5-1-2017

Advanced Applications Of Big Data Analytics

Sai Phani Krishna Parsa
University of Nevada, Las Vegas

ADVANCED APPLICATIONS OF BIG DATA ANALYTICS

by

Sai Phani Krishna Parsa

Bachelor's Degree in Computer Science and Engineering
Sree Nidhi Institute of Science and Technology, Hyderabad, Telangana, India
2014

A thesis submitted in partial fulfillment of
the requirements for the

Master of Science in Computer Science

Department of Computer Science
Howard R. Hughes College of Engineering
The Graduate College

University of Nevada, Las Vegas
May 2017

UNLV | GRADUATE COLLEGE

Thesis Approval

The Graduate College
The University of Nevada, Las Vegas

May 12, 2017

This thesis prepared by

Sai Phani Krishna Parsa

entitled

Advanced Applications Of Big Data Analytics

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science
Department of Computer Science

Justin Zhan, Ph.D.
Examination Committee Chair

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Interim Dean

Yoohwan Kim, Ph.D.
Examination Committee Member

Juyeon Jo, Ph.D.
Examination Committee Member

Ge Lin Kan, Ph.D.
Graduate College Faculty Representative

Abstract

Human life is progressing with advancements in technology such as laptops, smart phones, high speed communication networks etc., which helps us by reducing load in doing our daily activities. For instance, one can chat, talk, make video calls with his/her friends instantly using social networking platforms such as Facebook, Twitter, Google+, WhatsApp etc. LinkedIn, Indeed, etc., connects employees with potential employers. The number of people using these applications are increasing day-by-day, and so is the amount of data generated from these applications. Processing such vast amounts of data, may require new techniques for gaining valuable insights. Network theory concepts form the core of such techniques that are designed to uncover valuable insights from large social network datasets.

Many interesting problems such as ranking top-K nodes and top-K communities that can effectively diffuse any given message into the network, restaurant recommendations, friendship recommendations on social networking websites, etc., can be addressed by using the concepts of network centrality. Network centrality measures such as In-degree centrality, Out-degree centrality, Eigenvector centrality, Katz Broadcast centrality, Katz Receive centrality, and PageRank centrality etc., comes handy in solving these problems.

In this thesis, we propose different formulae for computing the strength for identifying top-K nodes and communities that can spread viral marketing messages into the network. The strength formulae are based on Katz Broadcast centrality, Resolvent matrix measure and Personalized PageRank measure. Moreover, the effects of intercommunity and intracommunity connectivity in ranking top-K communities are studied. Top-K nodes for spreading any message effectively into the network are determined by using Katz Broadcast centrality measure. Results obtained through this technique are compared with the top-K nodes obtained by using Degree centrality measure. We also studied the effects of varying α on the number of nodes in search space. In Algorithms 2 and 3, top-K communities are obtained by using Resolvent matrix and Personalized PageRank measure. Algorithm 2 results were studied by varying the parameter α.

Acknowledgements

I would like to express my sincere gratitude towards my advisor, Dr. Justin Zhan, for his continuous mentorship, motivation and support to drive me into research not only for my thesis but throughout my Master's program.

Furthermore, I would like to acknowledge Dr. Yoohwan Kim, Dr. Juyeon Jo, and Dr. Ge Lin Kan for being part of my thesis committee. I am thankful to Dr. Ajoy Datta, who has always been available to me whenever I needed guidance throughout my Master's program.

I am thankful to Vivek Gudibande, Rohit Raj, Priyanka Thota, Akshita Reddy, Sneha Goswami, Aditya Rajuladevi, Ashish Mukka, Ashish Tamrakar, Elliott Collin Ploutz, and Preveen Raj, who stayed with me and provided me with insightful suggestions and strength in tough times . I honestly thank all my fellow labmates, at Big Data Lab for their support.

I must thank my mom Varalakshmi Parsa and my brother Bhargav Parsa for their unconditional love, support, and care without whom I would not have been here.

Sai Phani Krishna Parsa

University of Nevada, Las Vegas

May 2017

Table of Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

Advancements in science and technology are narrowing the distance between people. Moreover, the availability of smartphones at cheaper prices catalysied this process. People are actively exchanging information on social networking platforms such as Facebook, Twitter, WhatsApp etc. A multitude of social networking websites started taking shape and are serving people in various ways. For instance, platforms such as Facebook, Twitter, WhatsApp etc., allows an individual to get in touch with others, who share similar idealogy. Platforms such as Meetup allows a group of people with similar idealogy to meet once in a while and allows to exchange ideas or share knowledge. Platforms such as LinkedIn, Indeed, Glassdoor are connecting employers to employees. Kickstarter platform allows individuals or groups of individuals to pitch their start up ideas and get funded from enthusiasts with same interests around the globe. Apart from these, there are tonnes of social networking platforms that enable an individual to interact with others at different levels.

As the number of individuals using these social networking platforms are increasing day-by-day, the amount of data generated from their interactions is also increasing exponentially. This has lead to a new trend in big data community researchers, to engage themselves in studying individual interaction networks, as these networks have proven to be excellent sources of hidden information patterns.

Researchers came up with intruiging techniques for answering complex questions such as product recommendations, friendship recommendations, web page ranking for efficient information retrieval etc., by using various techniques. Network theory concepts form the core of such techniques designed to uncover valuable insights. Especially In-degree centrality, Out-degree centrality, Betweenness centrality, Eigenvector centrality, Katz Broadcast centrality, Katz Receive centrality, and PageRank centrality etc., comes handy in solving these intruiging questions.

1.1 Objective

In this thesis, various network datasets were examined for answering interesting questions such as "what are the top-K nodes(users) that are suitable for broadcasting a viral message into the network?", "what are the top-K communities(groups of users) that are suitable for instantly spreading a viral message into distinct number of communities?". Different formulae for computing the strength of nodes and communities in a network graph were proposed by using various centrality measures such as Katz Broadcast centrality, Resolvent matrix and Personalized PageRank (PPR) measures. Moreover, the impacts of intracommunity connectivity and intercommunity connectivity on the strength of communities were studied as a part of this thesis work. This thesis work also includes the study of effects of parameters involved in Katz Broadcast centrality, and Resolvent matrix measure on the results obtained for top-K nodes and top-K communities. Furthermore, the results were eloborated for explaining the need for such measures.

1.2 Outline

In Chapter 1, a brief introduction of social networking platforms, and their advantages were discussed.

In Chapter 2, related works in the field of network theory, which address complex problems using network centrailty measures were discussed.

In Chapter 3, three different algorithms for computing the strength of top-K nodes, and top-K communities were proposed.

In Chapter 4, dataset characteristics and experimental results showing the effects of parameters in ranking top-K nodes and top-K communities were discussed.

In Chapter 5, proposed methods and their results were summaraized along with the possible extension of this thesis work.

Chapter 2

Background and Preliminaries

2.1 Related Work

Researchers have been conducting experiments towards identifying interesting patterns in large network datasets using network theory concepts. A significant amount of research has been carried out for the identification and ranking of top-K influential nodes that are capable of spreading viral information messages into a given network using various approaches such as Centrality theory [1], [2], [3], [3], [4], Diffusion models [5], Heat diffusion theory [6], Evidence theory [7] etc.

Li et al. [7] in their recent research paper published an evidence theory based method for identifying top-K nodes in a network of networks (NON). The central idea of their approach is to reduce a complex network(any) into a group of sub-networks. Kimura et al. in [8], proposed a novel method for the identification of influential nodes by combining bond percolation theory and graph theory. Doo et al. in [6], theorized an activity oriented influence model for social networks. They used heat diffusion concepts to characterize the influence propagation in real time social networks. Zhang et al. in [9], and Leung et al. in [10], came up with user preference based methods for identifying the top-K nodes. Zhang et al. in [9], identified top-K nodes using a two-staged approach called GAUP, which also includes the users preferences. In [10] Leung et al., proposed a MapReduce model for search space reduction by considering user-specified constraints for mining uncertain data. Jia-Lin He et al. in [11], proposed a community structure based influence maximization strategies in complex networks for identifying top-K nodes. In [12], Weiwei Liu et al. proposed a topic based novel appraoch for identifying the top-K nodes in a given network.

Network theory concepts can also be applied for identifying and ranking communities in large network graphs. In [13], Xie et al. came up with a unique spectral property based community

structure detection algorithm. In [14], Li et al. proposed a flooding time based approach for detecting influential communities in large networks. The amount of time taken to spread a given message from one node/community to the other node/community is known as flooding time [14], [15]. Another approach for detecting the most influential community is by identifying those nodes with radiates maximum information. Ma et al. in [16], proposed a similar method by using heat-diffusion processes. Another approach to identify influential communities is by computing the information diffusing power of boundary nodes. Boundary nodes in each community plays a vital role in information propogation to neighboring communities. In [17], Faisal et al. came up with a novel algorithm for boundary node detection in a cluster. This can be further extended for ranking influential communities. Sweeney et al. in [18] applied game theorey concepts for detecting communities in large datasets. Wu et al. in [19], described a new method, which uses distance centrality as a measure for detecting communities. Their approach is based on the most central nodes and their similarities with other nodes. In [20], Zhang and Wu proposed a core nodes approach for the local community detection.

Personalized PageRank (PPR) measure also gained importance for evaluating network topology. Larry Page and Sergey Brin in their research paper [21] proposed PageRank (PR) and Personalized PageRank (PPR) measures. PR is measure used to compute the global importance of vertices in a network. While, PPR measures the same for any vertex but with respect to a particular vertex instead of entier network. PPR considers nodal ties while computing the importance score of a vertex. This makes PPR more accurate measure than PR [22]. The applications of PR include efficient information retrieval for search queries and the applications of PPR include personalizing social search, product recommendations etc. Zhu et. al. in [23], proposed an incremental approach for PPR computation with accuracy awareness. In [24], Lofgren et. al., proposed a bi-directional search algorithm based on Frontier set. Other works on computing PPR include [25], [26], [27], [28], and [29].

2.2 Preliminaries

2.2.1 Degree Centrality

Degree centrality of a Vertex v is obtained as the count of distinct ties that v has with other vertices.

$$C_D(v) = d(v) \tag{2.1}$$

where *d(v)* in Equation (2.1) stand for the count of distinct ties that Vertex v has.

In the case of a directed network, there are two different degree centrality measures, which are in-degree and out-degree centralities [30]. In-degree and out-degree centralities of a Vertex v are obtained as the number of distinct ties that are directed towards it and the number of distinct ties that are directed outwards from it. Accordingly, the equations for them are as follows:

$$C_{in}(v) = d_{in}(v) \tag{2.2}$$

$$C_{out}(v) = d_{out}(v) \tag{2.3}$$

where $d_{in}(v)$ (in Equation (2.2)) and $d_{out}(v)$ (in Equation (2.3)) corresponds to the number of inward and outward ties of Vertex v.

2.2.2 Closeness Centrality

Closeness centrality for a Vertex v is obtained as the average shortest path value of all the shortest paths between v with respect to all others in the network [31]. Closeness centrality can be used as benchmark for measuring the time that a vertex takes to spread information into the network. It can be used to identify vertices that are capable of quickly spreading a rumor into the network. The equation for it is as follows:

$$C_c(v) = \sum_{j=1}^{N} \frac{1}{d(v, v_j)} \tag{2.4}$$

where C_c in Equation (2.4) gives the Closeness centrality of Vertex v.

2.2.3 Betweenness Centrality

Betweenness centrality measures the number of times a Vertex v acts as a connector along the shortest paths between any two other vertices in the network. By acting as a connector, any vertex has the power to govern the flow of information through it. Betweenness centrality was introduced by Linton Freeman. The betweenness of a Vertex v is given by the formula in Equation(2.5):

$$C_B(v) = \sum_{l=1, m \neq 1} \frac{g_{lvm}}{g_{lm}} \tag{2.5}$$

where g_{lvm} is all paths connecting vertices l and m through v_i; g_{lm} is the geodesic distance between the vertices l and m [1].

2.2.4 Eigenvector Centrality

Eigenvector centrality of a Vertex v represents its global influence in the network. It is based on the concept that if a vertex is connected to other highly connected vertices, then the current vertex in contex will gain a high influence value through these highly connected vertices.

2.2.5 Katz Centrality:

Katz centrality of a Vertex v_i measures its relative influence in the network graph by considering v_i's immediate neighboring vertices as well as vertices that are connected through these immediate neighboring vertices and so on. The Katz centrality of a Vertex v_i is computed as:

$$C_{Katz}(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} C_{Katz}(v_j) + \beta \tag{2.6}$$

where α is called damping factor [32] and its value is restricted by the condition $\alpha < 1/\lambda_1$. λ_1 is the largest eigenvalue for the adjacency matrix of the network [33]. Parameter β is called the exogenous vector and generally it is chosen to be $\mathbf{1}$ and it is used to ensure that each vertex has a minimum centrality so that it can be transferred to other nodes and so on.

The concept of using Katz centrality to rank the actors in a social graph was first proposed by Leo Katz in [34]. The very fact that a humans influence in his/her social group decreases as one moves further from his/her close connections to loosely connected distant members forms the base of Katz centrality. Katz centrality consideres all the possible walks in the network graph, irrespective of the fact that whether a given path is a shortest path or not. As the length of a path increases the influence of a vertex decreases and Katz centrality achieves this by utilizing the damping factor α which reduces the influence across longer paths. Consider a Node i, the immediate neighbors, i.e. walk of length *1*, are given the value α^1, whereas the farther neighbors, i.e. walk of length k, are assigned as α^k and so on [32].

$$(I - \alpha A)^{-1} = I + \alpha A + \alpha^2 A^2 + ... + \alpha^k A^k + ... = \sum_{k=0}^{\infty} \alpha^k A^k, \qquad 0 < \alpha < 1/\lambda \tag{2.7}$$

Equation (2.7) can be generalized for the entire graph as:

$$C_{Katz} = \beta(I - \alpha A^T)^{-1}.\mathbf{1} \tag{2.8}$$

where $(I - \alpha A^T)^{-1}$ is the resolvent matrix and $\mathbf{1}$ is a column vector of ones.

Resolvent matrix $[(I - \alpha A^T)^{-1}]_{ij}$ gives the influence of Node i on Node j. From the Equation (2.8) it is also evident that Katz centrality is dependent on α and β [15]. Benzi et al. in their paper [35], showed that different choices of α and β lead to different centrality values. If $\alpha \to 0+$, then Katz centrality reduces to degree centrality. The degree centrality of a node i gives importance to connections that are one step away starting from i. If $\alpha \to (1/\lambda)-$, then it reduces to eigenvector centrality, for example, if $\alpha = (1/\lambda)$ and $\beta = 0$, then Katz centrality is the same as eigenvector centrality. In short Katz centrality covers both the *local* and *global* influences of a node i, otherwise given by independently by degree and eigen vector centralities.

