5-1-2016

Top-K Nodes Identification in Big Networks Based on Topology and Activity Analysis

Sweta Gurung
University of Nevada, Las Vegas, guruns1@unlv.nevada.edu

TOP-K NODES IDENTIFICATION IN BIG NETWORKS BASED ON TOPOLOGY AND ACTIVITY ANALYSIS

By

Sweta Gurung

Bachelors of Engineering in Computer Engineering
Kathmandu University, Nepal
2012

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 2016

UNLV | GRADUATE COLLEGE

Thesis Approval

The Graduate College
The University of Nevada, Las Vegas

April 29, 2016

This thesis prepared by

Sweta Gurung

entitled

Top-K Nodes Identification in Big Networks Based on Topology and Activity Analysis

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

Laxmi Gewali, Ph.D.
Examination Committee Member

Fatma Nasoz, Ph.D.
Examination Committee Member

Darren Liu, Dr. P.H.
Graduate College Faculty Representative

Abstract

Graphs and Networks have been the most researched topics with applications ranging from theoretical to practical fields, such as social media, genetics, and education. In many competitive environments, the most productive activities may be interacting with high-profile people, reading a much-cited article, or researching a wide range of fields such as the study on highly connected proteins. This thesis proposes two methods to deal with top-K nodes identification: centrality-based and activity-based methods for identifying top-K nodes. The first method is based on the topological structure of the network and uses the centrality measure called Katz Centrality; a path based ranking measure that calculates the local influence of a node as well as its global influence. It starts by filtering out the top-K nodes from a pool of network data using Katz centrality. By providing a means to filter out unnecessary nodes based on their centrality values, one can focus more on the most important nodes. The proposed method was applied to various network data and the results showed how different parameter values lead to different numbers of top-K nodes. The second method incorporates the theory of heat diffusion. Each node in the network can act as the source of heat. The amount of heat diffused or received by the node depends on the number of activities it performs. There are two types of activities: Interactive and Non-Interactive. Interactive activities could be likes, comments, and shares whereas posting a status, tweets or pictures could be the examples of non-interactive activities. We applied these proposed methods on Instagram network data and compared the results with the other similar algorithms. The experiment results showed that our activity-based approach is much faster and accurate than the existing methods.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Justin Zhan, for constantly guiding, motivating and mentoring me throughout my thesis. His support and encouragement gave me the extra drive to complete my thesis.

I would also like to thank my committee members: Dr. Laxmi Gewali, Dr. Fatma Nasoz, and Dr. Darren Liu for their support. I am very grateful to Dr. Ajoy K. Datta for all his support and guidance throughout the masters program. I would also like to thank Dr. Yoohwan Kim for introducing me to academic research and scientific writing.

My sincere gratitude goes to my parents, sisters and brothers as they have always been the source of inspiration. Their love and support have always inspired me to work hard. I would like to thank my dear friends, Mr. Ashish Tamrakar and Mr. Prajwol Sangat for their constant guidance and support throughout the thesis.

Sweta Gurung

University of Nevada, Las Vegas
May 2016

Table of Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

1.1 Motivation

Many things in this world can be visualized through the concepts of Graphs. Application of graph and network theory can be visualized in every genre, be it neurology, genetics, transportation, sociology, or computation. A group of people, an intertwined connection of different high-class proteins, or highway structure can be excellent examples of complex and big networks. Extensive studies have been conducted on identifying characteristics and heuristics of these network types. Maximizing the functionalities of complex networks in an efficient way is the ultimate goal for all the individuals involved in these areas. One of the major concerns for the graph theory specialists is to find optimal solutions to enhance node interactions and performance in terms of running time.

With the emergence of many social networking platforms – for example Facebook, Twitter, and LinkedIn – a trend among scholars in the Big Data community is to engage in social networking. Such networks have proved to be excellent examples of very large and complex networks. Ever since the advent of OSNs, people all around the world have been more connected with one another. More than 1.79 billion people use these OSNs every day [Inc16b]. Facebook alone has 1.04 billion daily active users, as of December 2015 [Inc16a]. Data collected from these users play a vital role in the marketing world. Facebook uses these connections and activity data to efficiently categorize the users for marketing purpose or for extensive studies. An example of identifying special nodes include discovering who are the really popular friends or identifying most active users/groups in social networks. These special nodes can be very useful from marketing point of view. Targeting the most popular users in a network has proven to be helpful in maximizing information flow in a very short time. Such users can act as a hub or as a medium to optimize internal as well as external communication.

The trend of searching for important nodes is popular not only with regard to social networks but also in

many other fields. For instance, in biology, many biological networks such as protein-protein interaction networks, cell signaling networks, and gene regulatory networks. The identification of the important nodes plays a vital role in biological discovery. Recent research in these areas [KK16] [WLY14] [KS08] have used various forms of topological centralities such as weighted sum of loads eigenvector centrality (WSL-EC) [KK16] or motif-based centrality [KS08], to either capture the important proteins or identify important features of the genes. Using concepts from graph theory such as cliques formation, centralities, etc. along with data mining algorithms like K-means, Random Forest, Naive Bayes, etc., many scientists have been successful in identifying proteins involved in many life-threatening diseases: cancers, AIDS, and many others [LHL$^+$12][GO12].

1.2 Objective

Networks are made up of nodes and edges. Generally in specific types of research, the researchers have a tentative idea on what they are looking for and what they want to get as the output. During these types of studies, the researchers are concerned for a specific set of nodes instead of the entire node lists. However, having to surf through the entire list of nodes is time-consuming, when an only particular set of nodes having particular characteristics is of interest. Generally, in specific-types of research, the researchers have a tentative idea of what they are looking for and what results they want to achieve. Therefore, giving researchers an option to filter out unwanted lists of data – in the case of this study, out-of-the-scope nodes – will allow the search for the important nodes or top-K nodes much more efficient and desirable. There are various network reduction algorithms such as disparity filter [SBV09], k-core decomposition [KGH$^+$10] [AHDBV05], etc. which prune the unwanted edges on the basis of certain filter functions. Similarly, we can apply some filtering strategies on nodes so that the focus is only on the desired lists of nodes.

On the basis of the structure, various characteristics of the network are derived. Node and edge information help in calculating the size, diameter, connectivity, triangulation, degree distribution, edge distribution entropy, and clustering coefficient of the network. These properties define the topological features of the network which in turn are the basis for characterizing network flows. How information flows from one node to another, the pattern of the flow and the speed with which it flows within the network depend on the topology and the characteristics of the network.

Assume a scenario where a piece of information has to be spread throughout the network. Usually, most connected nodes in the network are considered to be the key spreaders or the network hubs [PSV01]. But, it is always not the case that these highly connected nodes would have high effect on the spreading process. On top of the high connectivity of the nodes, the topology of the network also plays a role in selecting the hubs. If a hub is located at the periphery of the network, then it is bound to have nominal effect on the information diffusion whereas a centrally located node with possibly low connectivity will have signifi-

cant impact on the spreading process as this leads to wide dissemination among large sets of nodes [KGH+10].

Centrality is a network measure which computes the nodes centrality relative to its neighboring nodes. Since network flow depends on its nodes centralities, this measure is used for identifying popular users in social networks, hubs in physical networks, the mostly-visited webpage on the internet, specific genes associated with diseases, etc. There are various forms of centralities. Degree Centrality, Betweenness Centrality, Eigenvector Centrality, and Katz Centrality are some of them. Further discussion on each of these will be done in Chapter 2.

Out of all these centralities, one of our algorithms for identifying top-K nodes uses Katz centrality. The first algorithm is based on network topology and constraints. The algorithm starts by identifying the user-defined constraints, and applies those constraints on the nodes to extract only those nodes that satisfy them. Giving an option to filter out unwanted lists of data (which in our case is out-of-the-scope nodes) will make the important nodes or top-K nodes much more efficient and desirable. Katz centrality was used as a measure of topological centrality that helps to discover the relative influence of each node on the network. Given the global Katz centrality, users were required to provide the desired centrality for initial filtering of the nodes. Once the candidate nodes' list was collected, the top-K nodes were identified based on their local influence (i.e., local Katz centrality) and on a global scenario (i.e. global Katz centrality).

The above algorithm finds the top-K nodes purely on the basis of network structure and such nodes can be an excellent example of efficient network hubs. But several other factors have to be considered if we are looking for nodes popularity or influences. The activities that nodes perform and the characteristics they possess, both play the vital role in defining popular nodes identification. There has been a trend in social networks, especially, of finding the most influential users. These influential users are then targeted for marketing purposes.

The second part of the thesis focuses on discovering popular nodes by evaluating their activities. We know that network flow is most effective when the connectivity is maximum. It is not practical for a big network to have 100% connectivity among the nodes. So, big networks can be divided into sub-networks such that the connectivity among the nodes within these sub-networks is maximum. Such sub-networks are called communities. So, finding the most influential nodes for each community is highly effective from the point of information diffusion.

The proposed method includes extracting top-K nodes for each community on the basis of the individual node's activity history. The activity records for the nodes are quantified using the theory of heat diffusion: Heat received by a node is determined by the level of communication (interactive activities) between the

node and its neighbors. The non-interactive activities decide the amount of heat to be diffused.

1.3 Outline

The following sections briefly describe each chapter of the thesis.

In chapter 1, we provided the preface to the research topic and the proposed methods for the topic in consideration. We also briefly discussed the scenarios where the proposed methods can be useful.

In chapter 2, we will provide the background knowledge for understanding the graph theories. We will discuss various social network concepts and the models for community detection. Furthermore, this section will include influence maximization models and diffusion theories.

In chapter 3, we will discuss various methods to find top-K nodes using topological centralities, information diffusion methods and activity analysis approach.

In chapter 4, we will discuss the two proposed methods: one on the basis of network topology and the other on the basis of the activities performed by the nodes. Detailed descriptions of various parameters of the algorithms will be provided along with supporting examples to assist in the understanding of the methods.

In chapter 5, we will briefly mention the datasets used for the experimentation and the preprocessing performed on them. The results of the experimentations of both the algorithms will be explained in detail with necessary support.

In chapter 6, we will conclude the thesis by summarizing our methods and results, followed by the ideas and suggestions to further extend the research.

Chapter 2

Background

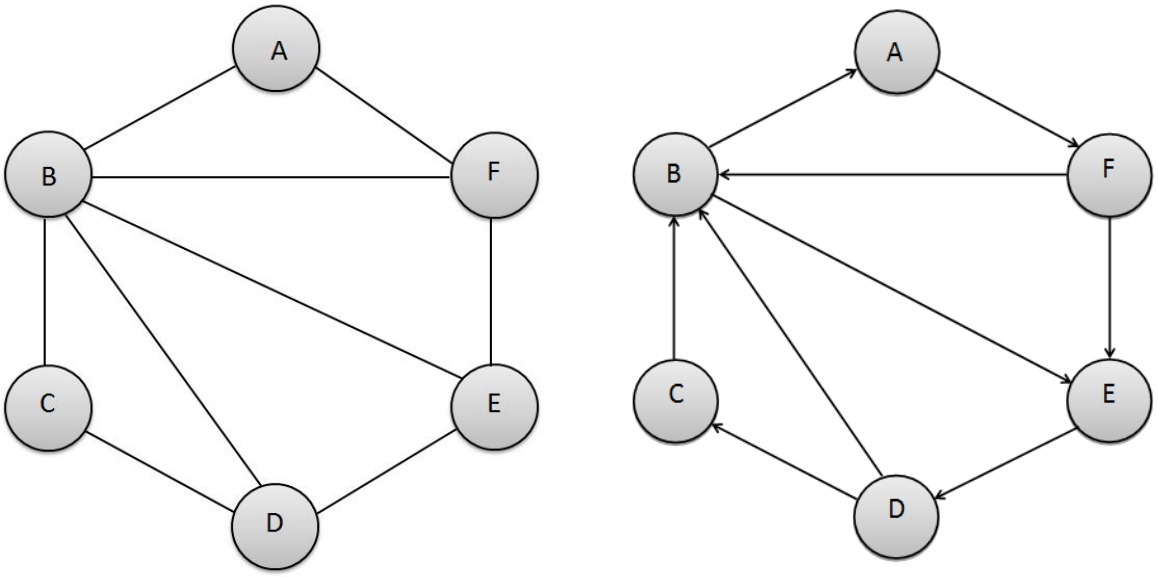
There are a few terms and terminologies in Graph Theory that must be cleared before going deep into the topic of discussion. This chapter discusses the fundamental concepts related to graphs and networks, followed by a brief introduction of social networks. The chapter proceeds into social influence and its maximization with information on influencers and influencees. Various methods for calculating and maximizing the influences are also discussed concisely. Finally, the chapter concludes by introducing the concept of graph centrality and its various types of measures which are frequently used for ranking purposes.

