed as:

$$EGr_i = \sum_{j \in N_i} Gr_j \quad (11)$$

Improved Gravity (IGr) centrality of a node i is assessed by using H-Index values of the node and its neighbor as their mass in the Gravity Centrality. It can mathematically be written as:

$$IGr_i = \sum_{j \in \varphi_i} \frac{h_i * h_j}{d_{ij}^2} \quad (12)$$

Further, Extended Improved Gravity can be characterized as:

$$EIGr_i = \sum_{j \in \varphi_i} \frac{h_i * h_j}{d_{ij}^2} \quad (13)$$

3.3 SPREADER SELECTION

The choice of k predominant diffusers dependent on the estimations of our method is done in an iterative manner as:

- The node that has the greatest centrality value is chosen.
- The chosen node, alongside its immediate neighbors, are eliminated from the list of diffusers.

This cycle is rerun for k occasions, and accordingly, a rundown of k spreaders is obtained, i.e., k most predominant nodes in that network. The thought backing the removal of a chosen node's neighbors from the set of diffusers depends on the way that, on the off chance that a node has a

high centrality value, at that point there is a greater likelihood that its neighbor's centrality value is going to be very high as well. Consequently to expand the scope or the proliferation scope of data, we consider just the disconnected spreaders.

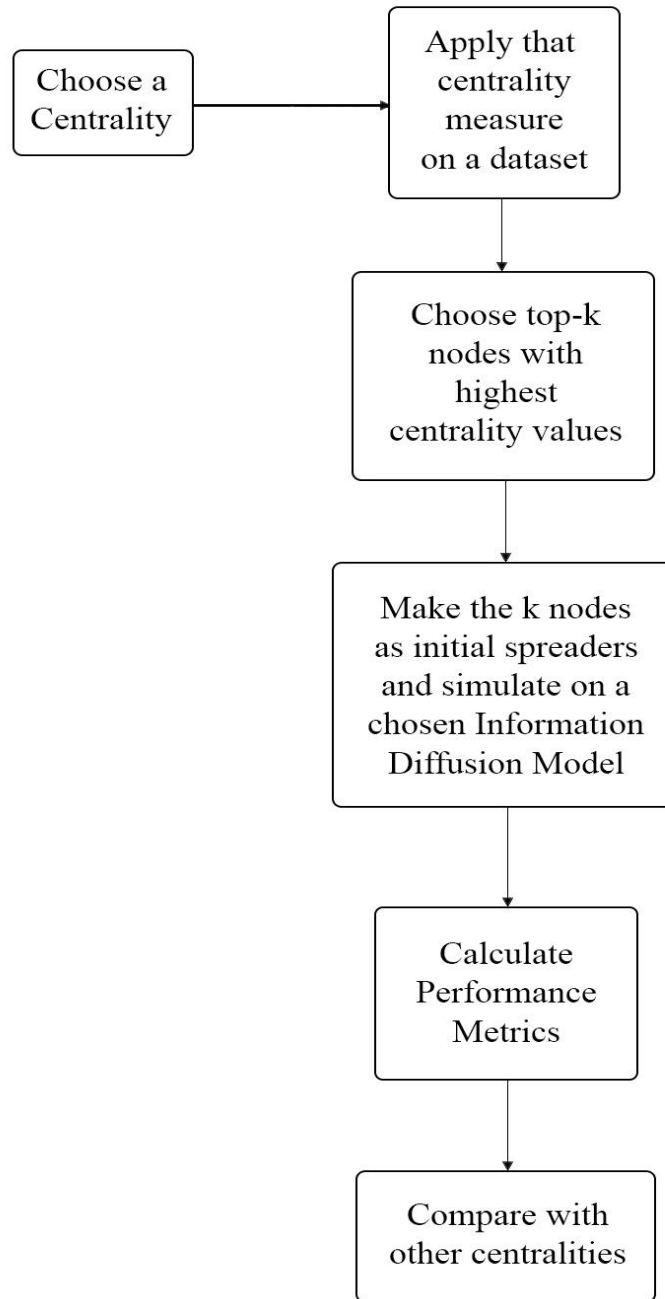


Figure 3: Flowchart for the steps in our methodology

3.4 SIR Model

To decide the information dissemination impact of different nodes in the network, we make use of the Susceptible-Infected-Recovered epidemic model (SIR model). The model accepts a populace of N people, classified along the lines of following three states:

- **Susceptible (S):** the individual isn't yet contaminated, accordingly being powerless to the illness;
- **Infected (I):** the individual has been contaminated with the sickness and he is fit for spreading the illness to the vulnerable populace;
- **Recovered (R):** after an individual has encountered the period of infection, he is considered as eliminated from the illness and he can't be contaminated again or communicate the sickness to other people (insusceptible to additional disease or death).

At first, the status of all nodes is susceptible aside from a set of m infected nodes already chosen as initial spreaders. At every time progression, an infected node attempts to contaminate one of its adjacent nodes with likelihood μ . Simultaneously, every infected node will recover with a likelihood β , if achieved, it will not be infected anymore and nor, at this point will infect other susceptible nodes. The cycle ends if no infected node remains in network. Here, we use $\lambda = \mu/\beta$ to depict the rate of infection, which is vital to the pace of infection and the final infected scale that generally are instruments demonstrate the propagation capacity of m root spreaders. Other than SIR model with restricted exposure, the presentation of techniques can likewise be assessed by full exposure SIR model and SI model that is generally used to assess the strategy on dissemination rate particularly in the beginning phase.

4. DATASETS AND PERFORMANCE METRICS

4.1 DATASETS

To test the performance of our method, we have executed it over different real-world datasets. The set of graphs selected are very varied regarding their nature, sizes and situation of utilization. The portrayal of the graphs is given in Table 2 where the given parameters give a brief glance at the graph properties.

Network	Type	Nodes	Edges	Average Clustering Coefficient	Description
Tech-pgp	Undirected	10680	24316	0.26598	Edge list for user network of PGP algorithm.
Jazz	Undirected	198	2742	0.617451	Jazz musicians network.
Facebook_combined	Undirected	4039	88234	0.6055	Anonymised Facebook friends lists.
P2P-Gnutella	Directed	6301	20777	0.0109	Snapshot graph for Gnutella peer-to-peer file sharing network

Table2: Overview of the used Network Datasets

4.2 PERFORMANCE METRICS

Each new research needs to demonstrate its fortitude by certain statistical data points. Different techniques are utilized to advance the proficiency of the new methodology. Also we have utilized a few measurements to contrast our outcomes and standard literature.

4.2.1 Infected Scale (F(t))

This metric is utilized to look at the behavior of a specific centrality measure during data dispersion. To put it simply, what fraction of nodes was dynamic at a given time? This metric definitely reveals to us how a specific centrality functions with respect to time. Some centrality may have a quick boost up, for example it initiates nodes at an uncommon degree however for a little time interval in particular, yet some centrality might not have a fast actuation rate yet it might enact increased count of nodes by and large. Here we are talking in casual terms while alluding centrality measure for data dissemination, despite the fact that there is no immediate relationship between them. We imply that centrality measure that assists us in identifying base nodes in the graph utilized for data dispersion utilizing SIR model.

The infected scale $F(t)$ at a given time t is expounded as the proportion of the aggregate of infected and recovered nodes at that given time to the exhaustive count of nodes present in the network.

$$F(t) = \frac{n_{I(t)} + n_{R(t)}}{N} \quad (14)$$

4.2.2 Final Infected Scale ($F(t_c)$)

This measure is utilized to think about the last level of dynamic nodes that are present in the network with the division of spreaders picked out as basenodes for SIR algorithm. The count of spreaders opted as basenodes influences the fraction of active nodes at termination and along these lines this metric helps us to see how the current initiation gets influenced by various number of spreaders. We can likewise think about different centralities utilizing this measurement by looking at their individual $F(t_c)$ values at various part of spreaders.