In the case of a directed network graph, there are two centrality measures, which are Katz Broadcast centrality and Katz Receive centrality.

Given a directed, unweighted graph $\boldsymbol{G = (V, E)}$ with adjacency matrix $\boldsymbol{A}$, Katz Broadcast and Katz Receive centralities of a vertex i are obtained as:

$$Katz_v^B = \beta(I - \alpha A)^{-1}.\mathbf{1} \tag{2.9}$$

$$Katz_v^R = \beta(I - \alpha A^T)^{-1}.\mathbf{1} \tag{2.10}$$

Clearly from the Equations (2.9) and (2.10), it is evident that we are considering row sums to obtain the outboundness of a node and column sums to obtain the inboundness of a node. In case of directed networks $[(I-\alpha A)^{-1}]_{ij}$ gives the broadcasting influence of i towards j and $[(I-\alpha A^T)^{-1}]_{ij}$ gives the receiving capacity of Node i when a message is triggered from Node j.

For the Figure (3.1), Katz Broadcast and Katz Receive centralities are computed as:

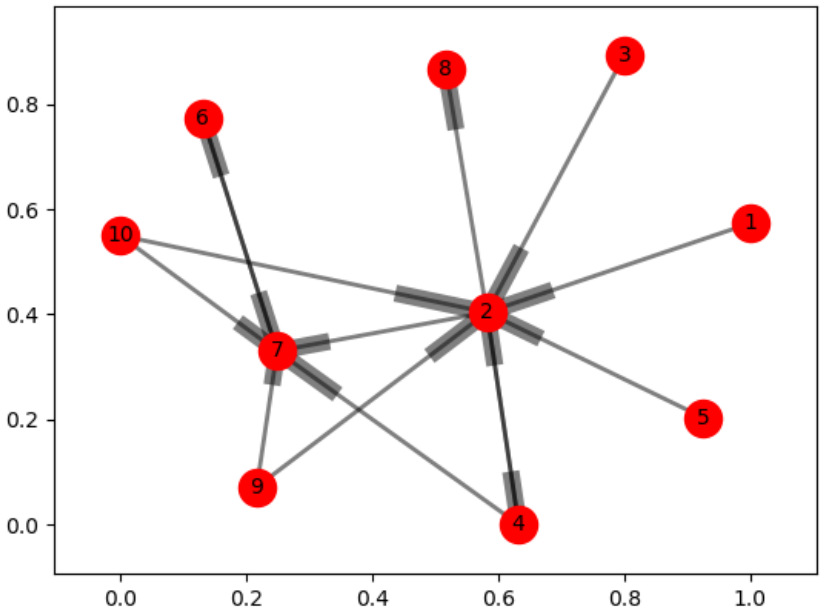


Figure 2.1: Sample Network Graph for Demonstrating Katz Centrality Computation

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

$$A^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$[I - \alpha * A]^{-1} = \begin{bmatrix} 1 & 3.063 & 0 & 2.604 & 0 & 14.754 & 17.357 & 2.604 & 0 & 0 \\ 0 & 3.604 & 0 & 3.063 & 0 & 17.357 & 20.420 & 3.063 & 0 & 0 \\ 0 & 3.063 & 1 & 2.604 & 0 & 14.754 & 17.357 & 2.604 & 0 & 0 \\ 0 & 3.063 & 0 & 3.604 & 0 & 17.357 & 20.420 & 2.604 & 0 & 0 \\ 0 & 3.063 & 0 & 2.604 & 1 & 14.754 & 17.357 & 2.604 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3.604 & 3.063 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3.063 & 3.604 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 3.063 & 0 & 2.604 & 0 & 17.357 & 20.420 & 2.604 & 1 & 0 \\ 0 & 3.063 & 0 & 2.604 & 0 & 17.357 & 20.420 & 2.604 & 0 & 1 \end{bmatrix}$$

$$[I - \alpha * A^T]^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3.063 & 3.604 & 3.063 & 3.063 & 3.063 & 0 & 0 & 0 & 3.063 & 3.063 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2.604 & 3.063 & 2.604 & 3.604 & 2.604 & 0 & 0 & 0 & 2.604 & 2.604 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 14.754 & 17.357 & 14.754 & 17.357 & 14.754 & 3.604 & 3.063 & 0 & 17.357 & 17.357 \\ 17.357 & 20.420 & 17.357 & 20.420 & 17.357 & 3.063 & 3.604 & 0 & 20.420 & 20.420 \\ 2.604 & 3.063 & 2.604 & 2.604 & 2.604 & 0 & 0 & 1 & 2.604 & 2.604 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where I is the Identity matrix, A is the Adjacency matrix of the given graph, A^T is the transpose of A, α value as 0.85 (λ_1 for the matrix A is 1), $[I - \alpha * A]^{-1}$ represents the Resolvent matrix for broadcasting ability and $[I - \alpha * A^T]^{-1}$ represents the Resolvent matrix for receiving ability. By choosing β value as $\mathbf{1}$ and solving for $\beta * [I - \alpha * A]^{-1}].\mathbf{1}$, and $\beta * [I - \alpha * A^T]^{-1}].\mathbf{1}$ Katz Broadcast and Katz Receive centralities are obtained as below:

$$Katz_{Broadcast} = \begin{bmatrix} 41.381381 \\ 47.507508 \\ 41.381381 \\ 47.048048 \\ 41.381381 \\ 6.666667 \\ 6.666667 \\ 1.000000 \\ 47.048048 \\ 47.048048 \end{bmatrix}$$

$$Katz_{Receive} = \begin{bmatrix} 1.000000 \\ 21.98198 \\ 1.000000 \\ 19.68468 \\ 1.000000 \\ 120.35736 \\ 140.42042 \\ 19.68468 \\ 1.000000 \\ 1.000000 \end{bmatrix}$$

2.2.6 PageRank Centrality

Named after its creator Larry Page, PageRank algorithm assigns a value to each vertex in a network as per the importance of that vertex in relation to the other vertices in the network [21]. For a Vertex u, F_u be the set of vertices that u points to, $N_u = |F_u|$ be the number of ties from Vertex u, B_u be the set of vertices that point to u, and c be a factor used for normalization such that the total importance score of all vertices is constant. PageRank of Vertex u is computed as:

$$PR(u) = (1 - c) + c \sum_{v \in B_u} \frac{PR(v)}{N_v} \tag{2.11}$$

where *PR(u)* is the PageRank of Vertex u, *PR(v)* is the PageRank value of Vertex v $\forall B_v$, and N_v is the number of forward links of each Vertex v $\forall$ B_v. This is the initial formula proposed by Page and Brin in [21].

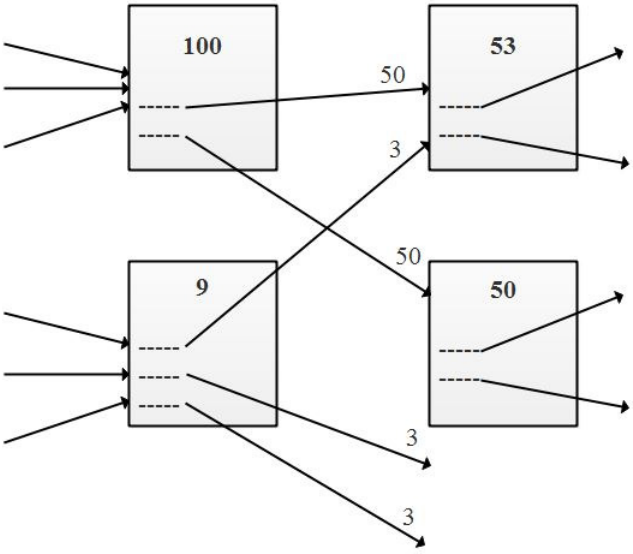


Figure 2.2: Sample Network of Web Pages Showing that PageRank Value of a web Page is Evenly Distributed Among its Outbound Links

From Figure (2.2), it can be observed that the rank of any vertex is obtained from its incoming vertices, and the rank of any vertex is distributed evenly among its outgoing ties. Page and Brin also discussed the possible problems with this simplified approach and proposed modifications for computing the PageRank of a vertex. Consider two Vertices u and v, that are pointing to each other only. Vertex w is pointing to Vertex u. Then, during the PageRank computation, rank is accumulated in the u and v loop only, and is never distributed to w. This is a sort of trap, which is called as a rank sink. To overcome the problem of rank sink in [21] Page and Brin proposed a modified formula for calculating PageRank values as:

$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + c * E(u) \tag{2.12}$$

such that c is is maximized and $|| R' || = 1$ *(*$|| R' ||_1$ *denotes the* L_1 *norm of R').*

In Equation (2.12), E(u) is a vector over all vertices that corresponds to a source of rank.

2.2.6.1 Power Iteration for Computing PageRank

The PageRank values of all the vertices in a network graph can be calculated by using Power Iteration approach. In this iterative approach, all the transition probabilities between the nodes are represented in the form of a matrix (known as Transition matrix). We will start with an initial

distribution of PageRank values. Generally an uniform distribution of PageRank values is used. The product of initial PageRank distribution and transition matrix gives the new PageRank values. These newly obtained values are used for computing the PageRank values in next iteration and so on. Let A be the transition matrix, such that $A_{uv} = \frac{1}{N_u}$, if Edge (u, v) exists, where N_u is the out-degree of Vertex u. Let π denote the initial distribution of PageRank values. As mentioned earlier, a uniform distribution is applied for the initial values, say $frac1|\ V\ |$, where $|\ V\ |$ is the number of vertices in the graph. Please note that π is a column vector and π^T is a row vector. The matrix multiplication of π^T and A gives the new PageRank values. As mentioned earlier, PageRank values are iteratively computed, until the values converge. Therefore, this multiplication process is repeated iteratively, each time considering new PageRank values and this process is repeated until the values converge.

$$\pi^{(1)T} = \pi^{(0)T}.A \tag{2.13}$$

where $\pi^{(0)T}$ is the initial PageRank distribution vector, $\pi^{(1)T}$ is the PageRank vector after first iteration. This process is repeated until the values converge.

$$\pi^{(2)T} = \pi^{(1)T}.A \tag{2.14}$$

$$\pi^{(k)T} = \pi^{(k-1)T}.A \tag{2.15}$$

$$\pi^{(k)T} = \pi^{(k)T}.A \tag{2.16}$$

where $\pi^{(k)T}$ is the PageRank vector and $\pi^{(k)T}$[u] represents the PageRank value of a Vertex u and so on.

2.2.6.2 Random Walk Perspective

Computation of PageRank values can be characterized by using random surfer approach. A random surfer starts his walk/tour from any Vertex i in the network graph and continues his walk/tour by randomly choosing one of the outbound vertices of i and so on. A random surfer may choose to start from any arbitrary point. All the vertices in the graph are given equal probability of getting selected for beginning a random walk/tour. The initial PageRank vector $\pi^{(0)T}$ captures the uniform distribution of a vertex getting selected for starting random walk/tour, the matrix A

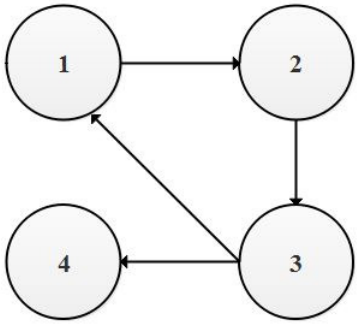


Figure 2.3: Sample Network Graph for Demonstrating PageRank and Personalized PageRank Computation

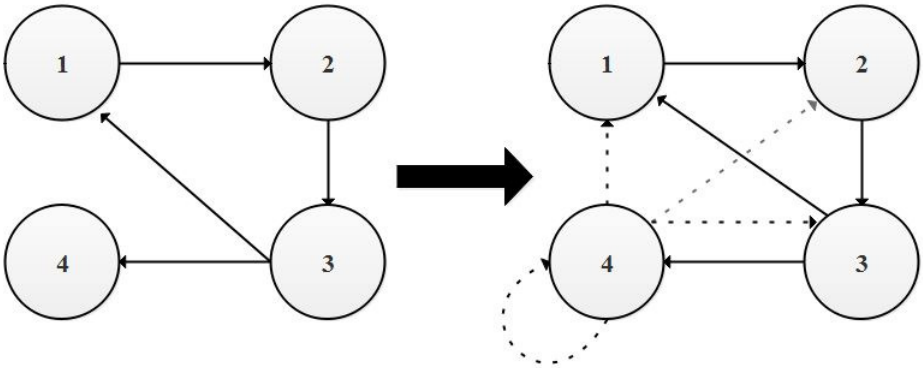


Figure 2.4: Sample Network Graph for Demonstrating Dangling Node Problem in Random Surfer Model

gives the probability of moving from one vertex to another vertex in the network. The PageRank values can be computed using Power Iteration method. The obtained PageRank values $\pi^{(k)T}[u]$ correspond to the probability of a random walker terminating the walk/tour at each Vertex u in network.

It is possible that a random surfer may get stuck at some vertex without any outbound ties. Such vertices without any outbound ties are termed as Dangling nodes. For example, consider Figure (2.3), where the random surfer choose to start his walk from Node *1*. From Node *1*, the surfer moved to Node *2*, and then to Node *3*, and then to Node *4*. At Node *4*, the surfer is stuck as there are no outbound edges for Node *4*. To address this problem, the transition matrix is modified as [36]:

$$S = A + d.w \tag{2.17}$$

where, d is the Dangling vector (a column vector) such that $d_i = 1$, if Node i is a Dangling node, else $d_i = 0$. $\mathbf{w}$ is a row vector of length $\mid V \mid$, containing uniform transition probabilities from Node i to all the nodes in the network [36]. After introducing transition probabilities for dangling nodes, the topology of above network graph is modified as shown in Figure (2.4).

