IDENTIFICATION AND ANALYSIS OF INFLUENCE IN SOCIAL NETWORKS: USER CENTRIC APPROACH

Thesis

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

SUMITH N.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA,
SURATHKAL, MANGALURU - 575025

JULY, 2017

If I have seen further
it is by standing on the shoulders of the Giants.
- Sir Issac Newton

Dedicated to
my beloved family, teachers and friends
who stood by me in this journey

DECLARATION

by the Ph.D. Research Scholar

I hereby declare that the Research Thesis entitled **Identification and Analysis of Influence in Social Networks: User Centric Approach** which is being submitted to the **National Institute of Technology Karnataka, Surathkal** in partial fulfilment of the requirements for the award of the Degree of **Doctor of Philosophy** in **Computer Science and Engineering** is a *bonafide report of the research work carried out by me*. The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.

CS12F07, Sumith N.
(Register Number, Name & Signature of Research Scholar)
Department of Computer Science and Engineering

Place: NITK, Surathkal.
Date: July, 2017.

CERTIFICATE

This is to *certify* that the Research Thesis entitled **Identification and Analysis of Influence in Social Networks: User Centric Approach** submitted by **Sumith N.**, (Register Number: CS12F07) as the record of the research work carried out by her, is *accepted as the Research Thesis submission* in partial fulfilment of the requirements for the award of degree of **Doctor of Philosophy**.

Prof. Swapan Bhattacharya and Dr. Annappa B.
Research Guides

Chairman - DRPC

Acknowledgements

It is a genuine pleasure to express my deep felt gratitude to all the people who have supported me.

I would like to express my deep felt gratitude towards my research supervisors **Professor Swapan Bhattacharya** and **Dr. Annappa B.** Professor Swapan Bhattacharya's belief in "love what you do and do what you love", has been a great motivator for me to give utmost priority to my work. Dr. Annappa B, as a dedicated and hardworking person has always inspired me to believe that I can achieve much more with dedication and clear vision. Both have encouraged free thinking in me, which has led to the successful completion of my work. Quality rather than quantity was their mantra that laid a good foundation for the publications during the research period. I thank them for their immense patience and insight during my research.

I am thankful to **Prof. S. Shrihari** , Department of Civil Engineering and **Dr. Aparna P.**, Department of Electronics and Communication Engineering, National Institute of Technology Karnataka Surathkal, for being a part of my Research Progress Assessment Committee (RPAC). Their valuable suggestions helped in shaping the research work.

I am thankful to the current head of the Department of Computer Science and Engineering, **Dr. Santhi Thilagam** and also the former heads of the department, **Dr. Annappa B.** and **Mrs. Vani M.**, who provided an independent working environment with all computing facilities required to carry out the research work.

I am also indebted to all the teaching staffs and my special thanks to **Dr. Shashidhar G. Koolagudi, Dr. Jeny Rajan and Dr. Manu Basavaraju**

for their support. My special thanks to **Mr. Dinesh Kamath** and I also appreciate all the support from the non teaching staffs.

I would like to express my sincere thanks to **Mrs. Saumya Hegde** and **Ms. Pushpalatha K.** for their academic and personal guidance and support, especially when my research seemed difficult and impossible. A word of gratitude to my other beloved friends. Without their support and care I would have not reached where I am today.

I express my gratitude to all my research colleagues; and specially **Mr. Vishnu, Mr. Likewin, Mr. Manoj Kumar, Mrs. Nagaratna, Mr. Sachin Patil** and **Mr. Ramteke P.B.**, for taking their time off and teaching me tools for the research work and documentation. I also thank them for reviewing my manuscripts and reports. The brainstorming sessions will remain as a cherished memory forever.

I humbly thank **National Institute of Technology Karnataka**, for providing financial assistance to participate in the conferences.

I express my gratitude to all the reviewers of the journals and conferences, for their valuable comments and expert opinions. Their comments helped in a big way to extend the experimental setup and evaluation of the presented methodology.

My heartfelt thanks to my English teacher, **Mrs. Nisha Shetty**, for her patience and interest in reading my work. She has been a wonderful person throughout and shared her knowledge that has refined the thesis.

My gratitude to all the researchers and professors who shared their knowledge during the discussions in the conferences and workshops. Their insight and advise has helped me to think out of the box. I also remember all my former teachers who have initiated my journey of knowledge and learning.

I strongly believe that, without the help, support and sacrifice of the loved ones, we cannot achieve success. I humbly thank my family for their constant support and belief in me, that has helped me complete my work. They have stood by me at all times and encouraged me to pursue my dreams. My deep appreciation to my mother, who has taken care of my family in my absence. My

sincere thanks to my husband for all the encouragements and just being there for me. My special thanks to my loving daughter for been a good child.

As I complete my research, I know in my heart that my father would have been proud of my achievement and has left behind his blessings, which will stay with me forever.

To conclude, every man has his own destiny. It is the Almighty who writes our destiny. I thank the Almighty for writing this research as a part of my destiny.

Place: NITK, Surathkal Sumith N.

Date: July, 2017.

Abstract

Social network is now becoming an indispensable part of the society. A large number of social networks play a vital role in the dissemination of information regarding various products, services, socioeconomic events etc. In addition, they influence the products one buys, places one visits; many a times whom one votes, events one attends etc. The countless ways in which social network affect lives, makes it important to understand its structure and investigate it further to make it an effective tool for various useful applications.

This research focuses on the influence maximization problem, which aims to fetch information propagation initiators, for the vast spread of information. However, picking the correct propagation initiators may not suffice for an effective optimal solution. Other aspects such as network structure and influence among users, have to be investigated.

In this work, a new model to map user's role during information propagation is presented. Along with this model, a holistic approach for influence maximization taking into consideration three aspects of social networks; i) network structure, ii) influence probability and iii) top influential users, is designed.

The first task is to fetch the sub set of users, who actively take part in the spread and adoption of information and opinions. This aspect is closely associated to target selection problem. The exponential and rapid growth of social networks in terms of users is a major challenge for its analysis. Due to the huge run time of popular influence maximization solutions, like the Greedy algorithm, distance, degree etc., it is difficult to evaluate its effectiveness in the enormous social networks. This research work addresses the scalability issue by reducing the social networks to smaller key components. This pruned network

comprises of probable adopters and spreaders of information, thus, making information propagation effective.

User influence plays an important role in social network analysis including influence maximization. Therefore, second task is to estimate user influence in social networks. In practice, influence probabilities have significant implications for applications such as viral marketing, poll prediction, political campaigns, recommendation system etc. Yet, predicting influence probabilities has not received significant research attention. In this research, Influx approach is devised to estimate user influence. This is further used to design a new variant of the independent cascade model, namely Influx-IC model. This model is used to predict the spread of information that is initiated by influential users.

The final stage is to fetch the top influential users in the social network, who can influence a vast population to adopt the information. To achieve this task, a new centrality metric is proposed. Based on this metric, two new heuristics are designed. Further, the heuristics employed with the estimated value of influence is used to predict the information diffusion in the social networks.

In the previous works the solution to influence maximization has been explored on either models, heuristics or estimating parameters such as influence. This research sets itself apart from its predecessors by identifying vital aspects that play an important role in estimating information diffusion. Further, this research proposes a holistic approach that solves influence maximization by amalgamating aspects of social network pruning, user influence and fetching top influential users. The combination of these aspects provide an effective and viable solution to predict the information diffusion in the social networks in the real world.

Keywords: social networks, user influence, information diffusion, models, NP-hard, centrality, graph simplification, estimation, heuristics.

Contents

List of Figures

List of Tables

Abbreviations and nomenclature

σ- Denotes the total number of users adopting the information.

α- the minimum activity rate

P-Set P denotes the probability of influence on every edge

$G(V, E)$- represents the social network

β- infection probability in SIR model or influence probability in RnSIR model

ACC- Average clustering coefficient

APL- Average path length

BICOT- Breadth independent cascade on timeliness

CELF- Cost effective lazy forwarding

CS-Topcent- compressive sensing Topcent

CTIC- Continuous time independent cascade

CTSIS-continuous time susceptible infected susceptible

DegRR- Round robin degree centrality

DOSIM- Double Oracle for Social Influence Maximization

FTPAS- Fully polynomial time approximation scheme

HEALER- Hierarchical Ensembling based Agent which pLans for Effective Reduction in HIV Spread

I: number of individuals who are infected at a given time.

IC- Independent cascade

IC-M- Independent cascade with meeting events

IC-N- Independent cascade with negative influence

ICOT- Independent cascade on timeliness

IRIE- Influence rank and influence estimate

LAIC- Latency aware independent cascade

LDAG- Local directed acyclic graph

LT- Linear threshold

LT-C- Linear threshold with colors

LT-M- Linear threshold with meeting events

LT-N- Linear threshold with negative influence

MIA- Maximum Influence Arborescence

MSS- Maximally influence set

MTS- Minimum target set

ORIE- Outdegree Rank with Influence Estimate

OSSUM- Online seed selection using unified model for signed network

p- influence probability

PMIA- Prefix excluding Maximum Influence Arborescence

R: number of individuals who have recovered at a given time.

RnSIR- Restrained Susceptible Infected Recovered model

Rn: number of individuals who restrain from activities at a given time

S: number of individuals who are susceptible to be infected at a given time.

SCC- Strongly connected components

SIR- Susceptible infected recovered model

SIS- Susceptible infected susceptible model

TAP-DIP- threshold activation problem with dynamic influence propagation

TSCM- Three step cascade model

α: Interaction rate of the individual.

β: Infection rate of the individual.

γ: Recovery rate of the individual.

Chapter 1

Introduction

You cannot teach a man anything; you can only help him discover it in himself.

-Galileo

The role of social network in the information diffusion process is discussed in this chapter. Further, it discusses various case studies, chosen from diverse domains, where in social networks has been employed as a medium to spread information to a vast population. The motivation to study the information diffusion is also explored. Furthermore, influence maximization is introduced.

1.1 Information diffusion in social networks

The Web 2.0 provides a range of applications that have huge impact on people. Social network is one such application that plays a very important role in connecting people across the world. It also plays an important role in the promotion of information, marketing, polls and has a huge impact on the economic growth of the country as well. Therefore, in recent years, the social network and methods of social network analysis have attracted considerable interest and curiosity among researchers. The term social economics, reflects the importance of social networks in economic transactions. In the era of cloud computing, the social media has proved to be more effective in business-related strategies (Rauch, 2007).

One of the popular rapid growing repository of massive data is the social

networks. The ability to store, track and analyze massive amount of data, depends on the technology. The availability of large scale data, prompts new research directions, computational frameworks and new opportunities to explore real world problems. The data in social networks are characterized by five Vs: velocity, volume, variety, veracity and value. These five characteristics make the analysis and application of social networks in various applications, a challenge. To deal with this, computational science is involved in social network analysis. This merger of social networks to computational science has created social computing. Social computing is defined as computational facilitation of social studies and human dynamism, as well as the design and use of information and communication technologies that consider social context (Surowiecki, 2005).

In decisions related to product adoption, the importance of user influence is quite evident. In many products or service choices, where some sort of standards are needed, or where individuals care about the compatibility of their product choices with those of his/her friend's, one cannot view the decisions of individuals in isolation. The term network externalities embodies such relationships. On one hand, consumers shift to new technology simply because their friends also have opted to do so. On the other hand, consumers stay in, on an inferior technology simply because it is pervasive, even when it is clear that some other technology is superior. Over a period of time, this results in interesting dynamics in product marketing and is used by firms to sell products. An early example of the explicit modeling of network structures, with some perspective on their influence on economic outcomes, came through the works of Myerson (2003) in the cooperative game theory literature. The game theory relies on the premise that the users can cooperate only when they are connected. People who can communicate can cooperate and generally cooperation leads to higher production. To understand these connections graphs were used. Thus, graph representations became an important part of game theory and social network analysis.

In recent years, there has been tremendous interest in the phenomenon of

influence exerted by users of a social network on other users and how it propagates in the network. It is observed that when a user sees his/her online friend performing an action such as joining some community, playing a game, bidding on article, sharing photos, writing comments, there is a high probability that he/she too repeats the same action. This is shown in figure 1.1, where user A reads an article and writes comments and within few minutes his/her social contact, user B also reads the same article.

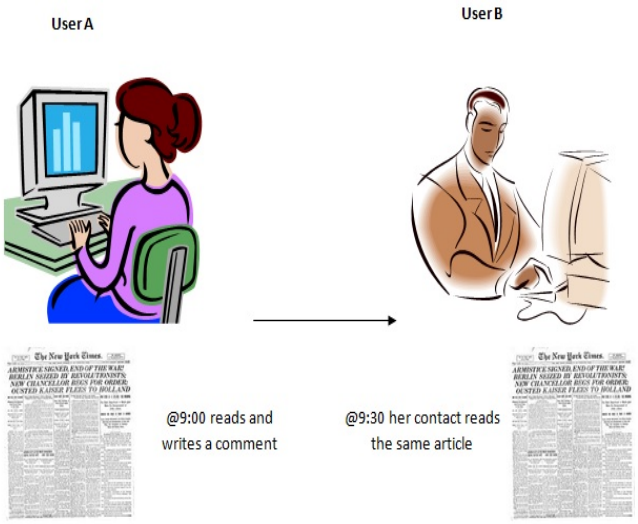


Figure 1.1: Influence in social networks

The users are definitely influenced by the action of the most popular contact. The most prominent application of such user influence is seen in viral marketing (Leskovec et al., 2007a). Enterprises use social network as a medium to enable better sales and promotions of their brands. One of the major decisions in marketing deals with the allocation of given budget among users who can influence their peers in such a way that eventually a chain of promotion messages are passed forward. As the result of this process, a vast sale of the product is possible in short span of time. This strategy is known as viral marketing, which targets the most influential users in the network. This activates the word-of-mouth chain-reaction of information spread, in such a way that, with a very small marketing cost a very large portion of the network can be reached.

However, selecting these key users, in a very large social networks is a time consuming and challenging task.

1.2 Viral marketing in the real world

With billions of users in social network, it has become the most powerful tool for marketing. User involvement has made viral marketing more dominant than the traditional marketing approaches. Viral marketing describes a strategy that encourages individuals to pass the message to others, creating a vast spread of information and influencing others to adopt it and propagate it further. The brand awareness is thus created by viral marketing with low cost and is more effective. The practice of viral marketing in the digital era has been around for more than a decade. The early adopters of viral marketing strategy are the HotMail, which grew to 12 million in 18 months and John West's Salmon Bear advertisement (Kirby and Marsden, 2006), to name a few. These campaigns were more successful then were expected to be. The low expenditure on popularizing products, has always attracted the enterprises towards social networks. In this section three popular cases across various domains, where viral marketing created a success story are discussed.

1.2.1 Case study 1: Fiesta ford movement

Ford had made several attempts, without much success, to market small car, since the discontinuation of their model Aspire in 1997. In 2009, Ford Motors launched *Fiesta Movement* campaign (McCracken, 2010) to promote sales. For six months, Ford gave 100 people a car to use and asked them to write their experiences on the social media. Consumers used their Fiestas for various activities including adventure trips. These consumers shared their experiences on Facebook, Twitter, YouTube and Flickr. The social media audience took great interest in these blogs and it soon resulted in massive sales of Fiesta.

The *Fiesta Movement* was the most successful social media marketing

experiment for the automotive world. The campaign news was all over the social media with 6.5 million YouTube views alone and 50,000 queries on the car from new customers. In the first week of the campaign, Ford sold nearly 10,000 cars. The Fiesta Movement cost the company a small expense as compared to the typical traditional TV campaign. In 2014, Ford used this strategy to introduce their latest Fiesta.

1.2.2 Case study 2: Why so serious?

In 2008, *Why So Serious?* campaign, an Augmented Reality Game (ARG) was launched to promote the movie, *The Dark Knight* (Treagus, 2014). Millions of users took interest in this campaign which was launched 15 months before the release of the movie. Over 10 million people participated in this campaign. Various games and rewards were announced all over the social media and participants took great interest in them. The ARG was thus able to maintain fan interest up to the release of the movie. Millions of blog/posts were seen on the social media which resulted in success of the ARG and lead to the success of the film, earning over US$ 1 billion in box office collections.

The Dark Knight Rises' promotion also saw a similar campaign. This time the participants were given graffiti to help the Gotham City Police Department find *Batman*. For every piece of graffiti found and tagged on social media, a frame of the trailer would be released. This marketing strategy, due to the massive fan interest led to the release of the trailer within few hours.

1.2.3 Case study 3: Ice bucket challenge

In 2014, to promote awareness on Amyotrophic Lateral Sclerosis (ALS), the *Ice bucket challenge* campaign was designed (Ganesan, 2016). In this challenge, a person has to pour a bucket of iced water over the head, film it and upload the same. A person who does not accept the challenge has to donate to ALS cause within 24 hours. After this challenge is accepted or a donation is made, it has to be passed to other three friends.

This campaign was popular on Facebook and Twitter with over 2.4 million tagged videos and 2.2 million tweets respectively. Due to this challenge, the views per month on Facebook grew to from 0.16 million views, to over 2.89 million views per month, resulting in huge donations to ALS. The ALS fund had received over $40 million from seven hundred thousand donors within 30 days. The ALS association had declared that the total donation received was around $100 million.

There are a number of similar successful cases where the social network was used to effectively promote information for various causes. User involvement in social networks is the driving force behind these successful campaigns. In the following sections, viral marketing is presented as an optimization problem and a solution is presented.

1.3 Motivation

With the advent of Web and Internet, real world off line social network has rapidly converted to online social networks. The popularity of social network has a great potential for many applications such as viral marketing, poll campaigning, recommendation system, crime detection, security flaws detection and so on. The users in a society are likely to be affected by the decisions of their friends. Thus, enterprises have invested in social networks for promoting new products and making profitable strategies.

Moreover, marketers have long known the commercial value of the influencers. Pushing a new product into the market requires that a circle of trust is well established in the social space. The potential consumers know that their influential friend has not shilled him/her for a freebie. Therefore, they are sure to go ahead with the recommendations. When the influencers promote new information about the service or product in the social networks, it rapidly spreads throughout the network. This phenomenon is similar to that of the epidemic virus spread in the population. Therefore, this marketing strategy is referred as viral marketing. Viral marketing is studied as an optimization problem in computer science popularly known as the influence maximization

problem (Domingos and Richardson, 2001). Although, the main objective of this research is to solve influence maximization problem, it also incorporates an effective strategy for information spread in social networks.

Influence maximization is one of the popular research topics of the decade and various solutions have already been proposed to solve it. However, each of the prominent approaches focuses on either the heuristics, parameter estimation or model formulation. These approaches, although theoretically very popular, often lack viability in the real world. The applicability of these solutions in the present scenario is questionable. Therefore, there is a need for new approaches that meet the current circumstances. To bridge the gap between theoretical formulation and real world applicability, a holistic approach is designed in this research work.

The major hindrance in implementing any strategy in social networks is its enormous size. Statistics reveal a huge growth in social network users from 0.97 billion in 2010 to 2.22 billion in 2016. Further, the growth is predicted to reach 2.72 billion by the end 2019 (Statista, 2016). With the rising number of users in social networks, the content that is generated by them also increases. Thus, the huge size of the social networks becomes a big challenge in its analysis and deployment of solutions. Also, the growing privacy concerns among social network users is a blockade in employing user data for precise outcomes. Furthermore, the user aspect is a missing component in most of the prior solutions. A user centric approach will fetch more reliable solution to information diffusion process.

Information diffusion in social networks has a huge impact on the society as well as on the enterprises. Therefore, emphasis is on developing a viable solution to solve it.

1.4 Influence maximization

Social networks plays an important role in the spread of information. One of the application of social networks is viral marketing. The enterprises have created

many success stories via viral marketing strategy. The key factor that contributes to these successful outcomes are those first few users who start the product campaign in the social networks. These initial users were picked by enterprises based on various criteria with the aim to create massive sales. Also, there is a monetary expense involved in picking these initiators, which includes giving freebies. Therefore, these individuals have to be picked with proper planing. Picking these individuals is known as influence maximization problem. The problem was first proposed by Domingos and Richardson (2001) and soon became the new research direction in the social network domain. The influence maximization problem is formally defined as follows:

Definition 1.4.1. Influence maximization problem: Given a budget k and a social network, which is represented as a graph $G = (V, E)$, where users are represented as nodes and edges indicate their relationships, the goal is to select a seed set of k users such that by initially targeting them, the expected influence spread (in terms of expected number of adopted users) can be maximized (Zhang et al., 2014).

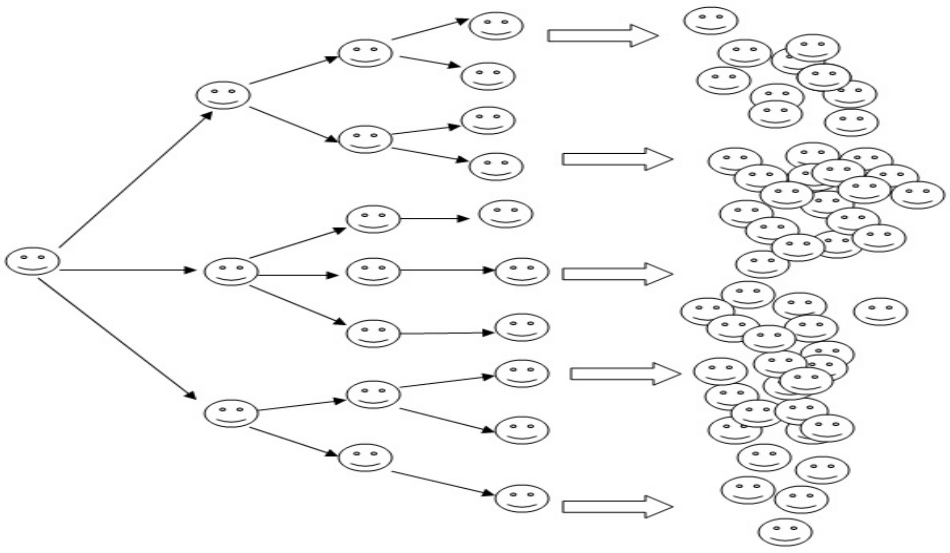


Figure 1.2: Information spread phenomenon in social network

The figure 1.2 shows the phenomenon of information spreading in a online social network. The propagation begins at a single user and spreads along the links.

The expected spread of information depends on the propagation, which is captured by the diffusion models. The solution to influence maximization is computationally NP-hard, in most of the models. Therefore, heuristics are developed to fetch results close to optimal value and also to reduce the run time involved in the computation.

1.5 Thesis overview

With the advent of Internet, one can track and predict communications and information propagation. The thesis aims at identifying and analyzing social influence in the social network, which is further used to pick a quality seed set of initial propagators in the network. The research work aims to provide contribution to influence maximization in the context of viral marketing on three aspects which are: (i) selecting social network users who contribute in the information diffusion process, (ii) estimating user influence in the social network and finally (iii) ranking users, with the aim to fetch the top influential users.

The thesis is organized as follows. **Chapter 1** discusses the role of social network in marketing and discusses the motivation for the research work. **Chapter 2** discusses the concepts and terminologies that are used throughout the thesis. Literature review and research problem are discussed in **Chapter 3**. This chapter is divided into three section, each covering various aspects involved in solving the influence maximization problem. Also, research gaps, motivation and objectives are discussed.

A new metric to evaluate users is discussed in **Chapter 4**. The thesis presents a new model to map the information spread process in social networks which is discussed in **Chapter 5**.

The thesis presents the solution to effective information spread in social networks in three aspects which are discussed in **Chapter 6, 7** and **8**. Specifically, **Chapter 6** discuses an approach to prune the social network to fetch an optimal social network. Such an optimal pruned social network should have all the properties that enable effective information propagation. In this chapter, methodology, validation techniques and results of pruning process is

discussed in detail. **Chapter 7** explains the approach to estimate user influence. Peer influence plays an important role in an effective spread of information. In majority of the existing works, the computation of influence is largely left unexplored. In the presented work, an approach to estimate user influence from interaction count between the pair of connected users is developed.

The third aspect of the solution towards influence maximization is discussed in **Chapter 8**. Ranking the social users is the theme of the thesis. Since fetching top influential users is a computationally expensive problem, various heuristics are developed. The majority of the heuristics are based on the topological aspects such as distance, degree and other centrality measures. Often the behavior aspects are not considered in order to rank users. On the contrary, in the proposed research work, the behavior aspect is used to rank the users. This chapter, discusses the new heuristic OutDegree Rank for fetching the top influential users. Finally, the conclusions and future works are discussed in **Chapter 9**.

1.6 Summary

This chapter laid down the foundations for the thesis work. It introduced social networks and its importance in the spread of the information. Various case studies discussed in this chapter uncover the real world application of the problem discussed in this research work. Further, the motivation emphasizes on the significance of carrying out the research work.

Chapter 2 discusses social network concepts, terminologies, diffusion models and datasets that has been used in this research work.

Chapter 2

Preliminaries

Everything must be made as simple as possible. But not simpler.

-Albert Einstein

This chapter explores concepts that are based on the graph theory and various terminologies that are used, and coined in this research work. The diffusion models that are used in this research are also explained. The structure of the social network and also datasets are discussed in this chapter. The chapter also highlights the difference between celebrity endorsement and influencer.

2.1 Graph theory concepts

Definition 2.1.1. Social graph: It is a graph G(V,E) of the underlying social network, where V represents the set of users and E represents the set of links.

Definition 2.1.2. Degree: The degree of a node v is the cardinality of the edges incident on v.

Definition 2.1.3. Average Clustering Coefficient (ACC): Clustering coefficient is a property of a node in a network. This metric represents the connectivity of the neighboring nodes. It is also a metric of transitivity of a graph. A node with a higher value of ACC is connected to the direct neighbors are more likely to be connected to the node, leading to a community structure in the graph. ACC is given as in Eq(2.1.1) and Eq (2.1.2).

$$C_i = \frac{|e_{jk} : v_j, v_k \in N_i, e_{jk} \in E|}{k_i(k_i - 1)} \quad . \ . \ . \ . \ . \ . \ . \ . \qquad (2.1.1)$$

where N_i is the set of immediate connected neighbors of v_i and k_i is the number of neighbors of a vertex(Yang et al., 2006).

$$ACC = \bar{C} = \frac{1}{n}\Sigma_{i=1}^{n} C_i \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \qquad (2.1.2)$$

Definition 2.1.4. Diameter: The length of the shortest path between the most distanced nodes of a graph is the diameter. It is also defined as the maximum eccentricity among the vertices of graph. Thus the diameter is calculated as in Eq (2.1.3).

$$Diameter(G) = max\{e(v) : v \in V(G)\} \quad . \ . \ . \ . \ . \ . \qquad (2.1.3)$$

When a graph has higher diameter, the nodes are not tightly linked (Rodrigue et al., 2009). In such a graph there are numerous outliers.

Definition 2.1.5. Average Path Length(APL): This measures the length of path between two distant nodes in the graph. It is given as in Eq(2.1.4).

$$\frac{2}{N(N-1)}\Sigma_{i,j\neq i} l_{ij} \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \qquad (2.1.4)$$

where l_{ij} is the distance from node i to node j (Chun-Ping et al., 2008). When a graph has smaller APL, the reachability of nodes is less time consuming and information diffusion is faster.

Definition 2.1.6. Modularity: Modularity measure indicates the partitions in the network. If a network is divided into many groups, for instance, it could provide evidence for a modular view of the network's dynamics, with different groups of nodes performing different functions with some degree of independence(Newman, 2006).

If nodes u, v have degrees d_u, d_v then any one of the m edges has probability $2\frac{d_u}{2_m}.\frac{d_v}{2_m}$ of connecting u and v. By linearity of expectation, the expected number

of edges between u and v is then $\frac{d_u.d_v}{2_m}$. Thus, the modularity of a clustering C is given as in Eq(2.1.5).

$$Q(C) = \frac{1}{2m}\Sigma_{u,v}(a_{u,v} - \frac{d_u.d_v}{2_m}).\delta(\gamma(u), \gamma(v)) \quad . \quad . \quad . \quad . \quad . \qquad (2.1.5)$$

where δ denotes the Kronecker Delta, which is 1 if its arguments are identical and 0 otherwise and $a_{u,v}$ are the entries in the adjacency matrix (Agarwal and Kempe, 2008). When a network has lower modularity, it will have fewer groups. This indicates that all nodes are well connected to form a large component and information spread is easy through the network.