2.1 Fundamentals of Graph Theory

Graphs are usually the most popular forms used for representing large groups of entities with high interconnectivity such as road networks, social networks, genetics, etc. They can be used to demonstrate various types of relationship models in many sectors such as biology, sociology, computer systems, etc. Graph Theory has been taught in many institutes as an independent course and covers vast topics related to it and its applications. Often, Networks and Graphs are referenced interchangeably.

Definition 2.1.1. A graph G, represented as $G = (V, E)$, is a set of vertices (or nodes), $v \in V$ which are connected to each other through a set of edges (or links), $e \in E$.

Nodes are the points on the graphs whereas edges are the means of connections among them, which may be either directed or undirected. Depending on the edge types, there are directed graphs and undirected graphs. Directed graphs are those graphs in which edges are directed from one node (source) to another (target). Undirected graphs have no such distinction; in other words, the edges are bidirectional. Sometimes, edges are provided with special values, $edge-weights$, either for quantifying the connection or for edge distinction.



(a) Example of Undirected Graph. (b) Example of Directed Graph.

Figure 2.1: Example of Directed and Undirected Graphs.

Graphs with edge weights are called weighted graphs.

A node may be connected to other nodes through one or multiple edges. There are various types of graphs based on their structures. Simple graphs are those graphs in which two nodes are connected to each other by at most one edge. If there are multiple edges between any two nodes, then such graphs are defined as multigraphs. There is a special type of graph called pseudograph which contains multiples edges as well as loops. A loop is an edge which connects a node to itself.

Besides these, there are also various classes of graphs. Some of them are listed below.

Regular Graph: A graph is said to be regular if each node has the same number of neighbors.

Complete Graph: In a complete graph, there exists a distinct edge between every possible pair of nodes. There are $n(n-1)/2$ edges in a complete graph with n nodes.

Connected Graph: A connected graph has no unreachable nodes. Every node can be reached through one or the other edge connections.

Bipartite Graph: A bipartite graph can be divided into two distinct and independent groups of nodes, U and V, such that every node in U is connected to one or more nodes in V.

Cycle Graph: In a cycle graph, the nodes are connected in a closed chain fashion.

As mentioned earlier, any type of network can be represented in a graph form for a variety of modeling purposes. For analyzing such network data, various graph measures can be applied. Some of the graph terminologies and measures are mentioned below.

Order, Size, and Diameter of a Graph: Number of nodes in a graph is called the order of the graph, denoted by $|V|$, whereas the size of the graph, $|E|$ represents the number of edges in it. The diameter of the graph, $d(v_i, v_j)$ is the longest shortest path between any two nodes, (v_i, v_j).

Degree of a node: Generally, for an undirected graph, the degree of a node is the number of edges incident on it. But in the directed graph, there are in-degree (number of incoming edges) and out-degree (number of outgoing edges) of a node. Mathematically, degree of a node can be represented using the adjacency matrix, A, of the graph, G:

$$d(v_i) = \sum_{j=1}^{n} a(i,j) \tag{2.1}$$

where $a(i,j) \in A$ and n is the number of nodes in G.

Adjacency matrix of a Graph, G of order n is a $n \times n$ square matrix where each element $a(i,j)$ denote whether there exists a link between nodes i and j.

Neighborhood of a node: The neighborhood of a node v is the set of all nodes which are incident or connected to v.

Walk: A walk in a graph is the sequence of nodes between any two nodes (inclusive). Walk of length k is a sequence of nodes and edges, $v_1, e_1, v_2, e_2, \ldots, e_{k-1}, v_k$. A path is a walk where no node occurs more than once in the sequence. A cycle is a closed path in which the start node and end node of the sequence are the same.

Definition 2.1.2. For a graph $G = (V, E)$, a walk of length m denotes a set of nodes $\{v_1, v_2, v_3, ..., v_m\}$ such that there exists an edge between v_i and v_{i+1}, $\forall 1 \geq i \geq m$.

Clique: In an undirected graph, a clique is a subset of nodes where each node is adjacent to one another.

2.2 Graph Centrality Concepts

In graph theory, centrality helps in identifying the most central or important nodes in a graph. Centrality is completely based on the structure of the graph. There are various types of centrality measures. Some of the measures are discussed below.

Degree Centrality

Degree Centrality is defined as the number of edges incident at each node. Higher the degree centrality, higher the connectivity of a node to the other nodes. Depending on the type of graph, degree centrality may vary. For an undirected graph, degree centrality of a node, v_i, is equal to the degree of the node, d_i, [ZAL14] and is given by:

$$C_d(v_i) = d_i \tag{2.2}$$

Since there is an in-degree and an out-degree for directed graphs, degree centrality can be of three forms:

$$C_d(v_i) = d_i^{in} \tag{2.3}$$

$$C_d(v_i) = d_i^{out} \tag{2.4}$$

$$C_d(v_i) = d_i^{in} + d_i^{out} \tag{2.5}$$

Eigenvector Centrality

Eigenvector Centrality is an extension of degree centrality. In the degree centrality, all node connections are credited of equal importance. But in real life, each node may have different importance. For example, a node connected to highly important nodes itself is an important node. Thus, Eigenvector Centrality provides a relative score to each node depending on the type of nodes (high-scoring and low-scoring) it is connected to. For a given graph $G = (V, E)$ containing n nodes, let A be the adjacency matrix of G and λ be the eigenvalue. Then Eigenvector Centrality is given by

$$C_e(v_i) = \frac{1}{\lambda} \sum_{j=1}^{n} d_i a_{j,i} C_e(v_j) \tag{2.6}$$

Equation 2.7 can be summarized as

$$C_e = \frac{1}{\lambda} A^T C_e \tag{2.7}$$

Closeness Centrality

Closeness Centrality is based on the geodesic path which is the shortest path between any two nodes, v_i and v_j [New10]. The average geodesic distance from node v_i to node $v_j \in (V - v_i)$ is given by:

$$g(v_i) = \frac{1}{n} \sum_j d_{ij} \tag{2.8}$$

where d_{ij} is the number of edges along the path from node v_i to node v_j. Thus, Closeness centrality is as follows:

$$C_c(v_i) = \frac{1}{g(v_i)} = \frac{n}{\sum_{j=1} d_{ij}} \tag{2.9}$$

Betweenness Centrality

Betweenness Centrality measures the importance of a node in connecting any two other nodes. How often a node falls in the path connecting two nodes defines the betweenness property of the node [New10]. Let g_{st} be the total number of geodesic paths from node v_s to node v_t and n^i_{st} be the number of geodesic paths from node v_s to node v_t that pass through node v_i. Then, betweenness centrality of node v_i is as follows:

$$C_b(v_i) = \sum_{st} \frac{n^i_{st}}{g_{st}} \tag{2.10}$$

2.3 Social Network Concepts

A social network is a structure consisting of socially relevant units (e.g. individuals or entities) which are connected by one or more relationships. These relationships such as friendship and trust, can vary over various types of social networks, which will be discussed later. Some common examples of social networks are Facebook, Twitter and LinkedIn. Several theories and models have been proposed and developed for analyzing the structure, linkage and patterns among the social entities. Social Network theories, models, and their applications are coherently termed as social network perspective and provide fundamental concepts related to the social networks [Car13].

Actor: Individual social units in a social network are referred to actors. These actors play distinct roles based on the network they belong to, for example, authors in journal collaborative networks, teachers, and students in school networks.

Ties: Ties define the way in which actors are linked. Based on the number of actors involved in the connection, ties can be called a dyad (tie between two actors), triads (tie involving three actors) and subgroups (involving multiple actors and all the ties among them). Based on how actors communicate with each other, some of the common ties prevalent in the social networks are as follows [Car13]:

- Behavioral Interaction, like sending messages, liking posts or posting comments on the posts.
- Physical Connection (e.g., actors belonging to the same groups and communities)

- Collaboration/Affiliation, (e.g. working in a same project or research)
- Formal Relations (e.g. hierarchical relations)

Relations: Actors usually have more than one tie with their fellow actors. Having multiples ties with other nodes generally mean richness in the connections. So, during social network analysis, relations, (defined as the set of ties among the actors) are quantified. Borgatti et al. identified four types of dyadic relations: similarities, social relations, interactions, and flows[BMBL09].

Similarities			Social Relations				Interactions	Flows
Location	**Membership**	**Attribute**	**Kinship**	**Other role**	**Affective**	**Cognitive**		
e.g., Same spatial and temporal space	e.g., Same clubs, Same events, etc.	e.g., Same gender, Same attitude, etc.	e.g., Mother of, Sibling of	e.g., Friend of, Boss of, Student of, Competitor of	e.g., Likes, Hates, etc.	e.g., Knows, Knows about, Sees as happy, etc.	e.g., Talked to, Advice to, Helped, Harmed, etc.	e.g., Information, Beliefs, Personnel, Resources, etc.

Figure 2.2: A typology of ties studied in social network analysis [BMBL09].

Network analysts consider these types of social networks as the primary building blocks of the social world and perform various analyses using fundamentally different perspectives, as opposed to the historical attribute-based perspectives [SC11]. Apart from the studies on the key network attributes such as actors' characteristics or the neighborhoods, they also conduct deep analysis to figure out the connectivity patterns.

Definition 2.3.1. Social Network Analysis (SNA) is a strategic analysis process which involves investigation of social structures using the formal network and graph theories. [OR02]

Social network analysis is used extensively in a wide variety of disciplines. Many researchers have been greatly interested in studying the different phenomena associated with the social networks. Some of the areas where social network analysis is rigorously used are customer analysis, network modeling, behavior and sentiment analysis, marketing, recommenders system, criminology, etc. [Gol13]. These types of analysis are usually done using some other concepts like diffusion, influence, business intelligence, etc in conjunction with the social network theory.

2.4 Information Diffusion

People are often affected by their surroundings. It is humans nature to be influenced by judgments, emotions, and behaviors of others. We call it social influence. Social influence occurs when someone's actions or opinions cause a change in others' lives or behaviors and this phenomenon of transferring influence from one person to another is called Information Diffusion. Let us take an example of a classroom. If a student A follows the classroom ethics and rules, then it is very likely that her friends also follow them. Another example can be from marketing. Viral marketing is a marketing technique which utilizes the social networking

platforms to create or increase brand awareness among the social network users. This is usually achieved by message exchange among the users or by a chain of influence through word-of-mouth. Many small-scale to large-scale companies have been adopting this efficient and cost-effective marketing technique by selecting certain groups of social network members who have larger audiences. These chosen members tend to have the high potentiality to influence a huge number of people. For example, if a cosmetic company wants to advertise it's new product, it will select a certain number of highly popular cosmetic enthusiasts and then send them samples of the product, hoping that they will, in turn, recommend the product to their followers. But, the hardest part of it is to select a small set of influential people with larger influence rate in a large social network.

Many diffusion models have been proposed in recent years for maximizing the influence. Influence maximization problem can be defined as finding smallest number of nodes, say k, that can influence maximum number of nodes. Some of the popular models are Independent Cascade model (IC) and Linear Threshold model (LT) by Kempe et al. [KKT15].

Linear Threshold Model (LT)

In this model, nodes are either active or inactive and each node $v_i \in V$ has a threshold, θ_i, whose value ranges between $[0, 1]$. θ_i represents the total amount of weight required from all the neighbors in order to activate node v_i. The model starts by assigning a certain set of active nodes and individual θ values. For the given sets of active nodes, the diffusion process starts in discrete step-wise fashion. At step t, all the nodes that were activated in the previous steps remain active throughout the diffusion process. At each diffusion step, an inactive node is activated by its neighbors if the total weight received from its active neighbors is greater than or equal to its threshold:

$$\sum_{v_j \Rightarrow v_j, v_i} w_{i,j} \geq \theta_i \tag{2.11}$$

Independent Cascade Model (IC)

In IC model, each edge connecting two nodes, v_i and v_j, has a probability value, $p_{i,j}$, associated with it. Just like LT model, the model starts with an initial set of active nodes. These active nodes begin the diffusion process. If at discrete step t, a node v_i is activated, then it is given a single chance to activate each of its inactive neighbors, v_j. The node v_i has a probability of $p_{i,j}$ to get its neighbor v_j activated. If v_i is successful in activating the neighbors, then $v_j s$ will be activated in the next step, $t + 1$. At step $t + 1$, the recently activated nodes get the chance to activate their neighbors. This goes on until no further activations are possible.