However, in real time a user might not follow the random surfer approach and doesn't keep on clicking from one link to another. Users in real time may choose to move to any page on the Internet by entering the URL of that page. In order to capture this real time user behavior Equation (2.17) is modified as:

$$G = \alpha + (1 - \alpha).\mathbf{1}.\mathbf{v} \tag{2.18}$$

where $\leq \alpha \leq 1$ is a scalar, $\mathbf{1}$ is a column vector of ones, $\mathbf{v}$ is known as Personalization vector (row vector) and it contains the probability distribution of a random surfer teleporting to any random page without clicking links, the matrix $\mathbf{G}$ is called as Google matrix, α is the probability of moving from one page to another by clicking links, and 1-α is the probability with which the random surfer teleports to a random page without clicking links. α and 1-α are interchangeably used. Larry and Brin performed extensive experiments, in which they used α=0.85 and $\mathbf{v} = (\frac{1}{n},.,\frac{1}{n})$. Assigning a uniform probability distribution means that the web surfer can choose any of the web pages randomly, when not selecting the outbound links of a node. This matrix G, is used in power iteration method, along with the initial distribution of PageRank values to obtain a steady values for PageRanks of all the nodes in the network. The results obtained using Equation (2.16) are same as that of the results obtained using Equation (2.18).

2.2.6.3 Example of PageRank Computation for Small Network Data

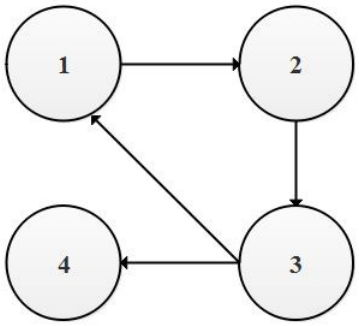


Figure 2.5: Sample Network Graph for Demonstrating PageRank and Personalized PageRank Computation

$$H = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$S = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

$$G = \begin{bmatrix} 0.05 & 0.85 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.85 & 0.05 \\ 0.45 & 0.05 & 0.05 & 0.45 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

$$Personalizationvector = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

$$InitialPR = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

$$PR = \begin{bmatrix} 0.21376215 & 0.26462229 & 0.3078534 & 0.21376215 \end{bmatrix}$$

2.2.6.4 Personalized PageRank

Personalized PageRank (PPR) is used to compute the reachability of all nodes in network with respect to a node. Mathematically the difference between PageRank and Personalized PageRank is that the Vector v in PageRank is populated with equal probability of moving to any random node by some means other than clicking the outgoing links. Whereas in Personalized PageRank, the Vector v can be manipulated in way that, the random surfer always moves to one node or a set of nodes of our interest, rather than moving to any node from all the nodes in the network. If we want the random surfer to move to a particular node, say Node i, then $\mathbf{v[i]} = 1$ and rest all are assigned as zero. The initial probabilities of a random surfer beginning the random walk at a any node can also be customised to begin the random walk from a particular Node i. The values obtained by using power iteration are the Personalized PageRank values of all nodes with respect to Node i.

2.2.6.5 Example of Personalized PageRank Computation

$$H = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$S = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

$$G = \begin{bmatrix} 0.05 & 0.85 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.85 & 0.05 \\ 0.45 & 0.05 & 0.05 & 0.45 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

$$Personalization vector with respect to vertex 1 = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

$$Initial PPR with respett to vertex 1 = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

$$PPR = \begin{bmatrix} 0.328 & 0.288 & 0.256 & 0.128 \end{bmatrix}$$

2.2.6.6 Applications of Personalized PageRank: Personalizing Search Results

Let us consider the scenario of social search where an User "A", who is interested in movies searches for another user named "John". If PageRank is applied then the top users with name "John" as their name and are having high global influence on other nodes, irrespective of their fields are retrieved. If Personalized PageRank is used, then all the users with "John" as their name and who are most influential in movies are retrieved.

Figure 2.6: Social Search Results when PageRank Measure is Used

Name	Description
John D	Actor|Host|Singer
Johnson A	Hollywood Actor|Producer
Johnny	Director|TV Series Producer
John Wright	Dancer|Play writer

Figure 2.7: Social Search Results when Personalized PageRank Measure is Used

2.2.6.7 Applications of Personalized PageRank: Product Recommendations

Let us consider the scenario of an e-Commerce website, recommending products to a customer "Doe" based on the history of items purchased by user "Doe" and other users who share similiar

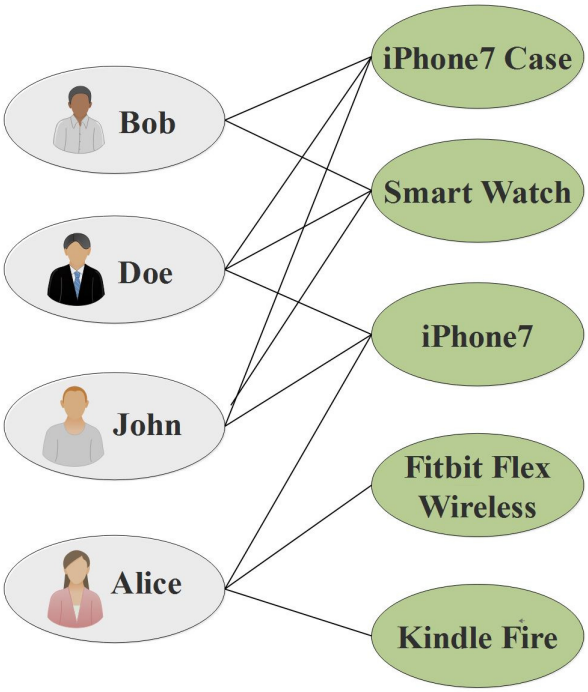


Figure 2.8: Bi-partite Graph for Demonstrating Product Recommendation Using Personalized PageRank

interestes as that of user "Doe". Customer-purchase-Product graph (with reverse edges) is used for this purpose. Customer-purchase-Product graph is a bipartite graph, where all the customers are represented in one set and all the products are represented on the other side. An edge exists between a customer and product only if the customer purchased the product. Since the firm wants to recommend products to an user "Doe", the random surfer will start from node "Doe". Interestingly, the random walk is very likely to touch the products purchased by "Doe", and other users who purchased those products, and also other products purchases by those users and so on. This walk is able to reach the users who are similar to "Doe" as they purchased common products. In addition, the walk can discovers frequently purchased products because they were purchased by the same users. Thus, PPR is used in this scenario to identify both similar users to "Doe" and products that might be of Does interest [37].

Chapter 3

Proposed System

3.1 Algorithm 1: Ranking Top-K Influential Nodes Using Katz Broadcast Centrality

In this section, we will discuss our first algorithm used for finding the top-K influential nodes. This work is an extension of topological analysis algorithm proposed by Sweta Gurung in her Master's thesis [32], where she used Katz centrality measure to analyze top-k nodes. In this section we evaluate her proposed algorithm with various datasets and analyze the results from two different perspectives. Given a network dataset, first the Katz Broadcast centralities of all the nodes are computed. Next, our algorithm checks whether each node satisfies two constraints for considering them into top-K node candidacy set. The first filtering constraint is user defined and it can be varied as per the users' choice. This constraint tests whether the Katz Broadcast centrality of a node is greater than that of user defined threshold value or not. The second filtering constraint tests whether the average of Katz Broadcast centrality values of a node and its immediate neighbors is greater than that of average Katz Broadcast centrality of the all the nodes in the network. The first filtering constraint helps the users' to focus only on the nodes of interest, while the second constraint is tested only for those nodes which satisfy the first one. Nodes, which satisfy both the constraints are included into top-K nodes set. Thus a finely refined set of nodes are returned to the user for executing the top-K query (where K is less than or equal to the number of nodes in top-K nodes set).

The first filtering constraint is denoted as *Const*, keeps the users' in control on the choice of nodes they are interested in. While the second filtering constraint prioritizes the nodes with more number of immediately highly connected nodes.

The average centrality values of a node and its neighbors is denoted as LAC_{Katz} (Local Average Centrality) and the average centrality value for the entire network is denoted as GAC_{Katz} (Global Average Centrality):

$$LAC(v_i) = \frac{C_{Katz}(v_i) + \sum_{j=1}^{n_i} C_{Katz}(v_j)}{n_i + 1} \tag{3.1}$$

$$GAC(G) = \frac{\sum_{i=1}^{n} C_{Katz}(v_i)}{n} \tag{3.2}$$

where n_i is the number of out-bound neighbors of v_i and n is the total number of nodes in the network.

Below is the Algorithm 1 for ranking top-K influential nodes in a network. The algorithm first computes the Katz Broadcast centrality values of each node present in the network. Then the algorithm tests whether each nodes' Katz Broadcast centrality value is greater than the user-defined *Const* or not. For each node satisfying the first constraint, the algorithm computes LAC value and checks if it is greater than GAC or not. If the second constraint is also satisfied, then the node is retrieved into top-K candidacy nodes list. Thus, the algorithm reduces the search space for running top-K nodes query for effectively broadcasting a given message into a network at a low cost.

Algorithm 1: Algorithm 1 for Ranking Top-K Influential Nodes Using Katz Broadcast Centrality

Input: Network Graph G, alpha α, betaβ, *Const*
for *each node* $v_i \in V$ **do**
    Calculate Katz Centrality, $C_{Katz}(v_i)$;
    if v_i *satisfies Const* **then**
        set $LAC(v_i) \leftarrow C_{Katz}(v_i)$;
        set $ngbrCount \leftarrow 0$;
        Find a list of its out-bound neighbors $v_j \in N^{out}$ and their $C_{Katz}(v_j)$;
        for *each* $v_j \in N^{out}$ **do**
            $LAC(v_i) \leftarrow LAC(v_i) + C_{Katz}(v_j)$;
            $ngbrCount \leftarrow ngbrCount + 1$;
        end
        $LAC(v_i) \leftarrow LAC(v_i)/(ngbrCount + 1)$;
        if $LAC(v_i) \geq GAC(G)$ **then**
            Return v_i and its $C_{Katz}(v_i)$;
        end
    end
end

3.1.1 Working on Karate Club Dataset

The Karate club dataset [38] consists of friendships between 34 members of a karate club at a U.S university in the 1970s. This undirected network dataset consists of 34 nodes and 78 edges. This dataset is obtained from Mark Newman network datasets repository.

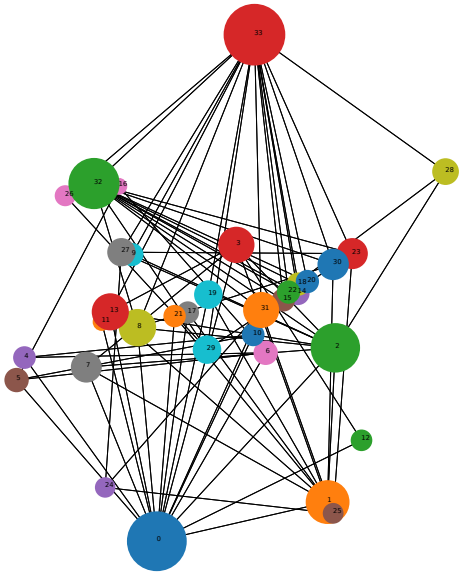


Figure 3.1: Algorithm 1 Working on Karate Club Membership Network

The karate club network visualization shown in Figure (3.1) was created by using D3.js (Data Driven Documents) [39]. Let $\mathbf{A}$ represent the adjacency matrix of this network. The largest eigenvalue λ obtained for A is 6.725 and to satisfy the constraint that $\alpha < \frac{1}{\lambda}$, α value is chosen to be less than 0.148. We also tested the algorithm by varying α value and with a β value kept constant as 1. Constraint *Const* is chosen as the sum of standard deviation of Katz centralities of all the nodes in network and average of Katz centralities of all the nodes in network. The second filtering constraint is set to the average of Katz centralities of all the nodes in the network. For all the datasets, the first filtering constraint uses the same formula as above and β value is set as 1. We studied the effect of α values on the number of nodes in search space and the ordering of top-K nodes.

Figure (3.2), clearly depicts that the filtering constraint are effective in filtering the unwanted nodes. The top-5 nodes obtained using the proposed algorithm are 33, 0, 32, 2 & 1 and the top most influencing node is 33, which in reality represents the president of the karate club, and the second most influencing node is 0, which in reality represents the instructor of the club. These are the most influential people of the club and they fought with each other, which eventually led to

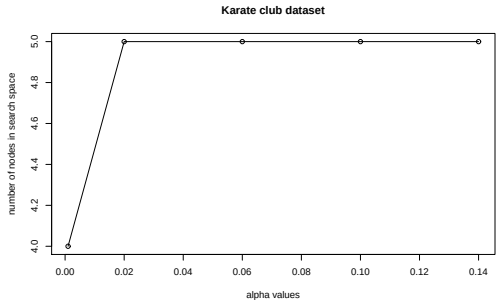


Figure 3.2: Graph Demonstrating the Relationship Between α Values and the Number of Nodes in Search Space

the seperation of the club into two factions aligned around the president and the instructor [40].

3.2 Algorithm 2: Resolvent Matrix Based Measure for Community Strength Detection

Consider a network of communities, and you are interested in spreading an important marketing message into the network at a low cost. The ideal approach for this problem is to identify important communities that are capable of spreading information effectively into other communities, instead of sending it to each and every community in the network. In this section, we propose a novel approach for computing and ranking them in the order of their broadcasting abilities. For each community we consider several parameters, which gives us insights about the communitys' capacity to broadcast a given message into the network.

For each community, inorder to compute strength we consider the following parameters: Firstly, we consider the communicability between the members of a community. This also accounts for community's connectivity, as communicability will be high if and only if the connectivity is high. For this purpose, we classify the members of a community as internal and boundary nodes. Internal nodes of a community are those nodes which do not share ties with nodes belonging to other communities. Boundary nodes of a community are those nodes which share atleast an edge with nodes beloning to other communities. As we are discussing abound spreading a viral marketing messages into different communities, boundary nodes play an important role as it is through these nodes that information diffuses into other communities. For spreading a rumor into the network, a message can be given to any node in the community. If that node is an internal node, then the message should reach the boundary nodes first and then it is trasmitted to other communities. If a community's internal nodes are not connected well to its boundary nodes, then there is a slight/ no chance of message being transmitted to the boundary nodes. In our strength formula, the first parameter considers this connectivity/communicability. The resolvent matrix $[(I - \alpha * A)]^{-1}$ measure gives the communicability between any two nodes (Resolvent matrix is used to compute Katz centrality). For a given graph we compute the resolvent matrix measure and get the score of each nodes' communicability to the boundary nodes of the same community. Sum of all such scores is taken and is represented as intracommunity communicability/connectivity.