Definition 2.1.7. Strongly connected components(SCC): SCC metric is related to modularity. Lower SCC implies fewer components, which indicate that the network is well connected. Tarjan's algorithm (Tarjan, 1972) is used to find the SCC.

Definition 2.1.8. Weighted graph: A graph G(V,E,W), where every edge has a weight associated with it. This weight can mean various things depending on what the graph represents.

2.2 Terminologies

Following terms are frequently used in the thesis.

Definition 2.2.1. Diffusion: According to Rogers Everett (1995), diffusion is the process by which an innovation is communicated through certain medium, over time among the members of a social system.

Definition 2.2.2. Influence: Social influence is defined as change in an individual's thoughts, feelings, attitudes, or behaviors that results from interaction with another individual or a group (Lisa, 2008).

Definition 2.2.3. Probability of influence: It is a value that reflects the chance that user u influences user v to adopt the information. It has a value in the range [0,1]. Mathematically it is formulated as follows.

For a undirected weighted social graph $G(V, E, P)$, $P = \{p_{i,j}, 1 \leq i, j \leq |V|\}$ where

(i) $e(i, j) \in E$

(ii) $\forall e(i, j) \in E, 0 \leq p_{i,j} \leq 1$

Definition 2.2.4. Contact edge: For a social graph, a contact edge $e(v_i, v_j) \in E$ is a edge connecting two users v_i and v_j by the underlying relation defined in the social network.

Definition 2.2.5. Interaction edge: For a given graph $G(V, E)$, contact edges which are used for interaction and $e(v_i, v_j) \in I$; $I \subseteq E$.

Definition 2.2.6. Activity: Any action performed by the *user* in the social network, for e.g., posts, likes, write blogs, comment, recommend, dig stories, vote etc.

Definition 2.2.7. Activity log: A log of the form $A(user, friend)$, maintained and readily available in the social network. An entry in this log indicates an interaction from *user* to $friend$.

Definition 2.2.8. Activity Rate log: A log of the form $AR(user, count)$, where *count* indicates the number of interaction of the *user* with his contacts. This log is computed from the activity log.

Definition 2.2.9. Metric: The cardinality of the set of users who adopt the information at the end of the diffusion process, is the metric to determine the influential rate of the seed set.

Definition 2.2.10. Seed set: In the social graph $G(V, E)$, a seed set is the set of initial adopters A, such that $A \subset V$, who have the capacity to influence the population to adopt the information.

Definition 2.2.11. Interaction graph: It is a weighted directed graph of the social network where the weights on the edges represents the number of interactions between the pair of users.

Definition 2.2.12. Influence graph: It is a weighted directed graph of the social network where the weights on the edges represents the user influence between the pair of users.

Definition 2.2.13. Contact Degree: In a social network graph, for a node v, its contact degree is referred to as the number of edges incident on it and is denoted as **Cd(v)**.

Definition 2.2.14. Interaction degree: In the Interaction Graph, for a node v, its interaction degree is referred to as, the number of edges incident on it and is denoted as **Id(v)**.

2.3 Diffusion models

Empirical study of diffusion in social networks began four decades ago with the works of Granovetter (1978). Currently, there are a variety of diffusion models arising from the economics and sociology communities. The most popular models are independent cascade model and linear threshold model, which are widely used in studying the information diffusion in social networks.

2.3.1 Independent cascade model

Cascading models can be better described with the probability value of a node u influences node v. This probability is represented as $p(u, v)$. The independent cascade model (Domingos and Richardson, 2001) is used to understand the process of information spread. The working of IC model is as follows. Suppose that node u is influenced (i.e. becomes active) at a time t. Then, u has an opportunity to influence every one of its neighbors v with probability *p(u,v)*. If u succeeds in activating v, then v is active from time *t+1* onwards. If not, u can never try influencing v in subsequent attempts. This process continues till no new node becomes active.

2.3.2 Linear threshold model

Linear threshold model is one of the models used to study information diffusion in social networks. In this model, each node v has a threshold θ_v, and for every $u \in N(v)$, (u, v) has a nonnegative weight $w_{u,v}$ such that $\Sigma_{u \in N(v)} w_{u,v} \leq 1$. Given the threshold and an initial set of active nodes, the process unfolds deterministically in discrete steps. At time t, an inactive node v becomes active if,

$$\Sigma_{u \in N^a(v)} w_{u,v} \leq \theta_v \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (2.3.1)$$

where $N^a(v)$ denotes the set of active neighbors of v. Every activated node remains active, and the process terminates if no more activations are possible. The threshold in this model is related to a linear constraint of edge weight. Given the thresholds in advance, the diffusion process is deterministic. The threshold θ selected by Kempe et al. are uniformly at random from interval[0,1] (Zhang et al., 2014).

2.4 Datasets

In this research, eight social network datasets of various sizes and properties are used. The description of these datasets are as follows:

1. **Email**

   The Email network is based on email data from a large European research institution for a period from October 2003 to May 2005 (18 months) generated by Leskovec and Krevl (2014). Given a set of email messages, each node corresponds to an email address. A directed edge between nodes i and j is created, if i has sent at least one message to j. This dataset has 265214 nodes and 420045 edges.

2. **HEP**

   The HEP dataset (Chen, 2009) is extensively used in many influence maximization studies. It is an academic collaboration network from the High Energy Physics-Theory section of arXiv form 1991 to 2003, where

nodes represent the authors and each edge in the network represents one paper co-authored by two nodes. It contains 15233 nodes and 58891 undirected edges.

3. **PHY**

   Arxiv of Physics collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers submitted to Physics - Theory category (Chen, 2009). If an author i co-authored a paper with author j, the graph contains a undirected edge from i to j. If the paper is co-authored by k authors this generates a completely connected (sub)graph on k nodes. The data covers papers in the period from January 1993 to April 2003 (124 months). It begins within a few months of the inception of the arXiv, and thus represents essentially the complete history. It contains 37154 nodes and 231584 edges.

4. **Wikivote**

   Wikipedia (Leskovec and Krevl, 2014) is a free encyclopedia written collaboratively by volunteers around the world. Using the latest complete dump of Wikipedia page edit history (from January, 3, 2008), all administrator elections and vote history data are extracted. This gave 2,794 elections with 103,663 total votes and 7,066 users participating in the elections. It contains 8275 nodes and 103689 edges.

5. **Twitter**

   The Higgs dataset (Leskovec and Krevl, 2014) has been built after monitoring the information spreading processes on Twitter before, during and after the announcement of the discovery of a new particle with the features of the elusive Higgs Boson on 4th July 2012. The messages posted in Twitter about this discovery between 1st and 7th July 2012 are considered. It contains 456626 nodes and 14855845 edges.

6. **Digg**

   Digg provides a social bookmarking service to over three million registered users. Digg2009 data set (Kristina, 2009) contains data about stories promoted to Digg's front page over a period of a month in 2009. It has two files, namely vote file and friend file. For each story, the list of all Digg users who have voted for the story up to the time of data collection and the time stamp of each vote are collected. The voters' friendship links are also retrieved. The semantics of the friendship links are as follows $user - id \rightarrow friend - id$ means that user-id is watching the activities of (is a fan of) friend-id. It contains 279392 nodes and 1730381 edges.

7. **YouTube**

   This is the data set crawled on Dec, 2008 from YouTube which is available at http://socialcomputing.asu.edu (Zafarani and Liu, 2009). YouTube is a video sharing site where various interactions occur between users. In particular, 30, 522 user profiles are crawled. For each user, his/her contacts, subscriptions and favorite videos are crawled. To avoid sample selection bias, we choose authors of 100 recently uploaded videos as seed set. This crawling reaches in total 848,003 users and 1,299,642 videos. However, not all users sharing all kinds of information. After removing those users, we have 15, 088 active user profiles and 76765 edges.

8. **Infectious**

   The infectious network (Isella et al., 2016), describes the face-to-face behavior of people during the exhibition INFECTIOUS: STAY AWAY in 2009 at the Science Gallery in Dublin. Nodes represent exhibition visitors; edges represent face-to-face contacts that were active for at least 20 seconds. Multiple edges between two nodes are possible and denote multiple contacts. The network contains the data from the day with the most interactions. It contains 410 nodes and 2765 edges.

2.5 Celebrity endorsement vs influencer

Marketers have long known the commercial value of influence. Two popular strategies used for product promotion are celebrity endorsement and influencer. One needs to differentiate these two strategies in order to understand its applicability.

Celebrity endorsement attaches the fame of the celebrity to the product. Seeing the celebrity using the product may result in huge sales (Geppert, 2016). However, the credibility and celebrity fame can not be viewed as one. There are many incidents where celebrities have openly refused to attach their credibility to the products they endorse (eg: Tobacco products). On the other hand, influencer creates a word-of-mouth advertising using people that are trusted in social space.

Communications with celebrities is a one way process, whereas, influencers are engaged in continuous communication and interaction, even before a product comes in the network. In a celebrity endorsement, he/she is a messenger of the products, produced by owner. The celebrity may or may not have any expertise on the product. Whereas, an influencer, is an enthusiastic salesman and creator of brand awareness. Therefore, it lends a certain credibility to the product, which often lacks in celebrity endorsement. Celebrity endorsement often come with heavy costs, where as, influencer marketing technique is not expensive. Influencer marketing dwells on trust factor whereas, celebrity endorsement achieves expected results till the fame of celebrity remains (ImpulseDigital, 2017).

In this research work, the celebrities are not considered to be influential users for endorsement of products. Instead, this research work fetches influencers among the users without accessing and storing their professional profiles. Thus, the presented approach reduces the memory usage and also computation time.

2.6 Summary

In this chapter various terminologies and definitions are introduced. Further, details on the information diffusion models like independent cascade and linear threshold models are provided. Also, description on the datasets used in this research work is given. The chapter ends explaining the difference between celebrity endorsement and influencer.

Chapter 3 unfolds various contributions towards influence maximization. The research problem is also discussed in chapter 3.

Chapter 3

Literature review

The more you know about the past; the more you are prepared for the future.

- Theodore Roosevelt

This chapter details various prominent works on influence maximization. It is divided into three subsections: approximation algorithm and heuristics, approaches for estimating influence, and strategies for pruning the network. It also discusses various research gaps, challenges, motivation and introduces the research problem.

3.1 Approximation algorithms and heuristics

The solution to influence maximization has been attempted in two streams. The first stream derives approximation guarantees for the solution under which Greedy algorithm (Domingos and Richardson, 2001), CELF (Leskovec et al., 2007b), CELF++ (Goyal et al., 2011b) have been proposed. The second stream of solutions arises from the structural concepts, where reducing the run time is the main concern. Also, aspects of temporal dynamics and work specifically on Twitter are also attempted by researchers.

3.1.1 Approximation algorithms

The seminal work on the study of information spread in social networks, with the perspective on data mining, has been proposed by Domingos and Richardson

(2001). Further, the same problem has been formalized using discrete optimization approach and has been classified as NP-hard (Kempe et al., 2003, 2005). Kempe et al. has focused on two fundamental propagation models, named Linear Threshold Model (LT) and Independent Cascade Model (IC). Although Kempe et al. proved that fetching influential users is NP-hard, they were able to derive approximation guarantee for their greedy algorithm which has been monotone and submodular. The greedy algorithm significantly outperforms the high degree and distance centrality heuristics giving 66% optimal spread. An approximation guarantee is also provided for the same. However, on the downside, the greedy algorithm uses Monte Carlo simulation to pick the initial seeds. The running time for the worst case of this algorithm is $O(n^2(m + n))$, resulting as an expensive process in terms of time, making its usage impractical for large social networks.

To reduce the run time of the greedy algorithm, Cost Effective Lazy Forwarding (CELF) (Leskovec et al., 2007b) has been proposed. CELF achieves near optimal placements and is 700 times faster than the simple greedy algorithm. Regardless of this big improvement over the basic greedy algorithm, CELF method still faced serious scalability problems. Goyal et al., have proposed CELF++ (Goyal et al., 2011b), to improve the CELF algorithm. CELF++ proposes a mechanism to avoid re-computation of the marginal gain with respect to the already selected nodes. The only concern with CELF++ is the increase in the memory usage when large social networks are analyzed. Although these are faster than the greedy algorithm, due to the use of Monte Carlo simulation, these cannot be implemented on large social networks (Shang et al., 2017). In the subsequent years, most of the solutions to influence maximization have been based on heuristics instead of approximation algorithms.

3.1.2 Centrality measures and structural concepts

A user's location within the social network, accounting to his structural properties, is the primary aspect in the information diffusion process. Based on

the network structure properties, the following solutions have been designed to improve the run time, but in turn have sacrificed the approximation guarantees.

The concept of evaluating users on the centrality scores such as; degree, betweenness, closeness and Eigenvector, are basically employed to determine influential users based on the location of the users (Hanneman and Riddle, 2005; Hinz et al., 2011; Hinz and Spann, 2008; Iyengar et al., 2011). Although some well-known global metrics such as betweenness centrality and closeness centrality, can give better results in information diffusion process, due to the very high computational complexity, they are not used in large social networks.

A method for efficiently estimating all the marginal influence degrees of a given set of nodes, on the basis of bond percolation and graph theory has been proposed by Kimura et al. (2009b). Improvements to the original greedy algorithm has been proposed through Mix greedy and New greedy algorithm (Chen et al., 2009). Picking seed nodes on degree centrality has been proposed by Chen et al. (2009) by introducing SingleDiscount and DegreeDiscount heuristics. These heuristics discount the degree of a node by already activated neighbors. Unlike Greedy algorithm, DegreeDiscount algorithm has no provable performance guarantee. In their following work, Maximum Influence Paths (MIP) has been proposed, which is a scalable heuristic to estimate coverage of a set under the IC model. Maximum Influence Arborescence (MIA) model and its extension, the Prefix excluding MIA (PMIA) model (Chen et al., 2010a) have also been designed to estimate information spread. The running time of PMIA is very sensitive to the clustering coefficient, the edge density and to the cascade size (Jung et al., 2011). Moreover, PMIA needs to maintain arborescence for each node, which consumes a huge amount of memory, making it unscalable on large social graphs.

Furthermore, various aspects of graph theory, such as community and paths, have been explored to solve influence maximization. In the LT model, through construction of Directed Acyclic Graphs (LDAG) (Chen et al., 2010b), a scalable heuristic for influence maximization has been proposed. However, this is a time

consuming process. Goyal et al., by proposing SIMPATH approach for LT model (Goyal et al., 2011c), has suggested that finding the vertex cover and enumerating all simple paths would solve the LDAG issue. Later, semi local heuristic has been proposed to pick influential users (Chen et al., 2012a), which is based on the local centrality. Local centrality is based on the nearest neighbor and next nearest neighbors of a node. This is unpractical in many real situations. For example, Zhuang et al. have observed that there are more than three million *following* relationships are newly added and three million are removed from Weibo network every day. In such a scenario, it is difficult to provide a fully observed network at any moment (Zhuang et al., 2013). Work of Mochalova and Nanopoulos (2013), has established an interplay between various centrality measures and the attitude of the network users during the diffusion process.

Wang et al., takes the community detection algorithm a step further, to mine the top influential users. Wang et al. have extended the near linear time community detection algorithm (Raghavan et al., 2007) and has proposed community based greedy algorithm (Wang et al., 2010) to mine top influential users in each community. Community detection algorithm is also seen in various works as a preprocessing task to fetch influential users (Chen et al., 2014; Shang et al., 2017; Zhang et al., 2013).

Heidemann et al. (2010), have proposed PageRank centrality, for fetching top influential users which is based on the popular PageRank approach (Page et al., 1999). Based on the similar concept, Lü et al. (2011), has proposed a random-walk-based algorithm, known as the LeaderRank to identify leaders in social networks. For LeaderRank to be workable, every individual should have a strong bidirectional connection in the network. LeaderRank, as well as PageRank has good performance for directed networks. However, they do not work well for undirected networks since, it degenerates to degree centrality in undirected networks.

Jung et al. (2011) has proposed Influence Rank and Influence Estimate (IRIE), an approach to rank the users and to estimate the user influence.

Nguyen and Zheng (2012) has designed an algorithm for belief propagation model on directed acyclic graph. This approach constructs DAG with at least one topological ordering where edges, going from a node of low rank to one with high rank, are allowed.

Finding contagious set in the expander graph (Coja-Oghlan et al., 2015) is another new direction in diffusion literature that could possibly be extended to find influential users. The prior works on influence maximization, assume the network to be static. However, Zhuang et al. have designed a solution for dynamic network and has proposed maximum gap probing (MaxG) (Zhuang et al., 2013). For this, Degree Weighted Round-Robin Probing (DegRR) has been designed, to probe each node with frequency proportional to their degrees. Since, probing is the additional task that has to be carried out before picking seed nodes, the running time increases. A similar strategy has been devised which includes a community detection procedure along with probing strategy (Han et al., 2017). Also, diffusion degree and maximum influence degree for independent cascade model has been explored by Pal et al. (2014). They design their algorithm that works on non uniform propagation probabilities, fetched from normal and uniform distributions.

The solution to influence maximization has recently come from a new domain referred to as the compressive sensing, that is predominantly used in sparse signals. Mahyar et al. design CS-TopCent (Mahyar, 2015), to identify top nodes when complete knowledge of the network topology is unavailable. The CS-TopCent approach is based on degree and betweeness centrality. Also, Belák et al. (2016) has discussed the information diffusion when partial network data is available, mainly due to privacy issue. Belak et al., referred to the nodes that were active and have participated in diffusion process, yet were hidden in the network as hidden nodes. The cascades created by these nodes have been termed as phantom cascades. Based on these two new terminologies, an approach has been proposed to study diffusion in hidden paths of the social networks. Their approach has been based on the degree discount heuristics.

Improved greedy algorithm (Wang, 2016) has been designed on on connected graphs for dynamic IC and dynamic LT models. A new type of influential nodes has been termed as super mediators by Saito et al. (2016). Super mediators are those nodes which when removed results in decrease of information spread. To identify super mediators, a new centrality referred to as super mediator degree has been proposed. Mining approach to solve influence maximization has been attempted by proposing InFlowMine. This is based on the content centered model (Subbian et al., 2016), which mines the information flow patterns using discrete events. These information flow patterns are used to solve influence maximization. The efficiency of the approach depends on efficiently keeping track of the frequency of the information flow paths. Content-influence behavior between different posts has to be determined for this approach. However, keeping track of streams for a long period of time can be time as well as a memory consuming process. Recently, He and Kempe (2016) and Chen et al. (2016), have come up with a new variant for influence maximization, known as robust influence maximization. He and Kempe, have found the top influential users in the setting where multiple influence functions have been used for the same model (He and Kempe, 2016). On the other hand Chen et.al. has discussed the solution to influence maximization, given an uncertainty in the parameter input. They have proposed LUGreedy algorithm to improve the existing Greedy algorithm.

This research explores the centrality concept on a different dimension. This work probes the nodal attribute of the user and design a new centrality measure to rank social network users.

3.1.3 Activities on Twitter

Twitter social network has been studied in the context of influence maximization. InfRank finds the influential users in Twitter network using the retweet activity (Jabeur et al., 2012). Also, InterRank approach that improves the formerly proposed Pagerank heuristic has been designed on topical similarity among the Twitter users (Sung et al., 2013). Further, works of Dubois and

Gaffney (2014) and Rudat and Buder (2015) have been designed on the contents of the tweets. These approaches have been designed for only Twitter social network and have been based on only one criteria; re-tweets, content or followers. The most recent work on twitter social network uses evidence theory and belief functions to fetch influential users. They claim this to be a working model in the scenario of uncertainty of data availability (Jendoubi et al., 2017).

The above discussed approaches were specifically used in Twitter networks. In contrast, this research proposes a strategy that is employable to any social networks in general.

3.1.4 Temporal dynamics

In the real world, diffusion process is often controlled by the temporal dynamics and cost. There are recent research contributions in this direction. Distance as a metric to analyze the social networks has also been proposed by Tang et al. (2009) with the aim to speed up the information diffusion. In their work, distance metrics to quantify and compare the delay of information diffusion processes taking into account the evolution of a network from a local and global view, has been proposed. The diffusion process has been studied for Latency Aware Independent Cascade (LAIC) model (Liu et al., 2012), which has incorporated the time latency to activate a new neighbor node. A similar aspect has been studied to design a heuristic to pick seed under the time constraint (Chen et al., 2012b). The work is designed for IC-M and LT-M model by specifically using MIA-M and LDAG-M heuristics, which are the extensions of previously designed MIA (Chen et al., 2010a) and LDAG (Chen et al., 2010b) heuristics. However, their work assumes that the dynamic network is fully observed, which may not be the case always. While these works were on discrete time, Rodriguez and Schölkopf (2012) designed the continuous time model. An optimal timing for promotions of products through incentives in a social network has been studied by Dayama et al. (2012) by proposing influence-exploit and exploit-influence strategies.

A different optimization problem to minimize time and number of seed nodes

for effective diffusion, in the context of information diffusion, is introduced by Goyal et al. Specifically, they have defined MINTSS, as the problem to find the smallest seed set and, MINTIME which aims to spread information with smallest propagation time (Goyal et al., 2013). Influence maximization problem has also been addressed as source selection problem by Saito et al., to select target nodes instead of source nodes for information diffusion (Saito et al., 2013).

Similar to that of Goyal et al., Minimum Target Set(MTS) problem has been proposed by Cordasco et al. (2015). Another approach on the constraints of time and budget has been designed by Cicalese et al. (2015) to fetch the seed set with minimum cost. This has been referred to as the Maximally Influencing Set (MIS) problem for a weighted, directed graph, on linear threshold model. Gargano et al. (2015), revisits the target selection problem and designs an approach to fetch smallest size seed set that potentially influences the entire network. The framework has been developed on a combinatorial model of influence spread under time window constraint. Also, Independent Cascade On Timeliness (ICOT) and Breadth ICOT (BICOT) models have been proposed by Han et al. (2016) to consider aspects of time acceptance ratio and breadth of influence spread. An interesting aspect of optimizing the rewards and cost involved in diffusion process has been studied by Kandhway and Kuri (2017).

Temporal dynamics of social network is explored in this research on the facet of user interactions. The time window of user activities are studied to devise a strategy to estimate user influence in social networks.

3.1.5 Diffusion models

There are also important contributions in the direction of diffusion models to study information spread. There is a close similarity between the epidemic spread and information spread in social networks. Therefore, the popular Susceptible-Infected-Recovered(SIR) (Johnson, 2009; Kermack and McKendrick, 1927), of epidemiology, is adopted to study information spread in social networks. The popular IC model and the LT model are the variations of SIR

model.

While SIR model does not allow the nodes (users) to become susceptible again after being infected, Susceptible Infected Susceptible (SIS) model (Kimura et al., 2009a), on the contrary, represents the phenomenon where an individual becomes susceptible multiple times. For predicting the information spread, SIS model allows nodes to be activated multiple times. Similarly, an approach within the stochastic framework of SIS model, namely the Continuous Time Susceptible Infected Susceptible (CTSIS), has been later proposed by Saito et al. (2010). The CTSIS allows continuous time delay and multiple activation of the same node. However, in the real world, users may not respond to the same information they receive the second time. Hence, SIS and its variants may not accurately model information diffusion in social networks.

A social network mining model namely DIFSoN (Tanbeer et al., 2012), discovers a group of influential friends from a large volume of social network data. Dynamic Independent Cascade Model (DICM) has been proposed for dynamic social networks (Wang, 2016). Models to represent information diffusion in the presence of negative influence is attempted by Jung et al. (2011), and has proposed the IC-N model. The Continuous Time Independent Cascade (CTIC) model (Tang et al., 2015), which is similar to works of Chen et al. (2012b) and Liu et al. (2012); and is also a variant of IC where a time parameter is associated with the activation of a node.

Linear threshold with colors (LT-C) (Bhagat et al., 2012) has been proposed to model the information adoption. LT-C aims at modeling scenarios where user may not adopt the product, yet may act as a bridge in the diffusion process and pass the information to his/her peer. The campaigns and the users have been considered as a bipartite graph in the bipartite influence model (Alon et al., 2012). A new model to map the diffusion of information in social networks, using Three Step Cascade Model (TSCM) (Qin et al., 2016) has been proposed to fill the gaps in IC model.

A new direction in influence maximization domain has also been seen in the

literature of autonomous agents and multi agent systems. Dhamal et al. (2016) study the diffusion as a two phase initiator selection process. Further, they also develop a budget splitting strategy for the same. Also, a real world implementation of influence maximization for the spreading of information on HIV disease has been done by Yadav et al. (2017). Also, the works of Yadav et al. (2016) and Wilder et al. (2017) are the new contributions in this domain. The HEALER (Hierarchical Ensembling based Agent which pLans for Effective Reduction in HIV Spread) model (Yadav et al., 2016) designs an adaptive software agent to pick influence users for spreading of information, where as DOSIM (Double Oracle for Social Influence Maximization) model(Wilder et al., 2017) is an algorithmic approach to solve influence maximization under uncertainty of the parameters of propagation.

Wang et al. (2017) propose stream influence maximization query, to retrieve top influential users by analyzing their contents. These stream queries are supported by their novel approach named as influential checkpoints and sparse influential checkpoints. Also, Pan et al. (2017), has designed dynamic influence propagation (DIP) model for dynamic social networks. They formulate the Threshold Activation Problem with Dynamic Influence Propagation problem (TAP-DIP), which asks for minimizing the seed set size and guaranteeing that the number of users who are influenced can reach a certain threshold within a time limit. Inclusion of rate change multiple times may be a concern for the implementation of TAP-DIP in enormous social networks. A summary of these prominent works is shown in Table 3.1.

This research work designs RnSIR model to map the information diffusion in social networks. Further, a new model named Influx-IC is proposed. This new model is a variant to the existing IC model and has edge weights that are estimated from user activities.

Table 3.1: Prominent works in influence maximization

Publication	**Contribution**
Domingos and Richardson (2001)	Network value of customers
Kempe et al. (2003)	NP hardness proof, ICM, LTM
Kempe et al. (2005)	Greedy Algorithm(GA) for DCM
Hanneman and Riddle (2005)	Eigen vector centrality
Leskovec et al. (2007b)	CELF for ICM
Hinz and Spann (2008)	betweeness centrality
Kimura et al. (2009a)	improve GA, bond percolation for ICM, LTM
Chen et al. (2009)	Mixed greedy, New greedy, DegreeDiscount, Singlediscount for ICM
(Tang et al., 2009) Saito et al. (2010)	temporal distance metrics CTSIS
Wang et al. (2010)	Community based GA for ICM
Heidemann et al. (2010)	Pagerank based approach
Chen et al. (2010a)	MIA, PMIA
Chen et al. (2010b)	LDAG
Hinz et al. (2011)	degree centrality
Iyengar et al. (2011)	degree centrality
Jung et al. (2011)	IRIE for ICM, IC-N
Goyal et al. (2011a)	propagation traces for Credit distribution model
Goyal et al. (2011b)	CELF++ ICM
Goyal et al. (2011c)	SIMPATH
Jabeur et al. (2012)	InfRank for twitter network
Chen et al. (2012a)	semi local centrality heuristic
Chen et al. (2012b)	IC-M and LT-M models
Nguyen and Zheng (2012)	Belief propagation in DAG for ICM
Tanbeer et al. (2012)	DIFSoN
Rodriguez and Schölkopf (2012)	InfluMax for CTMC
Liu et al. (2012)	LAIC-GA
Sung et al. (2013)	InterRank for twitter network
Zhang et al. (2013)	community detection
Zhuang et al. (2013)	dynamic IM, MaxG for ICM
Saito et al. (2013)	Target set selection for ICM, LTM
Goyal et al. (2013)	MINTSS, MINTIME for ICM, LTM
Coja-Oghlan et al. (2015)	d-regular graphs contagion set
Mochalova and Nanopoulos (2013)	centrality for ICM, LTM
Chen et al. (2014)	community detection
Pal et al. (2014)	Maximum influence degree for ICM
Dubois and Gaffney (2014)	topic based influencer in twitter network
Cordasco et al. (2015)	trees, cycles, MTS for LTM

Table 3.1: Prominent works in influence maximization

Publication	**Contribution**
Mahyar (2015)	CS-TopCent
Gargano et al. (2015)	Time constraint IM, TWC-TSS for SIR
Rudat and Buder (2015)	topic based influencer in twitter network
Cicalese et al. (2015)	cost bound MIS for LTM
Belák et al. (2016)	Phantom cascade ICM
Subbian et al. (2016)	InFLowMine for ICM
Saito et al. (2016)	super mediator for ICM
Wang (2016)	Improved GA for DICM, DLTM
Qin et al. (2016)	Three layer approximation approach, TSCM
Han et al. (2016)	ICOT and BICOT
Han et al. (2017)	community detection with probing strategy
Kandhway and Kuri (2017)	use of centrality measures
He and Kempe (2016)	Robust Influence Maximization
Chen et al. (2016)	LUGreedy algorithm
Yadav et al. (2016)	HEALER model based on software agents
Shang et al. (2017)	community detection
Wilder et al. (2017)	DOSIM
Wang et al. (2017)	Stream influence maximization
Pan et al. (2017)	TAP-DIP

3.2 Approaches to estimate user influence

Besides the diffusion models and heuristics, an important aspect in the solution to influence maximization is the user influence. In most of the prior works, a general strategy that has been followed to fix the value of user influence is to assign a constant value such as 0.01, 0.001 so on, as in trivalency model (Kempe et al., 2003, 2005; Leskovec et al., 2007b). This makes user influence uniform throughout the network. Yet another strategy, as in weighted cascade model, is to use the *degree* of the *node* to fix influence value i.e.,user influence is assumed to be $1/degree$. This makes influence uniform at every *node*, but non-uniform across the network. Thus, a perception is developed that the user influence is readily available as weight on the edges of the social networks. Therefore, the solutions to information spread in the initial years has left the estimation of user influence among users unexplored.