Definition 2.4.1. For a network graph G(V,E) and an initial set of active vertices, $A \subseteq V$, the set of nodes activated by the initial source of influence, A, at the end of diffusion process is called the influence spread of set A, which is denoted by $\sigma(A)$ [KKT15].

Chapter 3

Literature Review

A significant amount of research has been done regarding influential node identification and search space reduction. Most of the common approaches for selecting the influential or top-K nodes are usually based on the centrality theory; diffusion models, such as independent cascade model or the linear threshold model [KKT15]; the heat diffusion theory [DL14]; and the evidence theory [LZLD15]. Some of the literature on these topics are discussed as follows.

In order to extract the influential nodes, Kimura et al. [KSN07] came up with a method that used the theory of bond percolation along with graph theory. The purpose of their method was to maximize the influence for information diffusion. This method begins by finding a set of nodes, A, for initial activation, by using a greedy hill-climbing algorithm. Once a set of nodes are obtained, A is used to estimate the marginal gains, $\nabla\sigma(A)$ for the influence degree $\sigma(A)$ of the target set A.

At the time when most of the research on measurement for social network influence was being done by using topological connectivity of the networks' nodes, Doo and Liu [DL14] developed an activity-based social influence model which turned out to be more effective than existing topological-based models. The authors used the concept of heat diffusion to measure the influence diffusion among the nodes. Every interaction between any two nodes was labeled as heat diffused, $DH_i(\delta t)$, for outgoing edge activities and heat received, $RH_i(\delta t)$, for incoming edge activities. Activities between any two nodes, v_i and v_j, could be differentiated as interactive activities (IA_{ij}) such as comments, likes, etc. and non-interactive activities, (NIA_{ij}) – such as status updates or photo uploads, in the form of comments, likes, shares, posts, etc. Each of these activities was weighted as one point. Higher the number of non-interactive activities at node v_i, higher was the amount of heat collected at v_i and slower was the heat diffusion to its neighbors. Based on the values for $DH_i(\delta t)$ and $RH_i(\delta t)$, heat diffusion was calculated for each node, v_i. Finally, influence coverage, (IC_i) which is the list of nodes influenced by v_i, was generated. The top-K nodes were selected based on $|IC_i|$.

In an approach similar to the current study, Zhang et al[ZZC11] used a greedy algorithm, specifically a two-staged mining algorithm (GAUP) to locate the top-K nodes in social networks by considering the users' preferences. The initial stage involved estimating user preferences with a set of latent items for a specific topic by adopting the Latent Semantic Indexing (LSI) method. In the second stage, which was based on the Extended Independent Cascade Model, these estimations were used to maximize influences on the active nodes and then discover a selected set, S of top-K nodes.

Leung et al. [LMJ14] proposed an algorithm that reduced the search space based on user-specified constraints and used MapReduce model to discover interesting patterns from uncertain data that satisfied those constraints. The algorithm mined frequent singleton patterns followed by non-singleton patterns. The map function computed individual existential probabilities for each item in a transaction. The reduce function filtered the items that satisfied user-specified constraints, and computed the expected support, *expSup*, for each item; these were compared with the minimum support, *minSup*. Only those items whose $expSup \geq minSup$ were selected as the singleton pattern. Using the individual expSup, the non-singleton patterns were discovered.

In addition to these theories and methods, centrality has been widely used in many studies related to network analysis. Cupertino et al. [CZ12] came up with a network-based method that used Katz centrality to predict the pattern class to which the given group of invariant transformations of the same pattern belonged. Using another measure of network centrality called Principal Component Centrality (PCC), Ilyas et al. [IR11] identified a group of nodes – i.e., *social hubs* in the network that are at the center of influential neighborhoods – and compared their results with the nodes identified from a method using Eigenvector Centrality (EVC). To further enhance the usage of α-centrality, Ghosh et al. [GL11] introduced a normalized version of this centrality by generalizing a modularity maximization-based approach. Their method identified not just the local communities but also global ones.

In a recent paper by Li et al. [LZLD15], a method based on evidence theory was proposed to identify influential nodes in a network of networks (NON). Any complex network could be subdivided into sub-networks such as series of similarity networks from these single networks. For each of these individual networks, distance matrix, D which represents the similarity among nodes, was computed. This matrix D was further used to compute the similarity network that assisted in finding the Basic Probability Assignment (BPA). The nodes with a high similarity value along with other nodes in the fused similarity network were considered to be influential nodes in NON.

Chapter 4

Proposed Methods

4.1 METHOD I: Topology-Based Approach

This section includes background information on the Katz Centrality measures and how constraints could be applied to the filtering process.

4.1.1 Constraints-Based Mining using Katz Centrality

The popularity of nodes can be interpreted and characterized in many ways and many studies have been done in recent times to identify influential nodes. One approach to define node popularity is to congregate the ideas of an influential node with its popularity. Among various ways that influential nodes can be detected, one involves using topological state, i.e., node centrality. Many approaches are used to compute centrality: Degree Centrality, Betweenness Centrality, Closeness Centrality, Eigenvalue Centrality, Subgraph Centrality, Evidential Centrality, etc. Degree Centrality of node i measures the number of neighbors that i has. Betweenness Centrality measures the number of times a node acts as a bridge along a path between any two nodes. These centralities compute the local influence of a node i. On the other hand, Katz Centrality [Kat53] measures the relative influence of each node in a given network by taking into account the node's immediate neighbors as well as non-immediate nodes that can be connected through the immediate neighbors. Just like Subgraph Centrality and Total Communicability, Katz Centrality measures both local and global influences of a node on the entire network [BK14].

Definition 4.1.1. Given a graph $G = (V, E)$, where $V = \{v_1, v_2, v_3, ..., v_m\}$ is the set of m nodes or vertices and $E = \{e_1, e_2, e_3, ...\}$ is the set of edges; then, A is the adjacency matrix of the network G, denoting the immediate connectivity among the nodes. The Katz Centrality of a node, v_i, is given by:

$$C_{Katz}(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} C_{Katz}(v_j) + \beta \tag{4.1}$$

where α is a constant called the damping factor, usually considered to be less than the largest eigenvalue, λ i.e. $\alpha < 1/\lambda$ and β is a bias constant, also called the exogenous vector, used to avoid the zero centrality values [ZAL14, New10]. With $\alpha \geq \lambda$, the centrality tends to diverge [New10].

The concept of using Katz Centrality for filtering out the most central nodes may be more computationally efficient than other methods, especially for big directed networks [LMJ14]. Unlike other centrality measures, it takes into account the walks (i.e., alternating sequence of nodes and edges) instead of the more usual approach of the shortest (geodesic) path [HR05], such that the longer walks are penalized through the attenuation factor, α. 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 with the notion that k-step walk has α^k probability of being effective. Thus, the further the neighbors are from the node under consideration, the less is its influence on them. For instance, Katz centrality for node i considers all the walks starting from the node itself and penalizes the contributions of k-length walks by assigning α^k [BK14].

$$\begin{aligned}(I - \alpha A)^{-1} &= I + \alpha A + \alpha^2 A^2 + ... + \alpha^k A^k + ... \\ &= \sum_{k=0}^{\infty} \alpha^k A^k, \qquad 0 < \alpha < 1/\lambda\end{aligned} \tag{4.2}$$

It is clear from the Eq. 4.2 that the Katz centrality is a parameter dependent index, i.e., it depends on α and β. Their values play a decisive role in obtaining fluctuating Katz centrality values. Different choices of α and β lead to different centrality values resulting in different node rankings [12]. For instance, if $\alpha \to 0+$, then Katz Centrality reduces to Degree Centrality. 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. Hence, these parameters can be taken as a medium to tune between the rankings of nodes based on either local influence (short walks) or global influence (long walks) [BK14].

Equation 4.2 can be generalized for the entire graph as:

$$C_{Katz} = \beta(I - \alpha A^T)^{-1}.1 \tag{4.3}$$

where 1 is a matrix of ones. For example, for the graph shown in Fig. 4.1, its adjacency matrix is given by:

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix} = A_T \tag{4.4}$$

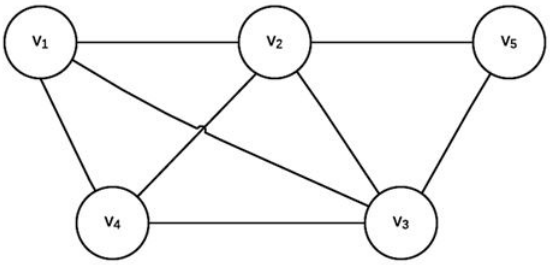


Figure 4.1: Sample network graph.

Since, the largest eigenvalue of A (λ) is 3.32 (such that $1/\lambda = 0.3$), we take $\alpha = 0.20$ and $\beta = 0.80$. Now, the Katz centralities for these 5 nodes are as follows:

$$C_{Katz} = \begin{pmatrix} 2.31 \\ 2.62 \\ 2.62 \\ 2.31 \\ 1.85 \end{pmatrix} \tag{4.5}$$

Here, nodes v_2 and v_3 have the highest centrality and thus have greater influence on the other nodes. It is clear from the Eq. 4.2 that C_{Katz} is directly proportional to α and β. However, the ranking of the nodes may vary depending on the choice of their values. For small graphs such as the one above, the parameters α and β have no effect on the nodes' ranking; but, for huge networks with larger node counts and higher connectivity, the choice of the α value may have a direct effect.

Constraints-based data mining [GR05] has been widely used to find frequent items or patterns in a given pool of data [LMJ14] [ZZC11]. For the scenario in this study, a similar approach was used, based on the user-specified constraints. Various types of constraints could be used in mining, such as knowledge type constraints, data constraints, dimension constraints, interestingness constraints, rule constraints, etc. Providing a means to apply certain constraints on the data allow the users to be specific in their search so that only those datasets satisfying the constraints are looked for in the database. These types of searches reduce the time as well as unnecessary computations; at the same time, only favorable and desired outputs are received. The constraint, $Const = min(node.C_{katz}) \geq 2.40$, denotes the user's interest in finding only those nodes whose Katz centrality (C_{katz}) is greater than or equal to 2.40. Here, $Const$ expresses the succinctness property.

Definition 4.1.2. A set is said to be a succinct set if it is formed as result of a selection operation, $\sigma_{a\theta b}$ where a and b are attributes and θ, a binary operation [GR05] [LMJ14]

ALGORITHM 1: Individual Katz Centralities

Read the network input file
Form the adjacency matrix, A, of the network
Get User desired values for α and β
for *each node* $v_i \in V$ **do**
    Calculate the Katz Centrality, $C_{Katz}(v_i)$
    Return each node, v_i and its respective, Katz Centrality, $C_{Katz}(v_i)$
end

So, each node and its neighbor(s) have their own Katz centrality. For example, in the above case, node v_2 has 4 neighbors, $[v_2, \{v_1 = 2.31, v_3 = 2.62, v_4 = 2.31, v_5 = 1.85\}]$. We can see that, neighbors v_1, v_4, v_5 have their centralities, $C_{katz} \leq 2.40$, which is the provided constraint. Now, the local Katz centrality of node, v_i, is calculated by summing all of its neighbors Katz centralities along with its own and then dividing by the (neighbor-count + 1):

$$\begin{aligned} LAC_{Katz}(v_i) &= \frac{C_{Katz}(v_i) + \sum_{j=1}^{n_i} C_{Katz}(v_{ij})}{n_i + 1} \\ &= \frac{LSC_{Katz}(v_i)}{n_i + 1} \end{aligned} \tag{4.6}$$

$$GAC_{Katz} = \frac{\sum_{i=1}^{n} C_{Katz}(v_i)}{n} \tag{4.7}$$

where n_i is the number of neighbors of a node, v_i and n is the total number of nodes in the network.