Secondly, the main intention of marketing is to send a message into diverse set of communities i.e., to maximize the reach into distinct communities in the network at a low cost. For any community, this is captured as its intercommunity connectivity. For a given community, its intercommunity connectivity score is determined by three factors: number of distinct neighboring

communities, number of ties that the boundary nodes of this community share with the nodes of other communities and the number of distinct neighboring nodes with which the boundary nodes share ties with. By knowing the number of ties shared with distinct nodes of other communities, we compute the average of edges shared with a node and then multiply it with the distinct number of communities that these nodes are spread over.

Lastly, the factors obtained in Step 1 and 2 are multiplied to get the strength of a community.

Given a community C_i containing **n** nodes N = $\{v_1, v_2, ..., v_n\}$, **k** number of boundary nodes which are B = $\{u_{1b}, u_{2b}, ..., u_{kb}\}$, and surrounded by **l** number of distinct neighboring nodes belonging to **L** number of communities with **t** number of ties, then the strength of a community can be obtained as:

$$Strength(C_i) = \sum_{v_i \in N} \sum_{u_{ib} \in B} [(I - \alpha A)^{-1}]_{v_i u_{jb}} * \frac{L * t}{l} \tag{3.3}$$

Algorithm 2 for computing strengths and ranking communities is as follows:

Algorithm 2: Algorithm 2 for Community Strength Computation and Ranking

Input: Network Graph G, set of distinct communities C

Compute resolvent matrix $[(I - \alpha A)^{-1}]$;
for *each community* $C_i \in C$ **do**
    $B = computeBoundaryNodes()$;
    $L = obtainNeighboringCommunitiesCount()$;
    $l = distinctNeighboringNodes()$;
    $t = computeTiesSharedWithOtherCommunities()$;
end
for *each community* $C_i \in C$ **do**
    $Strength(C_i) = \sum_{v_i \in N} \sum_{u_{ib} \in B} [(I - \alpha A)^{-1}]_{v_i u_{jb}} * \frac{L*t}{l}$;
end
Sort the communities in the descending order of their strength values;
return top-K communities;

3.3 Algorithm 3: Personalized PageRank Based Measure for Community Strength Detection

In Algorithm 2, for determining intracommunity connectivity we used resolvent matrix measure. But, in reality the communicability/closeness between any two nodes varies as per the category or topic of message in consideration. For example given a community with three nodes A, B and C. Node A might share strong ties with node B on a particular topic and may not share stong ties with C on the same topic. Resolvent matrix doesn't take this case into consideration and is static irrespective of the topic/category of the message. Personalized PageRank captures this exact essence and to gain better understanding of intracommunity ties, we use PPR score as a measure. For computing the strength of a community, we use the same paramters as in Algorithm 2, except for resolvent matrix measure instead we use PPR measure. The equation for computing a community's strength is as follows: Given a community C_i containing **n** nodes N = {v_1, v_2, ..., v_n}, **k** number of boundary nodes which are B = {u_{1b}, u_{2b}, ..., u_{kb}}, and surrounded by **l** number of distinct neighboring nodes belonging to **L** number of communities with **t** number of ties, then the strength of a community can be obtained as:

$$Strength(C_i) = \sum_{v_i \in N} \sum_{u_{ib} \in B} [PPR]_{v_i u_{jb}} * \frac{L * t}{l} \tag{3.4}$$

Algorithm 3 for computing strengths and ranking communities is as follows:

Algorithm 3: Algorithm 3 for Community Strength Computation and Ranking

Input: Network Graph G, set of distinct communities C

for *each community* $C_i \in C$ **do**
    $B = computeBoundaryNodes()$;
    $L = obtainNeighboringCommunitiesCount()$;
    $l = distinctNeighboringNodes()$;
    $t = computeTiesSharedWithOtherCommunities()$;
end
for *each community* $C_i \in C$ **do**
    $Strength(C_i) = \sum_{v_i \in N} \sum_{u_{ib} \in B} [PPR]_{v_i u_{jb}} * \frac{L*t}{l}$;
end
Sort the communities in the descending order of their strength values;
return top-K communities;

Chapter 4

Experimental Results

The experiments for Algorithm 1 were performed on 16 GB main memory in Intel Xeon(R) CPU E5-1607 @ 3.00 GHz x 4 on Windows 7 Operating system platform. The language used to write these algorithms was Oracle Java 1.7 using Jblas [41] and Graph-stream [42] packages. The algorithms were written using the Java data structures like Lists & Hashmaps.

The experiments for Algorithms 2 and 3 were performed on 64GB main memory in Intel Xeon(R) CPU E5-1630 v4 @ 3.70 GHz on Windows 10 Operating system platform. The language used to write these algorithms was Python using Networkx [43] and Numpy [44] packages.

4.1 Datasets

For all the algorithms, network datasets were collected from Mark Newmann datasets [38], SNAP Stanford Large Network Database Collection [45] & ILAB-Data Centre [46]. Table (4.1) summarizes dataset characteristics.

Table 4.1: Characteristics of Network Datasets Used for Experimentation

Dataset	Type	Number of Nodes	Connectivity
Facebook	Undirected	1034	53498
CA-GrQc	Undirected	5242	14496
CA-HepTh	Undirected	9877	25998
Epinions-I	Directed	1247	51558
Epinions-II	Directed	1799	61037

4.1.0.1 Facebook Dataset

Facebook is an on-line social networking platform [47], where nodes represents the users and edges represents the relation between the users. A total of 1034 nodes and 53498 connections are present in this dataset. The largest eigenvalue of the network is ≈ 123.215. Keeping the fact in mind that the value of α should be less than $\frac{1}{\lambda_1}$ (0.008 in this case) in mind, the values for the parameter α values are varied as 0.0005, 0.001, 0.0015, 0.002, 0.0025, 0.003, 0.0035 ... 0.008.

4.1.0.2 Coauthorship Network Dataset for General Relativity and Quantum Cosmology Category

CA-GrQc dataset covers the scientific collaboration between the authors who submitted papers "General Relativity and Quantum Cosmology" category between January 1993 and April 2003 (124 months) [48]. If an author **a** worked on a paper in collaboration with another author **b**, then the graph contains an undirected edge from a to b. A total of 5242 nodes and 14496 connections are present in this network dataset. The largest eigenvalue is ≈ 45.616. As, the value of α should be less than 0.021, α values are varied as 0.005, 0.01, 0.015 & 0.02.

4.1.0.3 Coauthorship Network Dataset for High Energy Physics-Theory Category

CA-HepTh dataset covers the scientific collaboration between the authors who submitted papers submitted to "High Energy Physics - Theory" category between January 1993 to April 2003 (124 months) [48]. If an author **a** worked on a paper in collaboration with another author **b**, then the graph contains an undirected edge from a to b. A total of 9877 nodes and 25998 connections are present in this network dataset. The largest eigenvalue is ≈ 31.03485. As, the value of α should be less than 0.0322, α values are varied as 0.03, 0.02, & 0.01.

4.1.0.4 Epinions Network Datasets

Epinions.com [49] is a general consumer review site, where the members can choose to"trust" each other or not. A Web of Trust is formed basing up all the trust relationships interact and then combined with review ratings to determine which reviews are shown to the user. For the purpose of experimentation we used both Epinions-I and Epinions-II datasets. Both the datasets are that of directed graph. In Epinions-I dataset a total of 1247 nodes and 51558 connections are present.

In Epinions-II dataset a total of 1799 nodes and 61037 connections are present. As the largest eigenvalues of these two networks are ≈ 83.751, α values should be less than 0.011. For both the datasets, α values are varied as 0.001, 0.004, 0.007, and 0.011 and results are analyzed.

4.2 Algorithm 1: Results and Discussion

For the purpose of experimentation of Algorithm 1, we considered Facebook, CA-GrQc, Epinions-I and Epinions-II datasets. For all the datasets, we analyzed the results from two different perspectives. Firstly, we analyzed the relationship between α values against the number of nodes obtained for each α value. Secondly, we compare the top-K results of our algorithm with the top-K results of Degree centrality algorithm using intersection similarity as a measure and analyze the significance of our algorithm.

Intersection similarity (Intersection distance) captures the notion of union minus the intersection. Previously, Benzi et al. used intersection similarity measure in their research in [2].

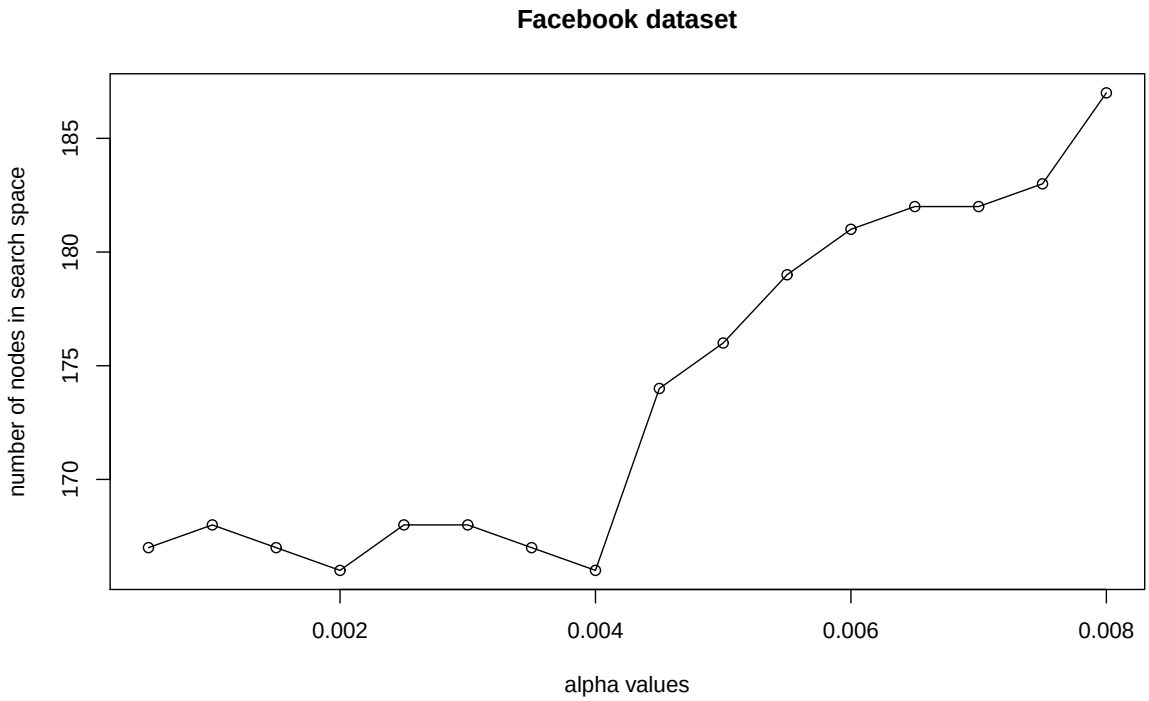
Let x_k and y_k be the top k ranked items in two ranked lists x and y respectively. Then the top k intersection similarity can be computed as

$$isim_k(x, y) := \frac{1}{k} \sum_{i=1}^{k} \frac{| \, x_i \Delta y_i \, |}{2i} \tag{4.1}$$

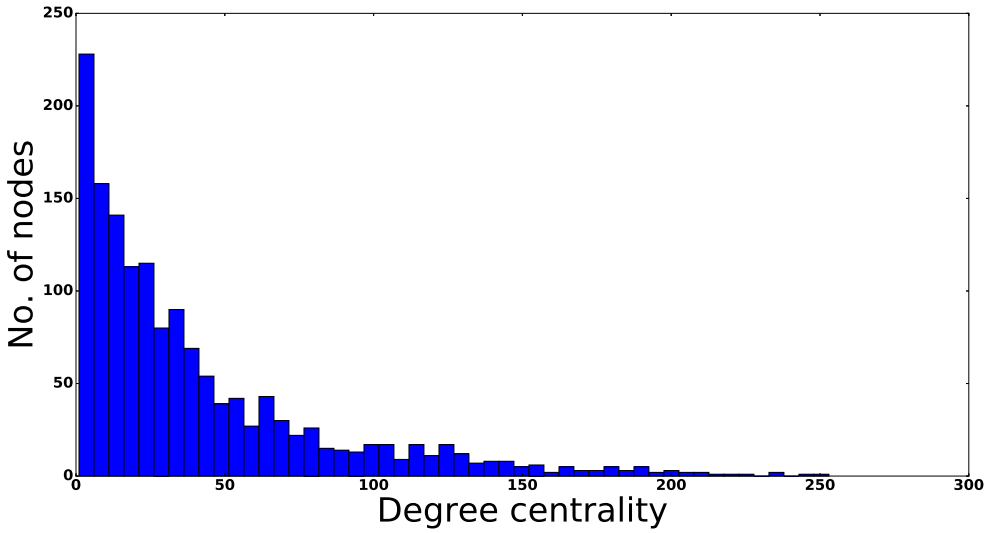
where Δ is the symmetric difference operator between the two sets. If the lists are identical, then $isim_k$(x, y) = 0 for all k. If the two sequences are disjoint, then $isim_k$= 1 [50], [2].

Figure (4.1(a)), shows the relationship between α values and the number of nodes in search space for Facebook dataset. The number of nodes in search space followed an increase- decrease pattern for α values between 0.0005 and 0.004. For α values between 0.004 and 0.008, the number of nodes in search space increased with an increase in α values. On the whole, there has been an increase in the number of nodes, with an increase in α value. This means with an increase in alpha values, the number of nodes that are capable of spreading an important marketing message into the network are increased. We can also draw further conclusions such as when α values are low, Katz Broadcast centrality tends to behave as degree centrality and there are less number of nodes in the search space, which indicates that there are a less number of highly connected actors in this network. But as α value is increased upto 0.008, the influence of those nodes which are connected to highly connected nodes also increased through these highly connected nodes.