Nevertheless, attempts were made to estimate the user influence in the later years. The set of past propagation, to learn the probabilities for the independent cascade model, has been attempted using expectation maximization (Saito et al., 2008). An attempt has been made to model relationship strength in online social networks using user profile information and interaction data (Xiang et al., 2010). Detailed user profile information is needed to quantify link strength and unavailability of such information makes the approach infeasible.

Goyal et al. (2010) has proposed three classes of models to estimate user influence. The first class of model assumes that the influence probabilities are static and do not change with time. The second class of model assumes that they are continuous functions of time. In the experiments it turns out that time-aware models are by far more accurate, but they are very expensive to learn on large data sets, because they are not incremental. Thus, they propose the third class named as discrete time model, where the joint influence probabilities can be computed incrementally based on propagation data. Spread restricted to topics has also been proposed as an initial stage in picking influential users (Weng et al., 2010). A new approach, to model global influence, has been proposed in the work of Yang et al. The linear influence model has been designed to estimate the user influence by tracking the memes (Yang and Leskovec, 2010). Recent work on estimating user influence in social network, uses a number of application dependent assumptions on fixing the edge weights (Wang et al., 2011). An attempt to measure influence in Twitter using re-tweets, indegree and mention metrics has been used in the work of Cha et al. (2010). K-shell decomposition method has also been used to estimate influence by Brown and Feng (2011).

The passivity of users, during the diffusion process, is another aspect that is explored to get a realistic solution to propagation process. Romero et al. (2011) developed influence and passivity algorithm which takes into account the state of a user. Their work is developed in the context of Twitter network. Influencer-influencee model (Mohite and Narahari, 2011) has considered an approach to fetch the influence probabilities from the users in the social network.

Their approach depends on the incentives given to the users to share the information. An approach has also been attempted where link strength is quantified by user profile data (Zhao et al., 2012). Estimation of activation probabilities using latent interaction such as profile visits has also been studied previously (Jiang et al., 2013). An approach to determine influence probabilities in the presence of various confounding factors that are generally unobserved has also been attempted (Fang et al., 2013). They emphasize that these confounding factors play a significant role in adoption probability predictions.

Many of the prior works have not distinguished between propagation and adoption of information. This leads to the assumption that every user who receives information will definitely adopt it. This assumption may cause erroneous outcome. However, Wang et al. (2013) has observed that all users may not forward the information and thus their work has distinguished between influence and propagation probability. A probability based algorithm has been proposed by Wang et al., to estimate influence which includes the degree as well as the activity rate of the user. Also, an attempt to approximate propagation probability has been achieved by processing data streams. This approach is similar to the one proposed by Goyal et al. (2011a). The streaming algorithms for Stream learning of Influence Probabilities (STRIP) has been developed by Kutzkov et al. (2013). This approach is built on probabilistic approximation, min-wise independent hashing function and streaming sliding windows.

Influence propagation in the context of large scale social network has been studied by Cordasco et al. (2015) to determine the probability of propagation of information. Information spread in the context of multiple influencers has been discussed in the work of Kasthurirathna et al. (2015). The user rationality is bounded by the availability of information or cognitive capacity. Thus user rationality is taken into account for estimating the information spread in the social networks. The interplay between the user interest intensity and dynamic influence has been explored for effective information diffusion in the work of Teng et al. (2015). Estimating influence in Twitter using belief functions has

been attempted in a recent work of Jendoubi et al. (2017). Table 3.2 summarizes the existing works and their contribution.

The approaches that are discussed here, require large data stream from users. Storing and processing these large streams may cause extensive resource consumption. In addition, fetching propagation data has consequential privacy issues. Therefore, devising methods respectful of the privacy of the social networks users, is also an important concern. Moreover, approaches based on in-depth user information may not be practical in the real world, due to the restrictions on accessing such data. Thus, implementing these approaches on large social networks remain a practical challenge.

In this research, an approach is designed to estimate user influence from user activities. This approach neither requires in depth user profiling nor unreasonable memory space. Thus, the presented approach is applicable across major social networks.

Table 3.2: Prominent works on estimating user influence

Publication	**Contribution**
Saito et al. (2008)	EM for ICM method
Kimura et al. (2009a)	SIS model
Xiang et al. (2010)	estimate weight on profile similarity
Goyal et al. (2010)	propagation traces Discrete time model
Yang and Leskovec (2010)	global influence
(Weng et al., 2010)	spread on topics
Cha et al. (2010)	influence in twitter network using re-tweets, mention and indegree
Brown and Feng (2011)	k-shell decomposition
Wang et al. (2011)	application dependent assumptions
Romero et al. (2011)	influence-passitivity algorithm(IP)
Zhao et al. (2012)	user profile similarity approach
Jiang et al. (2013)	latent interaction approach
Fang et al. (2013)	confounding factors
Wang et al. (2013)	Probability Based algorithm
Kutzkov et al. (2013)	STRIP
Cordasco et al. (2015)	effort needed to influence
Kasthurirathna et al. (2015)	multiple influencers based
Teng et al. (2015)	shortest effective sequence(ISES), Intensity dependence(ID) model
Jendoubi et al. (2017)	belief function

3.3 Approaches to prune the networks

In the real world social networks, scalability is a major concern in the implementation of influence maximization solutions. Pruning the social network will eliminate inappropriate components and simplify the network. This also aids in better visualization and analysis of the social networks. In this context, various approaches to prune the networks in general, is studied. Early attempts to prune the network using structural equivalence is carried out by Lorrain and White (1971). This strategy is used to simplify graphs by means of identifying sets of node with similar structural properties. Since then, structural equivalence is used for various applications including analysis of social networks (Breiger et al., 1975; Burt, 1987, 2009; Everett, 1985; Everett and Borgatti, 1988). Breiger et al. (1975) uses structural equivalence to cluster data hierarchically whereas, Everett et al. (1985, 1988), have proposed a new measure of structural complexity based on role similarity. Burt discusses the role of structural equivalence in the diffusion of technological innovation (Burt, 1987, 2009).

Lossy network simplification approach to simplify the graph by removing edges resulting in loss of connectivity has also been attempted (Zhou et al., 2010). A trade-off between simplicity and connectivity has to be made in this approach. Further, pruning strategy to maintain connectivity to retain important edges without losing connectivity (Zhou et al., 2012b), has also been developed. This work defines the best path function and based on this, edges are given priority. This approach is applicable to probabilistic graph, flow graph and distance graph.

For the flow network to reduce the computation time, source to link flow strategy (Misiolek and Chen, 2006) has been proposed. Triangular inequality strategy (Quirin et al., 2008a) has been used to reduce computation time of pathfinder algorithm. In their subsequent work, MST-Pathfinder (Quirin et al., 2008b) has been developed to retain edges on the minimum spanning tree, in various networks including social networks. Modularity concept (Arenas et al., 2007) has been proposed to reduce graph to core components with the constraint

of maintaining modularity. The cut sparsifier (Fung et al., 2011), based on connectivity concept for undirected graph, has also been proposed to understand the strength of connectivity of the graph. Serrano et al. (2009) and Foti et al. (2011), have focused on weighted networks and have selected edges that represent statistically significant deviations with respect to a null model. An application of pruning the graph on connectivity constraint has also been designed by Mathioudakis et al. (2011).

Pruning approaches are also applied to databases and relational schema. An interesting application of pruning has also been seen in the domain of relational schema for multiple databases (Guo et al., 2007). Collective inference technique (Shawndra Hill, 2007) produces networks of smaller size that facilitates improved performance through collective inference. Similarly, pruning method proposed by Singh (2005), is developed on the structural and descriptive concept. For the case of structural pruning, sub graph is pruned keeping all the hubs and/or brokers. In the case of descriptive pruning, authors have considered employee dataset and picked position of employee, tenure and age as attribute values to prune the graph. Nonavailability of such micro details about the social network users makes this approach infeasible in practice. Hence, this approach cannot be used in general scenario. The NTree approach, primarily used for subgraph query, also simplifies the graphs to obtain its core pattern. This is further used to analyze and reconstruct the tree (Lin and Bei, 2014). The summary of various pruning approaches is available in Table 3.3.

This section discussed prominent works, that are available on network simplification concept. In most of the cases, prior works use the structural properties of the graph during pruning process without understanding whether a link is used for communication or not. These methods cannot be used to simplify social networks, since social networks comprise of users, who not only have structural value but also nodal attributes. In social networks, the nodes play an important role and edges only serve as a medium of communication. Removing a connection edge from social graph may lead to the disturbance or loss of its

structural properties which may render the sub-graph unsuitable for specific application. Also, since accuracy declines with increasing erroneous removal of nodes and edges (Borgatti et al., 2006), care should be taken, such that, any application that uses the simplified network should not produce uncertain outcome. Moreover, any effective sparsification approach must retain and reflect the important structure in the network (Foti et al., 2011). These properties or structure that one aims to preserve, depends on the application of interest. In the case of information propagation process, the aim is to reduce the social network while improving the small world properties which facilitates information propagation.

This research presents an innovative approach to prune the social networks. The presented approach is based on the nodal attribute, which sets it apart from other techniques discussed earlier in this section.

Table 3.3: Prominent works on network pruning

Publication	**Contribution**
Lorrain and White (1971)	Structural equivalence
Singh (2005)	database pruning on description
Misiolek and Chen (2006)	Source to link flow
Arenas et al. (2007)	Modularity constraint
Guo et al. (2007)	database schema pruning
Shawndra Hill (2007)	database pruning on collective inference
Quirin et al. (2008a)	Triangular inequality
Quirin et al. (2008b)	MST pathfinder
Serrano et al. (2009)	weighted network pruning
Heidemann et al. (2010)	interactions to prune network
Zhou et al. (2010)	Lossy connectivity
Fung et al. (2011)	cut sparsifier
Foti et al. (2011)	weighted network pruning
Mathioudakis et al. (2011)	social network pruning
Zhou et al. (2012b)	connectivity based pruning
Lin and Bei (2014)	NTree

3.4 Research gaps

1. In the previous works, the entire network is assumed to participate in the diffusion process. However, in the real world only about 1% of the network users are involved in the diffusion process. Considering the entire network, results in huge run time, making the solution unscalable.

2. In previous works probability of influence is assigned either as (i) uniform constant value (ii) inverse of the in-degree of the node or (iii) random value. Thus estimating user influence needs further investigation.

3. The heuristics developed in the previous works are based on the topology of the social network. User attributes are often ignored while developing user ranking algorithm.

4. Although aspects of network structure, user influence and fetching key influential users are correlated, there has been no previous evidence for considering these in combination.

5. In prior works, information diffusion is measured in reachability. However, in real world, the metric to measure diffusion should be adoption. The aim should be to have effective spread of information which will end in adoption of information.

3.5 Research challenges

Various contributions in the directions of heuristics, algorithm, model formulation and approaches for estimating user influence are available to solve influence maximization. However, there are few concerns that need attention. The challenges addressed in this research work are summarized as follows. The size of social network has to be reduced to identify the probable spreaders from unlikely spreaders. Previous solutions developed to predict the information spread in the social network involves the entire set of users. However, studies

reveal that not all users actively participate in the spread of information (Jakob, 2012; Romero et al., 2011). For this reason, it is required to prune the social network to fetch those users who participate in the information spread and are probable adopters. Fetching the core component of the social network which facilitates in the spread of information has immediate implications. By segregating actual spreaders from unlikely spreaders, the large social network is reduced to manageable size for analysis. The resulting set of users are most probable in adopting the information(products), thus closely achieving market targets of the firm. Thus, the scalability issue can be solved.

Most of the prior works evaluate users on the number of connections or friends. For information propagation to be effective, user intrinsic attributes may play an important role. Finding an alternative metric to evaluate users is one of the aims of the research work.

The popular existing information propagation model such as SIR model, do not map the users transition through various phases during information propagation. There is a need to formulate this process for understanding information propagation in the real world.

Yet another concern is quantifying the user influence. The existing approaches assume a value to user influence. There are two drawbacks on using an assumed value in the solution. First, assuming uniform influence along all social ties can lead to overestimation of information dissemination as well as lead to selection of influential users that may not be optimal (Wilson et al., 2012). As such, an assumed value will only bias the outcome. Second, influence is a behavioral attribute that changes over time. Hence, this parameter should not be made constant. In addition, interaction intensities among users and also user's inclination in adopting information is important to predict influence probabilities. Also, there could exist additional factors such as structural features that help in information spread. These factors have to be considered while deriving at the solution for predicting spread in social networks.

This directs the research on developing approaches to estimate user influence.

Also, since estimating influence and fetching top influential users are not separate issues, use of realistic influence value can get results close to reality. The several attempts which are made in this direction is often resource expensive.

Finally, most of the heuristics to fetch the top influential users, are based on concept that are deeply rooted in graph theory and is based on the structural properties of the network. The nodal properties are often ignored while ranking users. For information diffusion process to be successful involvement of users is highly appreciated. In this regard, an approach to model information diffusion on the perspective of user is needed. Keeping these concerns in mind, this research develops a user centric approach to solve influence maximization.

3.6 Research motivation

Influence maximization problem that has been defined by Kempe et al. (2003) and its applicability to solve relevant real world problems has created a great enthusiasm among researchers. Social networks such as Google+, Friendster, Flickr, Facebook, Yahoo, Twitter so on, have served as a popular medium for fast and vast spread of information. However, these popular social networking sites, have grown from few users to billions of users. Statistics reveal a huge growth in social network users from 0.97 billion in 2010 to 2.22 billion in 2016. Further, the growth is predicted to reach 2.72 billion by the end 2019 (Statista, 2016). These numbers are sure to rise in the coming days clearly showing evidence to the fact that social networks are growing rapidly. With this rapid growth, comes the gigantic amount of data in various forms, posing a big challenge to data analysis. Further, it also makes the implementation of influence maximization algorithm infeasible and resource exhaustive. Therefore, there is a need to simplify the social networks. Moreover, it is often the case that only a small fraction of the network users actively participate in the information spread (Jakob, 2012; Romero et al., 2011). Fetching these users is a challenge. When the network is reduced to manageable size of active users, solution to influence maximization becomes scalable.

Further, prior works use a pre determined value of user influence. Assuming uniform information spread along all social ties can lead to overestimation of information dissemination as well as selection of non optimal influential users (Wilson et al., 2012). As such, an assumed value will only bias the outcome. In addition to this, user influence should reflect the changes that occur in the social network (AlFalahi et al., 2014). Thus, for a more realistic outcome, user influence should be estimated. Approaches discussed in this chapter have attempted to estimate the user influence. However, these approaches are infeasible under the constraints of resources availability and consequential privacy issues.

Since the social network is represented as a graph, the existing solutions for ranking users are also based on concept of graph theory. Most of the earlier solutions are based on the structural properties of the social network and the nodal properties are often ignored while ranking users. Information diffusion process depends not only on the structure of the network but, also on the user attitude towards propagation. When nodal properties are not involved in the solutions, in a practical scenario, the outcome may not be realistic. To make influence maximization realistic, the structure of the social network, seeding strategy, together with user influence have to be considered in entirety. In this work, amalgamation of these concepts is proposed as a contribution to influence maximization. The following sections discuss the research problem and objectives in detail.

3.7 Research problem

Given a directed unweighted social graph $G(V, E)$, a constant k and an activity log $\mathrm{A}(user_i, activitycount_i)$ of users, compute the probability of influence $P = \{p_{ij}, 1 \leq i, j \leq |V|\}$ such that $0 < p_{ij} \leq 1$ and develop a user centric model to fetch the seed set I, where $|I| = k$, that maximizes the spread $\sigma(I)$ in the social network.

This is mathematically formulated as:

$$\sigma(I) = E\left[|\varphi(p, I)|\right] \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3.7.1)$$

where,

$\varphi(p, I)$ is the function for computing the spread of seed set I

$|\varphi(p, I)|$ is the active nodes at the end of propagation.

E is a function that fetches the maximum value.

$\sigma(I)$ is the expected spread under the selected seed set I.

3.7.1 Research objectives

The objectives are as follows:

1. Analyze the role of users in a social network and find a suitable metric to evaluate them.

2. Develop a user centric model to represent the role of users in the information diffusion process in a social network.

3. Estimate information diffusion by,

   (a) Developing a method to prune the social network graph and show that the pruned graph thus obtained is an ideal substitute to the original social graph.

   (b) Developing a method to estimate the influence probability of a user.

   (c) Developing a seed selection algorithm and analyze the information spread.

3.7.2 Solution framework

This research identifies that the spread of information depends on the three factors: (i) network structure (ii) user influence and iii) seeding strategy. This research work addresses issues like scalability, parameter estimation and ranking users on new metric. The framework to solve influence maximization is developed in three stages as shown in figure 3.1.

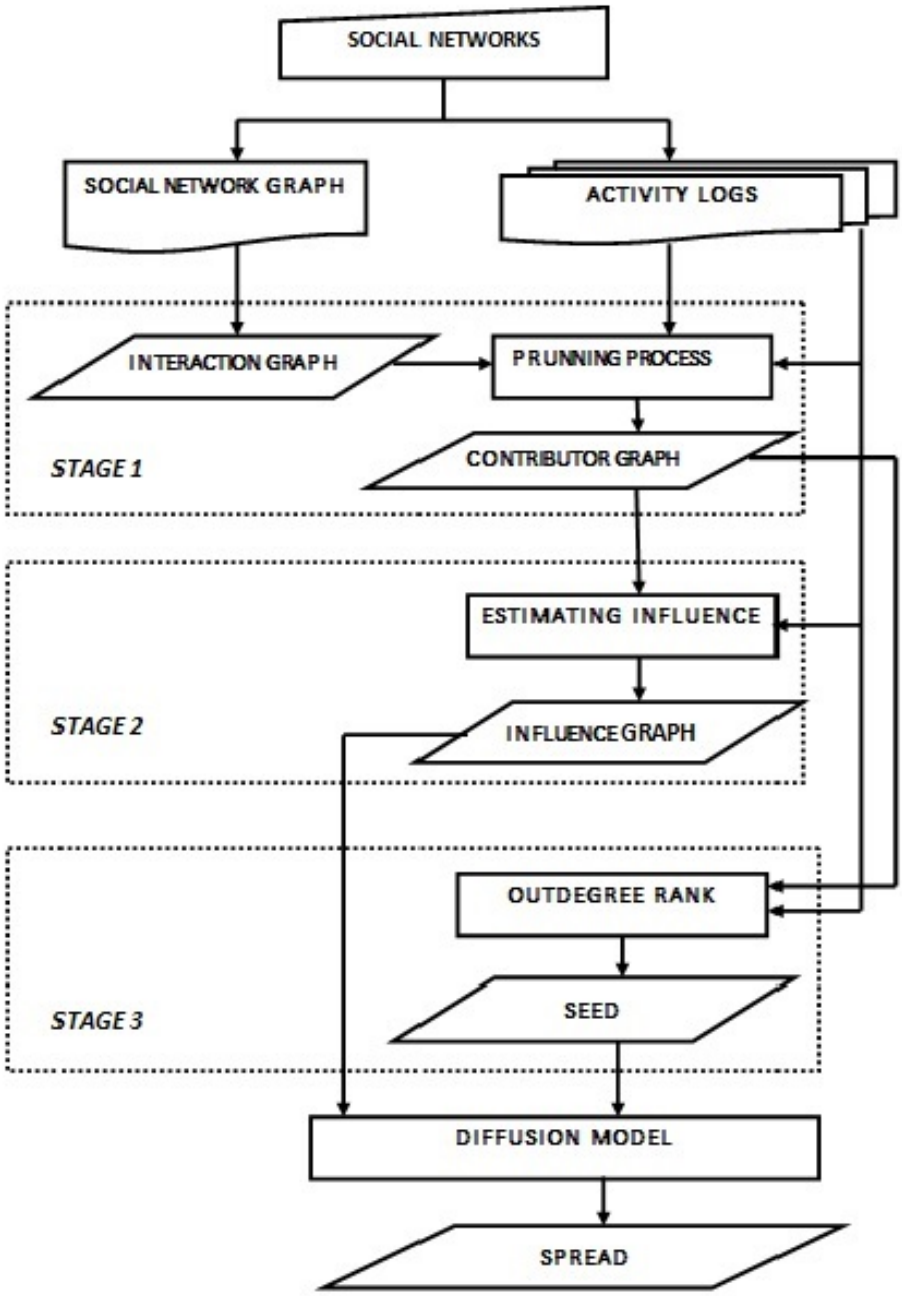


Figure 3.1: Proposed framework for influence maximization

3.8 Scope of the research work

This research work is in the context of picking the top influential users in social networks. The entire research work is on the premise of user activities which renders it a user centric approach. The user centric approach relies on the assumption that the amount of users' activities in the network determines their role in the network. However, due to the privacy concerns, detailed profiles are not available for real world applications. In this regard, an approach that uses minimum user data and yet is able to come up with a nearly accurate solution, is required. In this research, only activity logs are used which lists the activity as an event, without details on its type such as likes, posts, recommendations and so on. The focus is to know whether a pair of connected users interact well enough to influence each other. This research presents a method to determine the participation level of social network users.

In certain scenarios, availability of hardware resources may be one of the hurdles for implementing algorithms associated with social network. Therefore in this context, this research attempts to design a scalable solution to the influence maximization problem. The contributions made in this research work can be used in any environment, including the one where hardware is a major constraint.

This work further assumes that the content of the information is not altered by the forwarders of information. Thus, the aim of diffusion, which is maximizing the adoption, is not altered.

3.9 Research contributions

The contributions of this research are:

1. A new metric referred to as interaction count is introduced to estimate influence and rank users.
2. To closely depict the user's role in the diffusion process in social networks, a novel $RnSIR$ model, is designed.

3. The social networks are significantly pruned, that in turn addresses the scalability issue.
4. Influx, a new data driven approach to estimate user influence in social networks is devised.
5. A new diffusion model namely Influx-IC, is designed to predict the spread of information in the social networks.
6. A novel centrality referred to as Outdegree Rank is presented in this research. Further, the Outdegree Rank employs estimated influence to form Outdegree Rank with Influence Estimated(ORIE) heuristic and is also extended to include discount concept (Chen et al., 2009) to form ORIE-Discount heuristic. These are used to fetch the top influential users for information diffusion.
7. This research presents the combination of the aspects of network structure, parameter estimation and heuristics, as a contribution to influence maximization solution.

3.10 Summary

This chapter details an investigative survey on various approaches towards solving influence maximization. The aspects of heuristic, approximation algorithms, user influence estimation and also pruning approaches to reduce the networks in general are also discussed. The chapter highlights the research gaps that are investigated further in this research work. This chapter also presented the research problem and the objectives, that will be addressed to solve influence maximization. The scope of this research work is also discussed. The chapter ends listing the contributions made towards the influence maximization.

Chapter 4 explores in detail the need to develop a new metric to evaluate and rank users in the social networks. The social networks of various sizes are analyzed to find the new metric.

Chapter 4

Analysis of user's role in social network

True genius resides in the capacity for evaluation of uncertain, hazardous and conflicting information. -Winston Churchill

This chapter discusses the need to have a new metric to evaluate users for information diffusion. A brief discussion on various existing centrality measures are presented. Further, interaction count as a new metric to evaluate users in the social networks, for information diffusion is presented. Interaction count is used in the research at various stages, leading to the solution for influence maximization.

4.1 Background

Effective information initiators play an important role in information diffusion process. To select initiators, users are often evaluated on their structural properties. These structural properties are based on their position in the network and are termed as centrality. Centrality, measures the importance of a node/user in the network. In this section various centrality measures are discussed.

Robert (2008) has used centrality measures, such as degree centrality, closeness centrality and betweeness centrality, to quantify the importance of a node in social networks. The degree centrality assumes that, a node which has many direct connections, is at the center of the network and plays a very

important role in the information diffusion process. The second measure; namely the closeness centrality, focuses on how close a node is to all other nodes in the network. The betweeness centrality assumes that if a node is more often in the shortest paths between other nodes, it is more central to the network. Eigenvector centrality (Freeman, 1978) is yet another metric for measuring a node's popularity in a network. A node's eigenvector centrality is proportional to the sum of eigenvector centralities of all nodes directly connected to it. There are other metrics such as, PageRank (Brin and Page, 1998) and Hyperlink-Induced Topic Search (HITS) (Kleinberg, 1999), that rank the nodes individually based on their importance. In their basic form, PageRank and HITS, value a node according to the graph topology (Zhou et al., 2012a).

Hub in general terms is considered to be a center point in any activity. Concept of hub is prevalent in identifying key users in the network. Users who are in a hub position are characterized by a great potential for communication and interaction within network (Heidemann et al., 2010). Hub as a centrality measure, indicates the importance of certain key users in the network. However, in real world networks, users who are connected to large portion of the network, may not be very actively involved in the network. Also, previous study reveals that Twitter users who have the most number of followers are not the most influential users in the diffusion process (Yang and Leskovec, 2010). Here the concept of the hub fails to understand the structure of the social system. Hubs and brokers have been used for pruning the social network (Lisa, 2008). These approaches are considered as a structural pruning techniques, where all users who are hubs and/or brokers are retained in the sub-graph. However, such a sub-graph may not guarantee the diffusion and adoption of information.

Similar to identifying hubs, gatekeeper centrality approach (Narayanam et al., 2014) evaluates node on Shapley value (Shapley, 1953). Nodes that have the capability to disconnect graph are identified in this approach. In contrast, the approach designed in this research work aims at identifying nodes that are actively involved in network, thus facilitating information diffusion. Therefore,

gatekeeper strategy is not suitable for information diffusion application.

Everett and Borgatti (1999) have introduced the concept of group centrality measure and have used graph fragmentation to define it. White and Smyth (2003) have introduced personalization concept to understand a user's importance to a given subset in a social network. Further, Estrada and Rodríguez-Velázquez (2005) have introduced sub graph centrality that characterizes the participation of a node in all sub graph based on the spectral feature. Also, core centrality measure has been coined by Everett and Borgatti (2005) to evaluate the extent to which a social network revolves around a sub-network.

Role of user interactions in accurately evaluating socially enabled applications cannot be overlooked (Wilson et al., 2009, 2012). Therefore, a new metric to evaluate users is necessary. Traditional metrics have focused on topological characteristics of the social graph which are the underlying structures that capture explicit relationships between users. To better understand the true nature of relationships between users, recent works have shifted focus to measure observable social interactions. By examining records of interaction events across different links, the work has distinguished active relationships from dormant ones and has derived a more accurate predictive model for social behavior (Jiang et al., 2013).