A node can be considered a top-K node if $LAC_{Katz}(v_i) \geq GAC_{Katz}$. This condition removes those nodes that have higher degrees as well as a mixture of neighbors with much higher and much lower centralities.

4.1.2 Centrality-Based Method

The algorithm which was formulated to find the top-K influential nodes is based on the nodes' topological structures and uses the Katz Centrality as its base. Given a set of network data and user-specified constraints, the solution for finding the top-K nodes in the big network can be identified.

The algorithm begins by computing the Katz Centrality, $C_{Katz}(v_i)$, for each node v_i, based on the α and β values provided by the users; it computes the Katz centrality for every node, as shown below. Now, for every node v_i, its neighbors are listed along with their respective centralities. Using the neighbors lists, the Local Katz Centrality, $LAC_{Katz}(v_i)$, is calculated using the Equation 4.6, for only those nodes whose centralities are greater or equal to the desired centrality value provided by the user, *Const*. The nodes satisfying this *Const* are the candidates for the top-K nodes. Any node satisfying the centrality constraint, *Const* and has LAC_{Katz} greater or equal to GAC_{Katz} is included in the top-K nodes list.

ALGORITHM 2: Neighbours' Centralities

for *each node* $v_i \in V$ **do**
    Find a list of its neighbors, v_{ij}, and their Katz Centralities, $C_{Katz}(v_{ij})$
    Return each node, v_i, and the list of its neighbour's $C_{Katz}(v_{ij})$
end

ALGORITHM 3: Centrality-based Top-K Nodes (CTKN)

Const ← User Desired Katz Centrality for filtering purpose
$GAC_{Katz} \leftarrow \sum_{j=1}^{n} C_{Katz}(v_j)/n$
for *each* $(v_i, listof(v_{ij}, C_{Katz}(v_{ij}))$ **do**
    if (v_i *satisfies Const*) **then**
        $LAC_{Katz}(v_i) \leftarrow 0$
        $LSC_{Katz}(v_i) \leftarrow 0$
        $ngbrCount \leftarrow 0$
        for *each* $v_{ij} \in listof(v_{ij}, C_{Katz}(v_{ij}))$ **do**
            $LSC_{Katz}(v_i) \leftarrow LSC_{Katz}(v_i) + C_{Katz}(v_{ij})$
            $ngbrCount \leftarrow ngbrCount + 1$
        end
        $LAC_{Katz}(v_i) \leftarrow LSC_{Katz}(v_i)/ngbrCount$
        if $LAC_{Katz}(v_i) \geq GAC_{Katz}$ **then**
            Return v_i and its $C_{Katz}(v_i)$
        end
    end
end

4.2 METHOD II: Topology with Activity-based Method

Social Influences have been a huge topic lately, owing to the vast growth of networking platforms in the past few decades. Subscribing to the famous influencers' channels in the YouTube network or following the most popular Instagram accounts has always been the day-to-day activities to the social media fanatics. In the previous method discussed above, we strictly used the network layout to find the top-K nodes. But in a real scenario, many other factors play the roles in determining the top-K nodes. For example, in social networks like Facebook and Twitter, top-K nodes would mean popular/influential users or the most active users. These users are identified not just by their topology and connectivity but also by their activity histories and their inter-communications with their peers. Usually, users with higher activity levels tend to have higher influences on their neighboring friends than those who are less active, comparatively.

Users' popularity is widely measured in terms of the number of subscribers, followers or friends in the social networks. But the popularity metric also depends on the level of interactions between the two connected nodes. For example, suppose there is a node A with very high number of followers but its activity level is very low and there is another node B with small number of followers (compared to node A) but has a high level of interaction with its followers. Which of these two nodes should be considered popular or influential? There is a very high chance that node B could be more popular than A.

Before answering this question, we need to be clear about the types of activities that can occur within a social network. There are two types of activities: interactive activities (IA) and non-interactive activities (NA). Interactive activities occur between a pair of nodes. Non-interactive activities, on the other hand, occur strictly within the node itself. For example, Tim, an avid Instagram user, posts a photo in Instagram. The photo is liked and commented by Tim's followers. The act of posting a photo is a non-interactive activity whereas the likes and comments generated by the photos are in fact the interactive activities between Tim and his followers.

Doo et al. pointed out that influence of one node onto the other and vice-versa highly depends on the interaction between those pair of nodes. From the diffusion point of view, the influence can also depend on how active the node itself is. For example, without status updates or photo uploads by a Facebook/Instagram user, there would be less number of interactions between the user and his/her friends. So, some may argue that non-interactive activities are more important while others may focus more on the interactive activities to select the influential nodes.

The following figure 4.2 shows a typical social media network. Nodes in the graph are the users while the edges show the existence of connections among the users. The direction of an edge represents the relationship type, which in this case is a follower (a user who follows other users) and a followee (a user who is followed by other users). User *Bob* is followed by two users: *Mary* and *Mac*, and vice-versa. They have a total of 59 interactive activities performed among them. User *Ed* posted 33 photos/videos in his profile and those posts received 7 distinct comments/likes from his followers. So, the edge weights denote the number of interactive activities while the node weights represent the non-interactive activities.

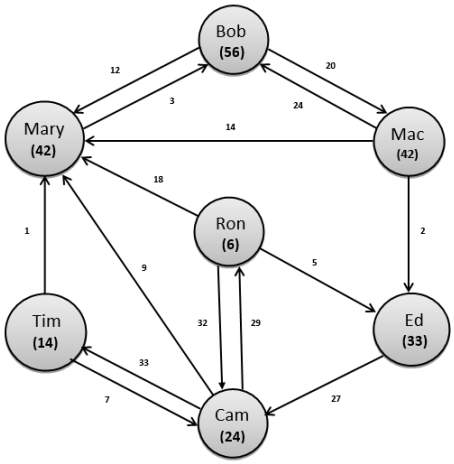


Figure 4.2: Social Media Network Example.

4.2.1 Heat Diffusion

Heat Diffusion, also called heat conduction, is a process of thermal energy exchange among particles within the body. Heat diffusion is governed by the Second Law of Thermodynamics which states that heat continues to flow from a hot body to a cold one until thermal equilibrium is reached. The amount of heat transfer during conduction is determined by thermal conductivity, α of the body. It is a measure of the ability of a body's material to conduct heat. Higher the value of α, higher the amount of heat transfer. We can relate this concept to network theory. A network as a whole can be considered as a body and nodes within it are referred as body particles.

Definition 4.2.1. Heat equation is a partial differential equation which involves three variables – two independent variables x and t, and one dependent variable $f = f(x, t)$.

$$\frac{df}{dt} = k\frac{d^2f}{dx^2} \tag{4.8}$$

where $f(x, t)$ is the heat at location x at time t, with initial heat distribution as $f(x, 0) = f_0$, k is the thermal diffusivity (α/(heat capacity per volume)) [LL05]

Heat Diffusion models have been applied in many areas and fields such as text classification, ranking, etc. Khurd et al. developed a kernel-based method for analyzing Diffusion Tensor Imaging (DTI) data using kernel Principal Component Analysis (kPCA) [KVD07]. In [ZCL05], Zhang et al. proposed a kernel based Negative Geodesic Distance (NGD) and proved that it can be effectively used as a kernel for support vector machines. Similarly, there exist many other studies which are focused on investigating the heat equations and their respective kernels. The heat kernel, denoted by $K_t(x, y)$ is the solution to the heat equation $f(x, t)$. It describes the distribution of heat diffusing from the heat source y at time t to the location at x, thus defining the connectivity between x and y [MYLK08].

Graphs and networks are the perfect examples of discrete objects and have non-continuous values associated with each node. Since the heat kernels help in recognizing the relationships between the two points (x, y), we can use the same logic for finding the relations among nodes within the graph. Kondor et al. came up with the idea of generating kernels on discrete objects and proposed a natural class of kernel on graphs called Diffusion Kernel [KL02].

4.2.2 Topology-Based Diffusion

Using the Kernel definition mentioned in [KL02], Ma et al. and Doo et al. proposed diffusion models for directed and undirected social networks [MYLK08], [DL14]. We will be focusing on the methods for directed graphs as our approach is also based on directed networks.

Let us take a directed social network graph $G = (V, E)$ where $V = \{v_1, v_2, ..., v_n\}$ is the set of n nodes and $E = \{e_i | \textit{there exists an edge from } v_i \textit{ to } v_j; v_i, v_j \in V\}$ be the set of all edges that exists in the graph. Let $f_i(t)$ be the heat at node v_i at time t, accumulated due to initial heat distribution of $f_i(0)$ at time $t = 0$. Let α denote the thermal conductivity of the network. At time t, heat is allowed to flow throughout the graph for a time Δt, as heat flows from higher temperature to lower temperature. So, during the time interval Δt, node v_i will diffuse $HD_i(\Delta t)$ amount of heat through its outgoing edges and at the same time, receives $HR_i(\Delta t)$ amount of heat from its incoming edges. So, the amount of heat at node v_i between time t and $t + \Delta t$ is given by:

$$f_i(t + \Delta t) - f_i(t) = HR_i(\Delta t) - HD_i(\Delta t) \tag{4.9}$$

For successful heat diffusion between any two nodes, v_i and v_j, Ma et al. considered several assumptions:

1. Each node in the graph has the ability to diffuse and receive heat.
2. $HD_i(\Delta t)$ and $HR_i(\Delta t)$ should be proportional to the time period Δt.
3. $HD_i(\Delta t)$ should be uniformly distributed among its neighbors through the respective outgoing edges, if exist.
4. $HR_i(\Delta t)$ should be proportional to the heat at the neighbor node v_j.

From the given assumptions, we can see that $HD_i(\Delta t)$ and $HR_i(\Delta t)$ depend on several factors such as heat conductivity α, number of respective neighbors, heat at receiving node, $f_i(t)$, heat at diffusing node, $f_j(t)$ and the time duration Δt. Therefore, their corresponding values are formulated as follows:

$$HD_i(\Delta t) = \alpha \Delta t f_i(t) \tag{4.10}$$

$$HR_i(\Delta t) = \alpha \Delta t \sum_{j:(v_j, v_i) \in E} \frac{f_j(t)}{d(v_j)} \tag{4.11}$$

In equation 4.11, $d(v_j)$ denotes the number of out-going edges from node v_j. It also represents the number of neighbors node v_j has. This means that heat at v_j is equally distributed among all of its neighbors. Substituting the values of $HD_i(\Delta t)$ and $HR_i(\Delta t)$ in equation 4.9, we get:

$$f_i(t + \Delta t) - f_i(t) = \alpha \Delta t \left(\sum_{j:(v_j, v_i) \in E} \frac{f_j(t)}{d(v_j)} - f_i(t) \right) \tag{4.12}$$

Representing the above equation in matrix form, we get:

$$
\begin{aligned}
& f(t+\Delta t) - f(t) = \alpha \Delta t H f(t) \\
\Rightarrow & \frac{f(t+\Delta t) - f(t)}{\Delta t} = \alpha H f(t)
\end{aligned}
\tag{4.13}
$$

where H is a $n \times n$ square matrix whose individual element $H(i,j)$ is denoted by:

$$
H(i,j) = \begin{cases} \frac{1}{d(v_j)}, & (v_j, v_i) \in E \\ -1, & i = j \text{ and } d(v_i) > 0 \\ 0, & \text{otherwise} \end{cases}
\tag{4.14}
$$

Taking limit $\Delta t \to 0$ in equation 4.13:

$$
\frac{d}{dt} f(t) = \alpha H f(t)
\tag{4.15}
$$

Solving equation 4.15 we get a matrix, $f(t)$ which consists of heat at every node at time t, along with the heat diffusion kernel represented by $e^{\alpha t H}$ as shown below.

$$
f(t) = e^{\alpha t H} f(0)
\tag{4.16}
$$

where initial heat source matrix, $f(0) = \{f_i(0) | v_i \in V\}$ at $t = 0$.

4.2.3 Activity-based Diffusion

Kernel-based approach has also been used for analyzing the network data with activity information. In [DL14], Doo et al. presented an activity based social influence model using the concepts proposed in [KL02, MYLK08] and showed that their approach is more effective than the topology-based approach as they not only considered the topology but also the node-by-node activities.