The Final Infected Scale can be elucidated as the fraction of the nodes that have recovered with respect to the total count of nodes present in the network.

$$F(t_c) = \frac{n_{R(t_c)}}{N} \quad (15)$$

4.2.3 Shortest Path Length (L_s)

This measurement gives us clear thought regarding how far the underlying spreaders(seed nodes) are from one another. A huge L_s value demonstrates that the seed nodes are a long way from one another and subsequently cover a bigger portion at first. This bigger inclusion by the seednodes brings about higher information dissemination (better results).

The value of L_s is contrasted with the fraction of spreaders to watch how the gap between different spreaders (therefore quality) increments with the expansion in amount. Analysis of L_s estimations of different centralities at various estimations of fraction of spreaders is done.

L_s is mathematically computed as:

$$L_s = \frac{1}{N(N-1)} \sum_{i,j \in V, i \neq j} d_{i,j} \quad (16)$$

Where, N represents total count of nodes in the network and $d_{i,j}$ represents the distance between i and j .

4.2.4 Kendall Tau (τ)

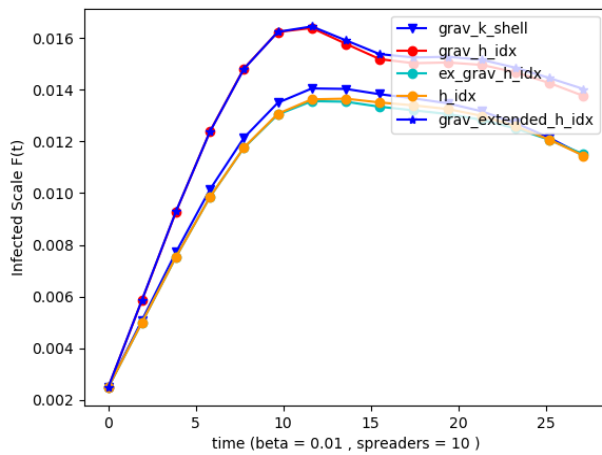
Kendall's tau relationship coefficient is utilized to measure the likeness between two rank lists. Assume there are two rank lists L_1 and L_2 containing positions $(r_1, r_2, r_3 \dots r_n)$ and $(m_1, m_2, m_3 \dots m_n)$ for nodes $(v_1, v_2, v_3 \dots v_n)$. We can have $n(n - 1) = 2$ pairings between components in a solitary position list. Couple (i, j) is supposed to be concordant or coordinating if in L_1 , $r_i < r_j$ suggests $m_i < m_j$ in L_2 . On the off chance that this condition doesn't hold true, we say that the sets are discordant. More concordant sets (n_c) infer that these two records have a comparative requesting of positions and consequently are comparable. More number of conflicting sets (n_d) infer that rundowns are not comparative. The estimation of Kendall's tau coefficient can go from -1 to 1, -1 shows extraordinary disparity and 1 suggest all out likeness. Executing SIR calculation on network nodes independently gives the extent of contamination of the concerned nodes. A position list worried about individual spreading capacity would thus be able to be shaped utilizing SIR positioning rundown. The subsequent rundown can be produced using the positioning dependant on the hypothesized calculation which needs to be tested. At that point we can utilize Kendall's Tau to assess the choice of nodes by the hypothesized calculation, in light of the SIR rank-list.

The measurement of Kendall-tau can be done as:

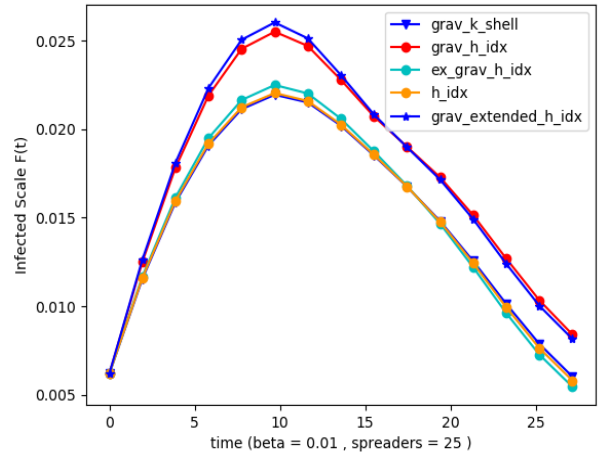
$$\tau = \frac{2(\text{concordant pairs} - \text{discordant pairs})}{N(N-1)} \quad (17)$$

RESULTS

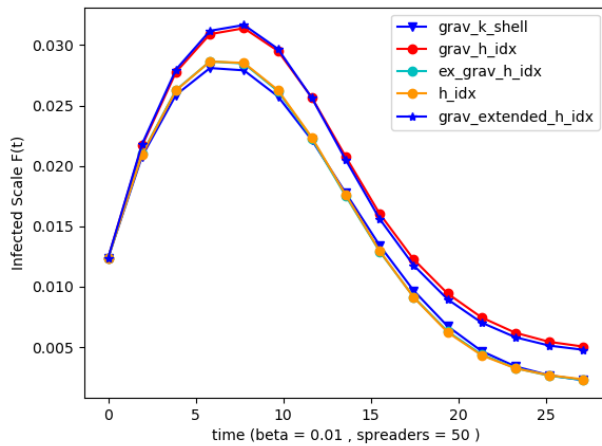
We assess the functioning of our proposed strategy Improved Gravity centrality with the current best in class centrality estimates like Hirsch index and gravity centrality. We also take a look at a few modifications of the algorithm such as Extended Improved Gravity and Improved Gravity Extended Hirsch centralities. The presentation of our calculation in contrast with different techniques is determined utilizing the performance metrics talked about above. Following are the plots of infected scale, according to eq. no. 11, $F(t)$ versus Time, final infected scale, $F(t_c)$ vs Spreaders, according to eq. no.12 and shortest paths, L_s vs Spreaders for various undirected and directed graphs where the estimation of β , i.e., the contamination likelihood is taken as 0.1. The SIR cycle is rehashed for 100 mock executions, and the estimations of contaminated scale acquired are found the middle value of over the quantity of simulations.