Realizing the importance of interactions in social networks for information diffusion process, this research work emphasizes the need to evaluate the social network users on their nodal properties.

4.2 New metric to evaluate users

Majority of the user interactions are latent and information diffusion also takes place through these. Previous studies also reveal that users with high number of friends are not correlated to popularity (Jiang et al., 2013). The degree concept is the most prevalent metric to identify the influential users. Traditionally, it is assumed that, higher the number of contacts a user has, the more the chance

he/she will get to influence the contact, in product adoption, information acceptance, information diffusion etc. However, weak ties may play a more prominent role in the dissemination of information in social networks (Bakshy et al., 2012). This gives a hint that it is not the number of contact links that make a user influential. Table 4.1 gives the description of the eight datasets that are analyzed to verify this claim. The interactions for the datasets are available in the standard repositories, with the exception of PHY and HEP datasets. For HEP and PHY datasets, interactions are developed on the power law distribution pattern (Clauset et al., 2009), as majority of the social networks exhibits this pattern.

Table 4.1: Dataset description

Sl.no.	Name	No. of nodes	No. of edges	No. of interactions
1	Email	265214	420045	4220430
2	Wikivote	8275	103689	1036890
3	HEP	15233	58891	588136
4	PHY	37154	231584	2315840
5	Digg	279392	1730381	3017020
6	Infectious	410	2765	17298
7	YouTube	15088	76765	2239440
8	Twitter	456626	14855845	563069

In this research work the following two terms: degree count and interaction count are defined and used.

Definition 4.2.1. Degree count: It is the number of edges incident on the node or in simple terms, it is the degree of the node.

Definition 4.2.2. Interaction count: It is the number of edges incident on the node which are used for interactions.

All the chosen eight datasets are analyzed to obtain on interaction count and their degree count. Since the datasets have a large number of users, only a part of this analysis is shown in figure 4.1 to figure 4.8.

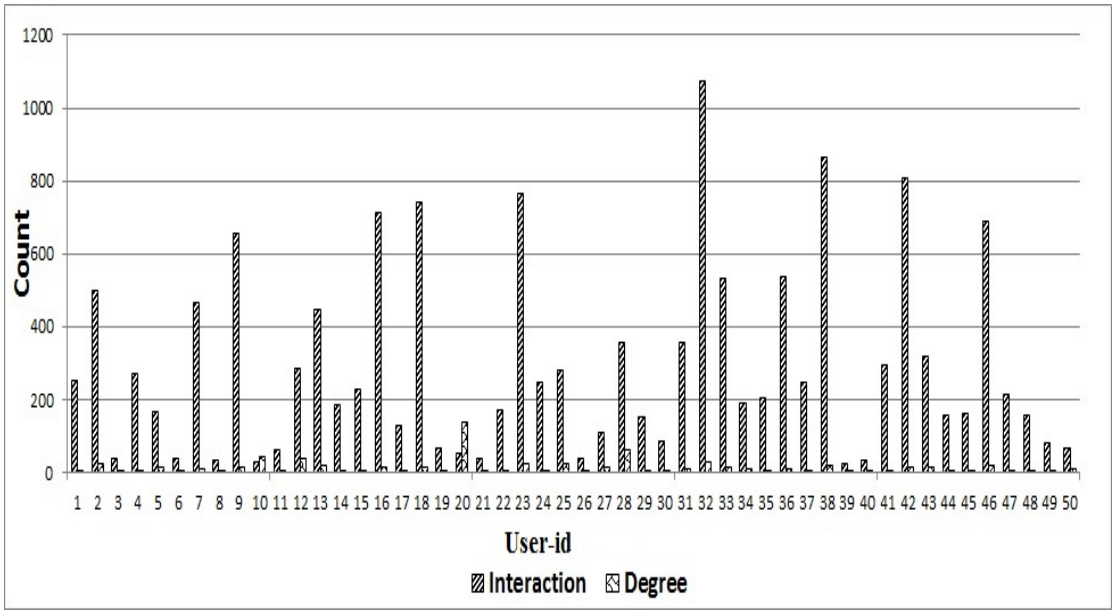


Figure 4.1: Interaction and Degree count of users in HEP dataset

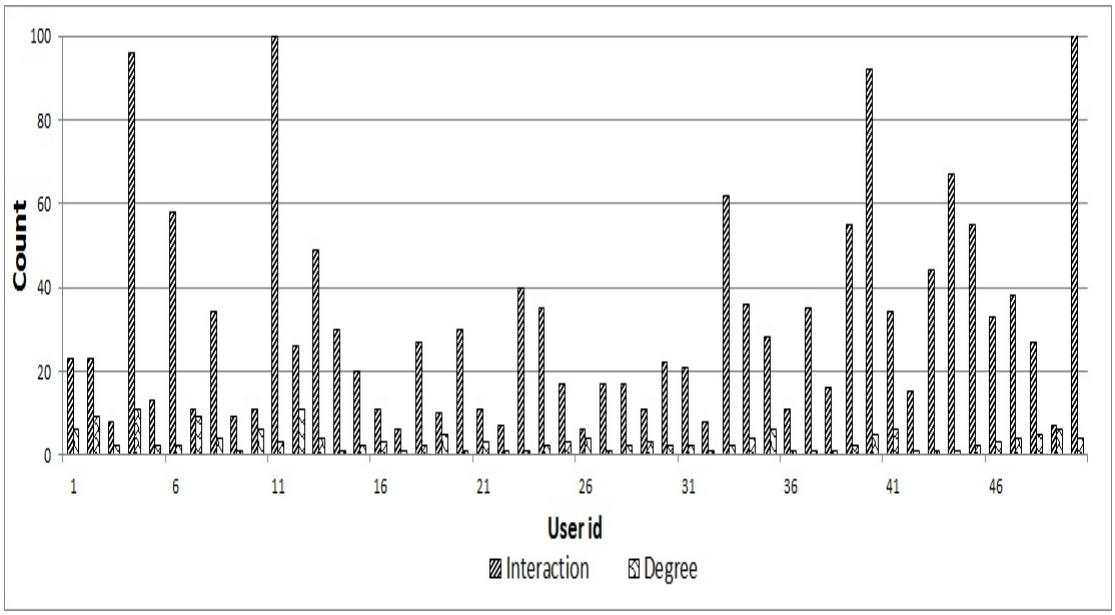


Figure 4.2: Interaction and degree count of users in PHY dataset

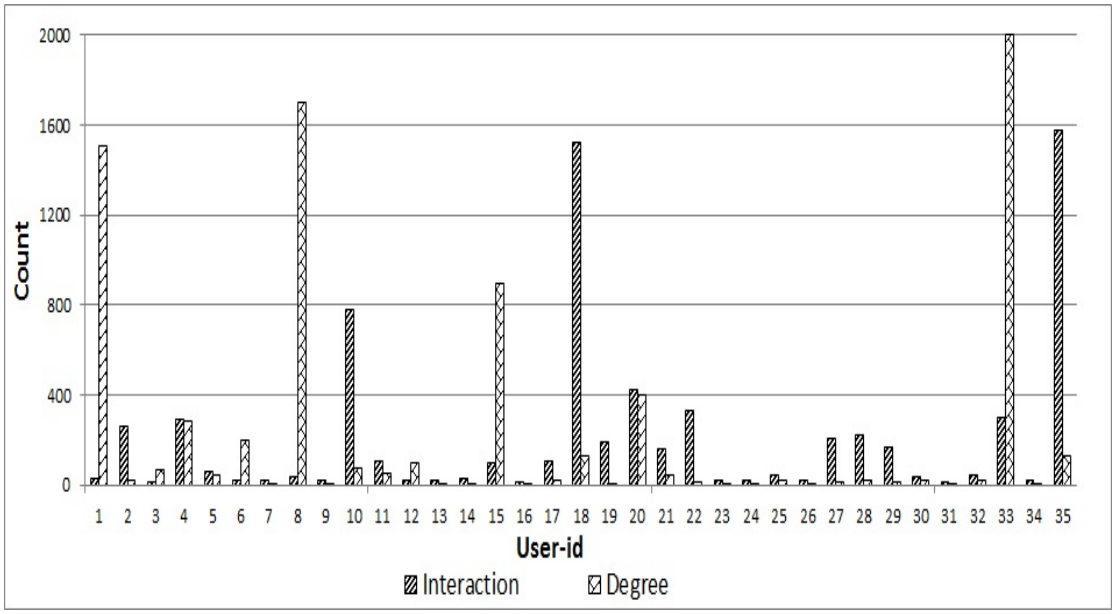


Figure 4.3: Interaction and degree count of users in Email dataset

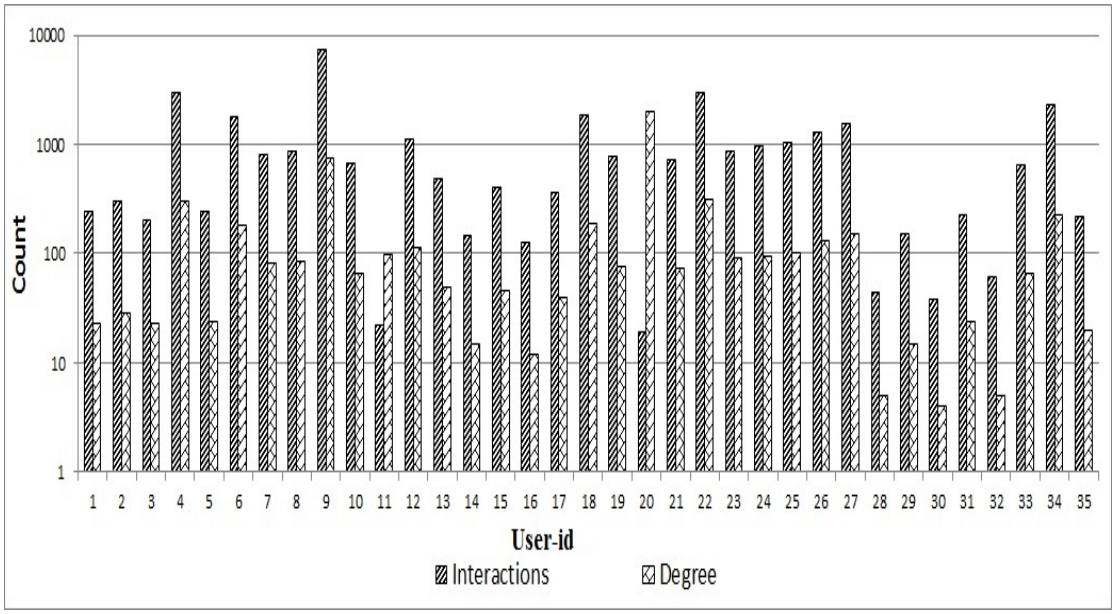


Figure 4.4: Interaction and degree count of users in WikiVote dataset

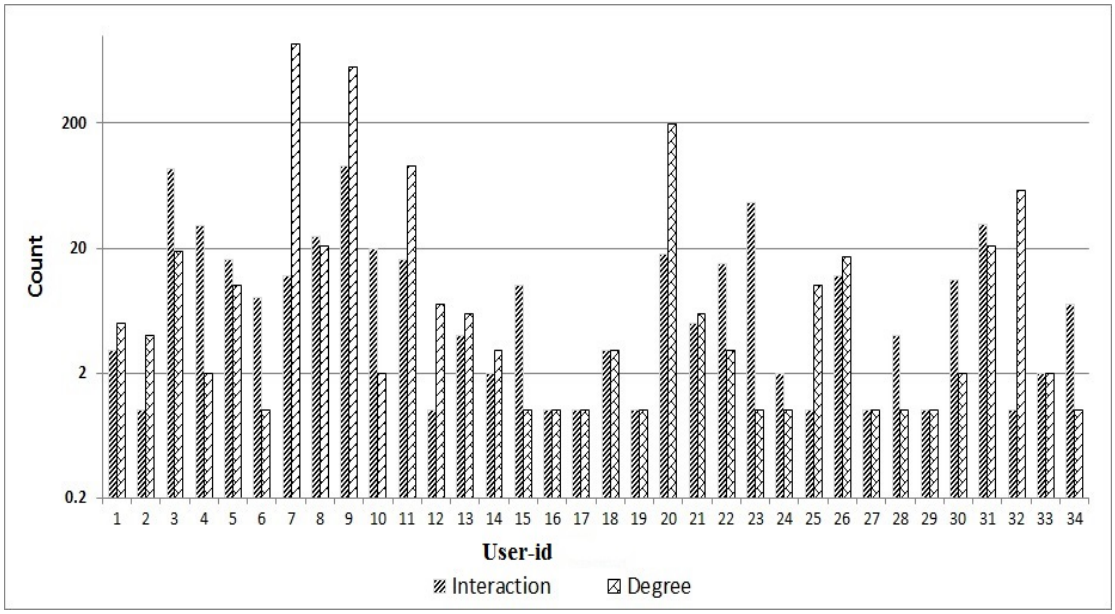


Figure 4.5: Interaction and degree count of users in Digg dataset

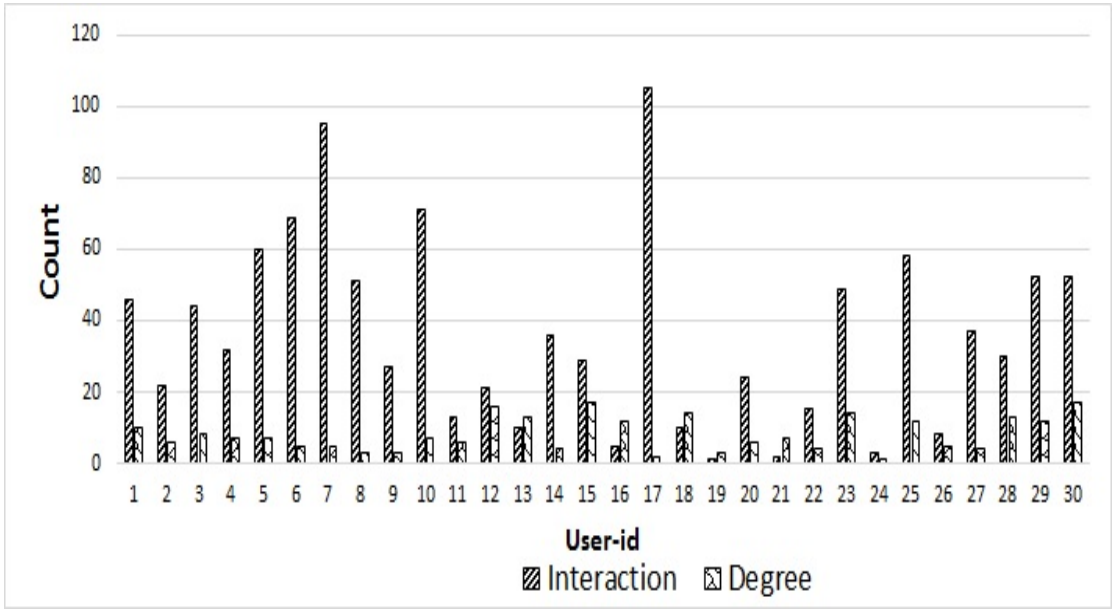


Figure 4.6: Interaction and degree count of users in Infectious dataset

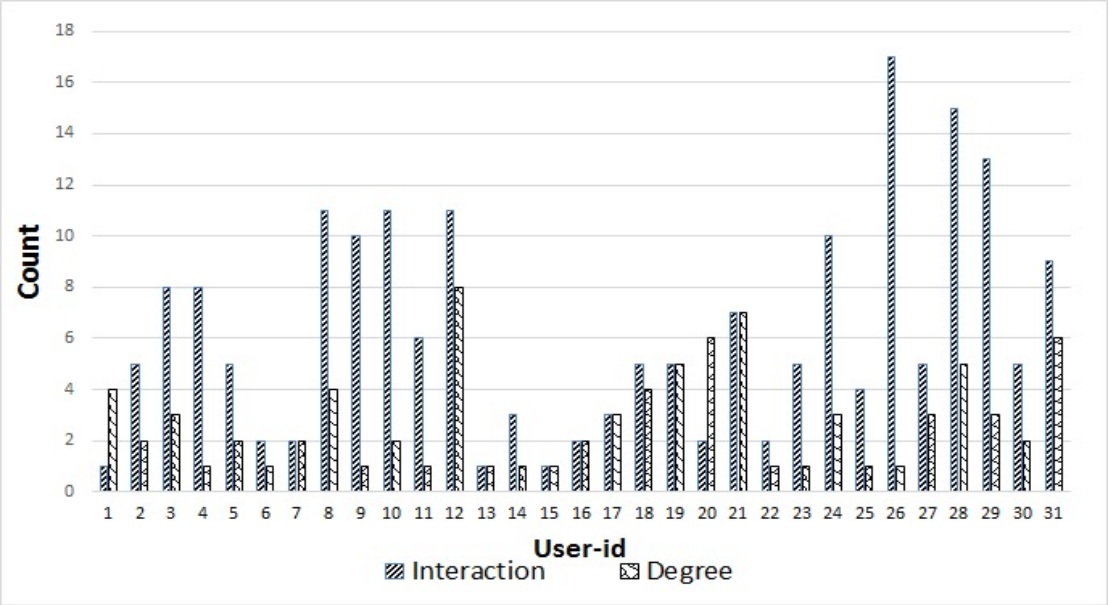


Figure 4.7: Interaction and degree count of users in YouTube dataset

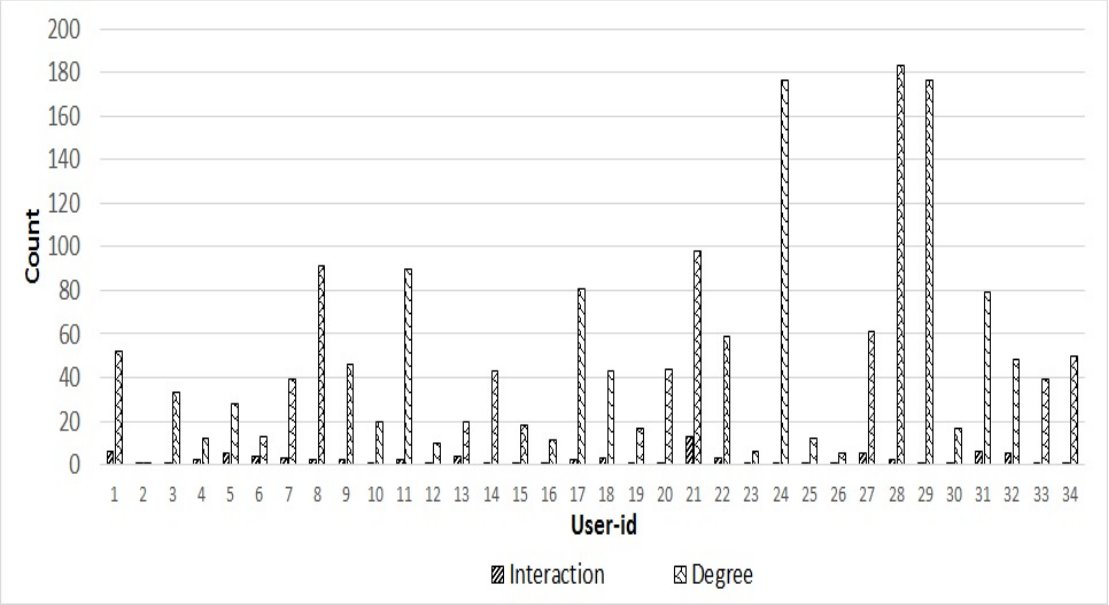


Figure 4.8: Interaction and degree count of users in Twitter dataset

From the study of the chosen datasets, one can infer that in most of the cases, the users who have high degree count, maintained interactions with only a small subset of their contacts. On the other hand, users who have a small degree count, had noticeably very high interactions with almost all their contacts.

Observation 4.2.1. In most of the cases, users with high degree count do not have high interaction count. Where as, users with low degree count have high interaction count.

Therefore, it is appropriate to evaluate a user in social network by his/her interaction count and bypass the use of the traditional approach of degree count. This research work concludes that, to evaluate users for information diffusion application, interaction count of users is a more suitable metric, when compared to the degree count.

4.3 Summary

This chapter has clarified that the traditional approach of picking users with high degree count may not lead to accurate results. It is observed that users with high degree count are not necessarily highly interactive. For information diffusion it is vital to pick users who are interactive rather than the ones with high number of friends. One of the findings of this research work is that, the interaction count is considered as a valid and appropriate metric to evaluate users.

Chapter 5 explains the limitations of the existing model, that is used to map the various stages of a user, during the information diffusion process. Also, it details the new model developed in this research work.

Chapter 5

User centric model of information diffusion

An equation means nothing to me unless it expresses a thought of God.

- Srinivasa Ramanujan

As identified in this research work, there is a need to develop a new model that represents the information diffusion process among users in social networks. The new model for representing information diffusion in social networks is presented here. This chapter also includes discussions on the existing models for information diffusion and the hindrance in using them for social networks analysis.

5.1 Background

The seminal work on spread of epidemic diseases is formalized through the Susceptible-Infected-Recovered (SIR) model (Kermack and McKendrick, 1927). In this model, the entire population has been divided into three groups; susceptible, infectious and recovered. This model has been successful in predicting the casualties of an epidemic outbreak. The SIR model provided a pedestal for understanding the dynamics of epidemic spread. Since then, various contributions to SIR are seen in epidemiology (Ganesh et al., 2005; Moore and Newman, 2000; Newman, 2006; Pastor-Satorras and Vespignani, 2001). SIR model has also been used to understand the computer virus in the network (Kephart and White, 1993).

Analysis of spread of information has also been investigated in the variants of SIR model such as Susceptible-Infected-Recovered-Susceptible (SIRS) (Gruhl et al., 2004), Susceptible-Infected-Recovered-Vaccinated(SIRV) (Wang et al., 2016), Susceptible-Accepted-Immunized-Disseminated(SAID) (Zhu et al., 2016) and the Independent Cascade Model(ICM) (Kempe et al., 2003; Saito et al., 2012). Since there is a close reassemble of epidemic spread to the information diffusion process in the social networks, the epidemic models have been used in social computing as well. However, the majority of models do not explore the aspect of user dynamics such as their inclination and interest to adopt information. Aspects associated with the user characteristics are often overlooked and are not appropriately modeled. Due to this, when these models are used, there is a gap between theoretical and observed results.

5.2 Susceptible infected recovered model

The SIR model, has been formalized to understand the epidemic outbreak in the population. It is a closed model and has three classes; Susceptible(S), Infected(I) and Recovered(R). It is assumed that initially, the entire population(N), before the outbreak of a epidemic virus is in susceptible state S. When an epidemic outbreaks, members in S class, who are susceptible to infection, get exposed to it, through their contacts and they enter the infected state I. In this state, they infect members of susceptible group. There is a sudden increase in the number of people who are infected. After a certain time elapse, the infected population gets healed and no longer spread infection. Thus the population recovers from the epidemic affect and it is in recovered state R. The only way a person can leave the susceptible group is to become infected, which may not hold true always. The only way a person can leave the infected group is to recover from the disease. Once a person has recovered, the person is no longer susceptible to the same disease.

The SIR model has been based on two assumptions. Firstly, SIR assumes homogeneous mixing of population, due to this, an individual is equally likely to

be infected by others. Hence, the infection probability parameter has been uniform and constant. In the information diffusion model, this parameter is referred to as the user influence. However, the degree of sparsity in interactions among the users will invalidate the homogeneous mixing concept. Secondly, it is assumed that, there is no inherited immunity and the entire population is in susceptible state, i.e., $S = N$. Due to this assumption, every individual is considered to be infected if contacted by an infectious person. In the context of information spread, it leads to the assumption that every user readily adopts and spreads the information. In social networks, these assumptions fail. Not all individuals are susceptible and majority of the social network users will restrain themselves from activities such as commenting, postings, forwarding, and so on. Also, the entire population do not mix with each other equally.

In this research work, Restrained-Susceptible-Infected-Recovered (RnSIR) model is developed to fill the gap seen in SIR model. The proposed model is able to make a clear distinction between the restrained and susceptible users in the network. This new model is able to represent the information diffusion in social networks, more accurately.

Observation 5.2.1. The SIR model is not suitable to accurately represent information spread in social networks.

5.3 Restrained-Susceptible-Infected-Recovered Model

In this research, to address aforementioned issues a new model named as RnSIR is designed. The RnSIR model is an extension to SIR model and represents the role of users during information diffusion in the social networks, as shown in figure 5.1. In addition to the three classes (S, I and R), present in the SIR model, the new model has a new class Rn, to represent the users who restrain themselves from network activities.

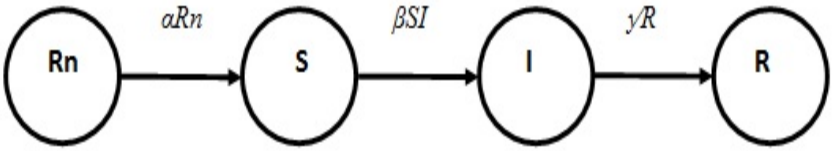


Figure 5.1: The RnSIR model

The RnSIR model is a discrete time probabilistic model, where α, β and γ are probabilistic variables and Rn, S, I and R are random variables. The parameters in this model are defined as follows:

Rn: number of individuals who restrain from activities at a given time

S: number of individuals who are susceptible to be infected at a given time.

I: number of individuals who are infected at a given time.

R: number of individuals who have recovered at a given time.

α: Interaction rate of an individual.

β: Influence of an individual.

γ: Recovery rate of an individual.

To understand the network dynamics, the rate of change, from one state to another, is given by the ordinary differential equations (ODE), as in Eq (5.3.1) to Eq(5.3.5) and are explained as follows: When the network is newly formed, all individuals are strangers to each other and remain in Rn state for certain time period (usually more than one time step). Therefore, initially the entire population is in Restrained state, i.e., $Rn = N$. As time progresses, individuals make new contacts and interact among these contacts and in the process generate contents in the network. As the number of interactions increases, individuals become open to others' ideas, likes and dislikes. The parameter α captures the interaction rate of individual, that would make a user susceptible. At a certain time step, α is high and at this stage, subset of individuals in state

Rn, enter state S. This is represented as in Eq(5.3.1).

$$\frac{dR_n}{dt} = -\alpha R_n \tag{5.3.1}$$

Once susceptible, an individual remains there for one time step. When there is outbreak of new information, an individual becomes ready to receive and act upon the information. Depending on the parameter β, at this point, influential users (part of state I) influence individuals in state S to adopt information. Thus, a subset of S i.e., $I \subset S$, adopts information. This quantity is represented as βSI in Eq(5.3.2).

$$\frac{dS}{dt} = \alpha R_n - \beta SI \tag{5.3.2}$$

Again, as time progresses, an individual recovers at certain rate γ and enters state R and does not spread the information further. This is represented as in Eq(5.3.3) and Eq(5.3.4) .

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{5.3.3}$$

$$\frac{dR}{dt} = \gamma I \tag{5.3.4}$$

Since $N = R_n + S + I + R$,

$$\frac{dR_n}{dt} + \frac{dS}{dt} + \frac{dI}{dt} + \frac{dI}{dt} = 0 \tag{5.3.5}$$

The RnSIR model, thus represents how a user moves from one state to another during the information diffusion process.

A similar explanation as the one available in (Shapiro and Delgado-Eckert, 2012) is used to explain the various stages of a user in the diffusion process under the RnSIR model. For a social network represented as $G(V, E)$, where V is a set of users, E is a set of edges, a state δ, for a user $u \in V$ is given by

$$\delta(V) = \{\varphi | \varphi : V \to \{Rn, S, I, R\}\}$$

Under the following constraints,

1. $\delta = \varphi_1(u) = Rn$, then $\varphi_2(u) = S$

2. $\delta = \varphi_1(u) = S$, then $\varphi_2(u) = I$

3. $\delta = \varphi_1(u) = I$, then $\varphi_2(u) = R$

4. $\delta = \varphi_1(u) = R$, then $\varphi_2(u) = R$

where, φ_2 follows φ_1. Thus, the states are considered to be a Cartesian product of $V \times \{Rn, S, I, R\}$.

5.4 Complexity of the model

The complexity of RnSIR model, in the context of information spread is discussed in the following theorems and corollary.