As mentioned earlier, activities can be either interactive or non-interactive. Let the number of interactive activities from node v_i to node v_j be represented by IA_{ij}. Similarly, NA_i represents the number of non-interactive activities at node v_i. Higher the number of interactive activities, higher the heat accumulated at node v_i as NAs act as the source of heat [DL14]. The accumulated heat is then slowly diffused among all the neighbors through the outgoing edges during diffusion process. The amount of heat diffused by v_i, denoted by $HD_i(t)$, is proportional to NA_i and is normalized by $MAX(NA)$. $MAX(NA)$ represents the largest number of non-interactive activities in V.

$$
HD_i(t) \propto 1 - \frac{NA_i}{MAX(NA)}
\tag{4.17}
$$

$$HD_i(\Delta t) = \alpha \Delta t f_i(t) \left(1 - \beta \frac{NA_i}{MAX(NA)}\right) \tag{4.18}$$

where β is the weight for non-interactive activities and is between 0 to 1. It is used in order to avoid the condition, $MAX(NA) = NA_i$.

Similarly, heat received by each node v_i from all of its neighbors $v_j s$ is represented by HR_i. This heat is proportional to the number of interactive activities that v_j has performed with v_i and is normalized by the total number of IAs of v_j performed with its neighbors.

$$HR_i(t) \propto \frac{IA_{ji}}{\sum_{k:(v_j,v_k)\in E} IA_{jk}} \tag{4.19}$$

$$HR_i(\Delta t) = \alpha \Delta t \sum_{k:(v_j,v_i)\in E} f_j(t) \frac{IA_{ji}}{\sum_{k:(v_j,v_k)\in E} IA_{jk}} \tag{4.20}$$

Now substituting these values in equation 4.9, we get the heat accumulated at each node v_i between time t and $(t + \Delta t)$ during the heat diffusion process.

$$\begin{aligned} & f_i(t+\Delta t) - f_i(t) = HR_i(\Delta t) - HD_i(\Delta t) \\ \Rightarrow\ & f_i(t+\Delta t) - f_i(t) = \alpha \Delta t \left(\sum_{k:(v_j,v_i)\in E} f_j(t) \frac{IA_{ji}}{\sum_{k:(v_j,v_k)\in E} IA_{jk}} \right) - \alpha \Delta t f_i(t) \left(1 - \beta \frac{NA_i}{MAX(NA)}\right) \end{aligned} \tag{4.21}$$

Using the same process as in Equations 4.13 and 4.15, we get:

$$\frac{d}{dt} f(t) = \alpha H f(t) \tag{4.22}$$

where H is a $n \times n$ square matrix such that:

$$H(i,j) = \begin{cases} \frac{IA_{ji}}{\sum_{k:(v_j,v_k)\in E} IA_{jk}}, & (v_j, v_i) \in E \\ -(1 - \frac{NA_i}{MAX(NA)}), & i = j \text{ and } d(v_i) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.23}$$

The figure 4.3 is extracted from [DL14] and shows the output of the heat diffusion based on the activities performed by each node. Here, the number of interactive activities is denoted by n and the number of non-interactive activities by $\underline{n}$. The $n \times n$ square matrix, H using the formula 4.23 is as follows:

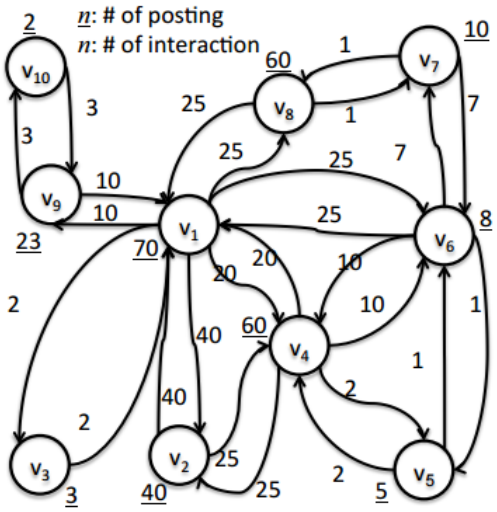


Figure 4.3: Example of a Social Network Graph [DL14].

$$H = \begin{pmatrix} -0.5 & 0.61 & 1 & 0.38 & 0 & 0.66 & 0 & 0.96 & 0.77 & 0 \\ 0.33 & -0.72 & 0 & 0.48 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.02 & 0 & -0.99 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.16 & 0.38 & 0 & -0.57 & 0.83 & 0.13 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.05 & -0.98 & 0.01 & 0 & 0 & 0 & 0 \\ 0.20 & 0 & 0 & 0.1 & 0.17 & -0.96 & 0.88 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.20 & -0.94 & 0.04 & 0 & 0 \\ 0.20 & 0 & 0 & 0 & 0 & 0 & 0.1 & -0.65 & 0 & 0 \\ 0.08 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.85 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.23 & -1 \end{pmatrix} \tag{4.24}$$

4.2.4 Extended Activity-Based Diffusion

The proposed method is an extension of the approach presented by Doo et al. [DL14]. Interactive Activities (IA) and Non-interactive Activities (NA) may have different priorities depending on each individual node. As mentioned earlier, NAs act as the heat source for the diffusion process. So the amount of heat at node v_i depends on its NA_i. However, from the diffusion point of view, interactive activities determine how much heat should be diffused to each of the neighbors. Therefore, some may argue that IA is more important than NA while others may consider NA more valuable than IA as it is the source of heat.

Two new parameters, ω_{IA} and ω_{NA} are introduced; the weights for interactive and non-interactive activities respectively. The amount of heat received by a node v_i from its neighbor v_j should be equal to the amount of heat diffused by node v_j to node v_i. Therefore, instead of using the $MAX(NA)$ as mentioned in equation

4.17 for calculating $HD_i(\Delta t)$, a new formula is derived, as shown below:

$$HD_i(t) \propto 1 - \frac{\omega_{NA} NA_i}{\omega_{NA} NA_i + \omega_{IA} \sum_{k:(v_i,v_k)\in E} IA_{ik}} \tag{4.25}$$

$$HD_i(\Delta t) = \alpha \Delta t f_i(t) \frac{\omega_{NA} NA_i}{\omega_{NA} NA_i + \omega_{IA} \sum_{k:(v_i,v_k)\in E} IA_{ik}} \tag{4.26}$$

Similarly, the weight ω_{IA} is added to the equation 4.20 but since it cancels out, it has no effect on the heat receiving process.

$$\begin{aligned} HR_i(\Delta t) &= \alpha \Delta t \sum_{k:(v_j,v_i)\in E} f_j(t) \frac{\omega_{IA} IA_{ji}}{\sum_{k:(v_j,v_k)\in E} (\omega_{IA} IA_{jk})} \\ &= \alpha \Delta t \sum_{k:(v_j,v_i)\in E} f_j(t) \frac{IA_{ji}}{\sum_{k:(v_j,v_k)\in E} IA_{jk}} \end{aligned} \tag{4.27}$$

Using the above formulas, the $n \times n$ square matrix, H becomes:

$$H(i,j) = \begin{cases} \frac{IA_{ji}}{\sum_{k:(v_j,v_k)\in E} IA_{jk}}, & (v_j, v_i) \in E \\ -(1 - \frac{\omega_{IA} IA_{ji}}{\sum_{k:(v_j,v_k)\in E} (\omega_{IA} IA_{jk})}), & i = j \text{ and } d(v_i) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.28}$$

The $n \times n$ square matrix, H for the network, as shown in the figure 4.3, looks like:

$$H = \begin{pmatrix} -0.996 & 0.615 & 1 & 0.351 & 0 & 0.581 & 0 & 0.961 & 0.769 & 0 \\ 0.328 & -0.992 & 0 & 0.439 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.016 & 0 & -0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.164 & 0.385 & 0 & -0.991 & 0.667 & 0.232 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.035 & -0.857 & 0.023 & 0 & 0 & 0 & 0 \\ 0.205 & 0 & 0 & 0.175 & 0.333 & -0.988 & 0.875 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.163 & -0.941 & 0.038 & 0 & 0 \\ 0.205 & 0 & 0 & 0 & 0 & 0 & 0.125 & -0.981 & 0 & 0 \\ 0.082 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.963 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.231 & -0.857 \end{pmatrix} \tag{4.29}$$

Activity-Based Top-K Algorithm

The proposed activity-based approach presented in this thesis is a modification to the algorithms proposed in [MYLK08] and [DL14]. For a given network with n nodes, we have to find k most influential nodes using

heat diffusion theory as discussed in section 4.2.4. For a given value k, the algorithm runs the loop k times, giving topmost influential node at each loop.

The major concern for applying heat diffusion concept is to find the heat source. Since every node is equally eligible to be a heat source, we first consider each node v_i as the only heat source in the graph. That is, at $k = 1$, every node v_i becomes the heat source and is given the initial heat of h_0 at $t = 0$, i.e. $f_i(0) = h_0$. All the other nodes except the heat source have initial heat, $f_j(0) = 0 \; \forall \; v_j \in (V - v_i)$. Now, v_i diffuses its heat through its outgoing edges and eventually covers all the nodes in the network. All the nodes that are influenced by the heat source v_i are listed in a set called InfluencedSet (IS_i). If the heat at node $v_j \in (V - v_i)$ at time t is greater than or equal to certain value θ, then we say that v_j has been influenced by v_i and is included in the InfluencedSet of v_i, (IS_i). The value θ is called Acceptance threshold and is defined as the capacity of each node to absorb heat. The node v_i with the largest IS_i is considered to be one of the most influential nodes and is included in the global influencer nodes list, V_{inflr}. In case of multiple nodes having the same maximum number of influenced nodes, we make a selection from those nodes. The node with the highest Katz centrality and maximum number of interactive activities is considered to be one of the most influential nodes.

There can be multiple heat sources at the time in the network. Even in reality, people are usually influenced by many other people. For example, before watching movies in Youtube or Netflix, one can go through the comments and ratings in order to find out whether the movie is worth watching. Based on such information, the final decision is made. Therefore, we consider multiple nodes as heat sources. When $k > 1$, all the nodes present in the global influencer nodes list, V_{inflr} along with individual $v_j \in (V - V_{inflr})$ are considered as heat sources.

Considering every node as the heat source may not be efficient from the computational point of view. Some nodes in the network may have very few connections or some may not even be connected to any nodes (outliers) at all. There is a very high chance of these nodes not making to the list of top-K nodes. So, such nodes are filtered out and not considered for heat source. This is done by using Katz Centrality, which measures the local and global influences of a node [BK15]. Any node v_j having its Katz centrality, $C_{Katz}(v_i) \geq Katz_{Th}$ is considered to be a heat source. $Katz_{Th}$ is the threshold value for desired Katz Centrality.

ALGORITHM 4: Activity-Based Top-K Nodes (ATKN)

Read network connectivity and activity data
Define Activity Weights, ω_{NA}, ω_{IA}
$\theta \leftarrow$ Heat Acceptance Threshold
$\alpha \leftarrow$ Heat Conductivity
$h_0 \leftarrow$ InitialHeat
Compute Katz Centrality of the network, $C_{Katz}(v_i \in V)$
$Katz_{Th} \leftarrow$ Katz Centrality Threshold
$V_{infld} \leftarrow \emptyset, V_{inflr} \leftarrow \emptyset$
$IS \leftarrow \emptyset$
$j \leftarrow 0$
while $j \leq k$ **do**
    $HeatDiffusion(t, V_{inflr}, h_0)$
    for *each* $v_i \in V$ **do**
        Find InfluenceSet $IS(v_i)$ that maximizes the set $\{IS(v_i) - V_{inflr} \cap IS(v_i)\}$
    end
    if *len(InfluenceSet)* > 1 **then**
        Find the node v_i with maximum $C_{Katz}(v_i)$ and IA_i
    end
    $V_{inflr} \leftarrow V_{inflr} \cup v_i$
    $V_{infld} \leftarrow V_{infld} \cup IS(v_i)$
end
Return V_{inflr}

ALGORITHM 5: HeatDiffusion(time t, set GlobalInfluencers, float InitialHeat)

$G_{inf} \leftarrow$ GlobalInfluencers
$h_0 \leftarrow$ InitialHeat
for *each node* $v_i \notin G_{inf}$ **do**
    if $Katz(v_i) \geq Katz_{th}$ **then**
        $H(t=0) \leftarrow 0$
        $H_{v_i}(0) \leftarrow h_0$
        for *each node* $v_j \in G_{inf}$ **do**
            $H_{v_j} \leftarrow h_0$
        end
        Compute Diffusion, $H(t) \leftarrow e^{\alpha t K} H(t=0)$
        for *each node* $v_j \in V$ **do**
            if $H_{v_j} \geq \theta$ **then**
                $IS(v_i) = IS(v_i) \cup v_j$
            end
        end
    end
end
Return $IS \leftarrow \{IS(v_i) | v_i \in V\}$

Chapter 5

Experiment and Analysis

5.1 Datasets

The network data was collected from the ILAB-Data Center [ila]. Three different sized networks were considered for this experiment: Facebook, Epinions, and Amazon. The algorithms 1, 2, 3 work for unweighted graph, however, weighted graphs can also be used. Since the weights have no significance in the algorithm, they can be discarded. Table 5.1 shows the information about the network data that we used during experimentation.