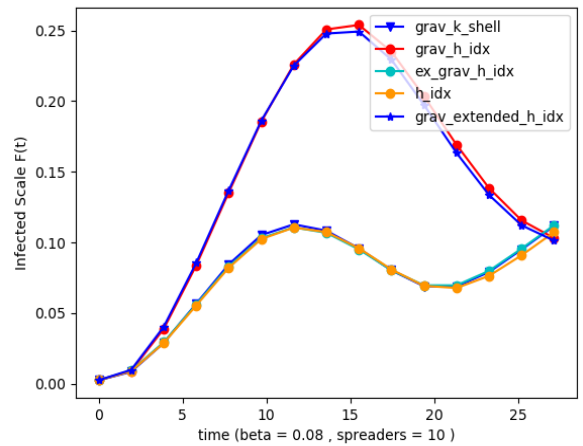
(a) $\beta = 0.01$, Spreader Nodes = 10



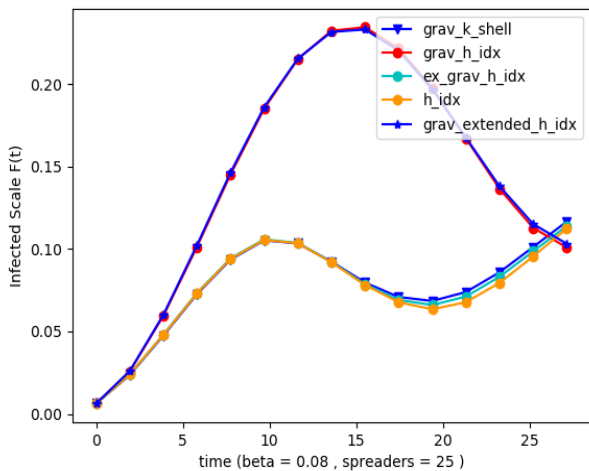
(b) $\beta = 0.01$, Spreader Nodes = 25



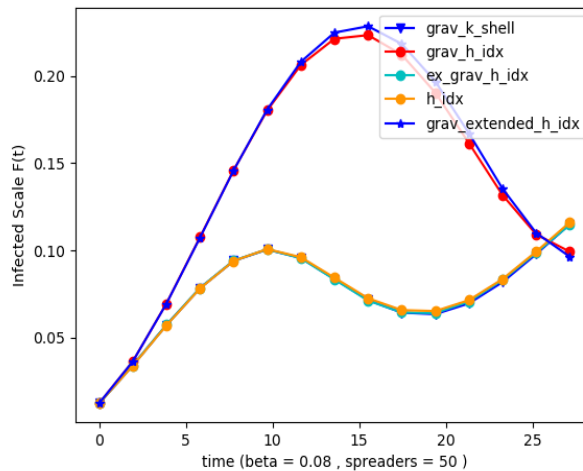
(c) $\beta = 0.01$, Spreader Nodes = 50



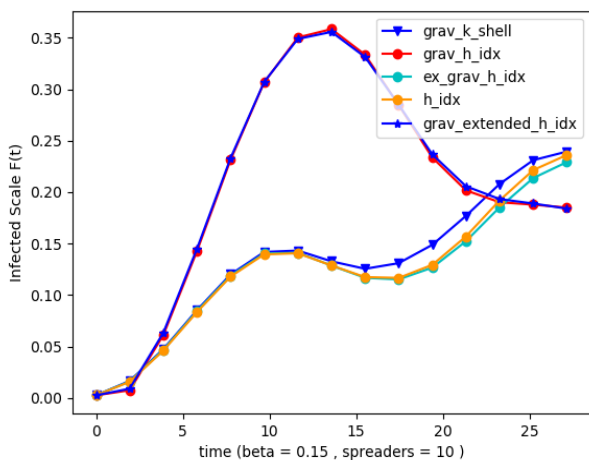
(d) $\beta = 0.08$, Spreader Nodes = 10



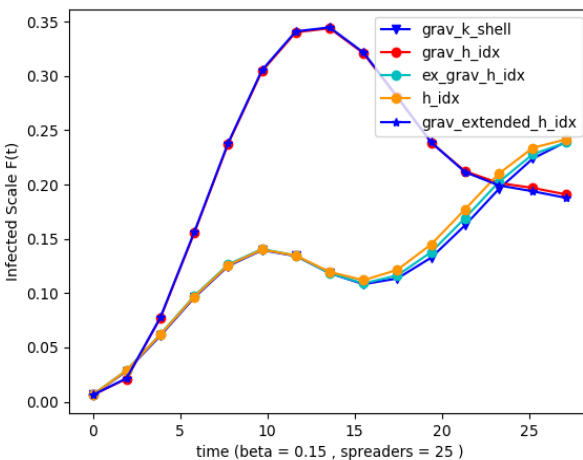
(e) $\beta = 0.08$, Spreader Nodes = 25



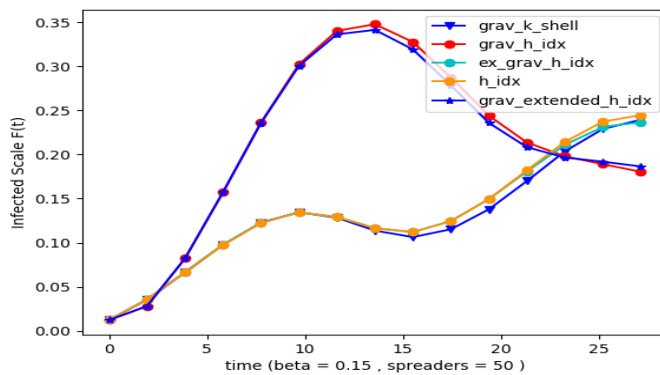
(f) $\beta = 0.08$, Spreader Nodes = 50



(g) $\beta = 0.15$, Spreader Nodes = 10

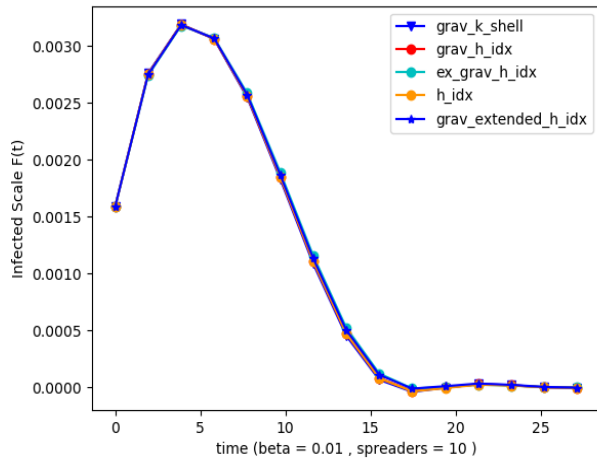


(h) $\beta = 0.15$, Spreader Nodes = 25

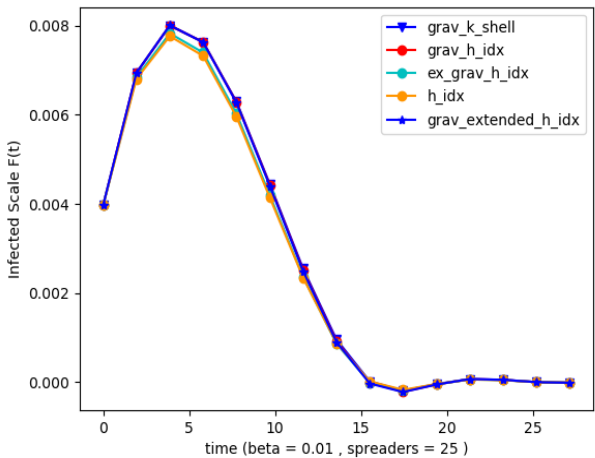


(i) $\beta = 0.15$, Spreader Nodes = 50

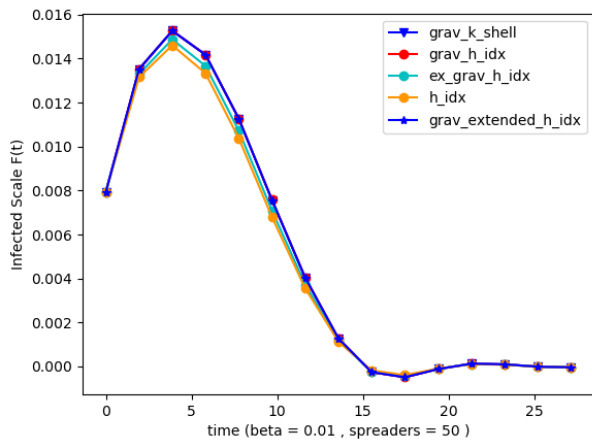
Figure 4: Infected Scale ($F(t)$) vs Time (t) for Facebook