Theorem 5.4.1. Finding probability of user influence, β, under *RnSIR* model is NP-hard.

Proof: Finding the infection probability under SIR model is NP-hard (Shapiro and Delgado-Eckert, 2012). The parameter, infection probability is also referred to as the probability of user influence in $RnSIR$ model. Using the reducibility principle, it may be argued, that the SIR model is a special case of RnSIR, where $Rn = \phi$. Therefore, finding the probability of user influence represented by parameter β, under RnSIR model is NP-hard as well. □

Theorem 5.4.2. Finding the number of activated nodes(a.k.a spread), represented as σ, at the end of diffusion process under RnSIR model is NP-hard.

Proof: The number of activated nodes at the end of diffusion process, initiated by the set I, is computed as given in Eq(5.4.1).

$$\sigma(I) := RnSIR(I, \beta) \tag{5.4.1}$$

where $\sigma(I)$ represents the number of activated nodes at the end of the diffusion process. The $\sigma(I)$ solely depends on the parameter, probability of user influence represented as β and constant I. With reference to Theorem 5.4.1, finding the probability user influence, β, under *RnSIR* model is proved to be NP-hard. When

probability of user influence cannot be determined in polynomial time, finding the number of nodes activated in the process is equally hard as well. Thus, finding the number of activated nodes, initiated by I, represented as $\sigma(I)$, is NP-hard under RnSIR model. □

Corollary 5.4.1. Computing the spread under *RnSIR* model is #P-Hard.

Proof: It is already proved in Theorem 5.4.2, that finding the number of activated nodes, also known as spread, under *RnSIR* model is NP-hard. It in turn asks, whether the number of such solutions are countable. Since the problem is NP-hard, the number of such solutions cannot be counted. Therefore, computing the spread under *RnSIR* model is #P-Hard. □

Observation 5.4.1. With the results of Theorem 5.4.1 and Theorem 5.4.2, one can conclude that probability of user influence and the spread of information can only be estimated.

5.5 Summary

This chapter discussed the limitations of SIR model in understanding the information diffusion in social networks. To fill the gaps seen in SIR model, this research presents a new model namely the RnSIR model. To this end, the RnSIR model distinguishes between restrained and susceptible users. Thus, the new model is more appropriate in representing the information spread in the social network. With this new modeling formalism, it is clear that the information diffusion process is able to distinguish between the probable spreaders from rest of the social network users. In a social network, it is these users, who participate in the diffusion process and should be employed for information diffusion applications.

Chapter 6 details the approach to solve the scalability issue, the methodology and verification of the proposed approach.

Chapter 6

Pruning the social network

Structure is more important than the content in the transmission of information.

-Abbie Hoffman

The time complexity of influence maximization increases drastically with the increase in the size of social network. Therefore, the scalability issue can be addressed by pruning the social network. This in turn, fetches the real contributors for diffusion. In this context, this chapter presents a method to prune the social network graph, the results and validation.

6.1 Background

An extensive study in information diffusion in Twitter reveals that the majority of users act as passive consumers of information and do not actively participate in the information diffusion. For eg., on an average, one in 318 tweets are re-tweeted. Passive participation is a major issue in diffusion process (Romero et al., 2011). Jakob (2012) observes that all large-scale, multi-user communities and online social networks that rely on users to contribute content or build services share one characteristic; most users do not participate in any of the network activities. Most of the contents come from very small percentage, usually only 1% of the network users. This discrepancy is referred to as 'participation inequality' and it typically follows a 90-9-1 rule in which users fall into one of three categories: lurkers, intermittent contributors and heavy contributors. This indicates that

almost 90% of the social network users fall under the category of lurkers. These users often lurk in background by browsing through web pages of other users. They do not contribute to the activities in the network. Due to this, one cannot be sure if these users propagate or use any information that has reached them. The next category of users is the intermittent contributors who form 9% of the population. These users occasionally participate in the network. There is a chance that these users propagate information. In the last category, who form 1% of the population of users are the heavy contributors. These users are involved in developing and propagating the contents or information in the network. They contribute to 99% of the total information in the network. Figure 6.1 shows the participation inequality that is seen in almost any social network. For effective information diffusion, it is important to identify the heavy and intermittent contributors, who participate in the diffusion process. This will reduce the size of the social network, which in turn results in better utilization of resources and lesser processing time.

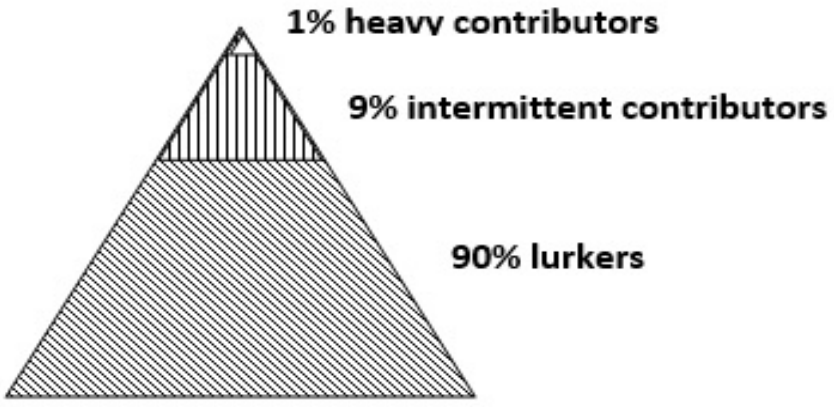


Figure 6.1: Participation inequality in social networks (Jakob, 2012)

A common practice followed during social network analysis is to view the social network as a graph. Once social network is represented as a graph $G(V, E)$, its properties are described using graph theory concepts. In such a graph $G(V, E)$, V is the set of users and E is the set of edges that define the underlying relationship. Link formed between a pair of users, indicate that two users are well connected in

terms of similar interest and ideas. It is observed that few of these links are more often used than others and these are the links that keep two individuals strongly connected. Also, these individuals have greater than average potential to influence each other when compared to any two randomly selected users. For this reason, identifying these links is important and the solution for this is attempted in this work.

6.2 Problem description

In social networks, the users play an important role. In most of the prior works, the structural properties of the graph are used to simplify it. However, randomly removing a connection edge from social graph may lead to disturbance or loss of its structural properties which may render the sub-graph unsuitable for specific application.

This work addresses the problem of pruning the social networks. The final outcome fetches a reduced social network that has active users, those who participate in the information diffusion process. The problem is defined as follows:

Problem 6.2.1. *Given a social graph G(V,E), an interaction threshold α and an activity log $A(v_i, activitycount_{vi})$, find the contributor graph $G_c(V_c, E_c) \subseteq G(V, E)$ such that*
(i) $v_i \in V_c$ if $v_i(activitycount_{vi}) > \alpha$, where α is the minimum activity rate.
(ii) $e(v_i, v_j) \in E_c$ if v_i and $v_j \in V_c$ and $e(v_i, v_j) \in E$.
(iii) maximize(S), where S is the small world properties.

The solution to this problem includes a sub-problem to fetch the *threshold* value that is used to identify a link as interactive. The sub-problem is defined as follows:

Subproblem 6.2.2. *Given an activity log $A(user, friend)$, of users of the underlying social network graph $G(V, E)$, find the threshold value α, based on the distribution pattern of interactions.*

This research work uses the nodal property for pruning social networks. The figure 6.2, shows how the new approach reduces the gigantic social network to a manageable size. On the left side of the figure 6.2, the original social graph is seen. The social network is labeled with the weights on its edges. This weight is the number of interactions between any pair of users (u, v). Specifically, the edges are labeled with two weights, one for each direction i.e., (u, v) and (v, u). The figure 6.2, also has a pruning threshold. On the right side of figure 6.2 is the pruned graph.

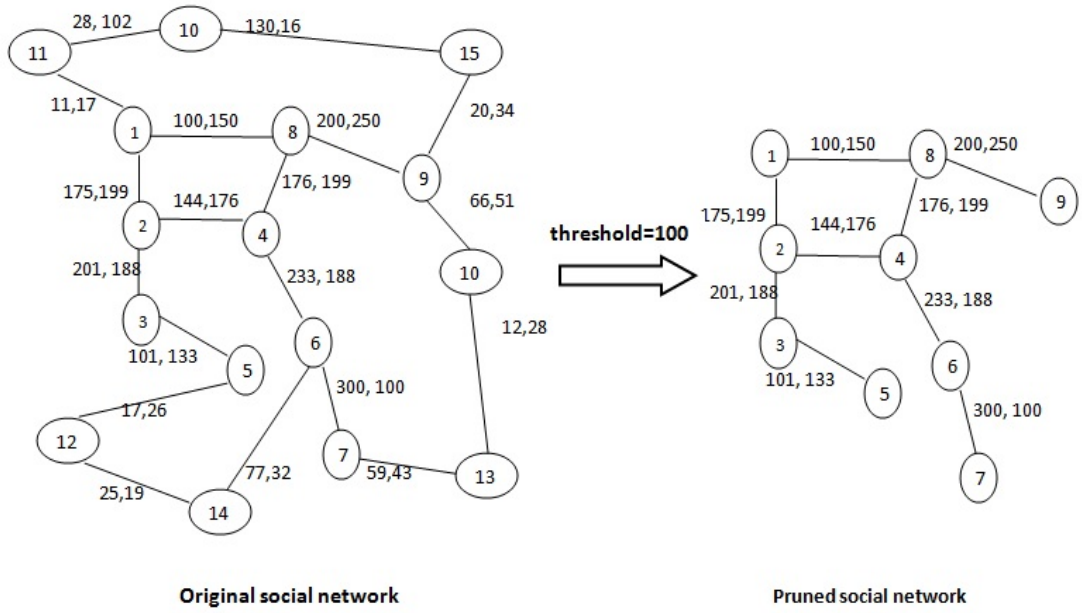
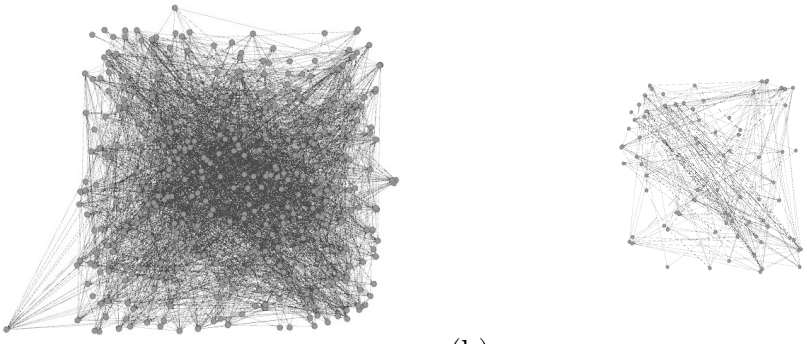


Figure 6.2: Social network and the sub graph of the social network

Figure 6.3 shows (a) a large social network and (b) its pruned social network of active users. The pruned network aids better analysis and visualization.



(a) (b)

Figure 6.3: (a) Original social network and (b)pruned social network

6.3 Proposed methodology

This section determines the threshold on the interaction count of users. Once the threshold is determined, it is used as a pruning parameter. The latter sections of this chapter detail the pruning approach and its applicability on social networks.

6.3.1 Computing the threshold

The interaction pattern of eight datasets namely Email, Wikivote, HEP, PHY, Digg, YouTube, Twitter and Infectious are shown in figure 6.4 to figure 6.11.

All the chosen datasets exhibit power law distribution pattern (Clauset et al., 2009) in their interaction count. The power law distribution is a mathematical relation between two variables and is used to model data where frequency of an event varies with respect to another associated variable of that event. Since the presence of power law distribution is seen in major social networks (Muchnik et al., 2013), the proposed approach is applicable on social networks. For the HEP and PHY datasets, interactions were not available and were synthesized[1] on powerlaw distribution pattern which can be produced using Algorithm1 or matlab tools.

[1]http://tuvalu.santafe.edu/ aaronc/powerlaws

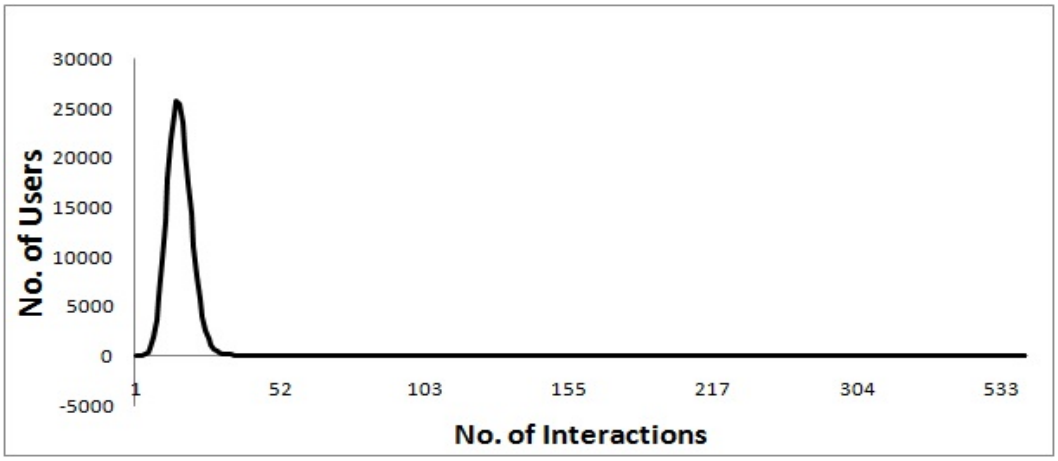


Figure 6.4: Interaction pattern of Email dataset

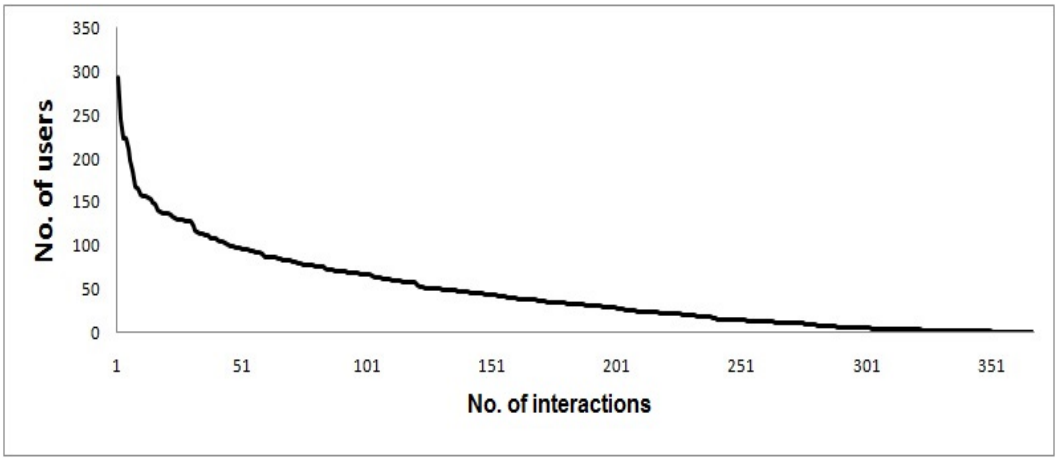


Figure 6.5: Interaction pattern of Infectious dataset

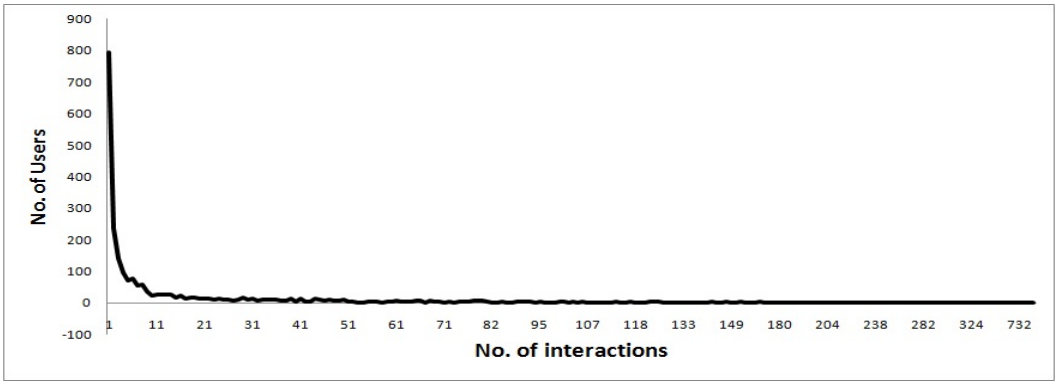


Figure 6.6: Interaction pattern of Wikivote dataset

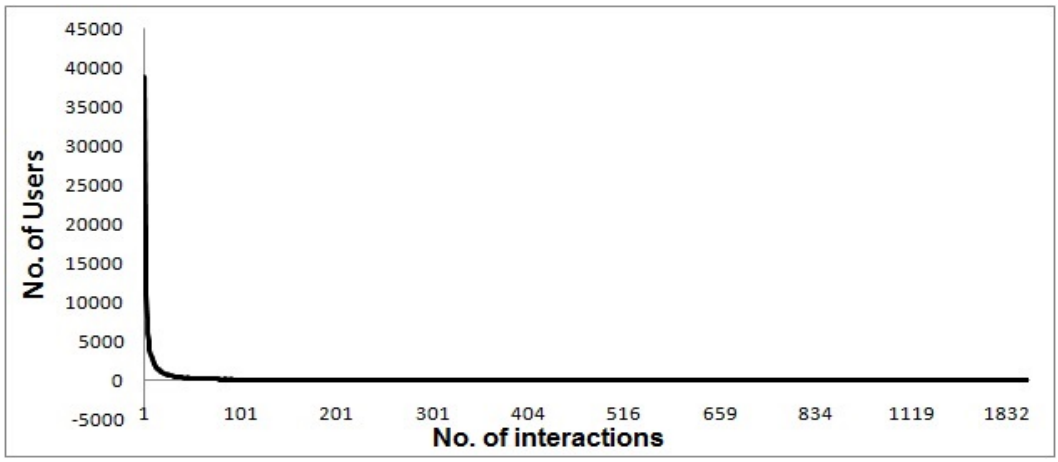


Figure 6.7: Interaction pattern of Digg dataset

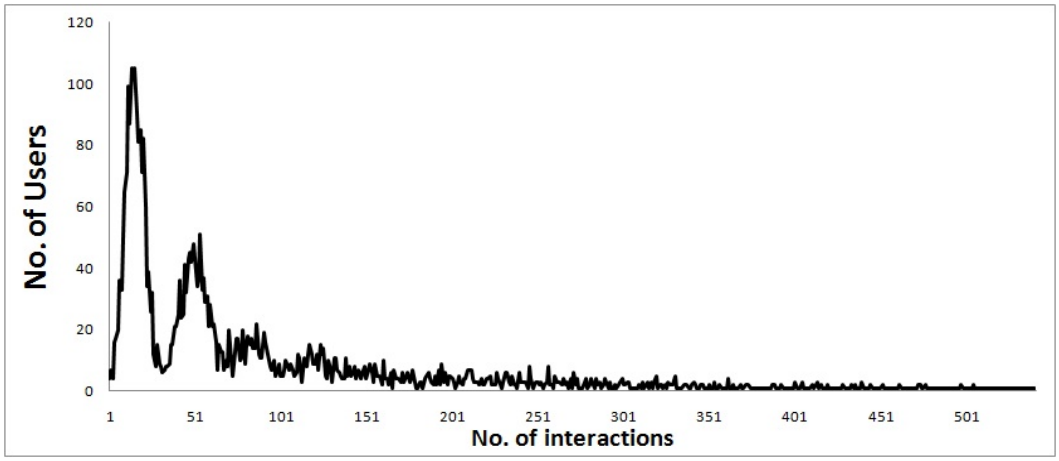


Figure 6.8: Interaction pattern of HEP dataset

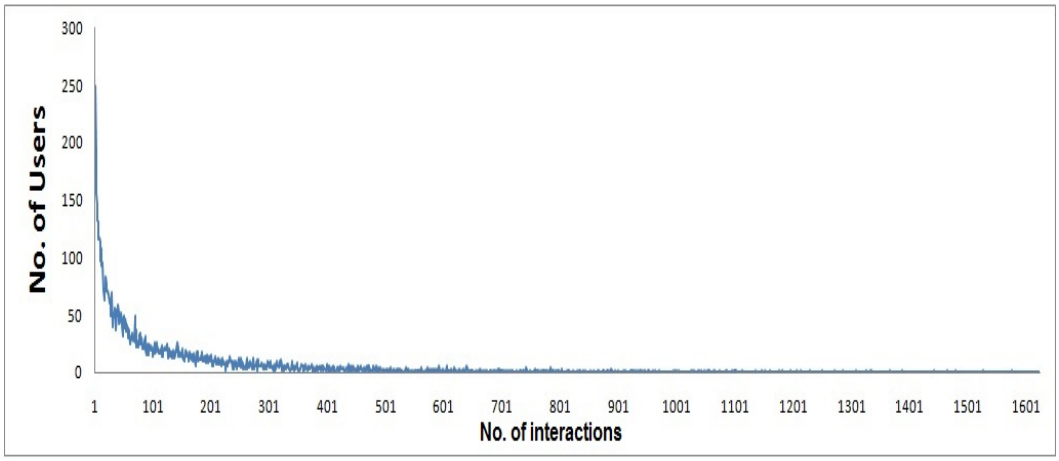


Figure 6.9: Interaction pattern of YouTube dataset

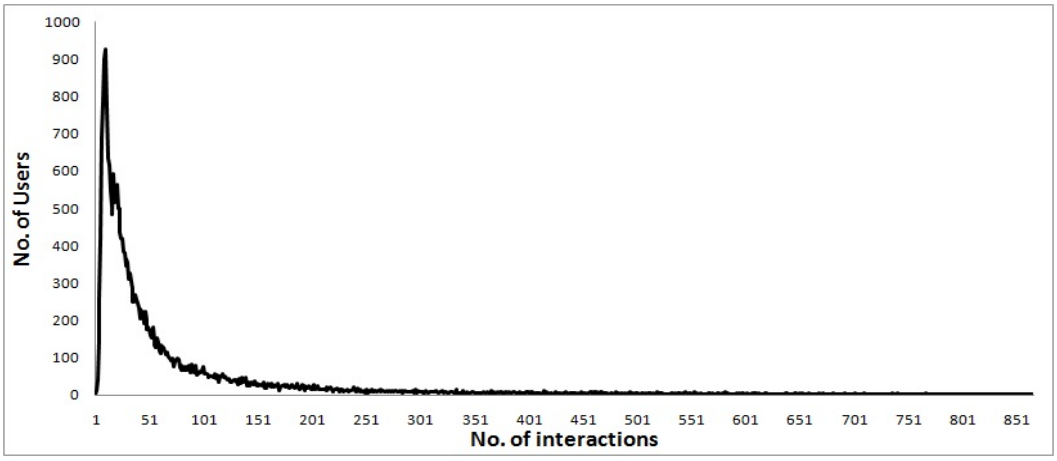


Figure 6.10: Interaction pattern of PHY dataset

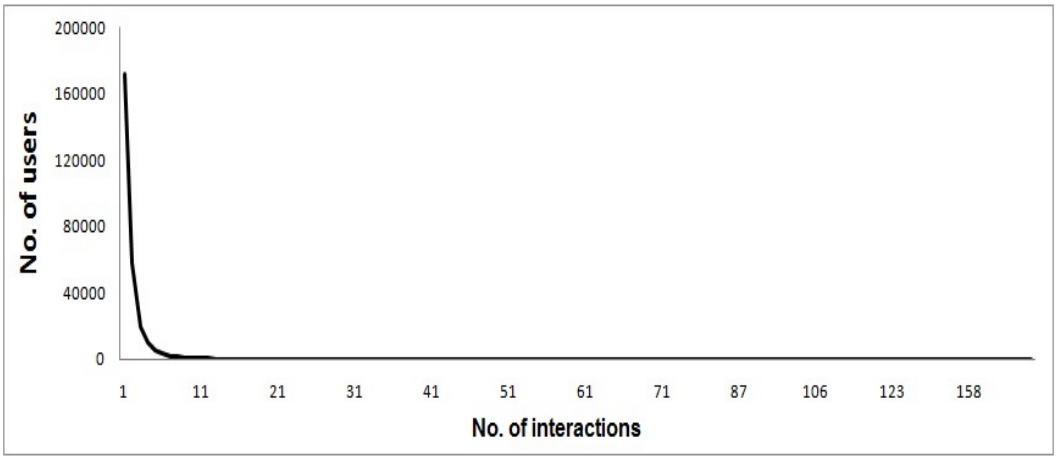


Figure 6.11: Interaction pattern of Twitter dataset

Observation 6.3.1. Social networks exhibit powerlaw distribution on interaction pattern of the users.

In a power law distribution pattern, *mean* is much larger than *median* and *mode*, i.e., $mean > median > mode$. Hence, the *mean* cannot be used as a measure of dispersion. For such a distribution a robust dispersion measure would be *Median Absolute Deviation (MAD)* (Pham-Gia and Hung, 2001). The median absolute deviation is more resilient to the outliers in the data than the standard deviation. In computing the median absolute deviation, a small number of outliers are irrelevant and will not skew the results, whereas, the standard deviation will be affected by outliers. Therefore, using the median absolute deviation method will avoid the need to subjectively eliminate outliers from datasets.

The median absolute deviation is computed by first determining the median for the given population. In the second step, the absolute value of the difference between each separate observation and the median is computed. The median absolute deviation is then determined by computing the median of the values computed in the second step. More precisely, for a univariate data $X_1, X_2, ..., X_n$, the MAD is defined as the median of the absolute deviations from the data median and is given as in Eq(6.3.1)

$$MAD = M(|x_i - M(x_j)|) \tag{6.3.1}$$

where $M(x_i)$ is the median of original observations, x_i where i=1 to n, is the original observation.

In Email dataset, there are 1,94,115 users who have performed 1 interaction. In statistical term, the frequency of 1 is 1,94,115. All datasets used in this research have such grouped data, where same values are repeated. Therefore, Eq(6.3.1) is rewritten as Eq(6.3.2) and used to find the threshold value α.

$$MAD = M(f_i * (|x_i - M(x_i)|)) \tag{6.3.2}$$

where x_i is the original observation of interactions, M is the median and f_i is the frequency of each original observation.

Using Eq(6.3.2) on the social network activity log, the threshold value α is determined. This is the parameter α, seen in RnSIR model, via which the user makes a transition from state Rn to state S. This threshold is the ideal count of interactions that is used to identify a user as a contributor from the rest of the network. Thus, the threshold value becomes the pruning parameter in the proposed pruning approach. As far as the distribution pattern matches the power law distribution, the decision of choosing median absolute deviation to determine the threshold on the interaction count is justified.

Furthermore, in this research work, the interaction count of the users are analyzed for a specific time period N, say from *t1* to *t2*. For the next analysis, one can take the interaction count between the interval *t2* and *tcurrent*. Thus, users who may have high interaction count in this particular time period is

included in the pruned graph. Thus, the approach presented here reflects the dynamic nature of the social network. Also, one can get a close look at the activity pattern in social network by varying the granularity of the time period.

6.3.2 Algorithm and data structures

This section details the algorithm and pseudocode for the methods discussed in the previous section. Algorithm 1 is the pseudocode for synthesizing data that follows the powerlaw.

The input to Algorithm 1 is the social graph $G(V, E)$. The function $randpower(1, n)$ generates n random numbers on powerlaw distribution. The list named as *powerlist* is available as the output. This list follows the powerlaw distribution.

Input: Social graph G(V,E)

Output: List powerlist

1 Initialize $power = \emptyset$, $powerlist = \emptyset$;
2 create random number list *power* as in step 3
3 power{} = randpower(1,n)// where $|V| = n$
4 Add each $e(u, v) \in E$ to *edgelist*
5 Repeat step 6 for each number r in *power*
6 $powerlist = powerlist \cup edgelist[r]$

Algorithm 1: Synthesize interactions from social network data

The steps for computing the threshold on the interaction count, from the activity log and further pruning the social network is shown in Algorithm 2. The inputs to Algorithm 2 are the social network $G(V, E)$ which is maintained as an edgeset file, pre computed threshold value α and an activity log of the form $A(user, friend)$.