Datasets for Facebook Sub-networks: I and II contained 1034 and 1495 nodes respectively, and 53498 and 61922 edges respectively. The networks are both undirected and unweighted. Since the largest Eigen values, λ, of both networks, are approximately 0.0081158, the values for the parameter, α, are incremented by 0.0005, starting from 0.008; in other words, $0.008, 0.0075, 0.007, 0.0065, 0.006, 0.0055, 0.005$ and so on. Similarly, directed and unweighted datasets for Epinions sub-networks: I and II as well as for the Amazon dataset contained 1247, 1799, and 1500 nodes, respectively, and 51558, 61037 and 6010 edges, respectively. Both Epinions datasets had the same, λ, i.e., 0.01194; moreover, the values $0.01, 0.0095, 0.009, 0.0085, ...$, etc. were considered as the α. For the Amazon network, the largest Eigenvalue was approximately 0.1999.

The second approach needs a network data with node-edge information along with the activity details.

Table 5.1: Network DataSet Summary.

Network	Type	No. of Nodes	Connectivity
Facebook-I	Undirected	1034	53498
Facebook-II	Undirected	1495	61922
Epinions-I	Directed	1247	51558
Epinions-II	Directed	1799	61037
Amazon	Directed	1500	6010

Table 5.2: Instagram Media Dataset Attributes.

Media	
mediaID	media ID
authorID	user ID who created the media
TS_upload	timestamp of media upload
tagset	set of tags assigned to the media
likes	no. of likes on the media
comments	no. of comments on the media

Table 5.3: Instagram User Dataset Attributes.

User	
follower	anonymized follower ID
followee	anonymized followee ID
likes	no. of likes posted by follower on followee's media
comments	no. of comments posted by follower on followee's media
timestamps	list of timestamps of the comments

The Instagram dataset used in the experiment is collected by Ferrara et al. [FIT14]. It consists of two files: media and users. The media dataset contains records of the form: the anonymized media ID, the anonymized ID of the user who created the media, the timestamp of media creation, the set of tags assigned to the media, the number of likes and the number of comments it received. The anonymized user network contains asymmetric relations (A follows B); each edge is associated with a number of likes (by A on media created by B), the number of comments and the list of comments timestamps. The Instagram network dataset has 1,686,349 media and 44,766 nodes with 677,686 edges. Due to computational limitations, the experiment was conducted on $1000 - 1500$ nodes.

5.2 Experiment Environment

5.2.1 Experimental Setup

In order to implement centrality-based approach 4.1, the codes were written in Java programming language using a linear algebra package called JAMA [Jama 2014] and Netbeans 8.0.2 IDE environment. The algorithms 1, 2, 3 were written using the JAVA data structures such as ConcurrentHashMap, TreeMap, and

Table 5.4: Instagram Network Data: Statistics.

Media		Users	
No. of Media	1,686,349	No. of Nodes	44,766
No. of Distinct Users	2,081	No. of Links	677,686
No. of Tags	8,919,630	Avg. In-degree	15.14
No. of Distinct Tags	269,359	Clustering Coefficient	0.041
No. of Likes	1,242,923,022	Diameter	11
No. of Comments	41,341,783	No. of Communities	151
		Network Modularity	0.578

Lists. For the activity-based approach 4.2, the algorithms 4, 5 were implemented using Python scripting language in PyCharm IDE. All the experiments were conducted on 64-bit Windows 7 Enterprise (Server Pack 1) with Intel(R) Xenon(R) CPU E5-1607 0 with 3.00GHz and 16GB RAM.

5.2.2 Constraints and Parameters

Centrality-based Parameters and Constraints

The driving factor for getting an efficient list of top-K nodes based on the Katz Centrality depends on the values selected for α and β. However, β does not have any direct influence on ranking. As mentioned earlier, the study by Benzi et al. [BK15] on centrality measures showed that the choice of values for these factors affected the node rankings as they tend to result in various Katz Centralities. The centrality-based algorithms developed in this study allow users to provide their own preferred values. Since α has to be lower than the greatest Eigen value (λ), users are given the option to choose any value up to $1/\lambda$. As for β, its default value generally is assumed to be 1. According to the parameterized matrix analyses conducted by [BK14], when the value of β was between 0.5 and 2, additional information on rankings was gathered; when β was greater than 2, the results were similar to those produced by eigenvector centrality. This analysis was conducted on exponential subgraph centrality and total communicability. Therefore, it was assumed to be the same for the case of Katz centrality, to be on the safe side.

Filtering on the computed centralities was accomplished based on the values of $Const$, provided by the users and of GAC_{Katz}. For the experiment in this study, the $Const$ was considered to be the same as the global average centrality. That is, all the nodes having centralities greater than or equal to the GAC_{Katz} were considered for top-K nodes ranking, followed by $Cons1 = (GAC_{Katz} + \sigma)$, where σ is the standard deviation of the network's Katz centralities. Users may or may not know what values to provide for the different parameters: α and β; and what centrality value should be suitable for filtering purpose. Therefore, some heuristic approaches, something similar to the recommendation systems [AT05], may be used to suggest the suitable numbers for these user-specified constraints based on the past results and outputs.

Activity-based parameters and constraints

In the activity-based method, there are multiple parameters such as thermal conductivity, (α), acceptance threshold, (θ), weights for interactive and non-interactive activities, (ω_{IA} and ω_{NA} respectively), and initial heat, (h_0), whose values directly affect the heat diffusion. From network's scenario, thermal conductivity is the property of a network that determines the amount of heat allowed to flow through it. Higher the thermal conductivity, faster the heat diffusion through the body. If $\alpha = 0$, then there is no heat diffusion. If $\alpha = 1$, then almost all of the heat presented at the heat source will be diffused through its outgoing edges

[DL14]. Therefore the acceptable range of values for α is $(0, 1]$. Likewise, higher the initial heat, h_0, at the heat source, v_i, higher the amount of heat received by node, $v_j \in (V - v_i)$ and higher the chance of node v_j, of being influenced by v_i.

The acceptance threshold, θ, determines the heat absorption capability of a node. Just like every material has a different heat absorption capability, every node can have different θ values. One way to define the individual θ values for the nodes is to assign a relative incremental number to each node based on its interactive and non-interactive activities. However, for simplicity, θ is taken as a constant value for all nodes. Node, v_i is said to be influenced by the heat source, v_j, if heat at the node, v_i, at time t, is greater than or equal to the provided θ value (i.e. $f_i(t) \geq \theta$).

The activity weights, ω_{IA} and ω_{NA}, determine which activity, either IA or NA is to be favored over the other. The total weights assigned to each type of activities equals to 100%. For example, if interactive activities, IA, are to be given more priority over the non-interactive activities, NA, then ω_{IA} should be greater than ω_{NA}.

5.3 Results and Discussion

Discussion on Centrality-Based Approach

Among the results of the experimentation, the change in α values had no effect on the computation speed of the algorithms. However, there were slight changes in the number of top-K values. Figures 5.2(a), 5.2(d), 5.2(e) and 5.2(f) show the results generated by using a user-specified constraint, $Const$, (i.e., the desired centrality value) were the same as those generated by GAC_{Katz}. Taking into consideration, $Cons1 = GAC_{Katz}$ resulted in a large number of nodes becoming the candidate nodes as $\alpha \to 0$. However, the case was just the opposite when $Const = (GAC_{Katz} + \sigma)$ was used. The number of top-K nodes increases as $\alpha \to 1/\lambda$ as shown in figures 5.2(b) and 5.2(c). This was because when α is increased from 0 to $1/\lambda$, centralities increased for all the nodes. The amount by which the individual centralities increases depends on how many of its neighbors' centralities increased and by how much. The change in α naturally induces a change in the Global Centrality which is why a fluctuation of the top-K nodes is seen in the graph.

Discussion on Activity-based Approach

The experiment was conducted using various value sets of $\{k, \alpha, \omega_{NA}, \omega_{IA}, \theta, h_0\}$. The parameter k represents the desired number of influential nodes. The effect of individual parameters on the influence maximization (i.e., the number of influenced nodes) is observed by keeping other remaining parameters constant and by applying the algorithms on multiple extracted datasets, dataset I with 1000 nodes and dataset II with 1500 nodes. Figures 5.3(a) and 5.3(b) show the effect of α on the number of influenced nodes. With increasing

value of α from 0.3 to 1.0, with constants: $k = 10/20$, $\omega_{NA} = 0.4$, $\omega_{IA} = 0.6$, $h_0 = 30$ and $\theta = 0.6$, the number of influenced nodes increased significantly for each dataset. Similarly, in two independent experiments, the number of influenced nodes increased when the θ value was changed from 0.3 to 1.0 and when h_0 was incremented from 25 to 45 as shown in the figures 5.3(c) 5.3(d) and 5.3(e) 5.3(f) respectively. On changing the weights of interactive and non-interactivity weights, no significant effect was observed as shown in the figures 5.4(a) 5.4(b). $(0.3, 0.7)$, $(0.5, 0.5)$ and $(0.1, 0.9)$ pairs of values were assigned to $(\omega_{NA}, \omega_{IA})$.

The activity-based top-K algorithm (ATKN) filters out nodes having Katz Centrality less than a certain threshold value i.e., $C_{Katz} < Katz_{Th}$. For the experiment, Katz Centrality threshold value was considered to be double of the 40^{th} percentile. This drastically reduces the number of global influencers and the number of heat sources at each loop. In addition to this, the proposed algorithm 4 strictly identifies the k influential nodes. This was done by adding additional logic. If there are multiple nodes that maximize the list of global influenced nodes, then the node with the highest $C_{Katz}(v_i)$ and the largest number of interactive activities IA_i is selected to be the influential node. In contrast to the proposed approach, the Minimizing Global Overlap algorithm (MGOA) presented in [DL14], there is no such additional logic to handle multiple nodes while calculating global influenced list maximization. Because of this, MGOA identifies more than k influential nodes. Figure 5.5(a) shows that the number of influenced nodes is consistent in both ATKN and MGOA algorithms. However, as the value of k increases from 15 to 20, the number of influenced nodes by MGOA become much greater than the k value. But, on comparing the execution times of both the algorithms, ATKN algorithm is much faster than MGOA, as it includes the filtering technique as mentioned above. This filtering technique discards all those nodes having less connectivity and lower influences on their neighboring nodes.

Table 5.6 shows the output of three algorithms mentioned above: CTKN, ATKN and MGOA. A small dataset of 50 nodes was extracted from the original Instagram dataset, as shown in figure 5.1 and the algorithms were run on that dataset. For Activity based algorithms, the parameter values used were $\alpha = 1$, $\omega_{NA} = 0.4$, $\omega_{IA} = 0.6$, $h_0 = 30$, $\theta = 0.6$. For Centrality based algorithm, the parameters α and β were 0.09 and 1 respectively. Few of the top-5 nodes are common in all the three algorithms. Nodes n_{24} and n_{49} did not make it to the top-5 list in CTKN. Even though the number of non-interactive activities of nodes n_8, n_{40} and n_{49} $(576, 288, 167, respectively)$ are much higher than that of n_{21} i.e. 118, they are not in the top-5 list in ATKN because their interactions with the neighbors are much lower than that of n_{21} as well as the n_{21} had maximum the influence on its neighbors, as shown in table 5.5.