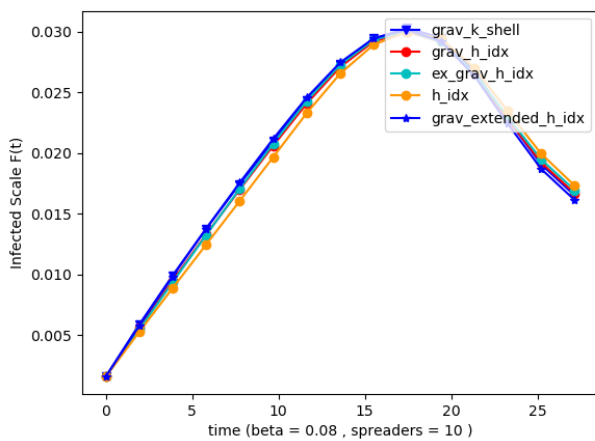
(a) $\beta = 0.01$, Spreader Nodes = 10



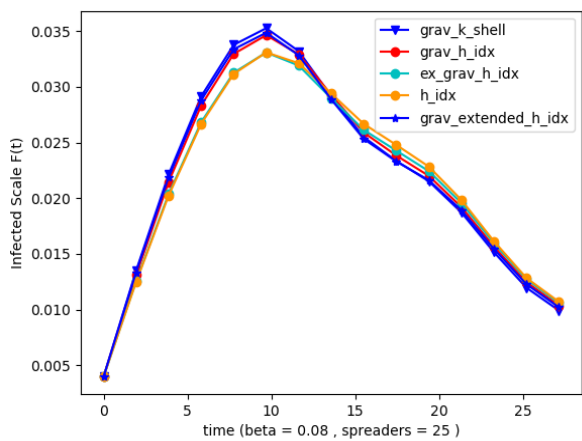
(b) $\beta = 0.01$, Spreader Nodes = 25



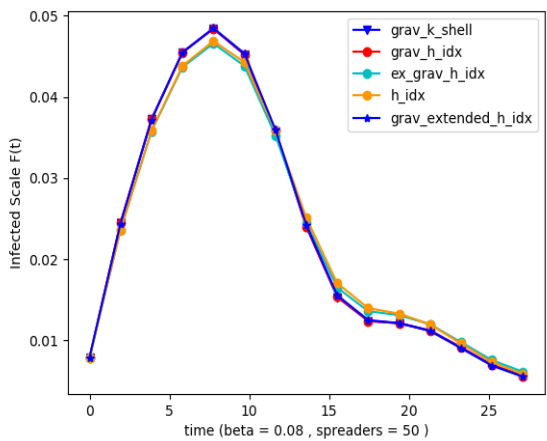
(c) $\beta = 0.01$, Spreader Nodes = 50



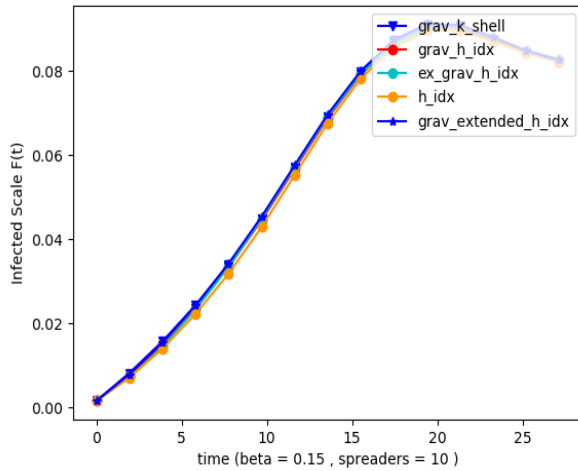
(d) $\beta = 0.08$, Spreader Nodes = 10



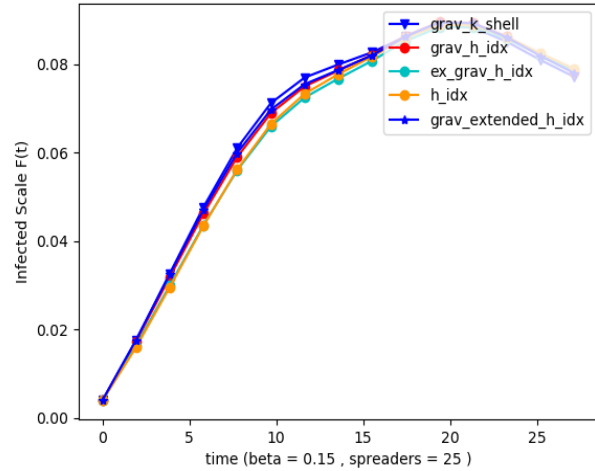
(e) $\beta = 0.08$, Spreader Nodes = 25



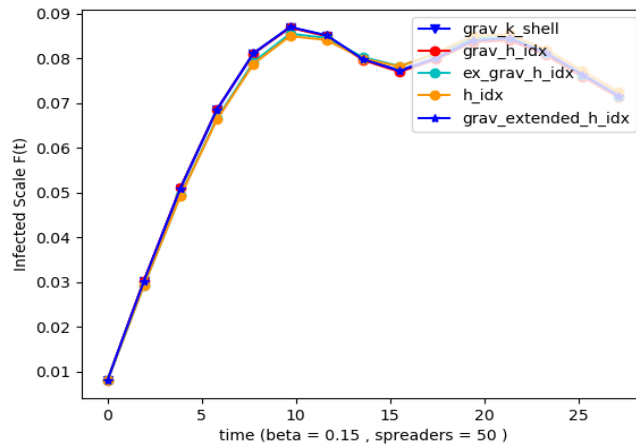
(f) $\beta = 0.08$, Spreader Nodes = 50



(g) $\beta = 0.15$, Spreader Nodes = 10

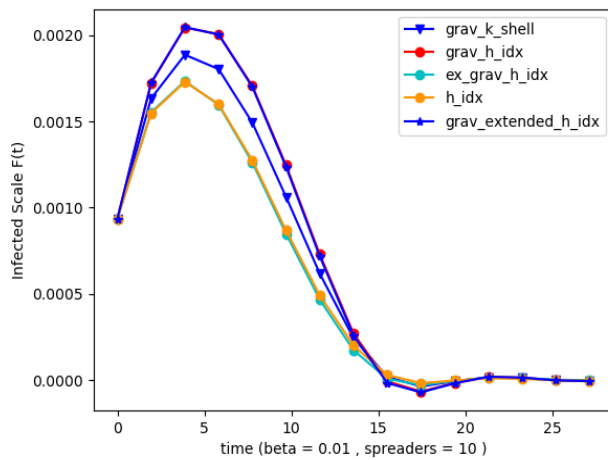


(h) $\beta = 0.15$, Spreader Nodes = 25

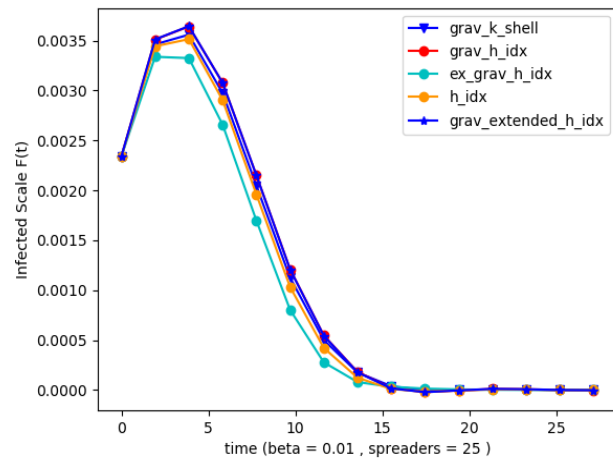


(i) $\beta = 0.15$, Spreader Nodes = 50

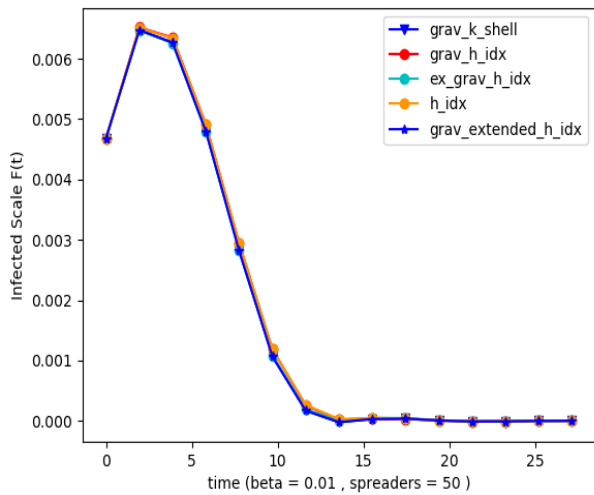
Figure 5: Infected Scale $F(t)$ vs Time t for P2P-Gnutella