Input: G(V,E), threshold value α, Activity log $A(user, friend)$

Output: Contributor graph $G_c(E_c, V_c)$

1 Initialize count=0, $G_c(E_c, V_c) = \phi$

2 For each *user* in activity log $A(user, friend)$

3 increment *count*

4 For each edge $e(u, v) \in E$

5 if *count* of both the end vertices, u and v, is greater than α

6 Add e(u,v) to $G_c(E_c, V_c)$

Algorithm 2: Pruning the social network

In Algorithm 2, steps 2 and 3 compute the interaction count of users. In steps 4 to 6, given the threshold value α as input, original social network is parsed. In this process, all those edges which have both their end vertices's activity count greater or equal to the threshold value are added to the pruned graph known as the contributor graph $G_c(V_c, E_c)$.

Thus, Algorithm 2 serves its intended purpose of creating pruned graph $G_c(V_c, E_c)$ of the original social network graph $G(V, E)$, having all those users who have activity count above the threshold value. As in step 6, by adding the edges to the pruned graph, which is also maintained as an edgeset file, the end vertices are implicitly added.

Complexity analysis: Steps 2 and 3 compute the number of activities of each user. For a activity log of k rows, this would take $O(k)$. Next, steps 4 to 6, add edges from the original social network to the pruned graph. If there are n edges in $G(V, E)$, then this would take $O(n)$. Thus, the complexity of the Algorithm 2 is $O(k + n)$, which is linear.

6.4 Results and analyses

This section details the results of the proposed approach. Further, the proposed approach is validated to prove its applicability to the information diffusion process in the social networks.

6.4.1 Pruning the social network graph

The datasets described in Table 4.1 are pruned, resulting to $G(V, E) \rightarrow G_c(V_c, E_c)$ as shown in the Table 6.1. The number of nodes, edges and the threshold value are in the columns $G_c(V_c)$, $G_c(E_c)$ and α respectively in Table 6.1.

Table 6.1: Value of *alpha*, V_c and E_c of pruned graph

Sl.no.	Dataset	G(V)	G(E)	α	$G_c(V_c)$	$G_c(E_c)$
1	Email	265214	420045	209	163	5305
2	Wikivote	8275	103689	660	410	11726
3	HEP	15233	58891	403	205	405
4	PHY	37154	231584	637	310	2934
5	Digg	279392	1730381	704	303	14243
6	YouTube	15088	76765	1544	545	4846
7	Twitter	456626	14855845	102	32	128
8	Infectious	410	2765	75	82	290

6.4.2 Verification of small world properties

In this section, the contributor graph is compared to the other representation of the social network, in terms of information propagation properties. These properties that are also termed as small world properties (Deyasi et al., 2014), are considered essential for information diffusion and are listed below.

1. Higher average clustering coefficient.
2. Lower diameter

3. Lower average path length
4. Fewer number of connected components to enable reachability and adoption of information.
5. Lower modularity

These properties are used to evaluate graphs. For eg., the average clustering coefficient as given in definition 2.1.3 is

$$ACC = \bar{C} = \frac{1}{n}\Sigma_{i=1}^{n} C_i \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6.4.1)$$

Using Eq(6.4.1), the original Twitter graph's ACC is 0.016 and its contributor graph has ACC 0.303. Thus, when graph properties are reduced to a rational number, it becomes convenient for comparing them. The following results discusses the outcome.

6.4.2.1 Outcome and discussion

The datasets are pruned and examined for the presence of small world properties. The results are shown in figure 6.12 to figure 6.18. The small world properties of the contributor graph are compared to the following representation of social network graphs.

1. The original social network graph.
2. Shortest Path Pruned **(SPP)** graph: Generated by pruning edges under constraint to keep edges on the shortest path which uses Dijkstra algorithm (Cormen et al., 2001).
3. **MeanAlpha**: The social network graph obtained using the *mean* of the activity count, as threshold to simplify it.
4. **MST-Pathfinder**: The pruned graph obtained by MST-Pathfinder approach (Quirin et al., 2008b).
5. **Contributor graph (CGA)**: The proposed approach of using median absolute deviation to determine the threshold.

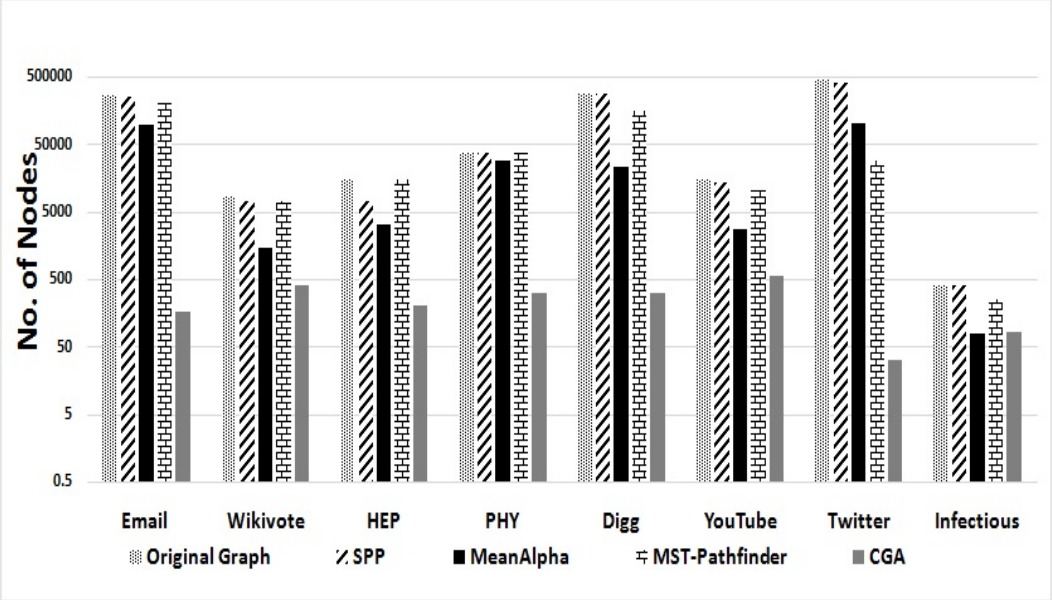


Figure 6.12: Comparison on number of nodes

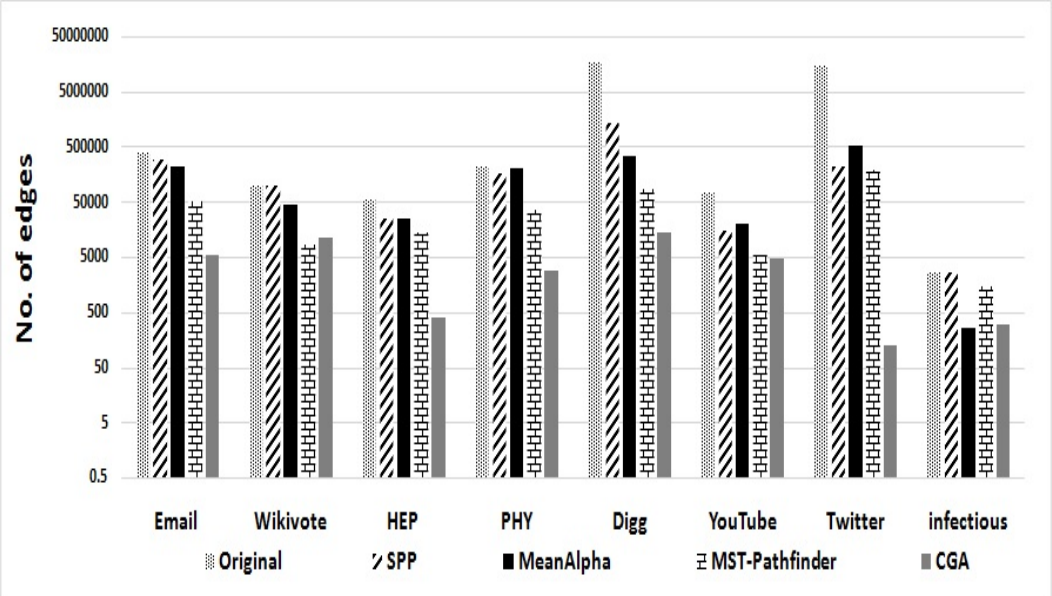


Figure 6.13: Comparison on number of edges

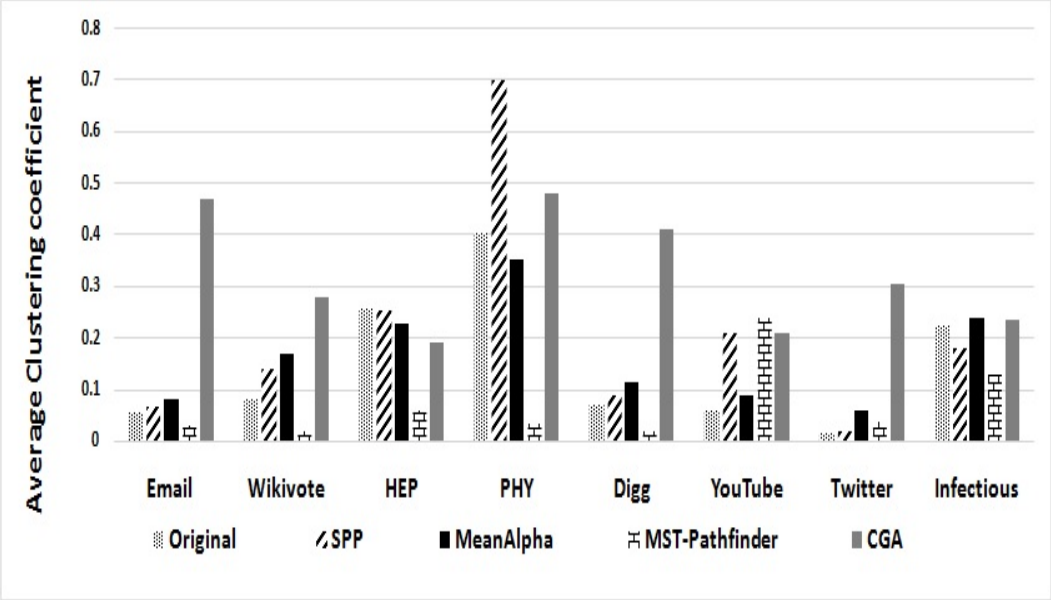


Figure 6.14: Comparison on average clustering coefficient

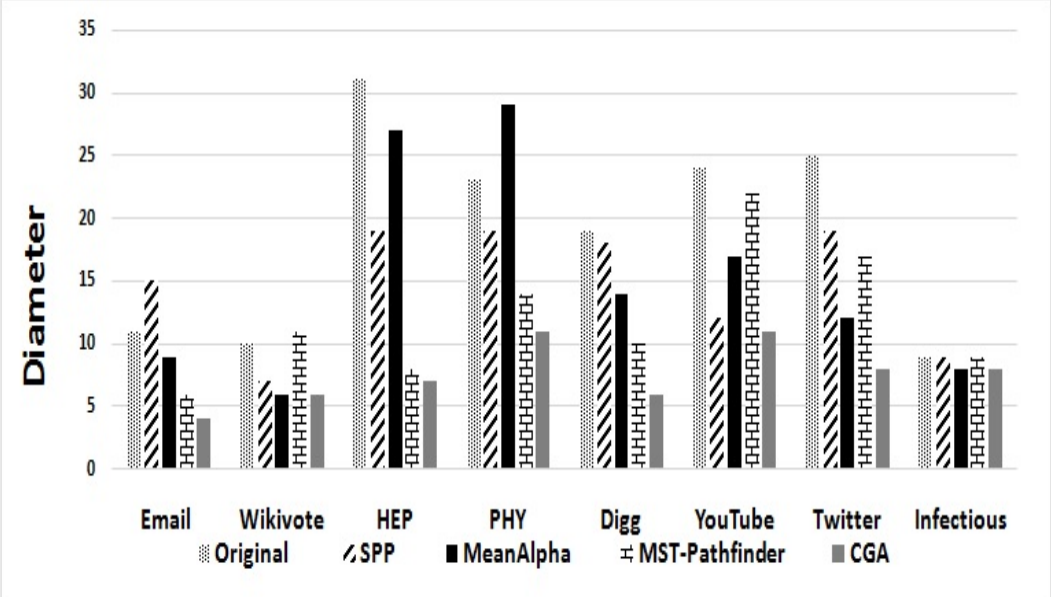


Figure 6.15: Comparison on diameter

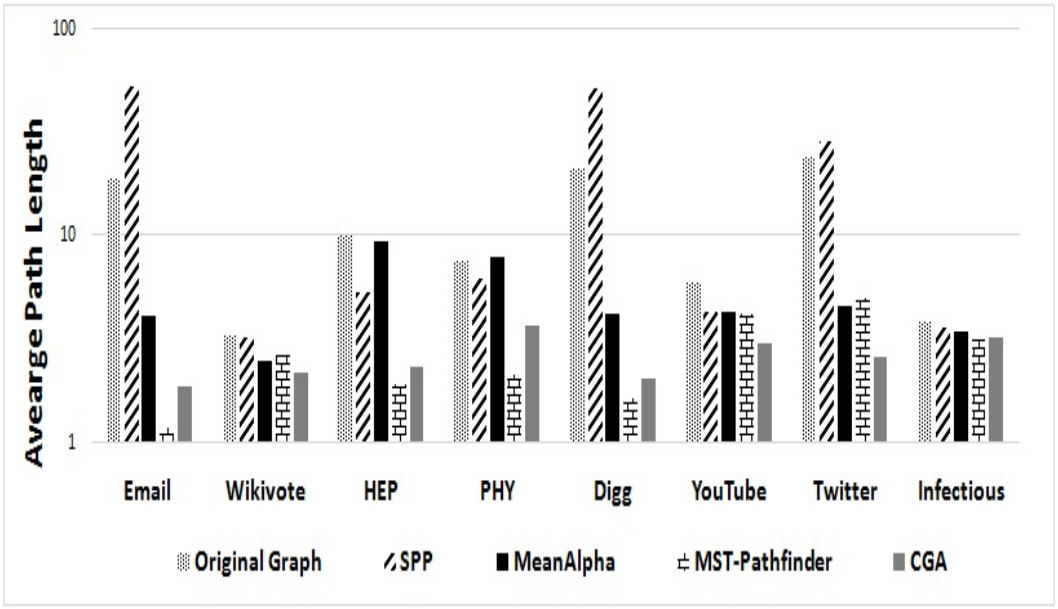


Figure 6.16: Comparison on average path length

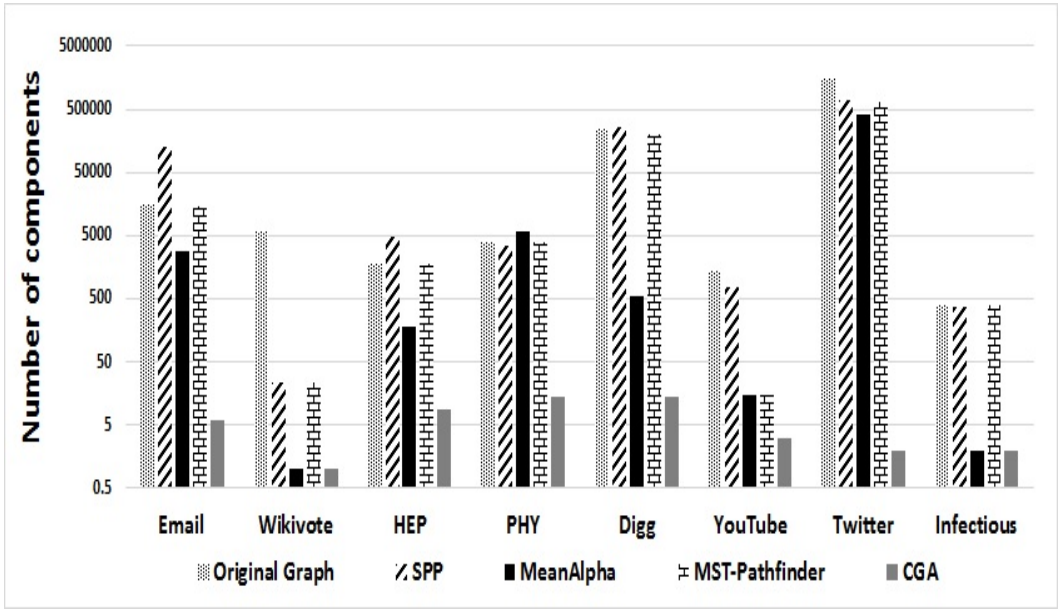


Figure 6.17: Comparison on number of components

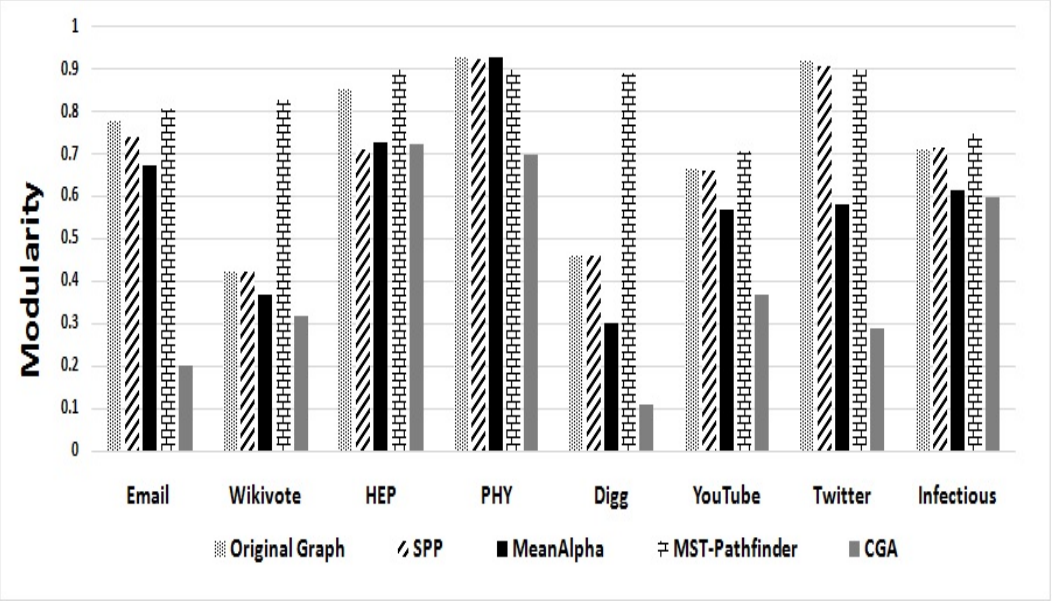


Figure 6.18: Comparison on modularity value

The analyses of the outcome are as follows:

- The small world properties of the original social networks were far more deviated from the expected values as compared to the pruned networks fetched by all the listed approaches.
- Pruning process would ideally want to retain the shortest paths to all nodes, thus resulting in the spread of information in less time. In the **SPP** method, only those edges which are part of the shortest path from every node to every other node are retained. It is basically an edge pruning method that uses Dijsktra's algorithm (Cormen et al., 2001). Once the original social network graph is pruned on the shortest path approach, it is verified for the presence of small world properties. Experimental results shown in figure 6.12 to figure 6.18, demonstrate that the **SPP** graph, although ideally an alternative to the original social network, does not improve propagation properties as compared to the contributor graph developed by the new approach.
- The effect of a different pruning parameter is investigated. In this case, the mean value of the interactions is selected as the pruning parameter and the

social network is pruned. Further, the pruned graph referred to as **MeanAlpha**, is investigated for the presence of small world properties. The results in figure 6.12 to figure 6.18 show that there is no significant gain in the small world properties in **MeanAlpha** approach, when compared to the new approach.

- The **MST-Pathfinder** approach (Quirin et al., 2008b), which prunes weighted undirected graphs is also implemented to verify the effectiveness of the new approach. For implementing MST-pathfinder, interaction count is employed as weight on the edges. The results are close to that obtained by the proposed approach. The **MST-Pathfinder** approach results in a pruned graph that has desirable values across average path length. However, it also results in lower average clustering coefficient and highly modular structure across all the social networks used in this research. Therefore, the **MST-Pathfinder** approach is also not suitable for fetching the subgraph which will be further used for information diffusion process.

- The contributor graph of the social network fetched from the new approach stands good in terms of small world properties in most of the cases.

Observation 6.4.1. In the presence of power law distribution pattern, median absolute deviation approach for determining the threshold is viable.

The experimental results support the claim that the proposed approach not only prunes the network but also improves the information propagation properties. Thus, the pruned graph is an ideal substitute to the original social graph for applications that are expensive in terms of resource usage. The graph size in terms of edges and nodes is drastically reduced. Yet, the properties of contributor graph are more desirable than the original social network and therefore can be used in its place for various applications where information propagation is involved.

6.4.3 Pruned social graph for information diffusion

Pruning is a data reduction operation with many applications. To demonstrate the effectiveness of the new strategy, the pruned graph is used in information spread process. In a social network context, information spread is associated with the influence maximization. Due to the complexity of the influence maximization algorithms, the run time increases enormously as the size of network increases. Due to this, the performance of these algorithms cannot be thoroughly evaluated in a large, real world social network. The effectiveness of these algorithms can be studied when the pruned graph is used.

6.4.3.1 Outcome and discussion

To show the effectiveness of using the presented approach, various pruning approaches and seeding strategies are analyzed in terms of percentage of information diffusion on IC and LT models (Kempe et al., 2003). The information diffusion under these two models are observed for three cases, (i) the original social network (ii) SPP and (iii) the contributor graph(CGA). The following seeding strategies are used for picking seeds, which are used in both the propagation models.

1. Degree: A simple heuristic that selects the k vertices with the largest degrees (Domingos and Richardson, 2001), to fetch the seed set S.
2. SingleDiscount: A simple degree discount heuristic where each neighbor of a newly selected seed discounts its degree by one (Chen et al., 2009).
3. DegreeDiscount: The degree discount heuristic for the IC model proposed by Chen et al. (2009), discounts the degree of a node by number of incident nodes already in seed set.
4. CELF: Cost Effective Lazy Forwarding has been developed by Leskovec et al. (2007b) to improve the Greedy algorithm(Kempe et al., 2003).
5. Distance: A simple heuristic that selects the k vertices with the smallest average shortest-path distances (Domingos and Richardson, 2001) to all

other vertices. The distance of two disconnected vertices is set to the number of vertices in the graph.

The results in figure 6.19 to figure 6.26 show the information diffusion in independent cascade model on the selected seeding strategies.

These results for independent cascade model are summarized as follows:

- For **HEP**, the proposed pruning approach diffuses information to 25.6% under DegreeDiscount heuristic, 25.5% under SingleDiscount, 25% under Degree heuristic, 25.52% under CELF and 24.9% under Distance heuristic.
- For **PHY**, the diffusion of 20.7% under DegreeDiscount, 20.4% under SingleDiscount, 17.6% under Degree heuristic, 29.8% under CELF and 17.9% under Distance heuristic is achieved.
- For **Email**, the percentage of diffusion is as much as 50% under DegreeDiscount, 48% under SingleDiscount, 46% under Degree heuristic, 50.23% under CELF and 47.7% under Distance heuristic.
- For **Wikivote**, the proposed pruning approach results in diffusion reaching 29% under DegreeDiscount, 28.7% under SingleDiscount, 28.4% under Degree heuristic, 28.9% under CELF and 28.4% under Distance.
- For **Digg**, the diffusion reached 45.58% under DegreeDiscount, 45% under SingleDiscount, 44% under Degree heuristics, 52% under CELF and 40.43% under Distance heuristic.
- Similarly, in **YouTube**, the diffusion reached 13.83% under DegreeDiscount, 13.8% under SingleDiscount, 13.6% under Degree heuristics, 13.2% under CELF and 13.3% under Distance heuristic.
- For **Twitter**, the diffusion reached 33.7% under DegreeDiscount, 33.4% under SingleDiscount, 32.9% under Degree heuristics, 33.8% under CELF and 31.3% under Distance heuristic.
- For **Infectious**, the diffusion reached 24.75% under DegreeDiscount, 24.6% under SingleDiscount, 24.5% under Degree heuristics, 24.6% under CELF and 24.4% under Distance heuristic.