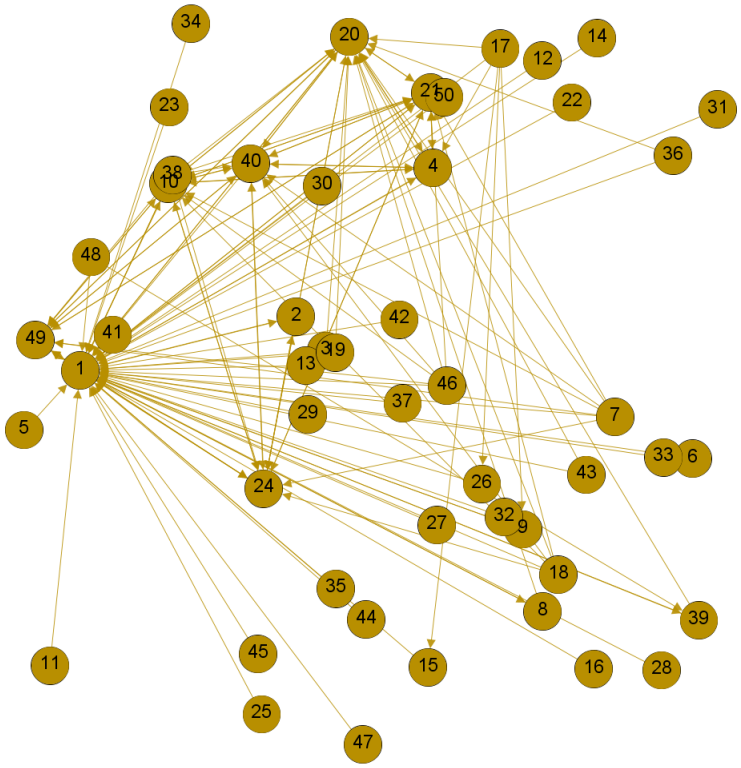


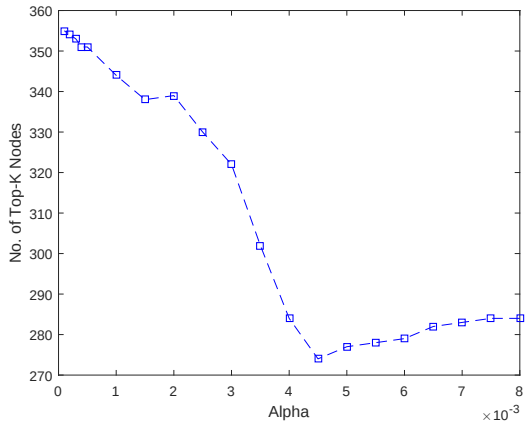
Figure 5.1: A small part of Instagram Dataset.

Table 5.5: Micro-analysis on Nodes.

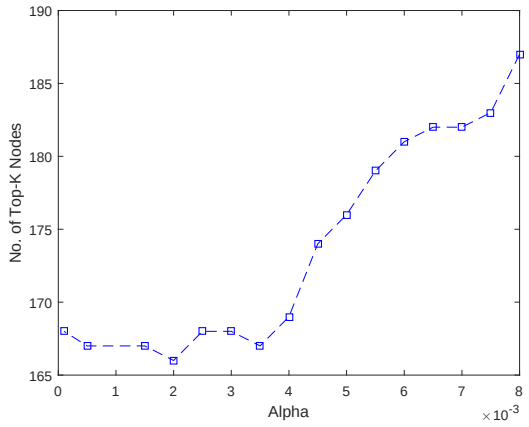
Nodes	No. of friends	No. of IA	No. of NA	Katz Centrality
n_8	2	2	576	0.07
n_{21}	7	31	118	0.35
n_{40}	7	15	288	0.30
n_{49}	6	15	167	0.32

Table 5.6: Comparison of CTKN, ATKN, MGOA.

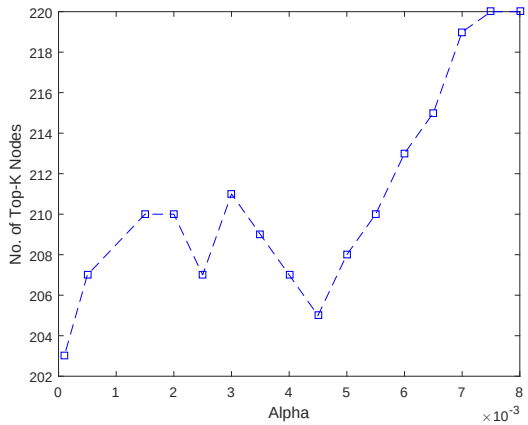
	Top-5 Nodes		
S.N	CTKN	ATKN	MGOA
1	n_1	n_1	n_1, n_{24}
2	n_{20}	n_{10}	n_8
3	n_{40}	n_{24}	n_{10}, n_{40}, n_{49}
4	n_{10}	n_{21}	–
5	n_{21}	–	–



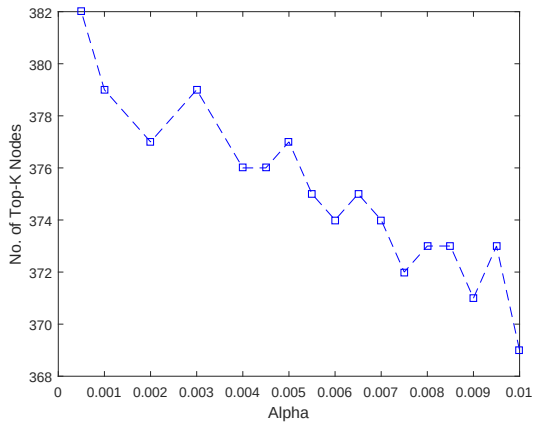
(a) Facebook-I with $Const = GAC_{Katz}$.



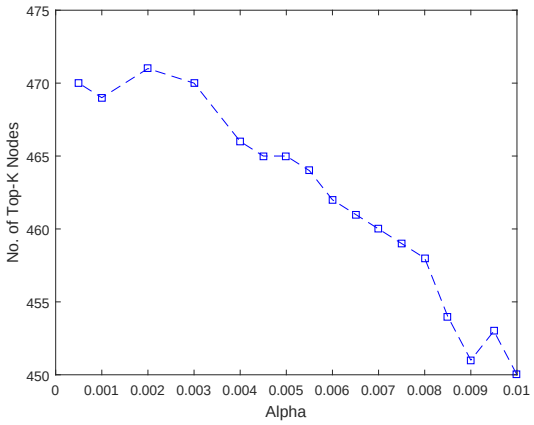
(b) Facebook-I with $Const = GAC_{Katz} + \sigma$.



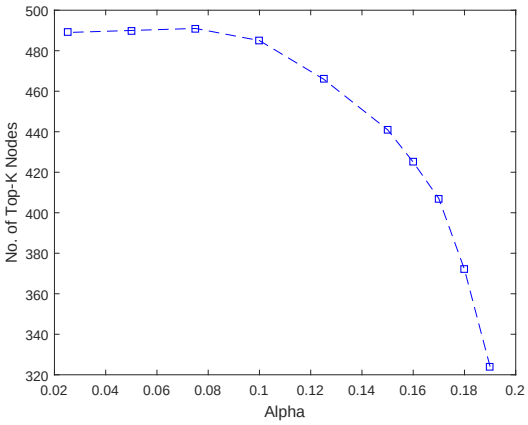
(c) Facebook-II with $Const = GAC_{Katz} + \sigma$.



(d) Epinions-I with $Const = GAC_{Katz}$.

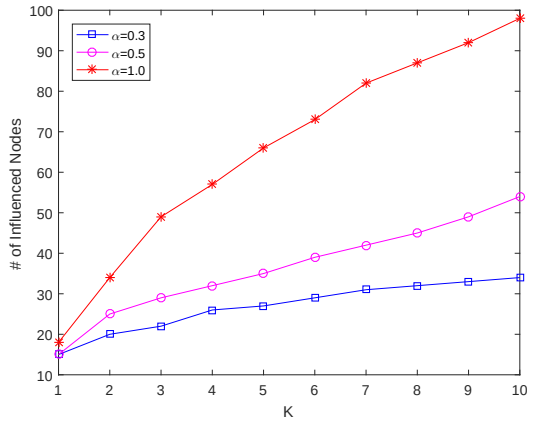


(e) Epinions-II with $Const = GAC_{Katz}$.

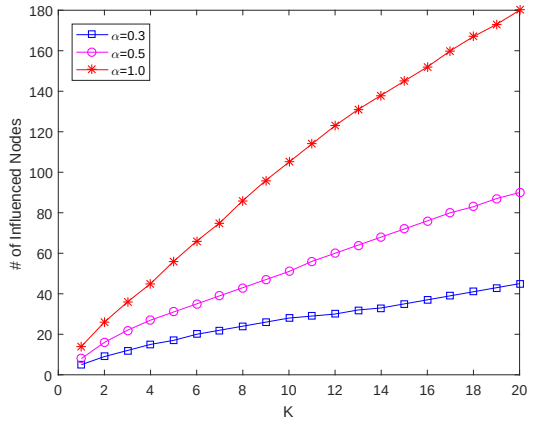


(f) Amazon with $Const = GAC_{Katz}$.

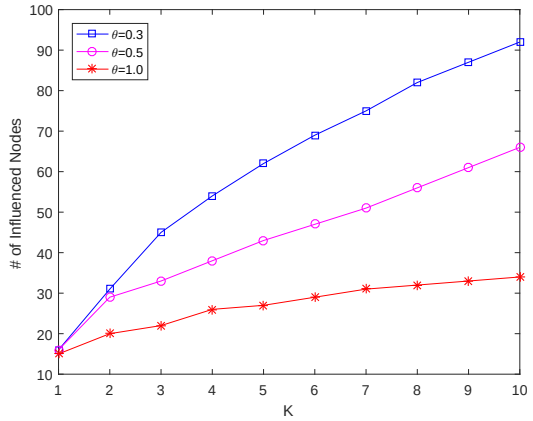
Figure 5.2: Centrality-Based Method: Experimental Results.



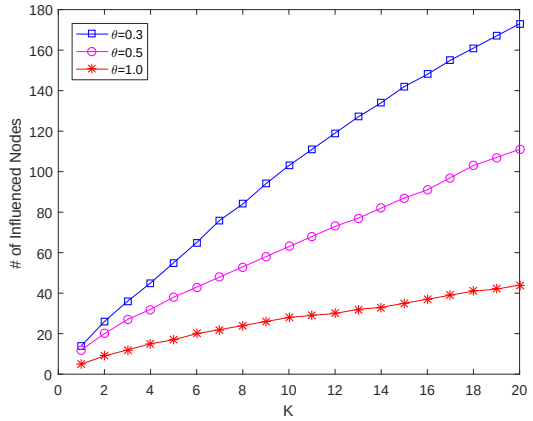
(a) Alpha (α) VS Number of Influenced Nodes with $k = 10$ and $n = 1000$.



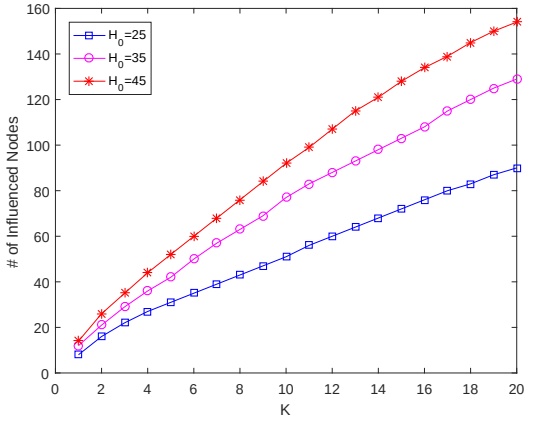
(b) Alpha (α) VS Number of Influenced Nodes with $k = 20$ and $n = 1500$.



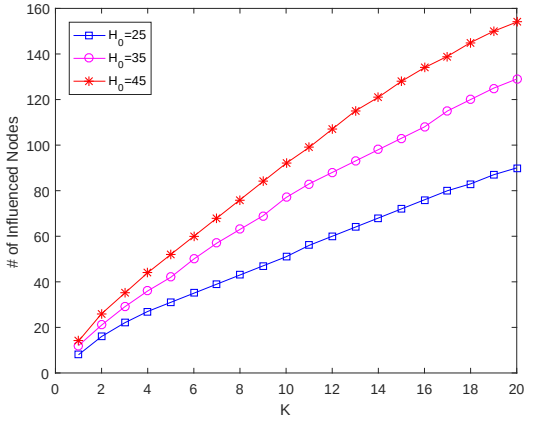
(c) Theta (θ) VS Number of Influenced Nodes with $k = 10$ and $n = 1000$.