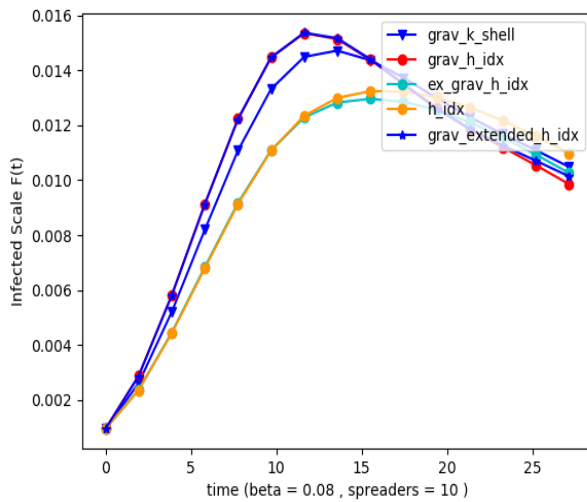
(a) $\beta = 0.01$, Spreader Nodes = 10



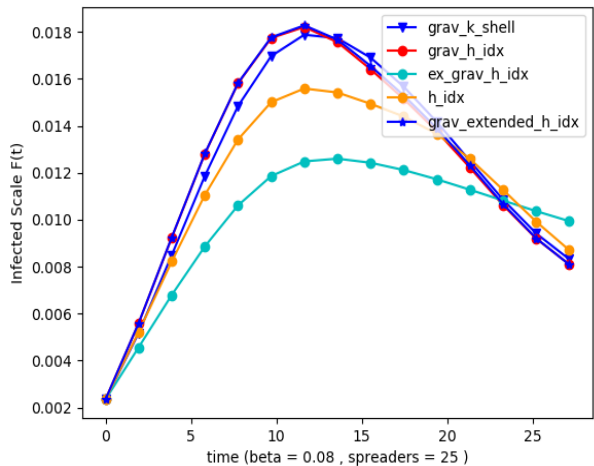
(b) $\beta = 0.01$, Spreader Nodes = 50



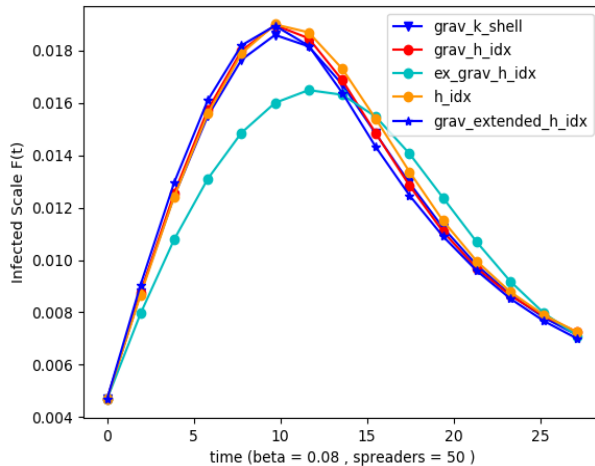
(c) $\beta = 0.01$, Spreader Nodes = 50



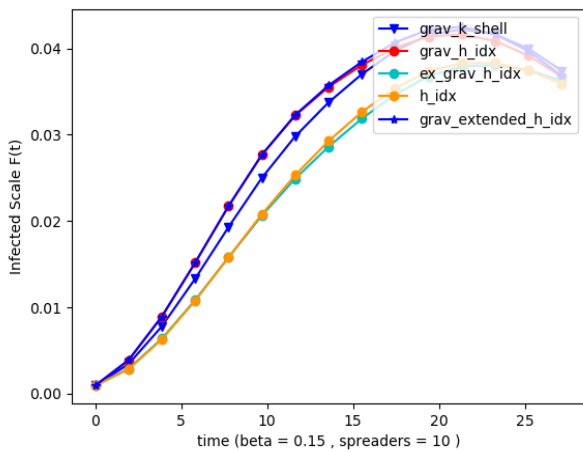
(d) $\beta = 0.08$, Spreader Nodes = 10



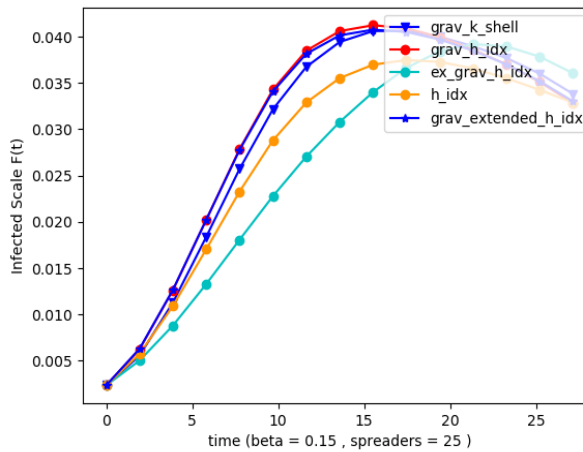
(e) $\beta = 0.08$, Spreader Nodes = 25



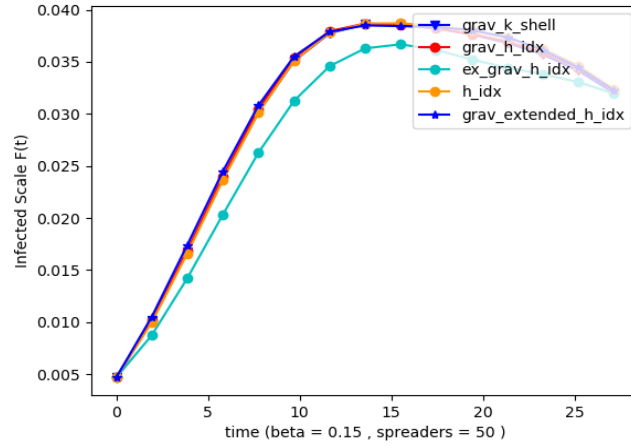
(f) $\beta = 0.08$, Spreader Nodes = 50



(g) $\beta = 0.15$, Spreader Nodes = 10

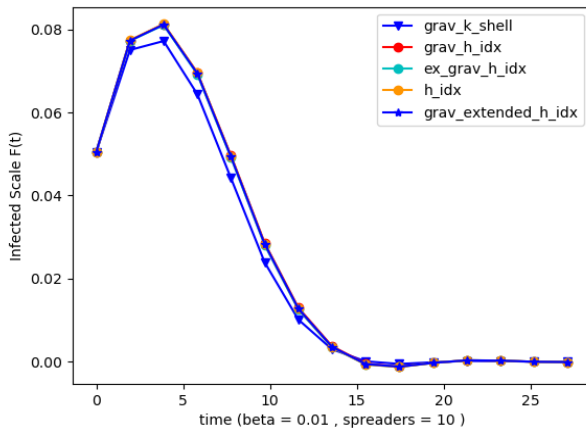


(h) $\beta = 0.15$, Spreader Nodes = 25

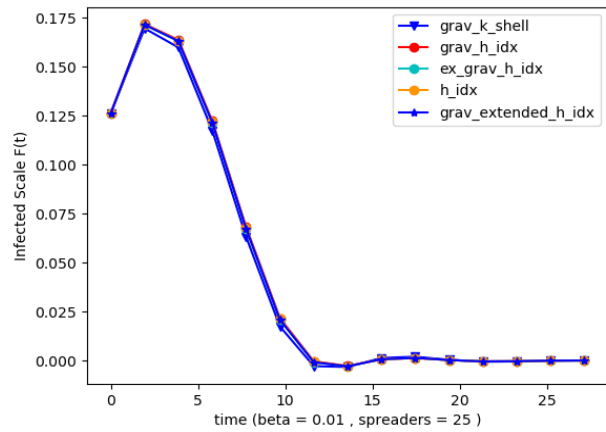


(i) $\beta = 0.15$, Spreader Nodes = 50

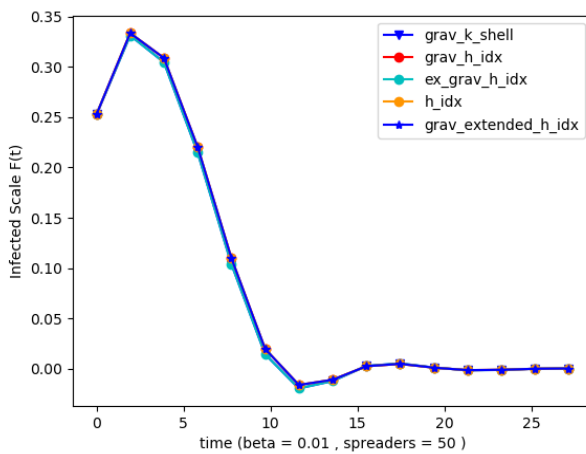
Figure 6: Infected Scale (F(t)) vs Time (t) for PGP