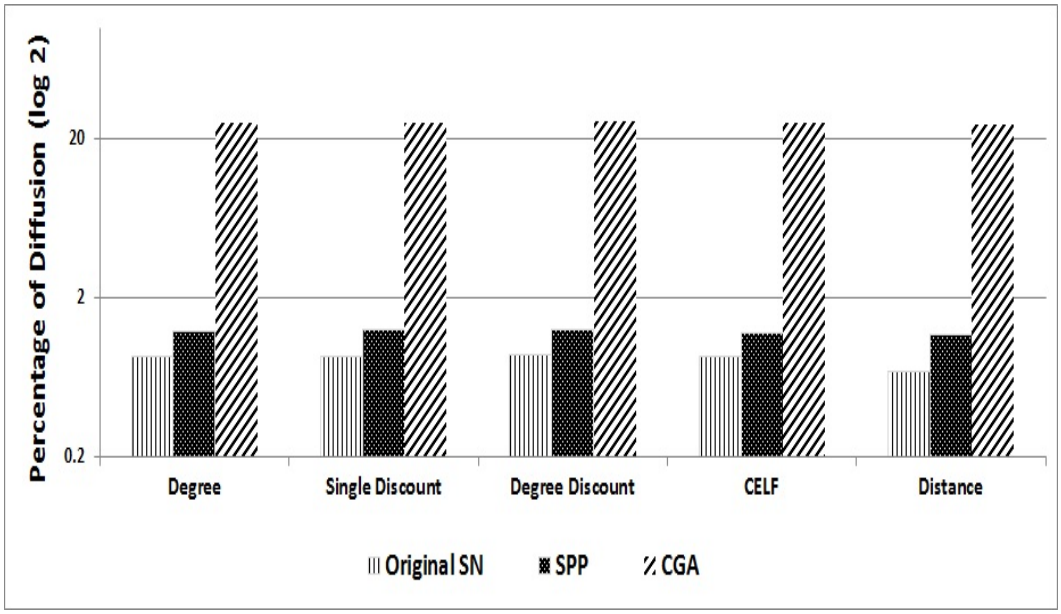


Figure 6.19: Percentage of spread in HEP under Independent Cascade Model

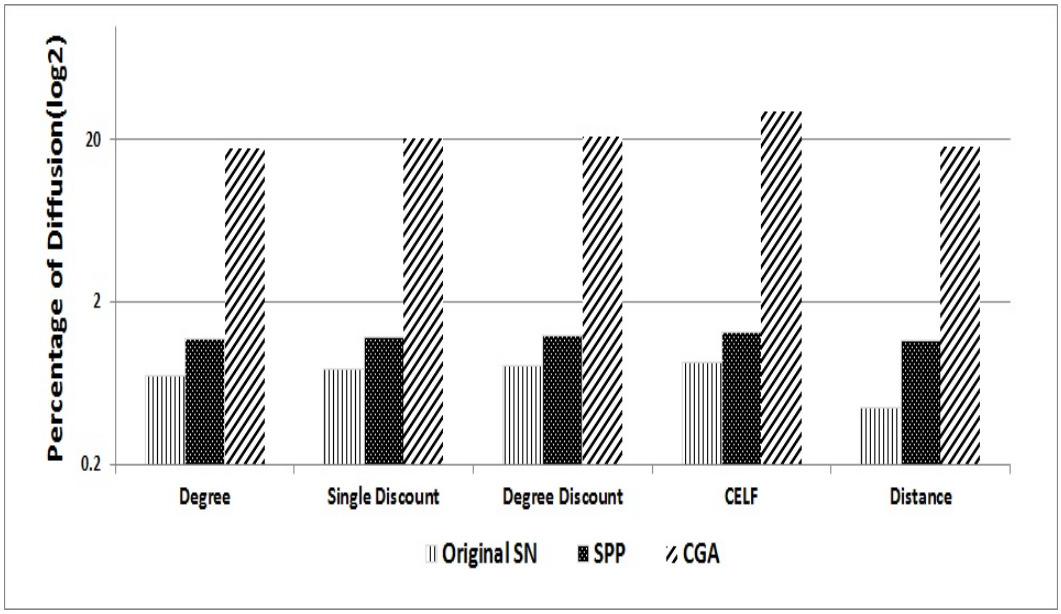


Figure 6.20: Percentage of spread in PHY under Independent Cascade Model

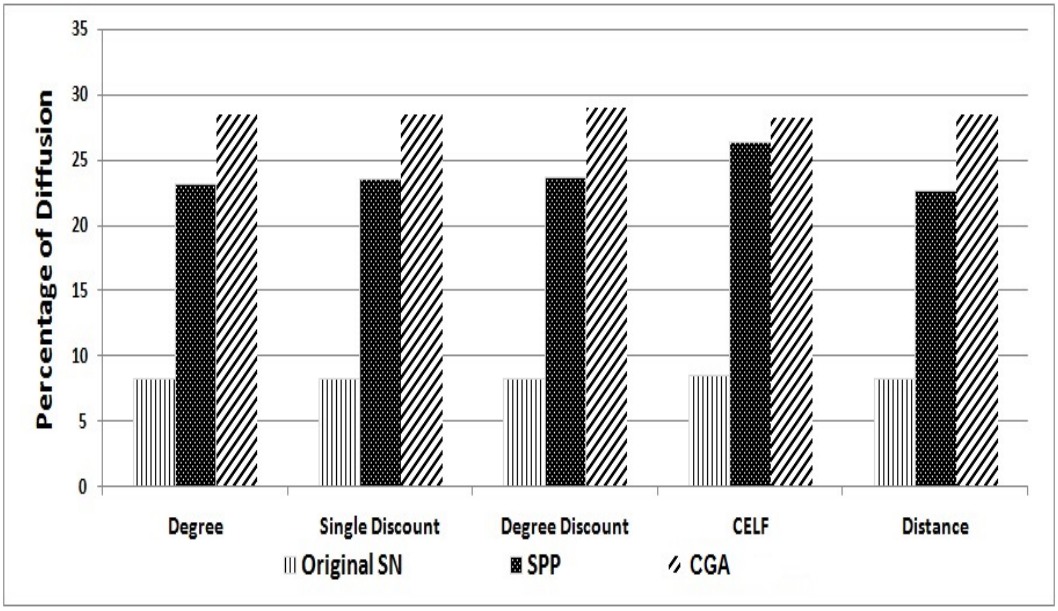


Figure 6.21: Percentage of spread in Wikivote under Independent Cascade Model

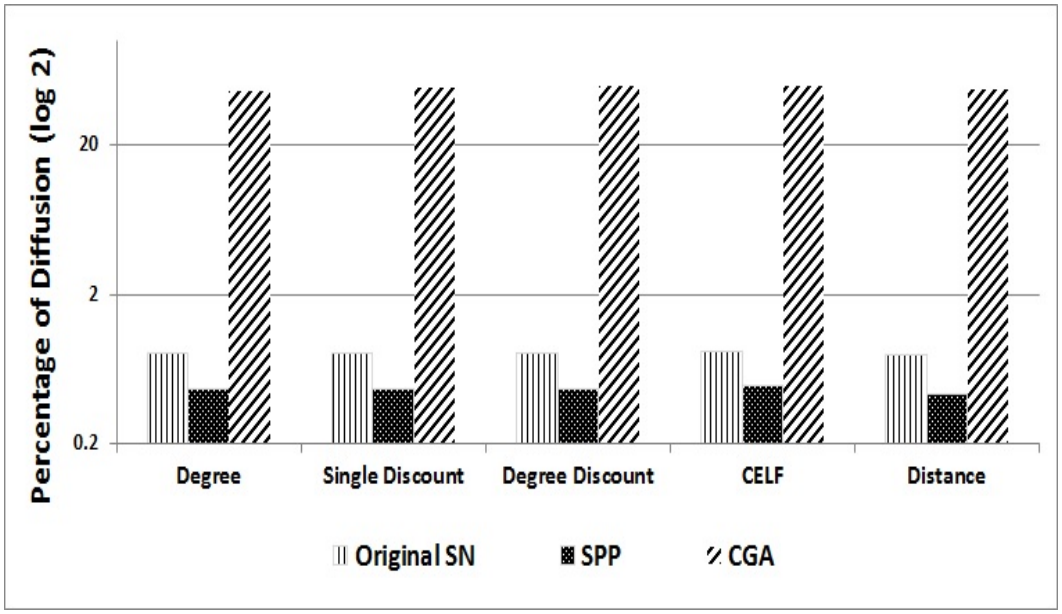


Figure 6.22: Percentage of spread in Email under Independent Cascade Model

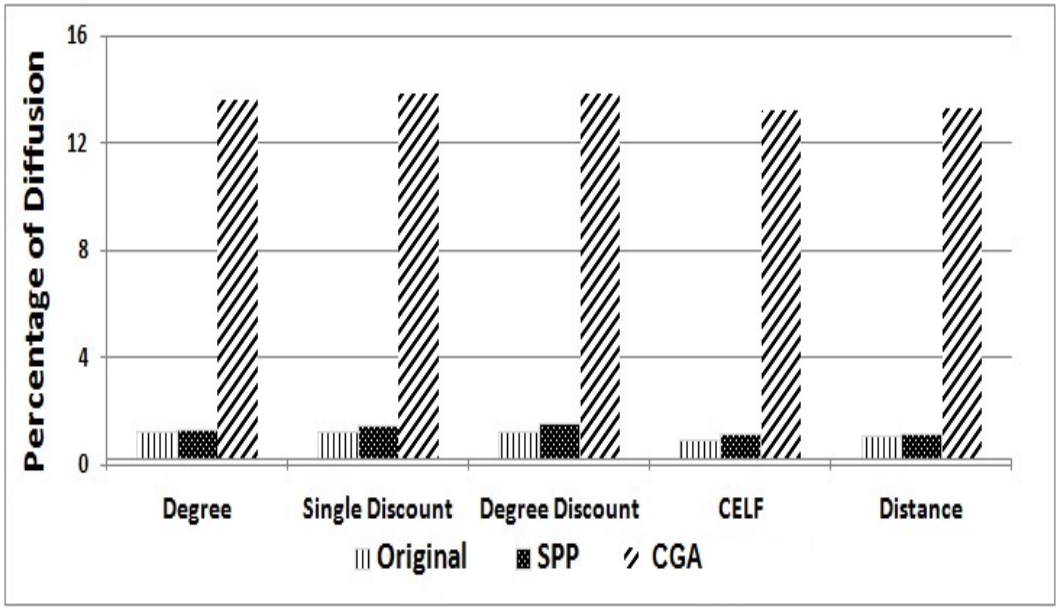


Figure 6.23: Percentage of spread in YouTube under Independent Cascade Model

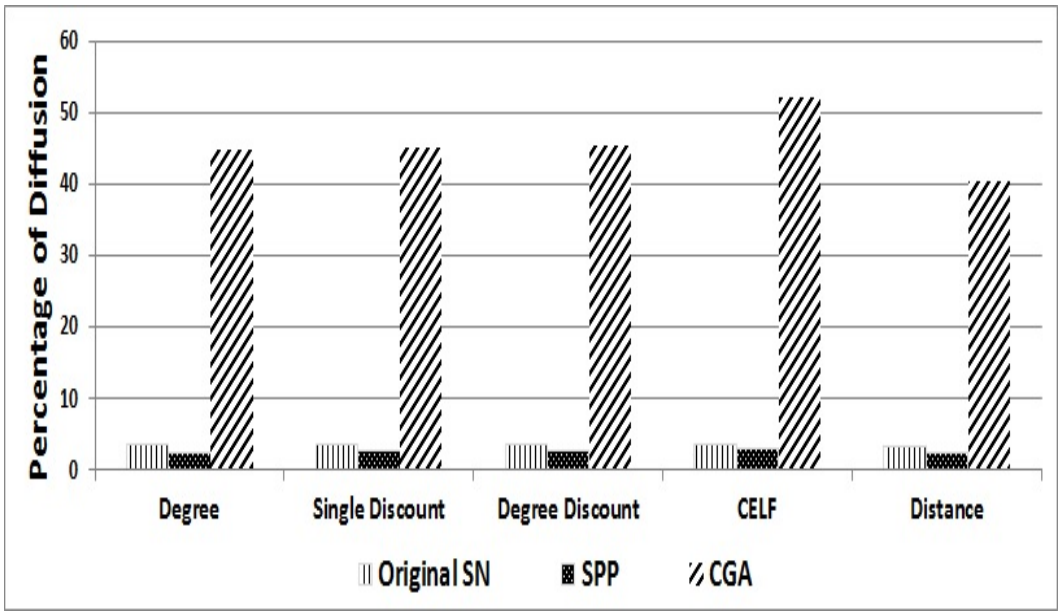


Figure 6.24: Percentage of spread in Digg under Independent Cascade Model

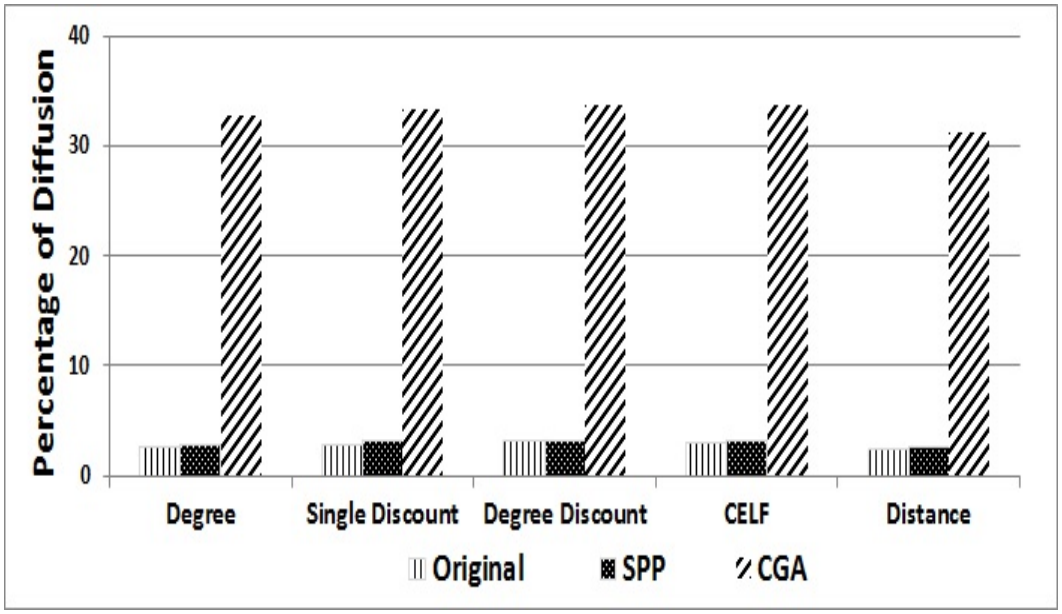


Figure 6.25: Percentage of spread in Twitter under Independent Cascade Model

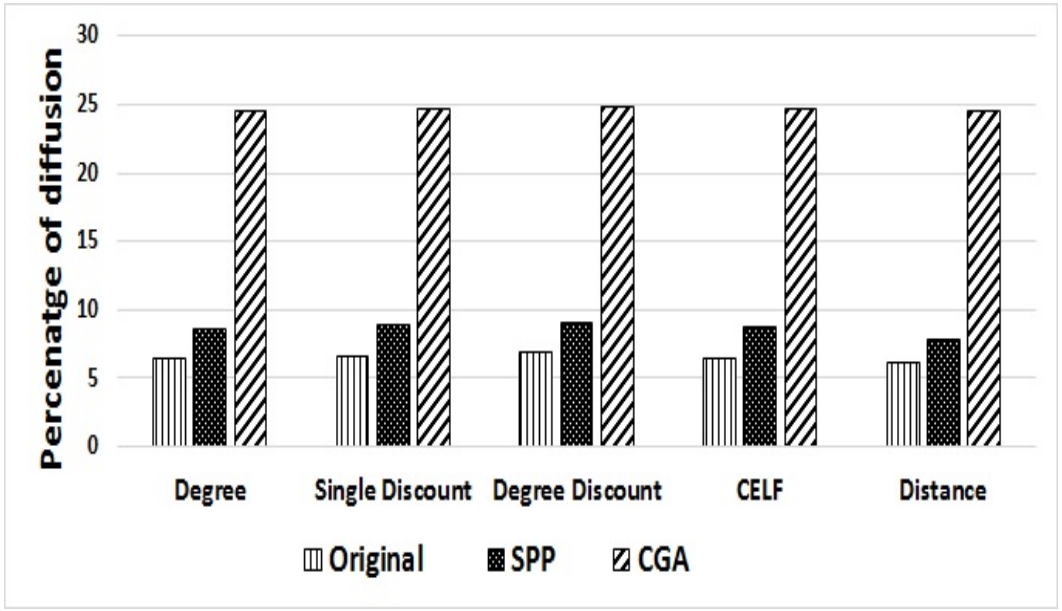


Figure 6.26: Percentage of spread in Infectious under Independent Cascade Model

The presented approach is also evaluated on the linear threshold model as shown in figure 6.27 to figure 6.34. The results are summarized as follows:

- For **HEP**, the proposed pruning approach produces diffusion reaching 79.9% on DegreeDiscount, 76.6% on SingleDiscount, 61.5% on Degree heuristic , 80.7% in CELF and 60.55% in Distance heuristic.
- For **PHY**, the proposed pruning approach produces diffusion reaching 64.78% on DegreeDiscount, 58.9% on Single Discount, 34.98% on Degree heuristics, 54.8% in CELF and 32.4% in Distance heuristic.
- For **Email**, the proposed pruning approach produces diffusion reaching 92.3% on DegreeDiscount, 91.9% on Single Discount, 91% on Degree heuristics, 87.3% in CELF and 85.5% in Distance heuristic.
- For **Wikivote**, the proposed pruning approach produces diffusion reaching 94% on DegreeDiscount, 93.4% on Single Discount, 93% on Degree heuristics, 87.3% in CELF and 79.4% in Distance heuristic.
- For **Digg**, the proposed pruning approach produces diffusion reaching 82.47% on DegreeDiscount, 81.7% on Single Discount, 80.5% on Degree heuristics, 65.34% in CELF and 64.77% in Distance heuristic.
- For **Youtube** the proposed pruning approach produces diffusion reaching 80.3% on DegreeDiscount, 79.4% on Single Discount, 78.7% on Degree heuristics, 73.8% in CELF and 74.3% in Distance heuristic.
- For **Twitter** the proposed pruning approach produces diffusion reaching 77.6% on DegreeDiscount, 75.6% on Single Discount, 71.8% on Degree heuristics, 73.9% in CELF and 69.3% in Distance heuristic.
- For **Infectious** the proposed pruning approach produces diffusion reaching 31.4% on DegreeDiscount, 30.36% on Single Discount, 24.39% on Degree heuristics, 28.65% in CELF and 25% in Distance heuristic.

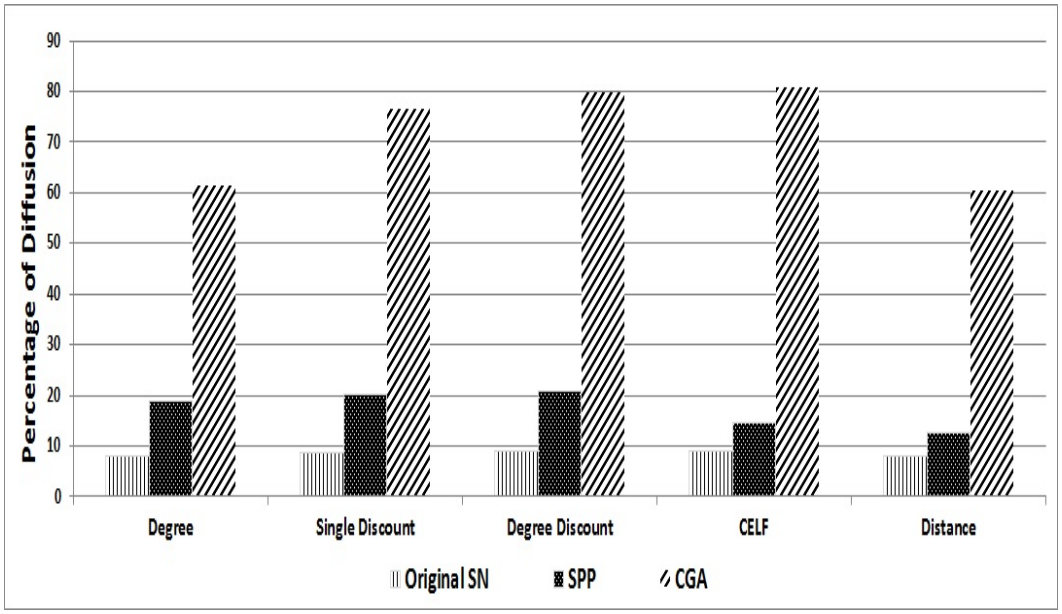


Figure 6.27: Percentage of spread in HEP under Linear Threshold Model

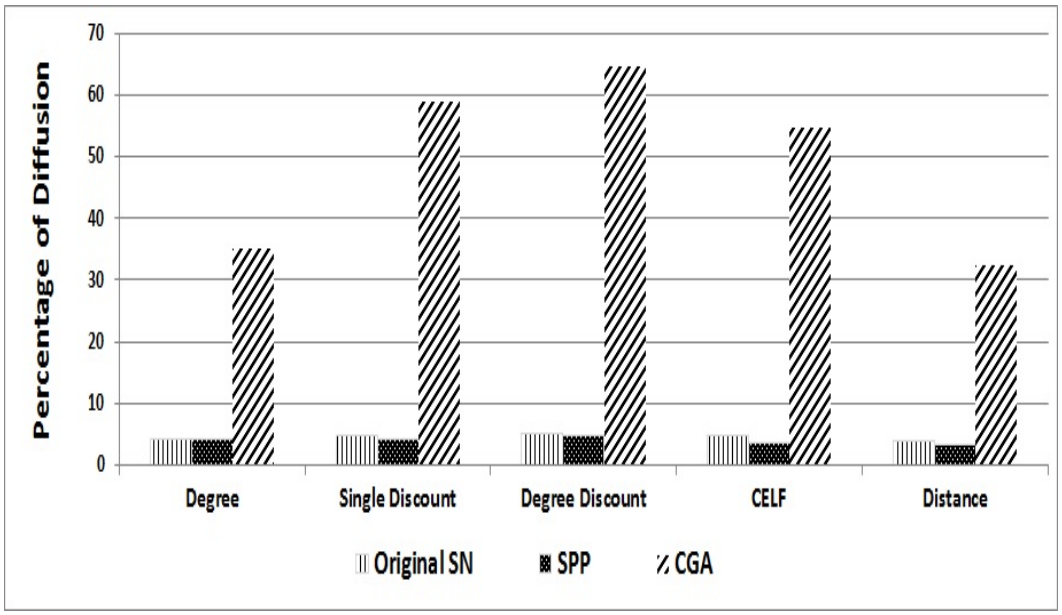


Figure 6.28: Percentage of spread in PHY under Linear Threshold Model

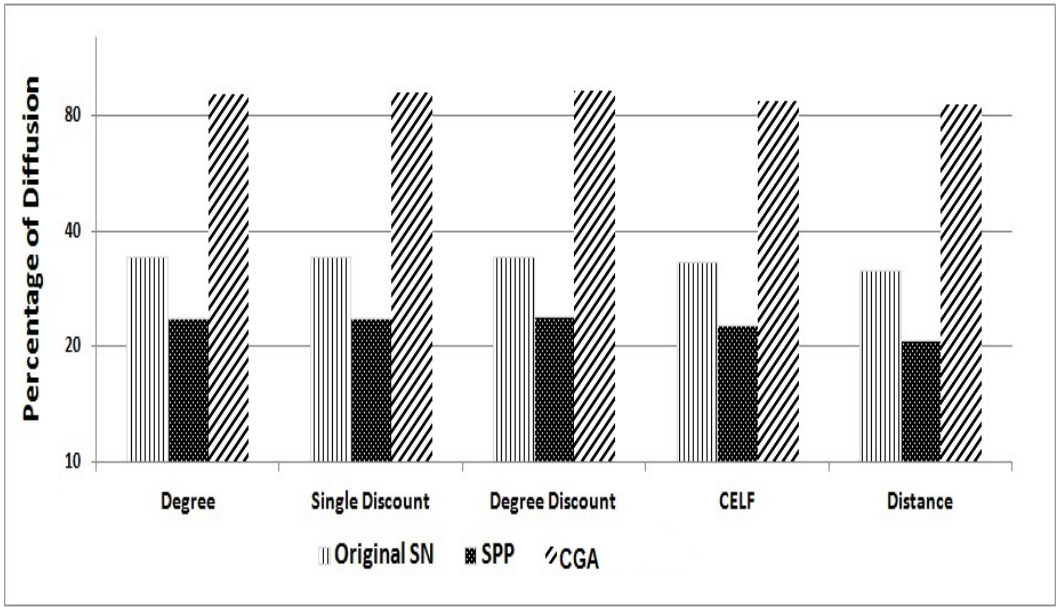


Figure 6.29: Percentage of spread in Email under Linear Threshold Model

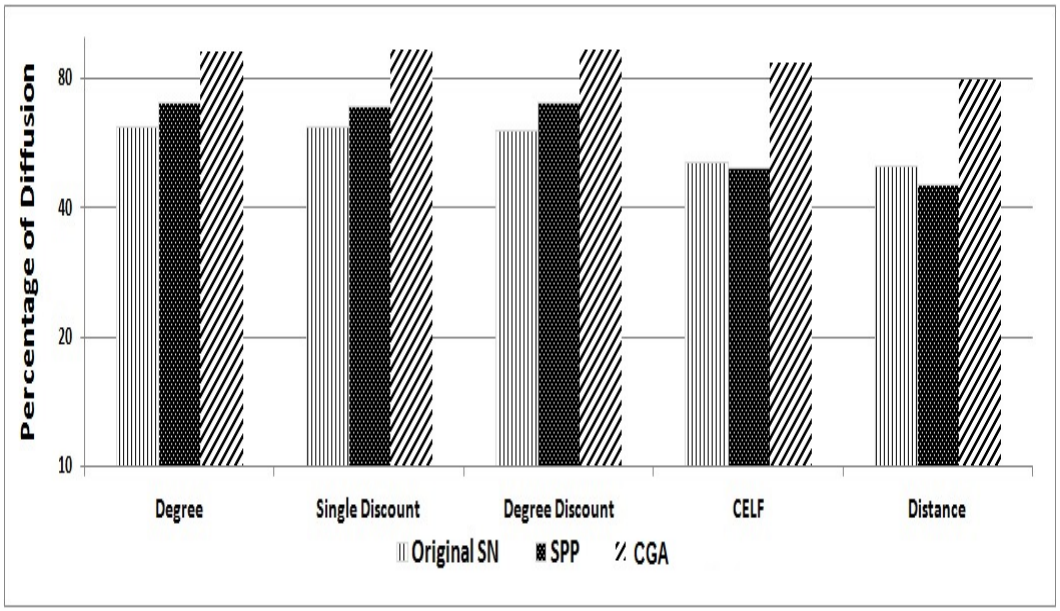


Figure 6.30: Percentage of spread in Wikivote under Linear Threshold Model

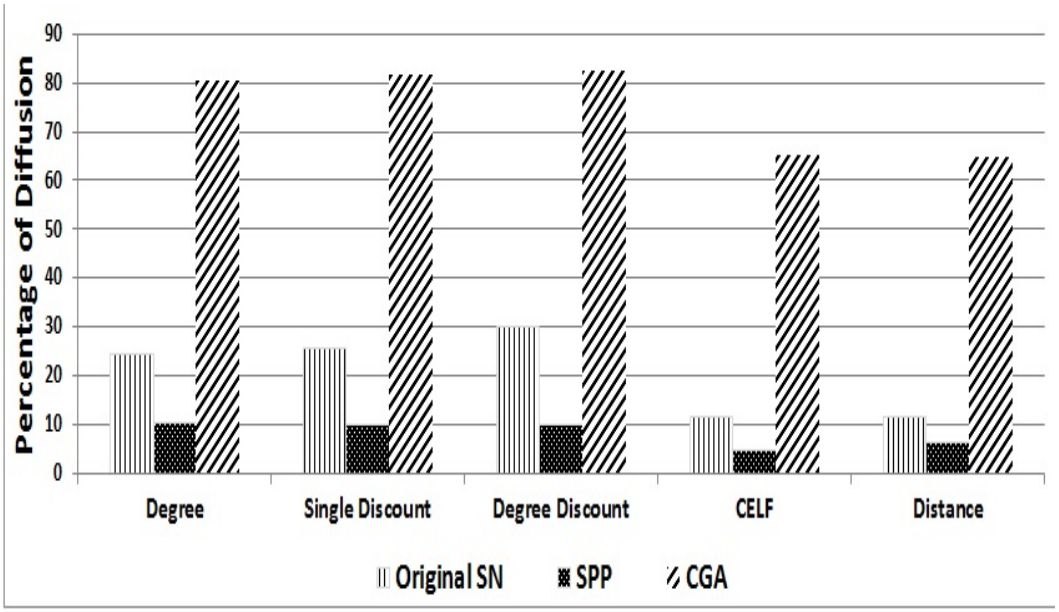


Figure 6.31: Percentage of spread in Digg under Linear Threshold Model

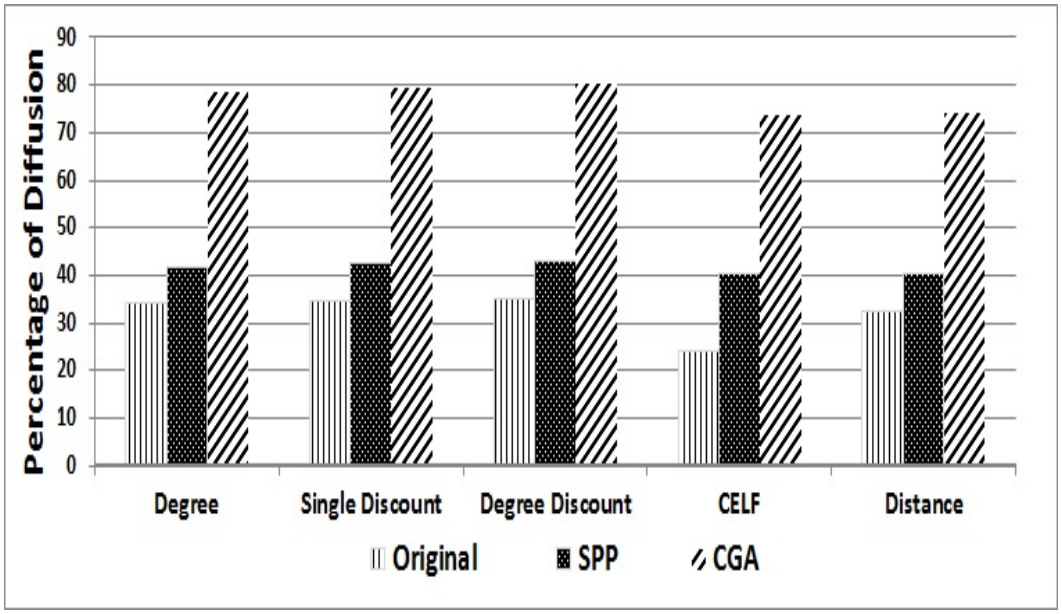


Figure 6.32: Percentage of spread in YouTube under Linear Threshold Model

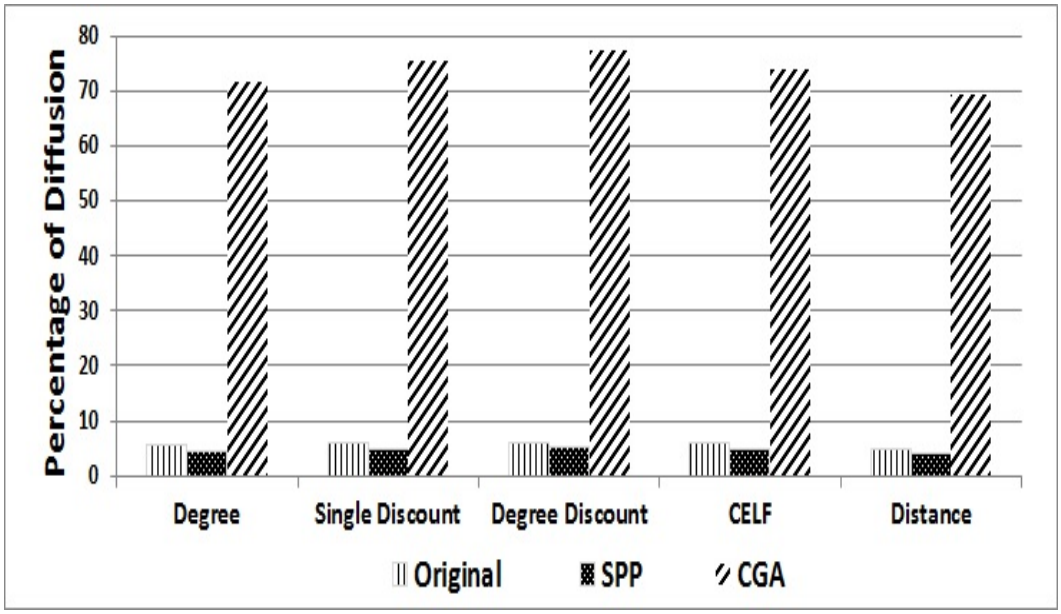


Figure 6.33: Percentage of spread in Twitter under Linear Threshold Model

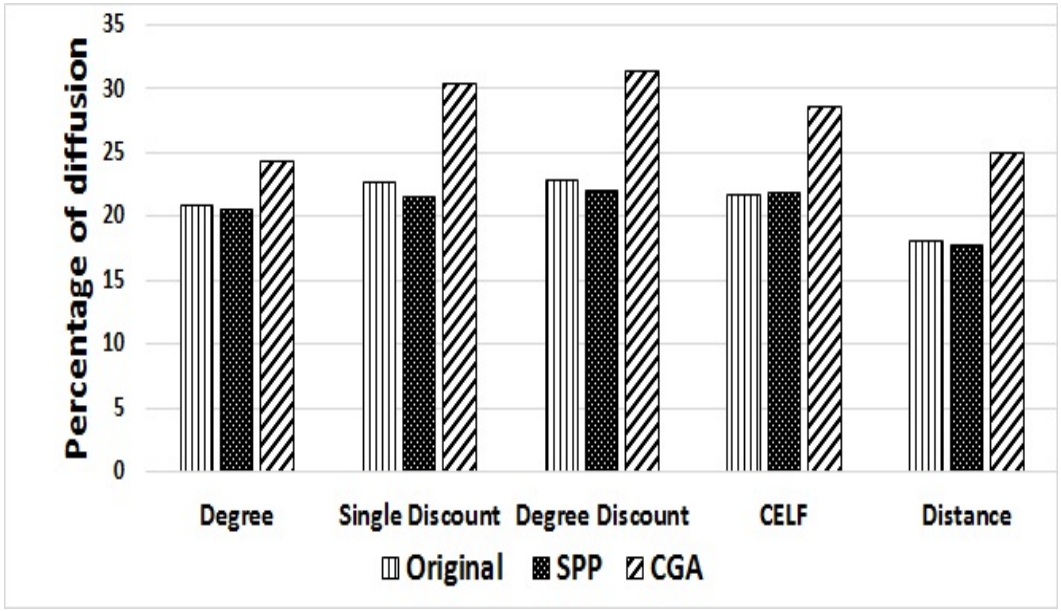


Figure 6.34: Percentage of spread in Infectious under Linear Threshold Model

Observation 6.4.2. The novel pruning approach developed in this research work improves the diffusion process in the independent cascade and linear threshold models.