Figure (4.2(a)), shows the relationship between α values and the number of nodes in search

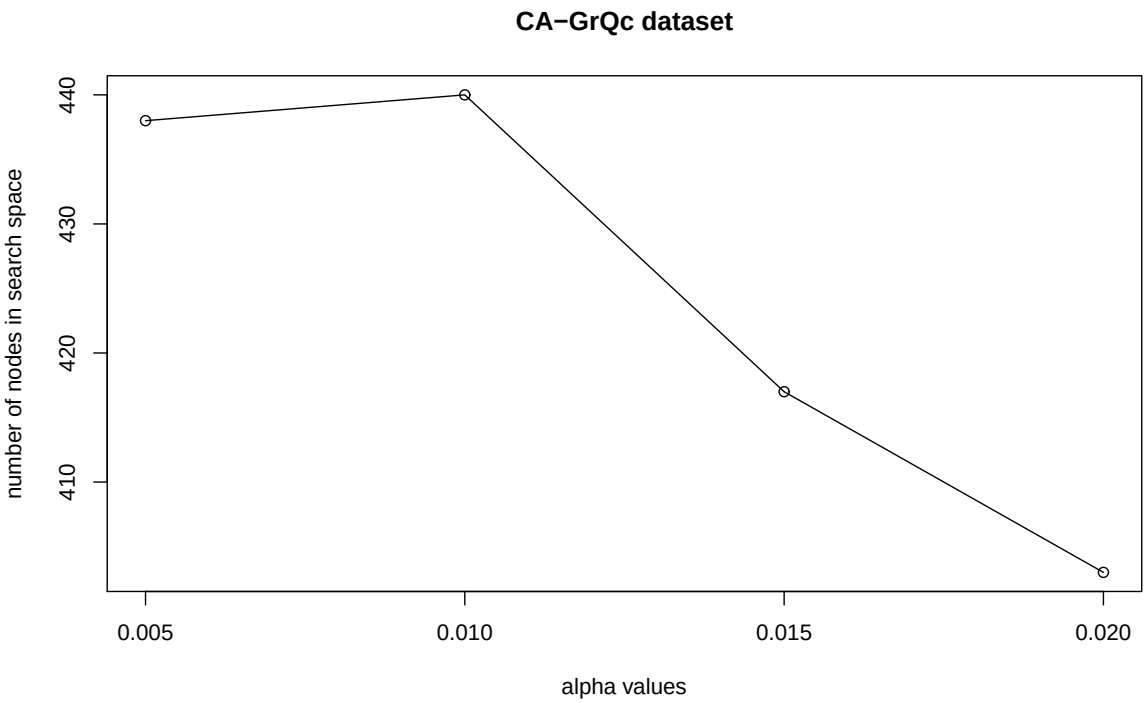


(a) Relation Between α and Number of Nodes in Search Space for Facebook Dataset

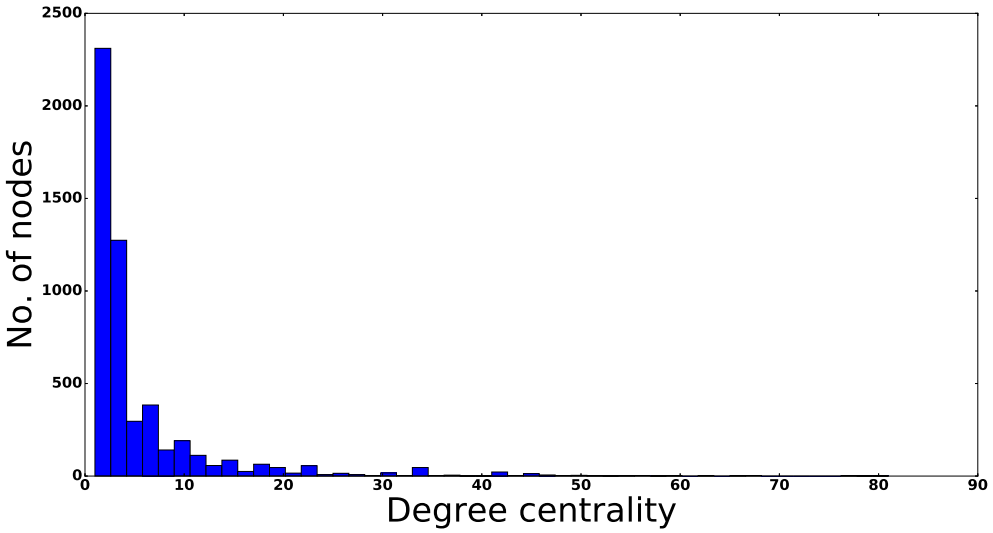


(b) Degree Centrality Frequency Distribution of Facebook Dataset

Figure 4.1: Experimental Results of Facebook Dataset

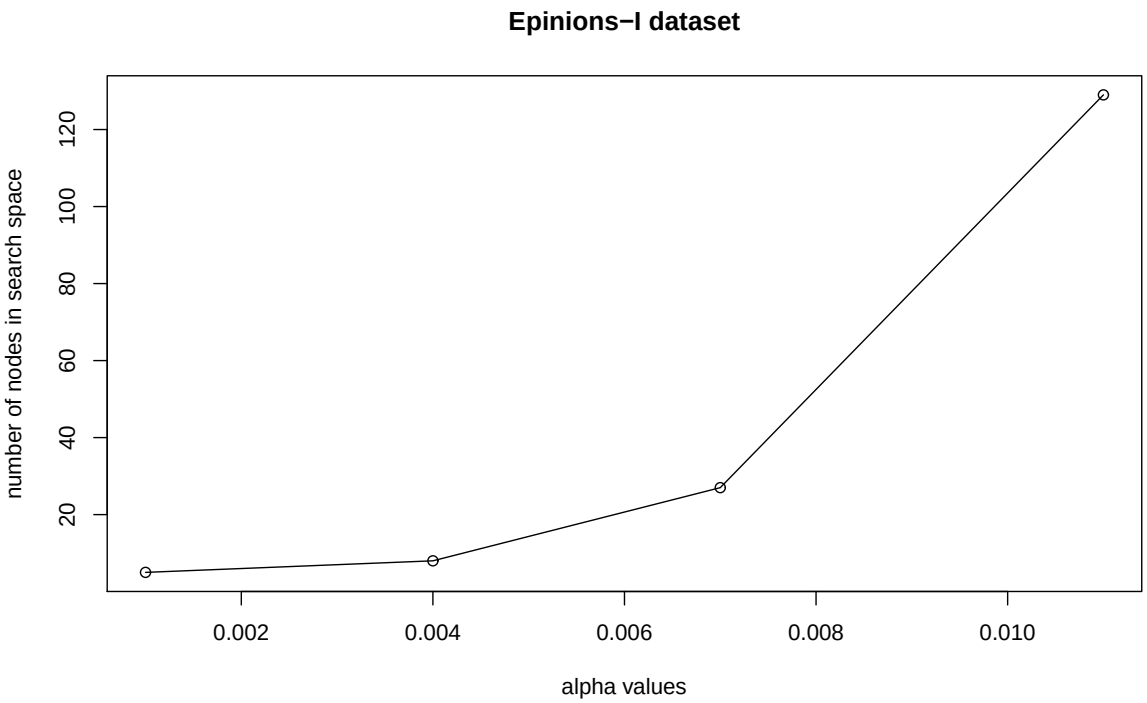


(a) Relation Between α and Number of Nodes in Search Space for CA-GrQc Dataset

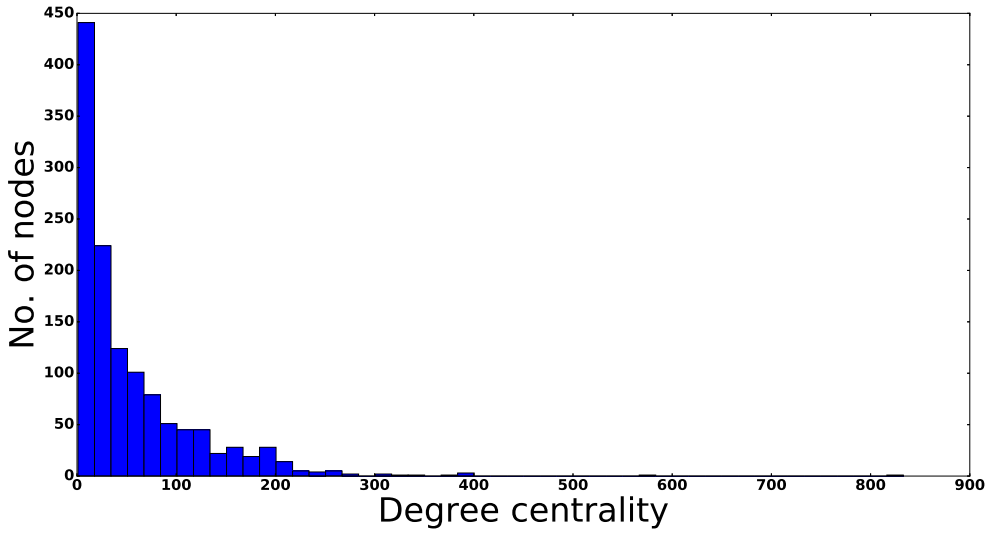


(b) Degree Centrality Frequency Distribution of CA-GrQc Dataset

Figure 4.2: Experimental Results of CA-GrQc Dataset

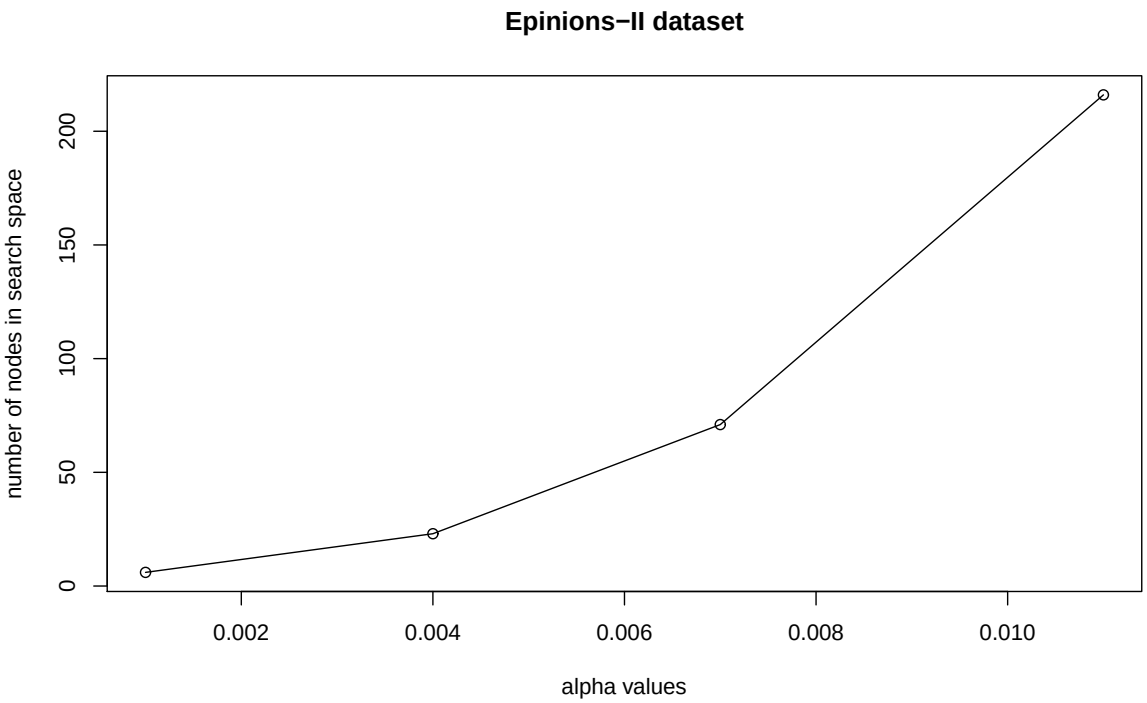


(a) Relation Between α and Number of Nodes in Search Space for Epinions-I Dataset

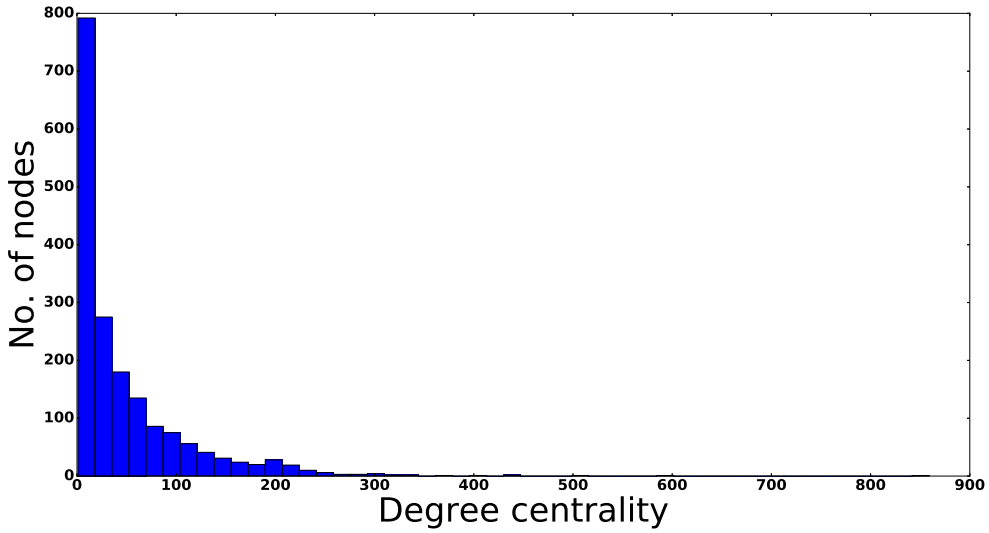


(b) Degree Centrality Frequency Distribution of Epinions-I Dataset

Figure 4.3: Experimental Results of Epinions-I Dataset



(a) Relation Between α and Number of Nodes in Search Space for Epinions-II Dataset