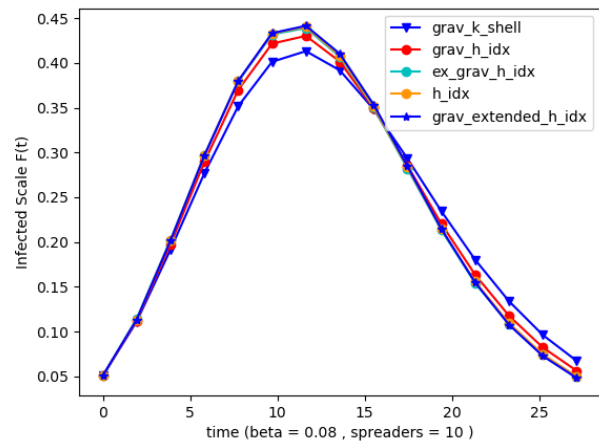
(a) $\beta = 0.01$, Spreader Nodes = 10



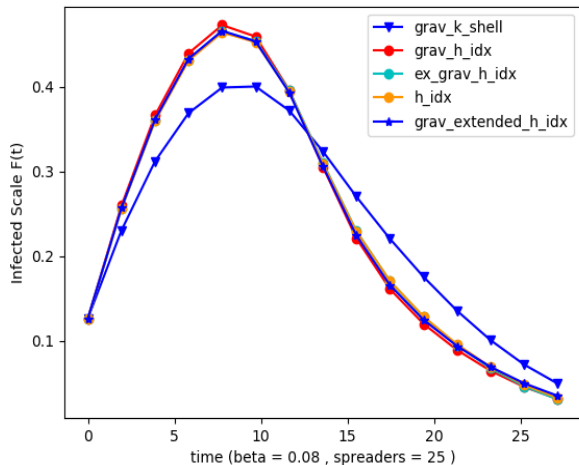
(b) $\beta = 0.01$, Spreader Nodes = 25



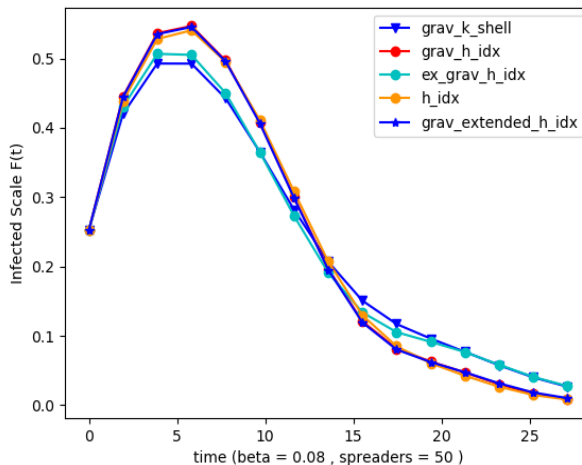
(c) $\beta = 0.01$, Spreader Nodes = 50



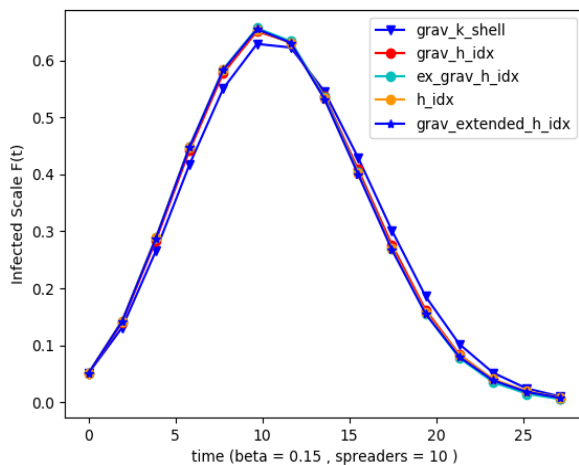
(d) $\beta = 0.08$, Spreader Nodes = 10



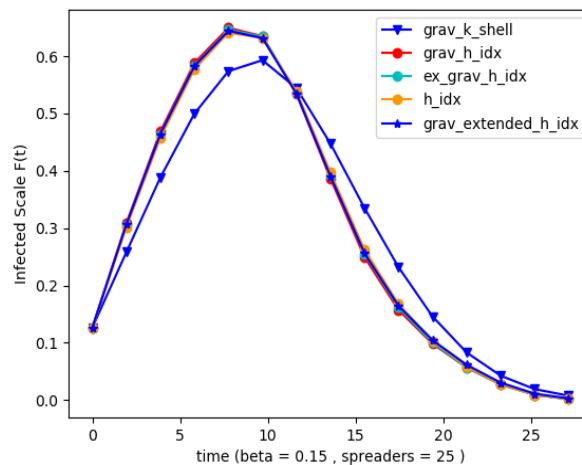
(e) $\beta = 0.08$, Spreader Nodes = 25



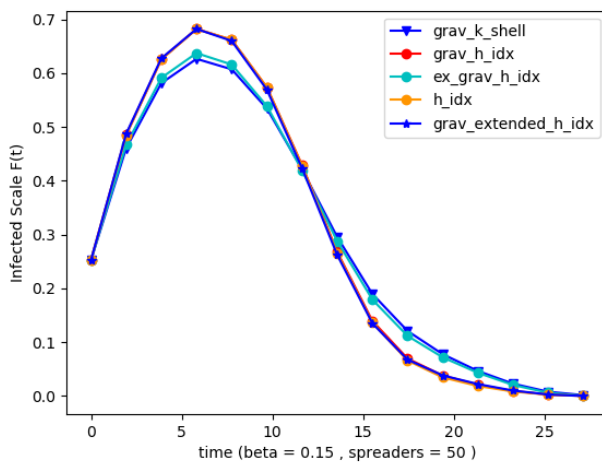
(f) $\beta = 0.08$, Spreader Nodes = 50



(g) $\beta = 0.15$, Spreader Nodes = 10

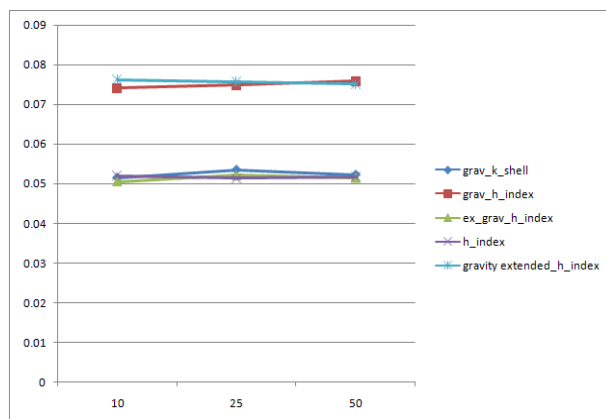


(h) $\beta = 0.15$, Spreader Nodes = 25

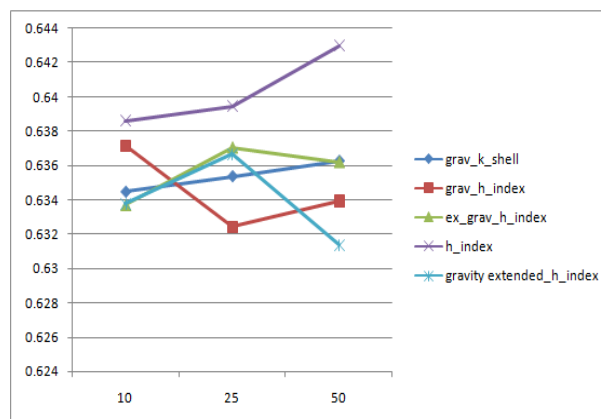


(i) $\beta = 0.15$, Spreader Nodes = 50

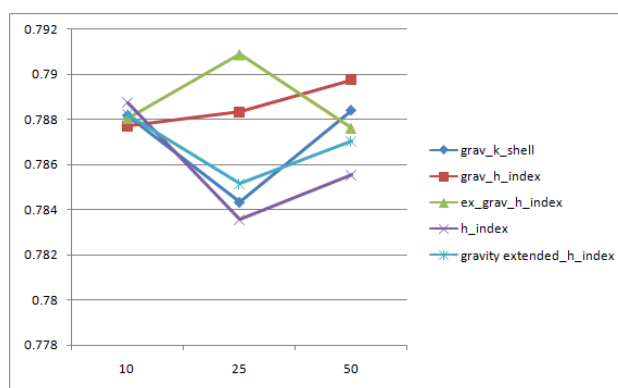
Figure 7: Infected Scale (F(t)) vs Time (t) for Jazz Network