(d) Theta (θ) VS Number of Influenced Nodes with $k = 20$ and $n = 1500$.

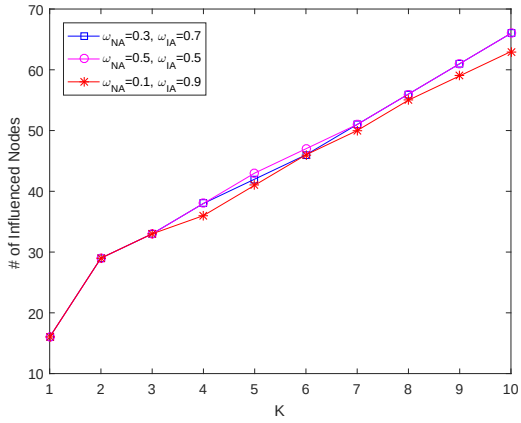


(e) Initial Heat (h_0) VS Number of Influenced Nodes with $k = 10$ and $n = 1000$.

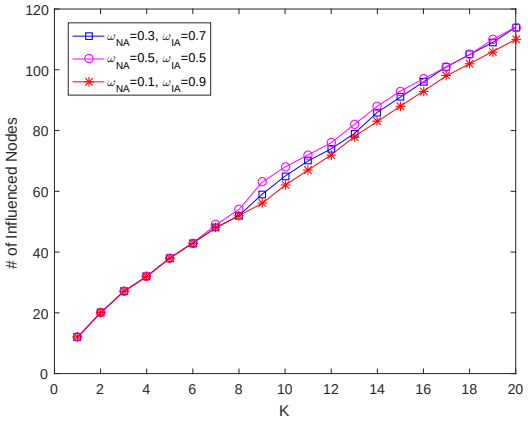


(f) Initial Heat (h_0) VS Number of Influenced Nodes with $k = 20$ and $n = 1500$.

Figure 5.3: Activity-Based Method: Experimental Results (Effect of Parameters on the Number of Influenced Nodes).

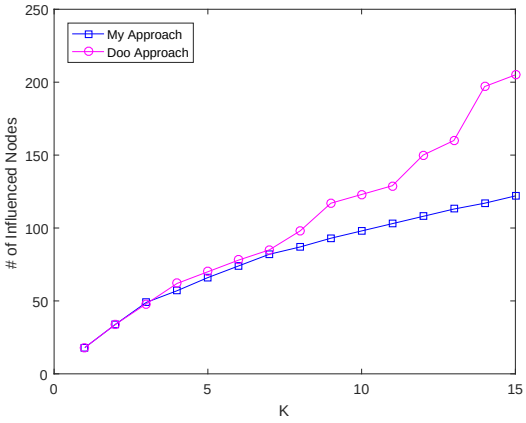


(a) Activity Weights VS Number of Influenced Nodes with $k = 10$ and $n = 1000$.

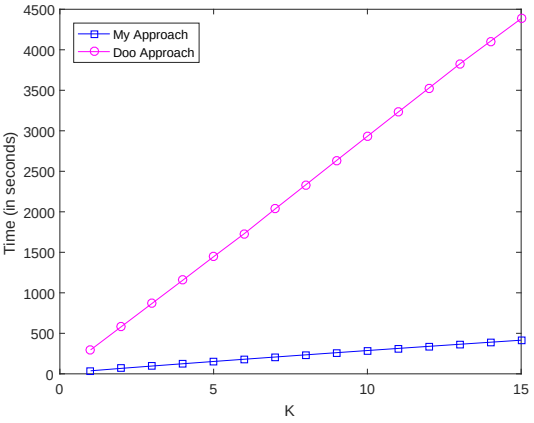


(b) Activity Weights VS Number of Influenced Nodes with $k = 20$ and $n = 1500$.

Figure 5.4: Activity-Based Method: Experimental Results (Effect of Weights on the Number of Influenced Nodes).



(a) Comparision on Influenced Nodes.



(b) Execution Time Comparision.

Figure 5.5: Experimental Results: Comparison between ATKN (proposed method) and MGOA (Doo's method) [DL14].

Chapter 6

Conclusion and Future Work

There has been a growing interest among graph specialists in the identification of most important nodes in big networks and centrality measures are one of the popular approaches for this purpose. The centrality-based algorithms in this study used the Katz centrality as a measure to discover the top-K nodes in the network. The centrality was computed using the user-preferred values for the parameters α and β. Nodes were filtered out on the basis of user-preferred centrality value. Furthermore, nodes satisfying the condition, $LAC_{Katz} \geq GAC_{Katz}$, were considered to be one of the top-K nodes. Usage of the centrality measures resulted in obtaining only those nodes that were important topologically. However other factors affect the nodes' importance, such as how active they were, the different types of activities they performed, their overall performances, etc.

The activity-based model (ATKN) was proposed to incorporate the nodes' activeness in determining the top-K nodes. This method used the concept of heat diffusion using the interactive and non-interactive activities. These activities play an important role in the amount of heat received and diffused during the diffusion process. With the help of Katz Centrality values, the nodes, $v_i \in V$, with $C_{Katz}(v_i)$ less than certain specified threshold value, would not be considered as heat sources; and hence, would not qualify to become one of the top-K nodes, as such nodes have lower connectivity and minimal influences on the neighboring nodes. In comparison with the existing activity-based model, ATKN was more efficient in terms of the execution time and also in identifying k most influential nodes.

In future, the experiments will be executed using the complete network datasets as it will give a much concise result for meaningful analysis. Future research will be related to incorporating activity analysis along with community detection algorithms for detecting the top-K nodes in each community within the network. Applying heat diffusion concept in the community structure would provide in-depth knowledge of nodes' behavior and their closed interactions with one another within the communities they belong. There are various community detection algorithms. Few of the most popular ones are InfoMap [BELR14],

GirvanNewman algorithm [GN02] and Louvain Method [BGLL08]. Many studies on community detection algorithms proved that InfoMap algorithm is one of the most efficient algorithms [OLC11] [LF09].

Bibliography

[AHDBV05] José Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. k-core decomposition: a tool for the visualization of large scale networks. *arXiv preprint cs/0504107*, 2005.

[AT05] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[BELR14] Ludvig Bohlin, Daniel Edler, Andrea Lancichinetti, and Martin Rosvall. Community detection and visualization of networks with the map equation framework. In *Measuring Scholarly Impact*, pages 3–34. Springer, 2014.

[BGLL08] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[BK14] Michele Benzi and Christine Klymko. A matrix analysis of different centrality measures. *arXiv preprint arXiv:1312.6722*, 2014.

[BK15] Michele Benzi and Christine Klymko. On the limiting behavior of parameter-dependent network centrality measures. *SIAM Journal on Matrix Analysis and Applications*, 36(2):686–706, 2015.

[BMBL09] Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. Network analysis in the social sciences. *science*, 323(5916):892–895, 2009.

[Car13] B.V. Carolan. *Social Network Analysis and Education: Theory, Methods & Applications*. SAGE Publications, 2013.

[CZ12] Thiago H Cupertino and Liang Zhao. Using katz centrality to classify multiple pattern transformations. In *Neural Networks (SBRN), 2012 Brazilian Symposium on*, pages 232–237. IEEE, 2012.

[DL14] Myungcheol Doo and Ling Liu. Extracting top-k most influential nodes by activity analysis. In *Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on*, pages 227–236. IEEE, 2014.

[FIT14] Emilio Ferrara, Roberto Interdonato, and Andrea Tagarelli. Online popularity and topical interests through the lens of instagram. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 24–34. ACM, 2014.

[GL11] Rumi Ghosh and Kristina Lerman. Parameterized centrality metric for network analysis. *Physical Review E*, 83(6):066118, 2011.

[GN02] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[GO12] Emre Guney and Baldo Oliva. Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PloS one*, 7(9):e43557, 2012.

[Gol13] Jennifer Golbeck. *Analyzing the social web*. Newnes, 2013.

[GR05] O Maimon Goethals and L Rokach. The data mining and knowledge discovery handbook, 2005.

[HR05] Robert A Hanneman and Mark Riddle. Introduction to social network methods, 2005.

[ila] Ilab: Interdisciplinary research institute. `http://www.ilabsite.org/`.

[Inc16a] Facebook Inc. Facebook newsroom - statistics, 2016. [Online; accessed 15-March-2016].

[Inc16b] Statista Inc. Social networks - statistics and facts, 2016. [Online; accessed 15-March-2016].

[IR11] Muhammad U Ilyas and Hayder Radha. Identifying influential nodes in online social networks using principal component centrality. In *Communications (ICC), 2011 IEEE International Conference on*, pages 1–5. IEEE, 2011.

[Kat53] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[KGH+10] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888–893, 2010.

[KK16] Muhammed Erkan Karabekmez and Betul Kirdar. A novel topological centrality measure capturing biologically important proteins. *Molecular BioSystems*, 2016.

[KKT15] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. *Theory OF Computing*, 11(4):105–147, 2015.

[KL02] Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th international conference on machine learning*, pages 315–322, 2002.

[KS08] Dirk Koschützki and Falk Schreiber. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene regulation and systems biology*, 2:193, 2008.

[KSN07] Masahiro Kimura, Kazumi Saito, and Ryohei Nakano. Extracting influential nodes for information diffusion on a social network. In *AAAI*, volume 7, pages 1371–1376, 2007.

[KVD07] Parmeshwar Khurd, Ragini Verma, and Christos Davatzikos. Kernel-based manifold learning for statistical analysis of diffusion tensor images. In *Information Processing in Medical Imaging*, pages 581–593. Springer, 2007.

[LF09] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

[LHL+12] Bi-Qing Li, Tao Huang, Lei Liu, Yu-Dong Cai, and Kuo-Chen Chou. Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network. *PloS one*, 7(4):e33393, 2012.

[LL05] John D Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. 2005.

[LMJ14] Carson Kai-Sang Leung, Richard Kyle MacKinnon, and Fan Jiang. Reducing the search space for big data mining for interesting patterns from uncertain data. In *Big Data (BigData Congress), 2014 IEEE International Congress on*, pages 315–322. IEEE, 2014.

[LZLD15] Meizhu Li, Qi Zhang, Qi Liu, and Yong Deng. Identification of influential nodes in network of networks. *arXiv preprint arXiv:1501.05714*, 2015.

[MYLK08] Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King. Mining social networks using heat diffusion processes for marketing candidates selection. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 233–242. ACM, 2008.

[New10] Mark Newman. *Networks: an introduction*. OUP Oxford, 2010.

[OLC11] Günce Keziban Orman, Vincent Labatut, and Hocine Cherifi. Qualitative comparison of community detection algorithms. In *Digital Information and Communication Technology and Its Applications*, pages 265–279. Springer, 2011.

[OR02] Evelien Otte and Ronald Rousseau. Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, 28(6):441–453, 2002.

[PSV01] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):32003203, 2001.

[SBV09] M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, 106(16):6483–6488, 2009.

[SC11] J. Scott and P.J. Carrington. *The SAGE Handbook of Social Network Analysis*. SAGE Publications, 2011.

[WLY14] Pei Wang, Jinhu Lü, and Xinghuo Yu. Identification of important nodes in directed biological networks: A network motif approach. *PloS one*, 9(8):e106132, 2014.

[ZAL14] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 2014.

[ZCL05] Dell Zhang, Xi Chen, and Wee Sun Lee. Text classification with kernels on the multinomial manifold. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 266–273. ACM, 2005.

[ZZC11] Yunlong Zhang, Jingyu Zhou, and Jia Cheng. Preference-based top-k influential nodes mining in social networks. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on*, pages 1512–1518. IEEE, 2011.

Curriculum Vitae

Graduate College
University of Nevada, Las Vegas

Sweta Gurung

Degrees:

Bachelor of Engineering in Computer Engineering 2012
Kathmandu University, Nepal

Thesis Title: Top-K Nodes Identification in Big Networks Based on Topology and Activity Analysis

Thesis Examination Committee:

Chairperson, Dr. Justin Zhan, Ph.D.
Committee Member, Dr. Laxmi Gewali, Ph.D.
Committee Member, Dr. Fatma Nasoz, Ph.D.
Graduate Faculty Representative, Dr. Darren Liu, Ph.D.