(b) Degree Centrality Frequency Distribution of Epinions-II Dataset

Figure 4.4: Experimental Results of Epinions-II Dataset

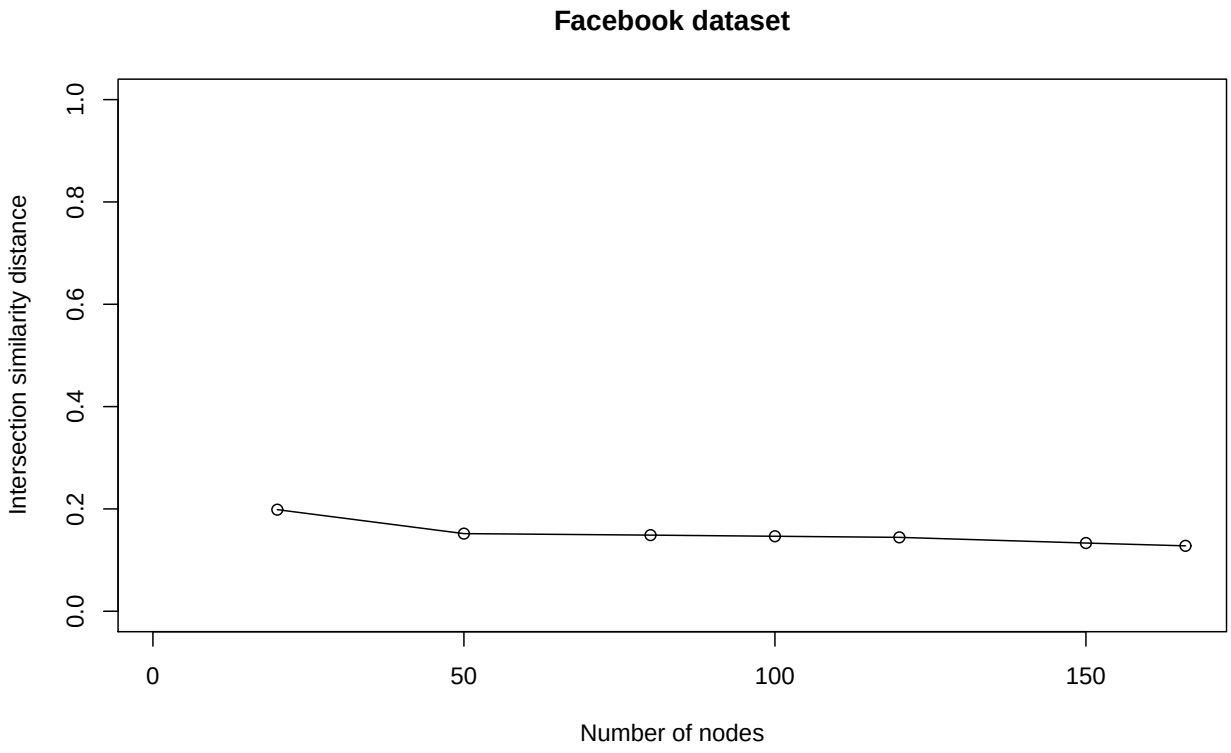


Figure 4.5: Intersection Similarity Distance Between Top-K Nodes Obtained Through Algorithm 1 and Degree Centrality Measure for Facebook Dataset

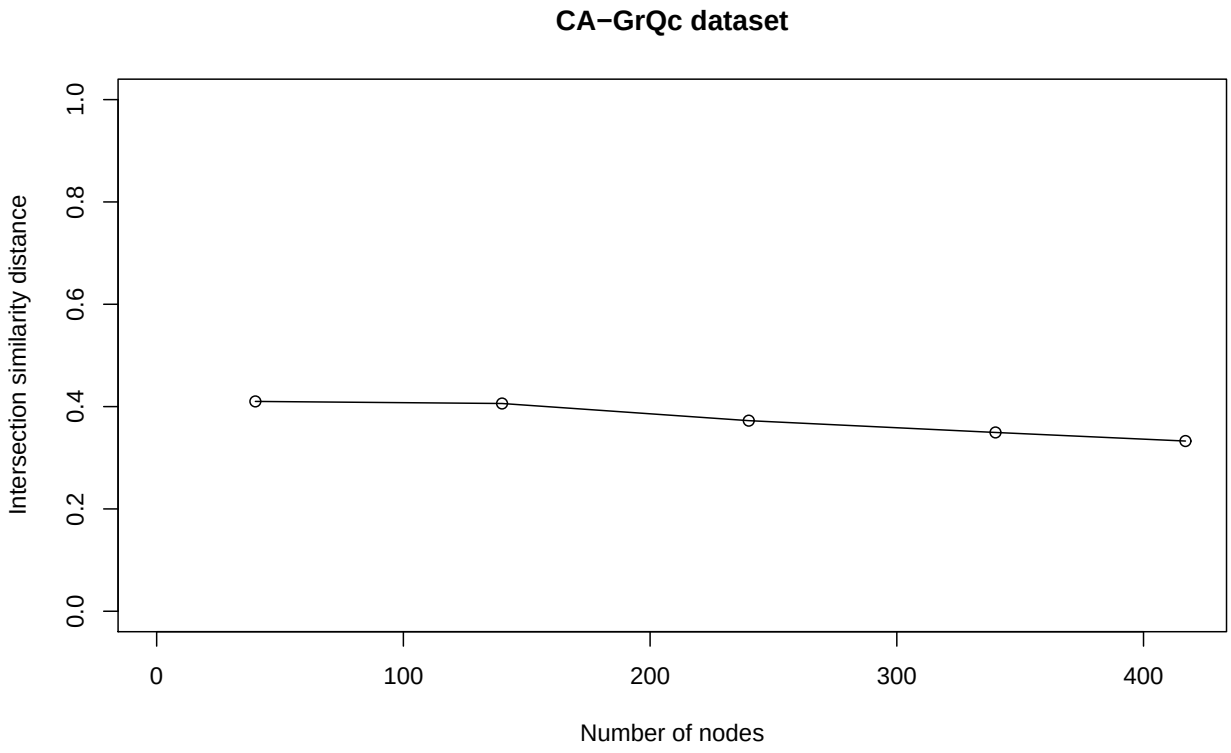


Figure 4.6: Intersection Similarity Distance Between Top-K Nodes Obtained Through Algorithm 1 and Degree Centrality Measure for CA-GrQc Dataset

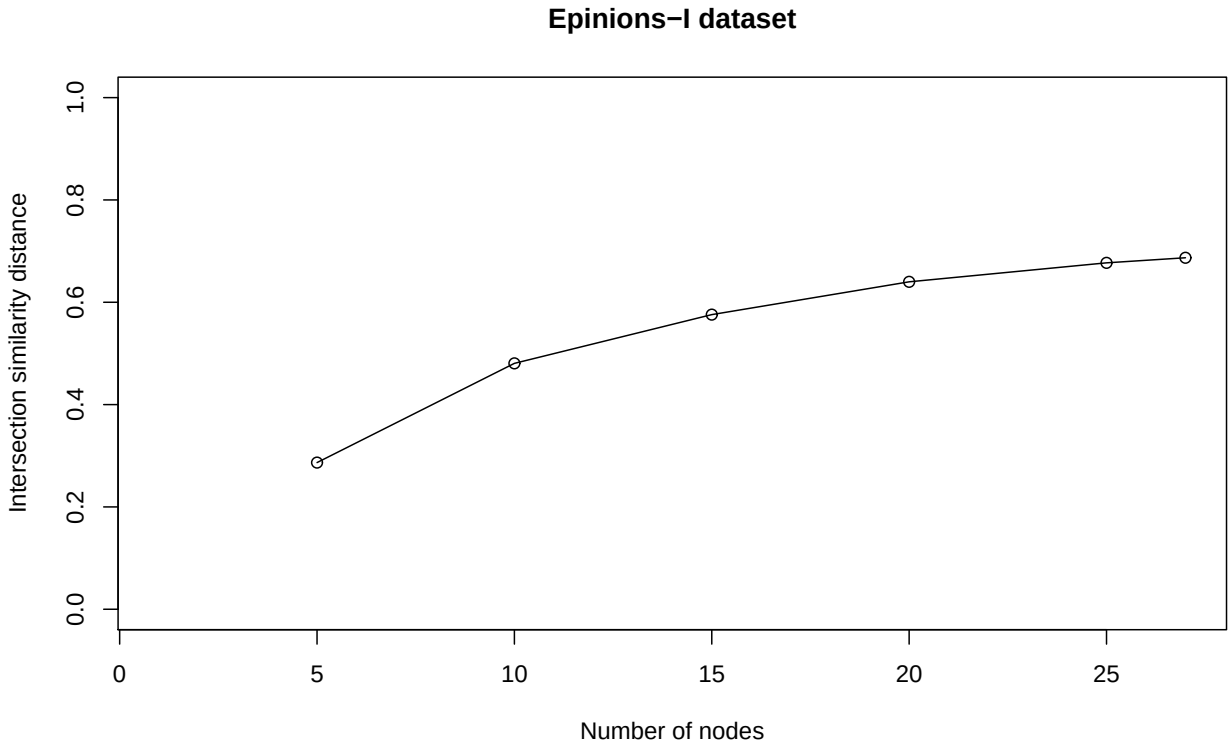


Figure 4.7: Intersection Similarity Distance Between Top-K Nodes Obtained Through Algorithm 1 and Degree Centrality Measure for Epinions-I Dataset

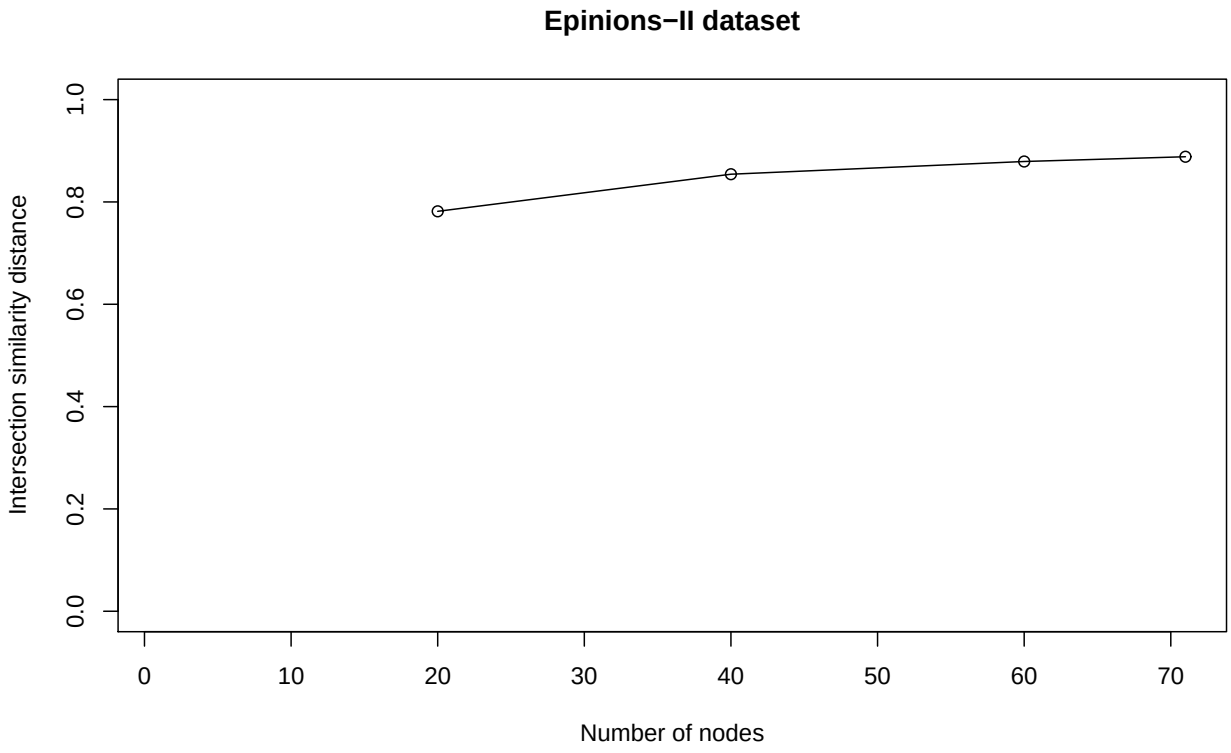


Figure 4.8: Intersection Similarity Distance Between Top-K Nodes Obtained Through Algorithm 1 and Degree Centrality Measure for Epinions-II Dataset

space for CA-GrQc dataset and unlike Facebook dataset, the number of nodes in search space decreased with an increase in *alpha* values (on the whole).

Figures (4.1(b)) and (4.2(b)) shows the degree distribution frequencies of Facebook and CA-GrQc datasets respectively. For Facebook dataset, a large number of nodes have small degree values or no degree values and yet there are a considerable number of nodes with high degree values. But, in case of CA-GrQc dataset a very large number of nodes have smaller degree values and the number of nodes with high degree values are negligible when compared to this. As mentioned before Katz centrality is a measure which captures both the local and global influences of a node. If the value of $\alpha \to 0^+$, then Katz centrality is approximately equal to that of Degree centrality. And as α values start moving from 0^+ to $\frac{1}{\lambda}^-$ Katz centrality values starts capturing the global influences of the nodes as well. As there are a very less number of nodes with high degree values, compared to the number of nodes with smaller degree values in case of CA-GrQc, the number of nodes that can exhibit local and global influence are very less than the number of nodes which can exhibit local influence (as $\alpha \to 0^+$). Hence, a decrease in the number of nodes in search space with an increase in α value. The converse of this can be observed in case of Facebook dataset, where there are a considerable number of nodes with higher degree values in comparison to those with a smaller degree values.

Figures (4.3(a)) and (4.4(a)), shows the relationship between α values and the number of nodes in search space for Epinions datasets. It can be seen that there is an overall increase in the number of nodes in search space with an increase in α value. The degree distribution frequencies for Epinions datasets, in Figures (4.3(b)) and (4.3(b)), are similar to that of Facebook dataset in Figure (4.1(b)). Hence, the relationship between α values and the number of nodes in search space is similar to that of in Facebook dataset.

Figures (4.5), (4.6), (4.7) and (4.8) shows the intersection similarity values for top-k nodes between degree centrality and our algorithm. For Facebook dataset, intersection similarity values are computed for top 20, 50, 80, 100, 120, 150 and 166 nodes, with α value as 0.004. It can be observed from Figure (5(a)), that in all the cases, intersection similarity values are around 0.2. For CA-GrQc dataset, intersection similarity values are computed for top 40, 140, 240, 340 nodes, with α value as 0.015. It can be observed from Figure (4.6), that in all the cases, intersection similarities are around 0.4. For Epinions-I dataset, intersection similarities are computed for top 5, 10, 15, 24, 25 and 27 nodes, with α value as 0.007. Intersection similarity values are increased with an increase in k value, with a maximum values around 0.6. For Epinions-II dataset, intersection

similarities are computed for top 20, 40, 60 and 71 nodes, with an α value as 0.007. On the whole, the intersection similarity values are around 0.8. Except Facebook dataset, experiments performed on the other datasets show that, intersection similarity values are morethan 0.4. This highlights the fact that there is a significant difference in the rankings produced by Degree centrality measure and our algorithm. Moreover, the top-3 nodes obtained in each case are same, but there is a considerable difference in rankings of the remaining nodes as our concept of giving importance to a node, based on Local Average Centrality proved to given importance to nodes with high local and global influence rather than nodes with high degree values. This confirms that the results obtained by using Degree centrality and our algorithm are different and both the approaches capture different perspectives in giving importance to nodes. Also, our approach gives more power to the user in choosing the parameters and narrowing the search space for running the top-K query. These results support our algorithm as a new method for ranking nodes in a given network.