(a) $\beta = 0.01$

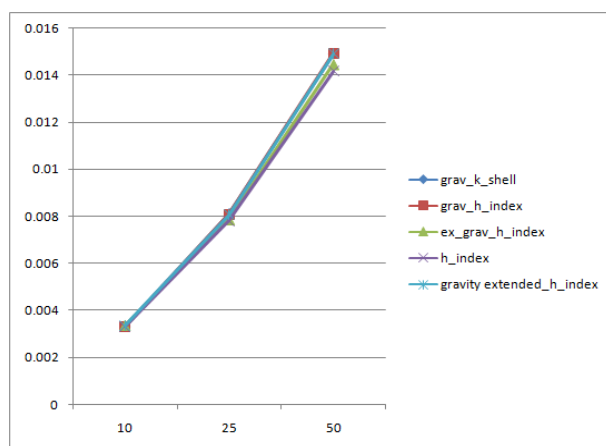


(b) $\beta = 0.08$

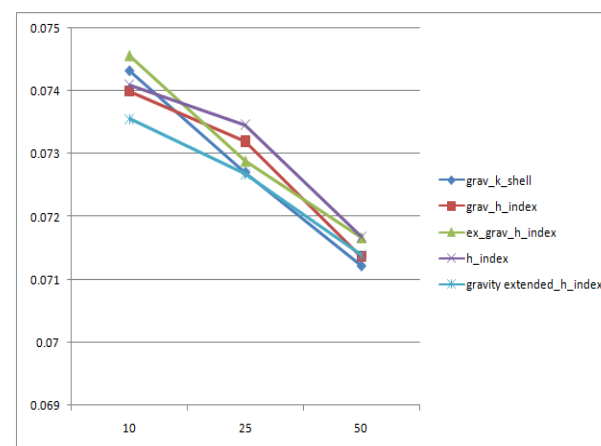


(c) $\beta = 0.15$

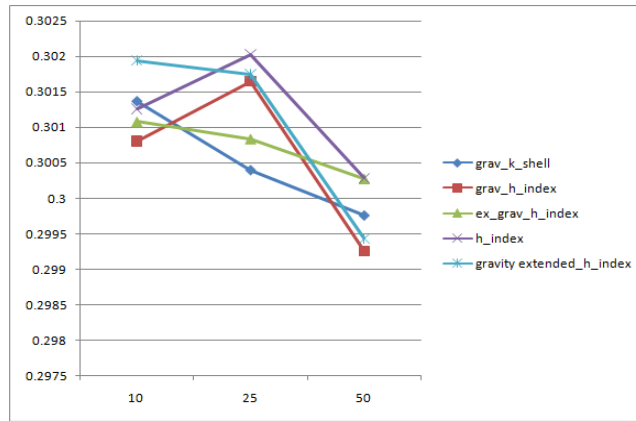
Figure 8: Facebook – Final Infected Scale $F(t_c)$ vs Spreaders



(a) $\beta = 0.01$

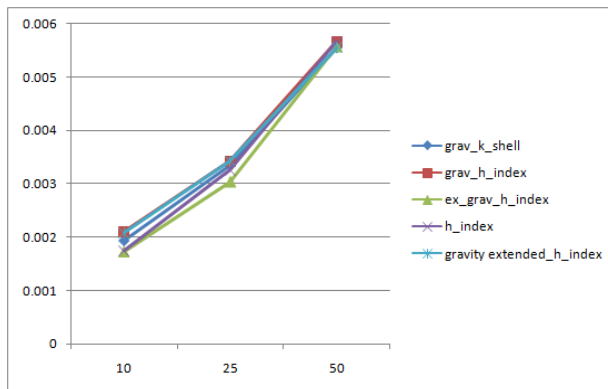


(b) $\beta = 0.08$

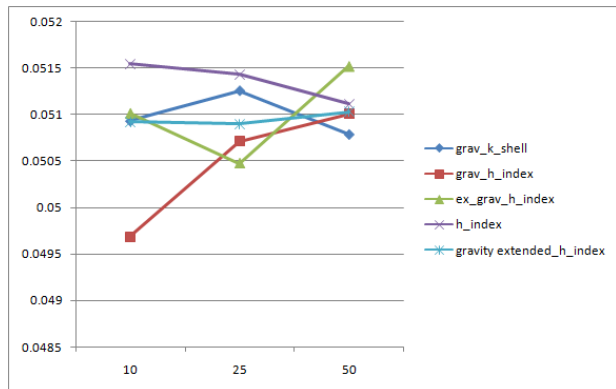


(c) $\beta = 0.15$

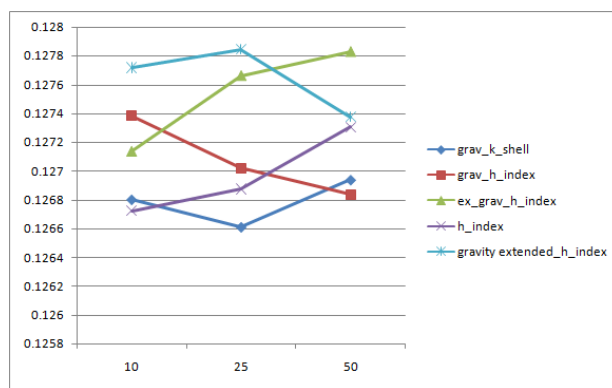
Figure 9: Gnutella – Final Infected Scale $F(t_c)$ vs Spreaders



(a) $\beta = 0.01$

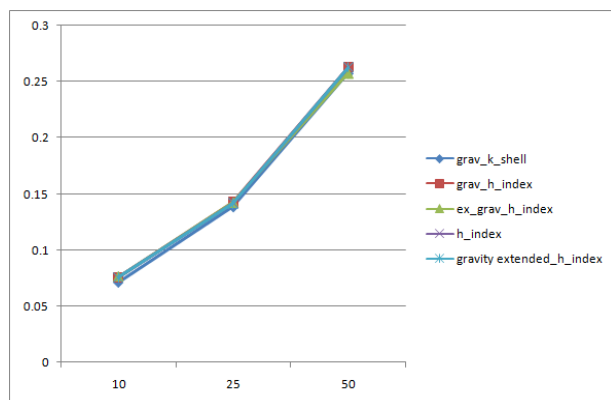


(b) $\beta = 0.08$

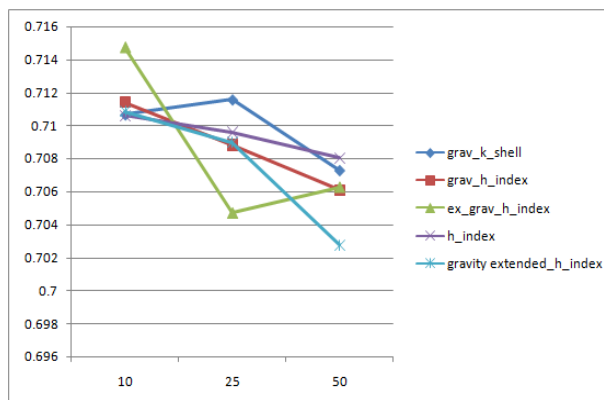


(c) $\beta = 0.15$

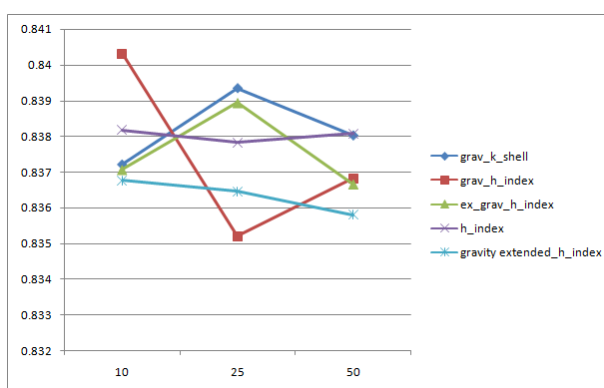
Figure 10: PGP – Final Infected Scale $F(t_c)$ vs Spreaders