4.3 Algorithm 2 and Algorithm 3: Results and Discussion

For testing both the Algorithms 2 and 3, we considered both the coauthorship network datasets i.e., CA-GrQc and CA-HepTh datasets collected from SNAP - Stanford Large Network Database Collection [45]. For CA-GrQc dataset α values are varied as 0.02, 0.015, 0.01, and 0.005 and for CA-HepTh dataset α values are varied as 0.03, 0.02, and 0.01. We have studied the impact of α values on top-K communities Strength values and their ranking.

For CA-GrQc dataset, when α value is set as 0.02, top-10 communities obtained are {1, 4, 3, 2, 6, 7, 9, 12, 13, and 18 }. In Table (4.2) (in all tables too) C stands for Community number, AD stands for Average Degree of nodes within community C, BN stands for Boundary Nodes in community C, CC stands for Community's Communicability(connectivity), NC stands for number of Neighboring Communities, AID stands for Average of Intercommunity Degree and ICC stands for InterCommunity Connectivity. Community number *1* is the largest community in the network. It has around 530 nodes. Clearly, from the tabluar values it is evident that Community number *1* is dominating other communities in all perspectives. For a fair comparison, the results for other communities in top-10 list were compared and evaluated against other communities. Consider Communities *2*, *3*, and *4*. As per Algorithm 2, Community number *4* stands first among the three, followed by Community number *3* and Community number *2*. Community number *4* has 117 boundary nodes (BN), where are communities *3* and *2* have 90 and 66 boundary nodes. The communicability of Community number *4* is highest than the other two and more over intercom-

munity connectivity is high for Community number *4*. Therefore, this led to a high strength value for Community number *4*. Communities *2* and *3* almost have same communicability scores but Community number *3* is surrounded by 23 distinct neighbors and this led to a high strength score for Community number *3* than Community number *2*. Communities *2* and *6* almost have same number of boundary nodes, but there's a considerable difference between their communicability scores. This led to a high strength value for Community number *2* than Community number *6*. Communities *6* and *7* also almost have same number of boundary nodes, but there is a huge difference in their communicability scores. Community number *7* is having more number of boundary nodes than Community number *9*, but Community number *9* has higher communicability score than *7*. But Community number *7* is surrounded by 20 distinct communities and this led to a higher strength for Community number *7* than Community number *9*. Though *9* and *12* have same number of boundary nodes and Community *12* has higher number of distinct neighboring communities, but still the communicability score of Community *9* dominates the strength values. The list is follwed by communities *13* and *18*.

Table 4.2: Top-10 Communities Obtained for CA-GrQc Network Dataset by Using Algorithm 2 with α Value as 0.02

C	AD	BN	CC	NC	AID	ICC	S
1	11.02099237	134	1188.532806	22	1.54679803	34.02955665	40445.24446
4	4.135278515	117	343.6467171	23	1.492063492	34.31746032	11793.08258
3	6.943620178	90	258.2605204	23	1.329341317	30.5748503	7896.27675
2	9.469230769	66	256.3660423	21	1.344262295	28.2295082	7237.087294
6	3.992673993	67	220.7227166	20	1.358208955	27.1641791	5995.751407
7	3.823529412	64	153.5518199	20	1.428571429	28.57142857	4387.194853
9	5.142857143	41	163.1969011	14	1.465116279	20.51162791	3347.434111
12	4.091954023	41	125.1378389	18	1.291139241	23.24050633	2908.266737
13	4.167741935	39	122.6259651	18	1.214285714	21.85714286	2680.253238
18	3.445544554	30	107.4995517	20	1.164556962	23.29113924	2503.787028

When α value is changed to 0.015, the top-10 communities obtained are same as that of when α value is 0.02. Table (4.3) contains the details about top-10 communities. But, it can observed that the communicability scores and strength values are decreased. Intracommunity connectivity is not impacted by the parameter α. As α value moves away from $\frac{1}{\lambda_1}$ the global influence will gradually starts to tend towards local influence.

Table 4.3: Top-10 Communities Obtained for CA-GrQc Network Dataset by Using Algorithm 2 with α Value as 0.015

C	AD	BN	CC	NC	AID	ICC	S
1	11.02099237	134	539.173674	22	1.54679803	34.02955665	18347.84108
4	4.135278515	117	325.7854878	23	1.492063492	34.31746032	11180.13055
3	6.943620178	90	248.0059327	23	1.329341317	30.5748503	7582.744265
2	9.469230769	66	211.3199069	21	1.344262295	28.2295082	5965.457044
6	3.992673993	67	209.6230512	20	1.358208955	27.1641791	5694.238106
7	3.823529412	64	146.9149649	20	1.428571429	28.57142857	4197.570424
9	5.142857143	41	151.9816223	14	1.465116279	20.51162791	3117.390486
12	4.091954023	41	118.6370062	18	1.291139241	23.24050633	2757.184093
13	4.167741935	39	116.8526768	18	1.214285714	21.85714286	2554.06565
18	3.445544554	30	103.220687	20	1.164556962	23.29113924	2404.127394

When α value is changed to 0.01, there is a slight change in the ordering of communities 6 and 2. Table (4.4) contains the results of top-10 communities. As α value is decreased futher, local influence starts to dominate the communicability score and this led for an increase in community number 6's communicability score. Overall, with a decrease in α there is a decrease in communicabilty and strength scores.

Table 4.4: Top-10 Communities Obtained for CA-GrQc Network Dataset by Using Algorithm 2 with α Value as 0.01

C	AD	BN	CC	NC	AID	ICC	S
1	11.02099237	134	415.1474506	22	1.54679803	34.02955665	14127.28369
4	4.135278515	117	309.7598829	23	1.492063492	34.31746032	10630.17249
3	6.943620178	90	238.6352166	23	1.329341317	30.5748503	7296.236022
6	3.992673993	67	199.5718027	20	1.358208955	27.1641791	5421.204192
2	9.469230769	66	189.6560877	21	1.344262295	28.2295082	5353.898082
7	3.823529412	64	140.8234063	20	1.428571429	28.57142857	4023.525895
9	5.142857143	41	142.2420328	14	1.465116279	20.51162791	2917.615649
12	4.091954023	41	112.6606925	18	1.291139241	23.24050633	2618.291536
13	4.167741935	39	111.5295835	18	1.214285714	21.85714286	2437.718039
18	3.445544554	30	99.23236613	20	1.164556962	23.29113924	2311.234857

When α value is changed to 0.005, the ordering of top-10 communities is same as that of when α value is 0.001. Table (4.5) contains the results of top-10 communities. Overall, there is a decrease in the communicabilty and strength scores with a decrease in α values.

For CA-HepTh dataset α values are varied as 0.03, 0.02, and 0.01. The top-10 communities when α value is set as 0.03 are {1, 2, 3, 6, 4, 5, 9, 8, 7, and 10}. Table (4.6) contains the results of top-10 communities. Community number *1* is the largest community in the network. From

Table 4.5: Top-10 Communities Obtained for CA-GrQc Network Dataset by Using Algorithm 2 with α Value as 0.005

C	AD	BN	CC	NC	AID	ICC	S
1	11.02099237	134	353.8520613	22	1.54679803	34.02955665	12041.42877
4	4.135278515	117	295.2388458	23	1.492063492	34.31746032	10131.84738
3	6.943620178	90	230.0000568	23	1.329341317	30.5748503	7032.217306
6	3.992673993	67	190.4053339	20	1.358208955	27.1641791	5172.204593
2	9.469230769	66	175.1793653	21	1.344262295	28.2295082	4945.227329
7	3.823529412	64	135.2047653	20	1.428571429	28.57142857	3862.993294
9	5.142857143	41	133.6483656	14	1.465116279	20.51162791	2741.345546
12	4.091954023	41	107.1342781	18	1.291139241	23.24050633	2489.854868
13	4.167741935	39	106.5951886	18	1.214285714	21.85714286	2329.866265
18	3.445544554	30	95.50144676	20	1.164556962	23.29113924	2224.337494

the tabe it is evident that Community number *1* dominated other communities in all aspects. For a fair comparison, remaining communities were compared with each other and the results are explained. Community number *2* is containing less number of boundary nodes than Community number *3*, but the communicability value of Community number *2* is higher than that of Community number *3*. Moreover, intercommunity connectivity of Community number *2* is greater than that of Community number *3*. Therefore, Community number *2* has higher strength than that of Community number *3*. Community number *3* clearly has more number of boundary nodes than Community number *6* and is also surrounded by more number of distinct neighboring communities. Therefore, Community number *3* is having higher strength than that of Community number *6*. Community number *4* is having less number of boundary nodes than that of Community number *5*, but the communicability value is high for Community number *4*. Eventhough Community number *4* is surrounded by less number of distinct communities than that are surrounded by Community number *5*, due to high intraconnectivity Community number *4* got higher strength than that of Community number *5*. Community number *9* has less number of boundary nodes than Community number *5*, but it has more number of distinct neighboring communitie than Community number *5*. But the intraconnectivity of Community number *5* dominated the strength value of Community number *9*. Though Community number *8* is having slightly higher connectivity than that of Community number *9*, it is surrounded by less number of distinct neighboring communities than taht of Community number *9*. Therefore, it stands after Community number *9* in the top-10 list. The list is followed by the communities *7* and *10*.

When α value is hanged to 0.02, the top-10 communities obtained are {1, 2, 3, 6, 5, 9, 4, 8, 7, and 10}. Table (4.7) contains the results of top-10 communities. When compared to the top-

Table 4.6: Top-10 Communities Obtained for CA-HepTh Network Dataset by Using Algorithm 2 with α Value as 0.03

C	AD	BN	CC	NC	AID	ICC	S
1	6.30162413	815	3845.779642	28	2.002970297	56.08316832	215683.5069
2	4.8	179	790.1971989	26	1.36318408	35.44278607	28006.79027
3	4.549248748	202	726.7088699	25	1.409703504	35.2425876	25611.10101
6	3.975708502	152	524.1011068	22	1.592741935	35.04032258	18364.67185
4	6.002336449	123	593.630283	20	1.456066946	29.12133891	17287.30866
5	4.234215886	153	492.3232998	24	1.362318841	32.69565217	16096.83137
9	4.152694611	111	389.8454439	25	1.47	36.75	14326.82006
8	4.230563003	92	399.615553	21	1.5	31.5	12587.88992
7	5.258169935	95	352.2145429	24	1.294117647	31.05882353	10939.36933
10	4.182926829	88	335.2440401	22	1.283505155	28.2371134	9466.323978

10 communities obtained when α value is 0.03, there is a slight change in the ordering of top-10 communities when α value is 0.02. Communities *5* and *9* dominated Community number *4*. In case of Community number *5*, it has high intraconnectivity and interconnectivity than Community number *4*. Whereas, in Community number *9*, a high interconnectivity dominated the strength of Community number *4*. Overall, there has been a decrease in the communicability values of all communities and also as communicability decreased, the strength of all communities decreased too.

Table 4.7: Top-10 Communities Obtained for CA-HepTh Network Dataset by Using Algorithm 2 with α Value as 0.02

C	AD	BN	CC	NC	AID	ICC	S
1	6.30162413	815	2919.149217	28	2.002970297	56.08316832	163715.1369
2	4.8	179	687.6789698	26	1.36318408	35.44278607	24373.25861
3	4.549248748	202	640.9837889	25	1.409703504	35.2425876	22589.92733
6	3.975708502	152	472.3361629	22	1.592741935	35.04032258	16550.81151
5	4.234215886	153	445.0475266	24	1.362318841	32.69565217	14551.11913
9	4.152694611	111	351.2981508	25	1.47	36.75	12910.20704
4	6.002336449	123	422.8957683	20	1.456066946	29.12133891	12315.29099
8	4.230563003	92	343.1370974	21	1.5	31.5	10808.81857
7	5.258169935	95	317.0311647	24	1.294117647	31.05882353	9846.614998
10	4.182926829	88	299.3275179	22	1.283505155	28.2371134	8452.145068

When α value is 0.01, the top-10 communities are same as that of when α value is 0.02. Table (4.8) contains the results of top-10 communities. Overall, there is a decrease in the communicability values of each community and this led to a decrease in the strength values of the communities.

Table 4.8: Top-10 Communities Obtained for CA-HepTh Network Dataset by Using Algorithm 2 with α Value as 0.01

C	AD	BN	CC	NC	AID	ICC	S
1	6.30162413	815	2382.064753	28	2.002970297	56.08316832	133593.7385
2	4.8	179	609.7515876	26	1.36318408	35.44278607	21611.29507
3	4.549248748	202	575.3777843	25	1.409703504	35.2425876	20277.80197
6	3.975708502	152	430.2217712	22	1.592741935	35.04032258	15075.10964
5	4.234215886	153	407.1764968	24	1.362318841	32.69565217	13312.90111
9	4.152694611	111	320.0423497	25	1.47	36.75	11761.55635
4	6.002336449	123	378.9504537	20	1.456066946	29.12133891	11035.54459
8	4.230563003	92	300.5822917	21	1.5	31.5	9468.34219
7	5.258169935	95	288.281791	24	1.294117647	31.05882353	8953.693274
10	4.182926829	88	271.4274726	22	1.283505155	28.2371134	7664.328324

Unlike Algorithm 2, Algorithm 3 uses PPR measure to compute intracommunity communicabilty. Table (4.9) shows the strength values of top-10 communities for CA-GrQc dataset. For Ca-GrQc dataset, the number of communities obtained are 379, out of which only 80 communities contain atleast 10 nodes. The largest community is Community number *1* with 524 members. From the table, it is evident that Community number *1* dominated all the other communities in all aspects, because it is the largest community in the network. For a fair comparison, communities *2*, *3* and *4* communities which contain 377, 337 and 260 nodes were compared. From the table, it can be seen that Community number *2* is having the highest average degree for the nodes with in the community. But, the comminication capacity of Community number *2* is less than those of communities *4* and *3*. This is beacause communities *4* and *3* have more number of boundary nodes that act as bridges to carry information into other communities. Moreover, communities *4* and *3* are driving information into 23 distinct communities, where as Community number *2* is capable of driving information into 21 distinct communities only. Thus the ordering of communities *4*, *3* and *2* are justified. CC and ICC values for these three communtiies show that Community number *4* is top among these three, where as Community number *3* stands second in the list, where as Community number *2* is in last position. Similarly, communities *3* and *4*, *6* and *7*. Both these sets have same number of neighboring communities which are 23 and 20. But in case of communities *3* and *4*, the communication capacity and intercommunity connectivity of Community number *4* is greater than that of Community number *3*. Thus the ranking of Community number *4* before Community number *3* is justified. Where as, between communities *6* and *7*, Community number *6* has a larger CC value compared to Community number *7*, but ICC value for Community number *6* is slightly lower than that of Community number *7*. But this doesn't affect the ranking

of Community number *6* as it has a higher CC than Community number *7*. The same reasoning can be extended to other communities in top-10 list as well. Overall, there has been a decrease in strength as the CC value tends to decrease, even the same trend can be seen in ICC values except for Community number *6*.

Table 4.9: Top-10 Communities Obtained for CA-GrQc Network Dataset by Using Algorithm 3

C	NC	BN	CC	AD	AID	ICC	S
1	22	134	528.603048	11.02099237	1.54679803	34.02955665	17988.12737
4	23	117	80.77957644	4.135278515	1.492063492	34.31746032	2772.149909
3	23	90	56.53926046	6.943620178	1.329341317	30.5748503	1728.679425
2	21	66	43.9666455	9.469230769	1.344262295	28.2295082	1241.15678
6	20	67	33.46895159	3.992673993	1.358208955	27.1641791	909.1565954
7	20	64	19.25017256	3.823529412	1.428571429	28.57142857	550.0049304
9	14	41	15.70058298	5.142857143	1.465116279	20.51162791	322.0445161
12	18	41	8.154229229	4.091954023	1.291139241	23.24050633	189.508416
13	18	39	7.102461137	4.167741935	1.214285714	21.85714286	155.2395077
11	17	45	5.636477931	4.226804124	1.217391304	20.69565217	116.6505868

Table (4.10) contains the top-10 communities for CA-HepTh dataset. For CA-HepTh dataset, the top-10 communities obtained are {1, 2, 3, 6, 5, 4, 8, 9, 7, and 10 }. When compared to the ordering of top-10 communities obtained using Algorithm 2, there is a slight change in the ordering of the top-10 communities. This suggests that there is difference in the perspectives used for ranking communities using Algorithm 2 and Algorithm 3.

Table 4.10: Top-10 Communities Obtained for CA-HepTh Network Dataset by Using Algorithm 3

C	AD	BN	CC	NC	AID	ICC	S
1	6.30162413	815	3662.786388	28	2.002970297	56.08316832	205420.6655
2	4.8	179	134.5142799	26	1.36318408	35.44278607	4767.560846
3	4.549248748	202	91.87821537	25	1.409703504	35.2425876	3238.026054
6	3.975708502	152	48.56004891	22	1.592741935	35.04032258	1701.559778
5	4.234215886	153	42.10539279	24	1.362318841	32.69565217	1376.663277
4	6.002336449	123	35.42243831	20	1.456066946	29.12133891	1031.548831
8	4.230563003	92	28.6953939	21	1.5	31.5	903.9049078
9	4.152694611	111	19.28098296	25	1.47	36.75	708.5761239
7	5.258169935	95	20.12306091	24	1.294117647	31.05882353	624.9985976
10	4.182926829	88	11.75829817	22	1.283505155	28.2371134	332.020399

Chapter 5

Conclusion and Future Work

With an exponential rise in the amount of data being generated, the interest to analyze such massive datasets is also increasing. This thesis study comprises of three different algorithms for computing strength of nodes and communities to rank them. Algorithm uses Katz Broadcast centality based measure to identify top-K influential nodes. This is a user centric algorithm, where the parameters such as α and β can be adjusted by the user to select the best fit top-K nodes. Algorithms 2 and 3, are for computing the strength of communities in order to rank and retrieve the top-K communities that are capable of effectively spreading a viral marketing message into other communities. Both the algorithms consider intracommunity and intercommunity connectivity in ranking the communities.

As Katz Broadcast centrality measure and Resolvent matrix measure are parameter (α) dependent, experimental results are carried out by varying (α), in order to facilitate the choice of α value. Evaluation is carried out by considering the Degree centrality distribution of the network and experimental results shows the relation between Degree centrality frequency distribution and the choice of α value. Moreover, experimental results show that the number of nodes obtained in search space are decreased by a factor of atleast 75 percent. This shows the effectiveness of the proposed algorithm. The same has been applied to Algorithm 2, and experiments are carried out by varying α values to understand the concept of local and global influence on communities strength values. Experimental results are evaluated and a detailed explanation has been provided to facilitate the understanding of the relation between communities ranking and α values. Algorithm 3 is parameter independent and the results and a detailed explanation has been provided for the experimental results.

In future, Algorithms 2 and 3 can be studied on diverese datasets to understand the behavior of communities in both directed and undirected datasets. Another area for improvising all the

algorithms is to incorporate real time data with network topology for finding topic/category wise top-K communities. Advanced techniques such as Deep Learning can also be applied to understand any hidden properties of network topologies.

Bibliography

[1] E. Yan and Y. Ding, "Applying centrality measures to impact analysis: A coauthorship network analysis," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 10, pp. 2107–2118, Oct. 2009. [Online]. Available: http://dx.doi.org/10.1002/asi.v60:10

[2] M. Benzi and C. Klymko, "Total communicability as a centrality measure," *CoRR*, vol. abs/1302.6770, 2013. [Online]. Available: http://arxiv.org/abs/1302.6770

[3] M. Borassi, P. Crescenzi, and A. Marino, "Fast and simple computation of top-k closeness centralities," *CoRR*, vol. abs/1507.01490, 2015. [Online]. Available: http://arxiv.org/abs/1507.01490

[4] M. Aprahamian, D. J. Higham, and N. J. Higham, "Matching exponential-based and resolvent-based centrality measures," *Journal of Complex Networks*, vol. 4, no. 2, p. 157, 2016. [Online]. Available: +http://dx.doi.org/10.1093/comnet/cnv016

[5] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '03. New York, NY, USA: ACM, 2003, pp. 137–146. [Online]. Available: http://doi.acm.org/10.1145/956750.956769

[6] M. Doo and L. Liu, "Extracting top-k most influential nodes by activity analysis," in *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, Aug 2014, pp. 227–236.

[7] M. Li, Q. Zhang, Q. Liu, and Y. Deng, "Identification of influential nodes in network of networks," *CoRR*, vol. abs/1501.05714, 2015. [Online]. Available: http://arxiv.org/abs/1501.05714

[8] M. Kimura, K. Saito, R. Nakano, and H. Motoda, "Extracting influential nodes on a social network for information diffusion," *Data Mining and Knowledge Discovery*, vol. 20, no. 1, p. 70, 2009. [Online]. Available: http://dx.doi.org/10.1007/s10618-009-0150-5

[9] J. Zhou, Y. Zhang, and J. Cheng, "Preference-based mining of top-k influential nodes in social networks," *Future Gener. Comput. Syst.*, vol. 31, pp. 40–47, Feb. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.future.2012.06.011

[10] C. K. S. Leung, R. K. MacKinnon, and F. Jiang, "Reducing the search space for big data mining for interesting patterns from uncertain data," in *2014 IEEE International Congress on Big Data*, June 2014, pp. 315–322.

[11] J.-L. He, Y. Fu, and D.-B. Chen, "A novel top-k strategy for influence maximization in complex networks with community structure," *PLoS ONE*, vol. 10.12, 2015. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0145283

[12] W. Liu, Z. H. Deng, L. Cao, X. Xu, H. Liu, and X. Gong, "Mining top k spread sources for a specific topic and a given node," *IEEE Transactions on Cybernetics*, vol. 45, no. 11, pp. 2472–2483, Nov 2015.

[13] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Comput. Surv.*, vol. 45, no. 4, pp. 43:1–43:35, Aug. 2013. [Online]. Available: http://doi.acm.org/10.1145/2501654.2501657

[14] R.-H. Li, L. Qin, J. X. Yu, and R. Mao, "Influential community search in large networks," *Proc. VLDB Endow.*, vol. 8, no. 5, pp. 509–520, Jan. 2015. [Online]. Available: http://dx.doi.org/10.14778/2735479.2735484

[15] J. Zhan, V. Guidibande, and S. P. K. Parsa, "Identification of top-k influential communities in big networks," *Journal of Big Data*, vol. 3, no. 1, p. 16, 2016. [Online]. Available: http://dx.doi.org/10.1186/s40537-016-0050-7

[16] H. Ma, H. Yang, M. R. Lyu, and I. King, "Mining social networks using heat diffusion processes for marketing candidates selection," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ser. CIKM '08. New York, NY, USA: ACM, 2008, pp. 233–242. [Online]. Available: http://doi.acm.org/10.1145/1458082.1458115

[17] S. M. Faisal, G. Tziantzioulis, A. M. Gok, N. Hardavellas, S. Ogrenci-Memik, and S. Parthasarathy, "Edge importance identification for energy efficient graph processing," in *2015 IEEE International Conference on Big Data (Big Data)*, Oct 2015, pp. 347–354.

[18] P. J. McSweeney, K. Mehrotra, and J. C. Oh, "A game theoretic framework for community detection," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ser. ASONAM '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 227–234. [Online]. Available: http://dx.doi.org/10.1109/ASONAM.2012.47

[19] L. Wu, T. Bai, Z. Wang, L. Wang, Y. Hu, and J. Ji, "A new community detection algorithm based on distance centrality," in *2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, July 2013, pp. 898–902.

[20] T. Zhang and B. Wu, "A method for local community detection by finding core nodes," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ser. ASONAM '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 1171–1176. [Online]. Available: http://dx.doi.org/10.1109/ASONAM.2012.202

[21] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, no. 1-7, pp. 107–117, Apr. 1998. [Online]. Available: http://dx.doi.org/10.1016/S0169-7552(98)00110-X

[22] M. Pirouz and J. Zhan, "Optimized relativity search: node reduction in personalized page rank estimation for large graphs," *Journal of Big Data*, vol. 3, no. 1, p. 12, 2016. [Online]. Available: http://dx.doi.org/10.1186/s40537-016-0047-2

[23] F. Zhu, Y. Fang, K. C.-C. Chang, and J. Ying, "Incremental and accuracy-aware personalized pagerank through scheduled approximation," *Proc. VLDB Endow.*, vol. 6, no. 6, pp. 481–492, Apr. 2013. [Online]. Available: http://dx.doi.org/10.14778/2536336.2536348

[24] P. A. Lofgren, S. Banerjee, A. Goel, and C. Seshadhri, "Fast-ppr: Scaling personalized pagerank estimation for large graphs," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: ACM, 2014, pp. 1436–1445. [Online]. Available: http://doi.acm.org/10.1145/2623330.2623745

[25] S. Banerjee and P. Lofgren, "Fast bidirectional probability estimation in markov models," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 1423–1431. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969239.2969398

[26] P. Lofgren, "Efficient algorithms for personalized pagerank," *CoRR*, vol. abs/1512.04633, 2015. [Online]. Available: http://arxiv.org/abs/1512.04633

[27] P. Lofgren, S. Banerjee, and A. Goel, "Bidirectional pagerank estimation: From average-case to worst-case," *CoRR*, vol. abs/1507.08705, 2015. [Online]. Available: http://arxiv.org/abs/1507.08705

[28] ——, "Personalized pagerank estimation and search: A bidirectional approach," *CoRR*, vol. abs/1507.05999, 2015. [Online]. Available: http://arxiv.org/abs/1507.05999

[29] P. Lofgren and A. Goel, "Personalized pagerank to a target node," *CoRR*, vol. abs/1304.4658, 2013. [Online]. Available: http://arxiv.org/abs/1304.4658

[30] C. F. Klymko, "Centrality and communicability measures in complex networks: Analysis and algorithms," Ph.D. dissertation, Laney Graduate School, Math and Computer Science, Emory University, http://pid.emory.edu/ark:/25593/f90wn, 2014.

[31] B. E. Furht, *Handbook of Social Network Technologies and Applications*, 1st ed. Springer US, 2010.

[32] S. Gurung, "Top-k nodes identification in big networks based on topology and activity analysis," Master's thesis, Ubiversity of Nevada, Las Vegas, 5 2016.

[33] D. Liu, S. Trajanovski, and P. Van Mieghem, "Random line graphs and a linear law for assortativity," *Phys. Rev. E*, vol. 87, p. 012816, Jan 2013. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.87.012816

[34] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953. [Online]. Available: http://dx.doi.org/10.1007/BF02289026

[35] M. Benzi and C. Klymko, "On the limiting behavior of parameter-dependent network centrality measures," *ArXiv e-prints*, Dec. 2013.

[36] R. S. Wills, "Googles pagerank:the math behind the search engine," 2006. [Online]. Available: http://www.cems.uvm.edu/~tlakoba/AppliedUGMath/other_Google/Wills.pdf

[37] Intuitive explanation of personalized pagerank. [Online]. Available: https://blogs.oracle.com/bigdataspatialgraph/entry/intuitive_explanation_of_personalized_page

[38] Mark newman datasets. [Online]. Available: http://www-personal.umich.edu/~mejn/netdata/

[39] M. Bostock, V. Ogievetsky, and J. Heer, "D3 data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2011.185

[40] W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, pp. 452–473, 1977.

[41] jblas: Fast linear algebra for java. [Online]. Available: http://jblas.org

[42] Graph-stream. [Online]. Available: http://graphstream-project.org/

[43] Networkx. [Online]. Available: http://networkx.github.io/

[44] Numpy. [Online]. Available: https://pypi.python.org/pypi/numpy

[45] Snap datasets. [Online]. Available: https://snap.stanford.edu/data/index.html

[46] Ilab datasets. [Online]. Available: http://www.ilabsite.org/?page_id=12

[47] Facebook. [Online]. Available: https://www.facebook.com

[48] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, Mar. 2007. [Online]. Available: http://doi.acm.org/10.1145/1217299.1217301

[49] Epinions. [Online]. Available: http://www.epinions.com/

[50] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," *SIAM Journal on Discrete Mathematics*, vol. 17, no. 1, pp. 134–160, 2003. [Online]. Available: http://dx.doi.org/10.1137/S0895480102412856

Curriculum Vitae

Graduate College

University of Nevada, Las Vegas

Sai Phani Krishna Parsa

Degrees:

Bachelor Degree in Computer Science and Engineering 2014

Sree Nidhi Institute of Science and Technology, Hyderabad, Telangana, India

Thesis Title: Advanced Applications of Big Data Applications

Thesis Examination Committee:

Chairperson, Dr. Justin Zhan, Ph.D.

Committee Member, Dr. Yoohwan Kim, Ph.D.

Committee Member, Dr. Juyeon Jo, Ph.D.

Graduate Faculty Representative, Dr. Ge Lin Kan, Ph.D.