(a) $\beta = 0.01$

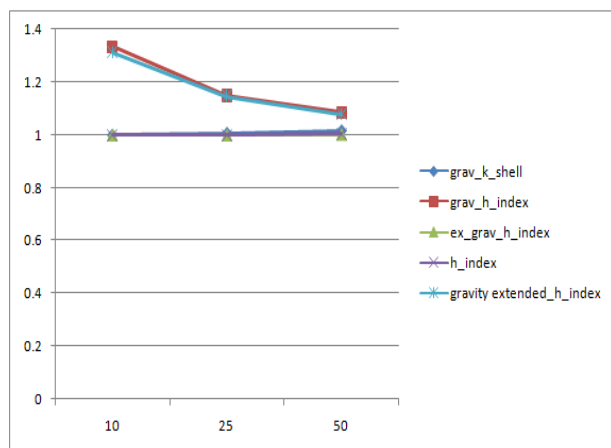


(b) $\beta = 0.08$

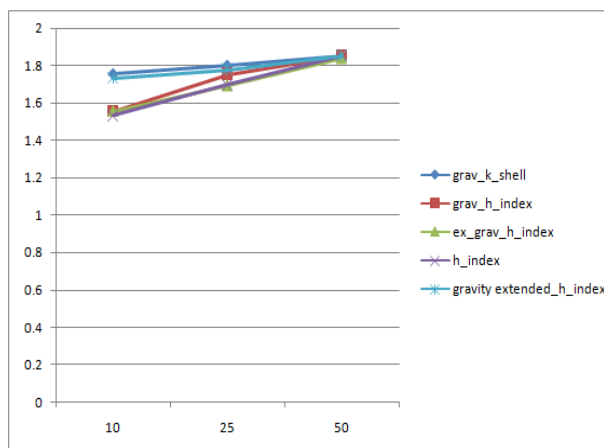


(c) $\beta = 0.15$

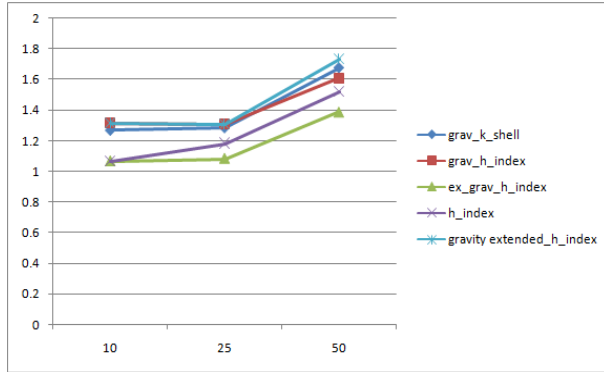
Figure 11: Jazz – Final infected scale $F(t_c)$ vs Spreaders



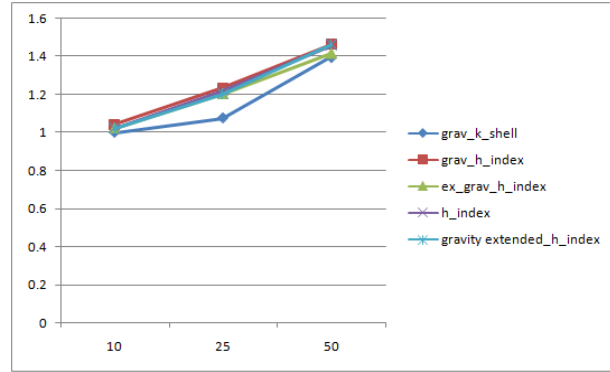
(a) Facebook



(b) Gnutella



(c) PGP



(d) Jazz

Figure 12: Shortest Path (L_s) vs Spreaders for different datasets

6. CONCLUSION AND FUTURE WORK

There is a rapid growth in social networks and there is a consistent increase in its user base with new people joining the bandwagon everyday in this era of rapid digitalization. With a base of users constantly expanding in the social networks, specialists have attempted to pick out the users that have a fair amount of influence in the social networks utilizing different properties of the user.

In this undertaking we have put forward another centrality measure which aims to categorize the users in the social network, specifically, Improved Gravity centrality, which consolidates both the internal and external properties of the node for a better categorization of nodes. We have contrasted our outcomes with other centrality measures, some regular and some later explores. We utilized SIR Epidemic Model (Susceptible-Infected-Recovered) for looking at the information diffusion by the k most predominant hubs chose utilizing positioning gave by the centrality measure.

We saw that our proposed technique performed in a way that is better than other methodologies in a large portion of the cases, and this methodology may end up being useful in other related territories of in social networks research.

Nothing can ever be perfect, subsequently we concur that there is extent of additional improvement in this proposition. Because of multifaceted nature of the calculation and restricted computational force available to us, we were unable to apply the proposed way to deal with an enormous network containing billions of nodes. Results may additionally improve if the covering network structure of the hub is thought of. Utilization of game-theory model for data dispersion may end up being predominant technique than SIR. Utilization of transformative calculations for finding the network structure is developing region of interest. We can prepare the model to naturally change the estimation of boundaries a and b in real condition of Improved Gravity centrality. This work might be stretched out to a time-varying social network to approve the outcome on ongoing social network. Utilization of node's embedded features in finding the impact of the node can also be looked at.

7. REFERENCES

- [1] Ma, Ling-Ling & Ma, Chuang & Zhang, Hai-Feng. (2015). "Identifying influential spreaders in complex networks based on gravity formula." *Physica A: Statistical Mechanics and its Applications*. 451. 10.1016/j.physa.2015.12.162.
- [2] Ahajjam, Sara & Hassan, Badir. (2018). "Identification of influential spreaders in complex networks using HybridRank algorithm." *Scientific Reports*. 8. 10.1038/s41598-018-30310-2.
- [3] Yang, Li & Qiao, Yafeng & Liu, Zhihong & Ma, Jianfeng & Li, Xinghua. (2018). "Identifying opinion leader nodes in online social networks with a new closeness evaluation algorithm." *Soft Computing*. 22. 10.1007/s00500-016-2335-3.
- [4] Bhat, Neeraj & Aggarwal, Nipun & Kumar, Sanjay. (2020). "Identification of Influential Spreaders in Social Networks using Improved Hybrid Rank Method." *Procedia Computer Science*. 171. 662-671. 10.1016/j.procs.2020.04.072.
- [5] Razaque, Abdul & Rizvi, Syed & Jaro Khan, Meer & Almi'ani, Muder & Al-rahayfeh, Amer. (2019). "State-of-Art Review of Information Diffusion Models and their Impact on Social Network Vulnerabilities." *Journal of King Saud University - Computer and Information Sciences*. 10.1016/j.jksuci.2019.08.008.
- [6] Qingcheng, Hu & Gao, Yang & Ma, Pengfei & Yin, Yanshen & Zhang, Yong & Xing, Chunxiao. (2013). "A New Approach to Identify Influential Spreaders in Complex Networks." *Acta Physica Sinica -Chinese Edition-*. 62. 99-104. 10.1007/978-3-642-38562-9_10.
- [7] Liu, Qiang & Zhu, Yuxiao & Deng, Lu & Zhou, Bin & Zhu, Junxing & Zou, Peng. (2017). "Leveraging local h-index to identify and rank influential spreaders in networks." *Physica A: Statistical Mechanics and its Applications*. 512. 10.1016/j.physa.2018.08.053.
- [8] Zang, Wenyu & Zhang, Peng & Zhou, Chuan & Guo, Li. (2014). "Discovering Multiple Diffusion Source Nodes in Social Networks." *Procedia Computer Science*. 29. 443-452. 10.1016/j.procs.2014.05.040.
- [9] <https://www.sciencedirect.com/topics/computer-science/information-diffusion>
- [10] <https://www.centiserver.org/centrality/>

- [11] <https://www.statisticssolutions.com/kendalls-tau-and-spearman-rank-correlation-coefficient/>
- [12] <https://snap.stanford.edu/data/index.html>
- [13] <http://networkrepository.com/>