This research presents an approach which examines the nodal attribute of the social network user and develops a criterion to retain a subset of users to form a pruned graph of the social network. Thus, a large social network is reduced to an optimal social network. The results demonstrate the effectiveness of the new approach. The contributor graph exhibits the presence of small world properties significantly better, when compared to original graph and graphs pruned on other approaches. The efficacy of the new approach is demonstrated in the diffusion process under independent cascade and linear threshold models on various seeding strategies. The results support the claim that small world properties play a significant role in information diffusion process. Although, this research discusses pruning of enormous social network in the context of information diffusion, this approach can be tried on any applications where propagation of information plays a vital role. Thus, this research explores a new strategy to deal with the enormous and rapid growth of the data in social networks.

6.5 Summary

This chapter presented an approach to prune the social network which in turn addresses the scalability issue seen in large social networks. The approach presented in this research, not only retains but also improves its information propagation properties. The new approach is evaluated on various metrics and compared with other approaches. The results show that use of the contributor graph results in better diffusion in the social network.

Chapter 7 discusses the user influence in the context of information diffusion in the social networks.

Chapter 7

Estimating user influence in social networks

The theory of probabilities is at bottom nothing but common sense reduced to calculus -Pierre Laplace

This chapter addresses the problem of estimating the user influence in the social networks. It also discusses the motivation to develop such an approach. The novel method named as *Influx*, along with the methodology and algorithm are discussed here. Further, the approach is substantiated through lemma and theorem. The results are also discussed in the later sections of this chapter.

7.1 Background

The popularity of social networks have resulted in many interesting applications such as viral marketing, recommendation systems and so on. In these applications, user influence also known as influence probability, plays an important role. In social networks, influence probability is defined as the probability a user can influence his/her friends to adopt an information immediately or in foreseeable future. Such probabilities are central to fundamental issues in social network analysis including influence maximization. In practice, influence probabilities have significant implications for applications such as target marketing, poll prediction, political campaigns and so on. Yet, predicting influence probabilities has not received sufficient research attention.

The solution to the influence maximization problem starts with the weighted undirected social graph $G(V, E)$ where V is the set of users and E represents the set of edges. The weight on the edge represents the influence probability. In reality, the social graph is readily available, whereas, the edge weight i.e., influence probability is not. This has lead to the use of a pre assigned value for user influence in the solution to influence maximization, which may not reflect the real world scenario. There are two reasons why the use of such an assumed value may not be an ideal setup in the solution. First, assuming uniform information spread along all social ties can lead to overestimation of information dissemination as well as lead to selection of influential users who may not be optimal (Wilson et al., 2012). As such, an assumed value will only bias the outcome. Second, influence is a behavioral attribute that changes over time. Hence, this parameter should not be made constant. In addition, interaction intensities among users and also users' inclination in adopting information is important to predict influence probabilities. These factors have to be considered while deriving a solution for predicting information spread in social networks.

There are several attempts made to estimate user influence (Fang et al., 2013; Goyal et al., 2010; Jiang et al., 2013; Kasthurirathna et al., 2015; Kimura et al., 2009a; Kutzkov et al., 2013; Mathioudakis et al., 2011; Romero et al., 2011; Saito et al., 2010, 2008; Teng et al., 2015; Wang et al., 2013; Xiang et al., 2010; Yang and Leskovec, 2010) which are discussed in Chapter 3. However, these approaches are resource expensive. Moreover, they require accurate and in-depth user profile details, which in most of the cases are unavailable. Hence, there is a need to design an approach to estimate user influence without bypassing user privacy. This research presents a new approach to estimate user influence, keeping in mind the restrictions on the availability of social network data. The new approach uses data related to user activities which are readily available in the action log repositories. By quantifying the influence among users via their interaction count, the outcome is made realistic.

7.2 Preliminaries

In this section, a case study of using various values of influence is discussed. Further the properties of influence as a relation are explored.

7.2.1 Case study

Most of the literature assumes that influence probability is readily available as weights on the edges of the social graph. For this reason computation of influence probability is largely left unexplored. The HEP dataset is studied under five values of influence, to analyze the impact of an assumed value of influence on the outcome. The values considered in each of these cases are: (i) 0.01, (ii) 0.05, (iii) 0.1, (iv) 0.5 and (v) 0.9. These different values of influence, in a similar experimental setup with the initial seed set obtained from Degree, Distance, Singlediscount and Degreediscount heuristics (Chen et al., 2009), are used.

Figure 7.1 shows that there is an increase in the spread of information when a higher value of user influence is chosen. Also, from the existing works, it is not evident why a particular value of influence is preferred over the other values. To accurately predict the spread, it is important to take into consideration both the seed set, as well as, influence among users. Moreover, with an assumed value of influence, the outcome may deviate. This is the motivation to develop an approach to estimate the user influence.

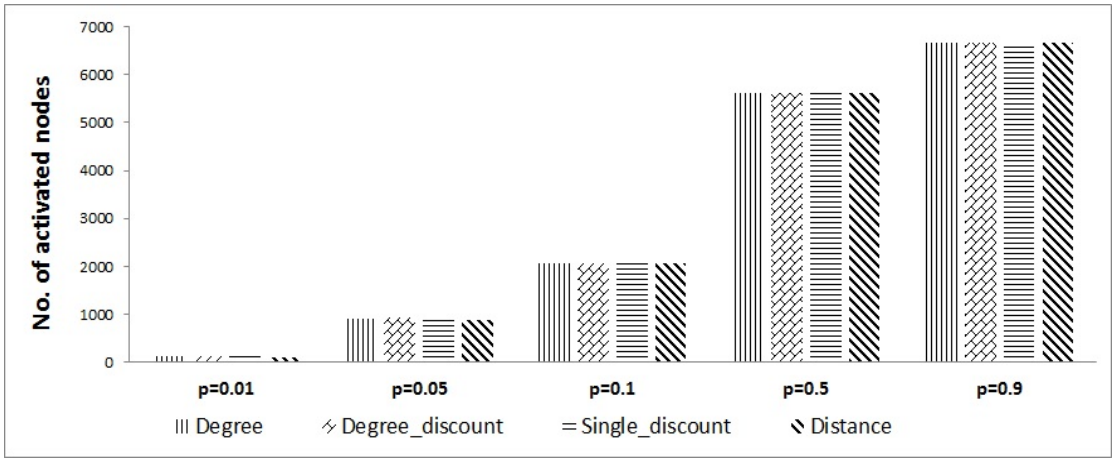


Figure 7.1: Information spread with various value of user influence

7.2.2 Properties of influence relation

User influence is considered as a relation between two connected users in the social network. Such a relation has the following properties.

1. Reflexivity: (a,a) relation holds true. It is a trivial case when a node influences itself to perform an action.
2. Symmetry:$(a, b) \rightarrow (b, a)$ does not hold true. Node A can influence B, but node B may not influence node A to perform an action.
3. Transitivity: (a, b) and $(b, c) \rightarrow (a, c)$ holds true. Influence propagates in the network.

7.3 Problem description

Since the seminal work of Saito et al. (2008), various contributions were done towards estimating influence. To accurately predict information spread and evaluate performance of seed selection algorithms, it is important to estimate the user influence. In this section a novel approach named as *Influx*, developed to estimate user influence is discussed.

One can establish a relation between the influence probability P, seed set I and the information spread initiated by I represented as $\sigma(I)$ as follows:

$$\sigma(I) := \textit{diffusionmodel}(P,I) \tag{7.3.1}$$

To rephrase, the spread of information is considered as a function of probability of influence and the seed set I. It is reasonable to believe that a user can only influence his/her friends to the extent he/she maintains interactions with them. On this surmise, user influence is estimated.

Problem 7.3.1. *For an interaction graph $G_I(V_I, E_I)$, representing the underlying social network and an activity log A of users, find*

$$P = \{p_{uv}, \forall u, v \in V_I, e(u, v) \in E_I\}$$

under the following constraints

(i) $p_{uv} \in P$ *iff* $e(u, v) \in E_I$

(ii) $0 < p_{uv} \leq 1$

such that for a seed set I, $\sigma(I) = Max\,[|\varphi(I, P)|]$.

The constraint (ii) eliminates the influence value 0. Thus, instead of using a randomly assumed constant, the proposed approach estimates the value for the set P of m values, one for each of the m edges in the social network graph.

7.4 Proposed methodology: Influx

Consider a scenario where a user A has five contacts B, C, D, E and F. For illustration, let us assume the number of interactions of A with B, C, D, E and F are 30, 40, 50, 60 and 30 respectively. Since the number of interactions of A with his/her neighbors are different, the probability that A influences each of his/her neighbors is different. Therefore, it is more appropriate to use interaction intensities while estimating user influence.

In this research the problem of estimating influence is addressed as follows: The strength of an edge reflects the intensity of the interactions through the edge. The strength of the edge is represented as y_{ij}, which is the number of interactions from v_i to v_j. Since the interactions on either sides are not in equal proportions i.e., $y_{ij} \neq y_{ji}$ the probability of influence is also not symmetrical i.e., $p_{ij} \neq p_{ji}$.

Thus, in this research, user influence is quantified by using the interaction count of a user as follows.

$$p_{u,v} = \frac{y_{u,v}}{\Sigma_{s=\{n \in N\}} y_{u,s}} \tag{7.4.1}$$

where, N is the set of nodes incident on node u and $y_{u,s}$ is the number of interactions of the node u to the incident node n. The normalization process, sets the value of p in the range (0,1], according to the definition of probability. Since the input to the solution of the problem is the interaction graph, having only edges which are used for communication, number of interactions between any users cannot be 0, hence $p \neq 0$.

Using Eq(7.4.1), each edge now has estimated non uniform probability of influence as its weight instead of an assumed value. Thus the interaction graph is converted to a weighted graph where edge weights represents the user influence. This new graph is referred to as *Influence graph* and this approach is named as the *Influx*. With this approach, in the given example, $p_{A,B}$ is 0.147. Weights on other edges are similarly calculated. This scenario is shown in figure 7.2.

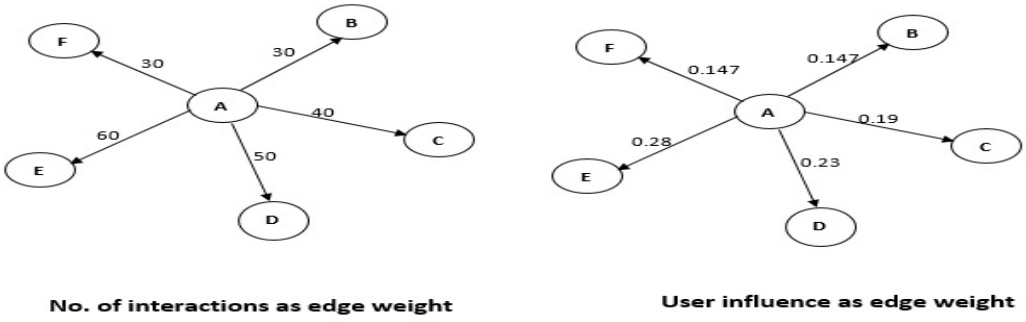


Figure 7.2: Social network with influence probability on the edges

7.4.1 Algorithm and proof of the concept

Algorithm 3 details the steps to estimate the user influence. The input to algorithm 3 is a directed interaction graph $G_I(V_I, E_I)$ of the given social network, and the activity log of the users $A(u, v, y_{uv})$ which contains the number of interaction of the node u with its neighbor v represented as y_{uv}. The output is the set $P = \{p_{uv}$ for $\forall e(u, v) \in E_I\}$ representing user influence on each edge.

Complexity: The execution time of algorithm 3 depends on the number of edges in the input graph. If m is the number of edges and q the size of neighbourset, then the time complexity of algorithm 3 is $O(\bar{m}.q)$, which is linear.

The proof for the correctness of the *Influx* approach is given via Lemma 7.4.1 and Theorem 7.4.2.

Input: $G_I(V_I, E_I)$ and activity log A(v,u, $y_{u,v}$)

Output: $G_I(V_I, E_I, P)$ where $P = \{p_{uv}$ for $\forall e(u,v) \in E_I\}$

1 Initialize $\Gamma_{u,s} = 0$

2 For each $e(u,v) \in E_I$ repeat steps 3 to 6

3 Compute N=neighbourset(u)

4 For each $s \in N$

5 Compute total number of interactions $\Gamma_{u,s}$ as in step 6

6 $\Gamma_{u,s} += y_{u,s}$

7 $p(u,v) = \frac{y_{u,v}}{\Gamma_{u,s}}$

Algorithm 3: Influx algorithm

Lemma 7.4.1. *As the number of interactions between a pair of nodes* u *and* v *increase, the probability of influence* p_{uv} *on the edge* e(u,v) *also increases.*

Proof: From Eq.(7.4.1), assume $y_{u,v} \to n$, where n is the total number of interactions of the node u; which is represented by $\Sigma_{s=1}^{k} y_{u,s}$. This happens when almost all the interactions of a user u is with user v. Then $p_{uv} \to 1$. Hence the proof. □

Theorem 7.4.2. *Given a diffusion model M and social network represented as a weighted directed graph* $G_I(V_I, E_I, P)$*, the spread under the seed set I of k nodes for* $P = \{p_i = 0.01, \forall e(u,v) \in E_I\}$ *is* $\sigma(I)$*. The spread with estimated influence probability* p_h *under the same model M and seed set S for* $P = \{p_1, p_2, p_3, ..., p_m\}$ (estimated values)*is* $\sigma(I*)$*. Then,* $\sigma(I*) \geq \sigma(I)$ *when there is high interactions among the users.*

Proof: For the information to spread in the network there should exist a path from seed nodes $s_i \in I$ to the target nodes $v \in V - S$. This is $path(s_i, u)$. The expected spread under a given propagation model M with an uniform influence probability p_i is given in Eq(7.4.2).

$$\sigma(I) = \Sigma_{u \in V-I} \;\; p_i \; . \; path(s_i, u) \tag{7.4.2}$$

where $path(s_i, u)$ is 1 if there is a path or 0. If l represents the number of $path(s_i, u) = 1$, Eq(7.4.2) is rewritten as in Eq(7.4.3).

$$\sigma(I) = \Sigma_{u \in V-I} \ p_i \ . \ l \tag{7.4.3}$$

Furthermore, the spread under the same propagation model M and seed set I, taking into consideration the estimated influence probability p_h is $\sigma(I*)$. Let us assume that of the previous l paths available from I to u, where $u \in V - I$; $l - 1$ paths have p_i(set to 0.01 or 0.001 or 0.001) and one path has p_h. Eq (7.4.3) is re-written as in Eq(7.4.4).

$$\sigma(I*) = \Sigma_{u \in V-I} \ p_i \ . \ (l-1) \ + \ p_h \ . \ 1 \tag{7.4.4}$$

As the nodes on this path are highly interactive, using lemma 1, $p_h \rightarrow 1$ on this path. Therefore $p_h \geq p_i$, when users are highly interactive. Thus it follows $\sigma(I*) \geq \sigma(I)$. Hence the proof. □

7.5 The Influx-IC diffusion model

This research designs a variant of IC model, namely *Influx-IC* model. This new model uses non uniform probability of influence estimated through *Influx* approach to predict the information spread in the social network. In the traditional IC model, all nodes are influenced by a uniform probability usually $p = 0.1$ or 0.01. In contrast, the *Influx-IC* model assigns every connected edge with an estimated value of probability of influence. Similar to the IC model, nodes in *Influx-IC* has a single chance to influence its neighbor. Once influenced, the nodes remain active, till the end of diffusion process. The diffusion process ends when there are no more nodes to be influenced. The diffusion process in *Influx-IC* is demonstrated in figure 7.3 to figure 7.4.

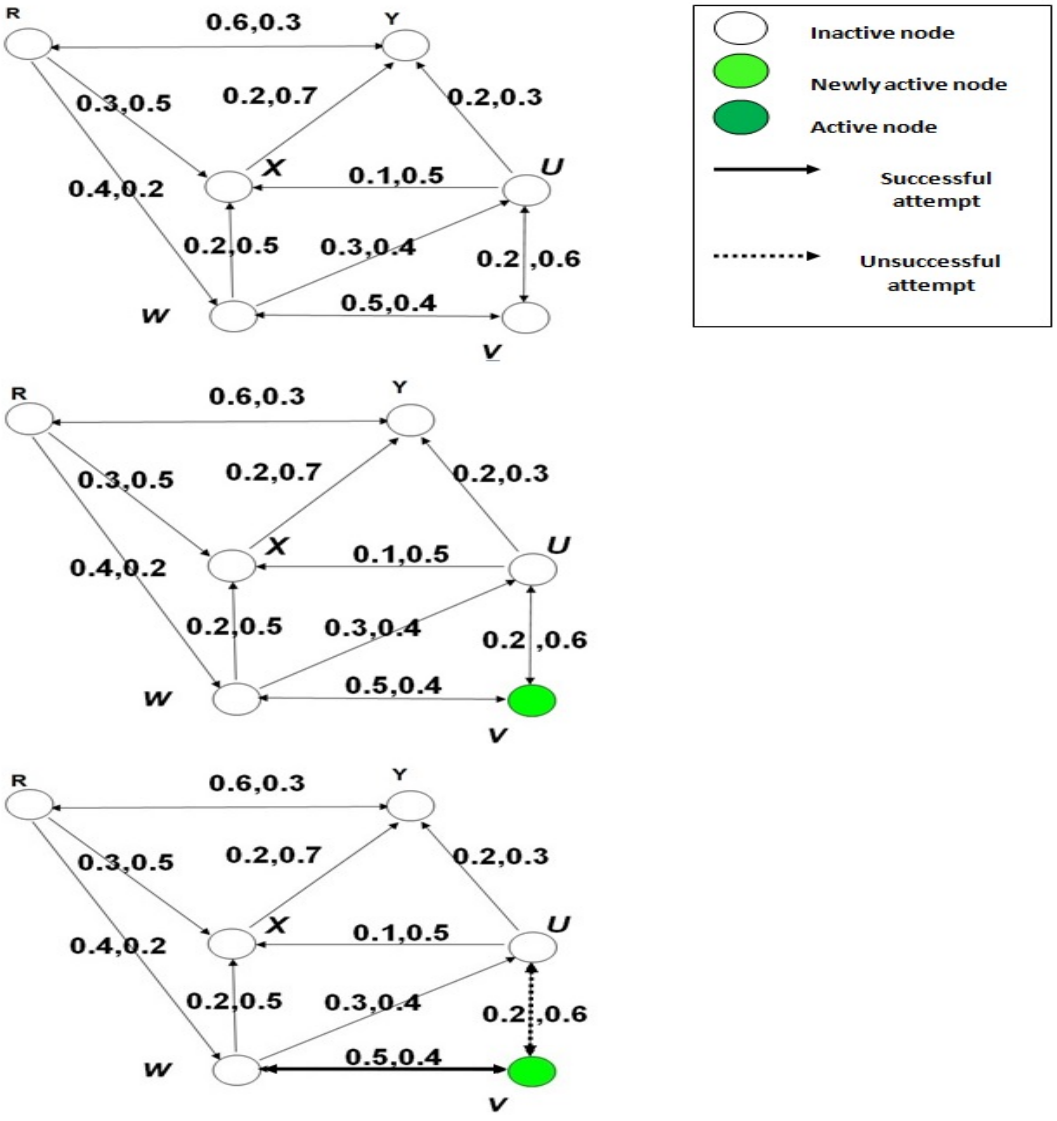


Figure 7.3: Diffusion at time *t1 to t3*

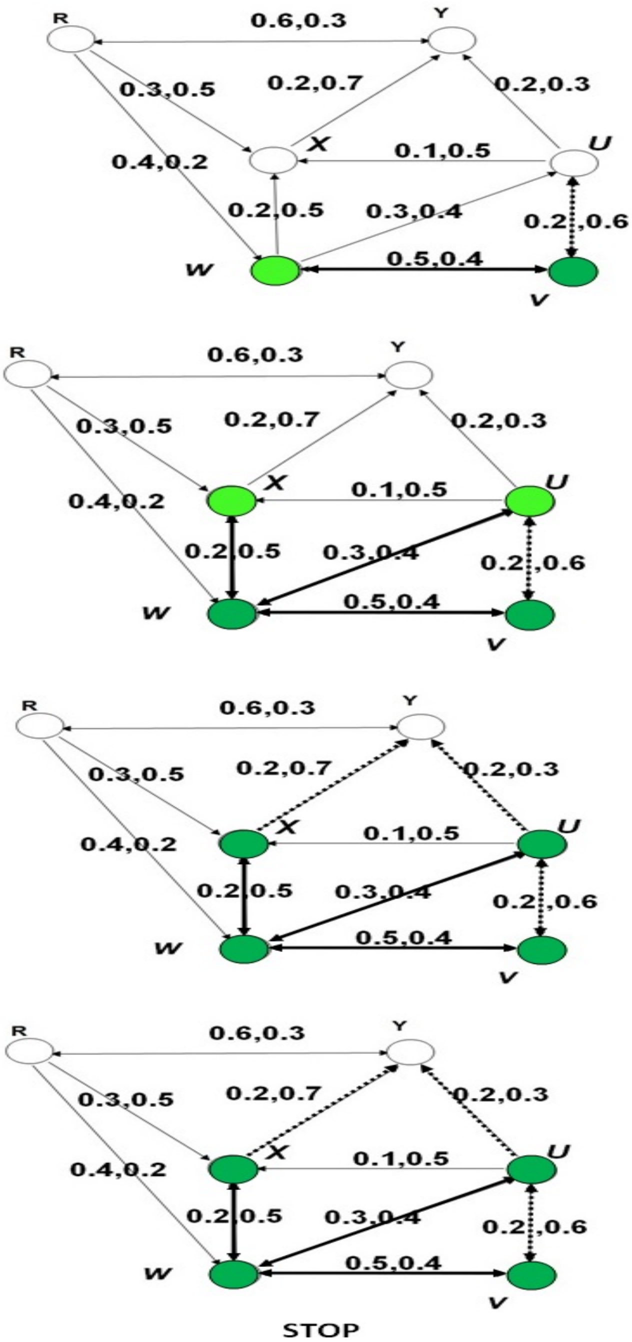


Figure 7.4: Diffusion at time *t4 to t7*

7.5.1 Results and analyses

7.5.1.1 Experimental setup

The *Influx-IC* model is used for predicting information spread in the chosen datasets. To evaluate its performance, the following influence models are used.

1. **RNUDp-IC**: The influence probabilities, generated from an uniform distribution, are randomly assigned to all edges of the graph.
2. **RNNDp-IC**: The influence probabilities, generated from a normal distribution, are randomly assigned to all edges of the graph, .
3. **Trivalency model (TVM)**: The influence probability value of 0.01 is assigned to all edges of the graph.

The first two approaches assign non uniform user influence and the TVM assigns a constant uniform weight on all the edges.

The standard seed selection algorithms such as: degree, distance, degreediscount and single discount, for various seed set size k, are investigated to understand the impact of the *Influx* on the estimation of information spread.

7.5.1.2 Outcome

The diffusion process is evaluated in the chosen diffusion models, with 1000 iterations and various budget constraint k, i.e., number of seeds. The results are shown in figure 7.5 to figure 7.11 and performance gain is available in table 7.1 to table 7.7.

Overall, the information spread predicted by the *Influx-IC* is higher than RNNDp-IC, RNUDp-IC and TVM, for each of the standard algorithms such as degree, distance, singlediscount and degreediscount. These results show that the *Influx* approach developed to estimate influence, yields better outcome since, it reflects the interactive nature of the users. Where as the existing approach, predicts the spread not accounting to the user inclination.

Table 7.1: HEP- Influx-IC Performance gain(in %)

-	Degree	Distance	SingleDiscount	DegreeDiscount
RNNDp	20	15	27.4	30
RNUDp	18	14	24.75	26.8
TVM	20	15.6	28.5	30

Table 7.2: PHY- Influx-IC Performance gain(in %)

-	Degree	Distance	SingleDiscount	DegreeDiscount
RNNDp	8.9	14	18.43	23.17
RNUDp	4.8	8.9	9.8	12.6
TVM	8.6	14	20	25.53

Table 7.3: Wikivote- Influx-IC Performance gain(in %)

-	Degree	Distance	SingleDiscount	DegreeDiscount
RNNDp	15	13.15	13.2	15.11
RNUDp	40.4	35.52	39.7	40.6
TVM	34.47	28	33.26	35.8

Table 7.4: Infectious- Influx-IC Performance gain (in %)

-	Degree	Distance	SingleDiscount	DegreeDiscount
RNNDp	29.8	29.6	28	27.7
RNUDp	29.7	29.6	27.9	27.7
TVM	29.36	30	29.39	32

Table 7.5: YouTube- Influx-IC Performance gain(in %)

-	Degree	Distance	SingleDiscount	DegreeDiscount
RNNDp	9.6	15.7	10.5	10.6
RNUDp	3.15	6.3	6.3	6.1
TVM	19	24.21	22.1	19.6

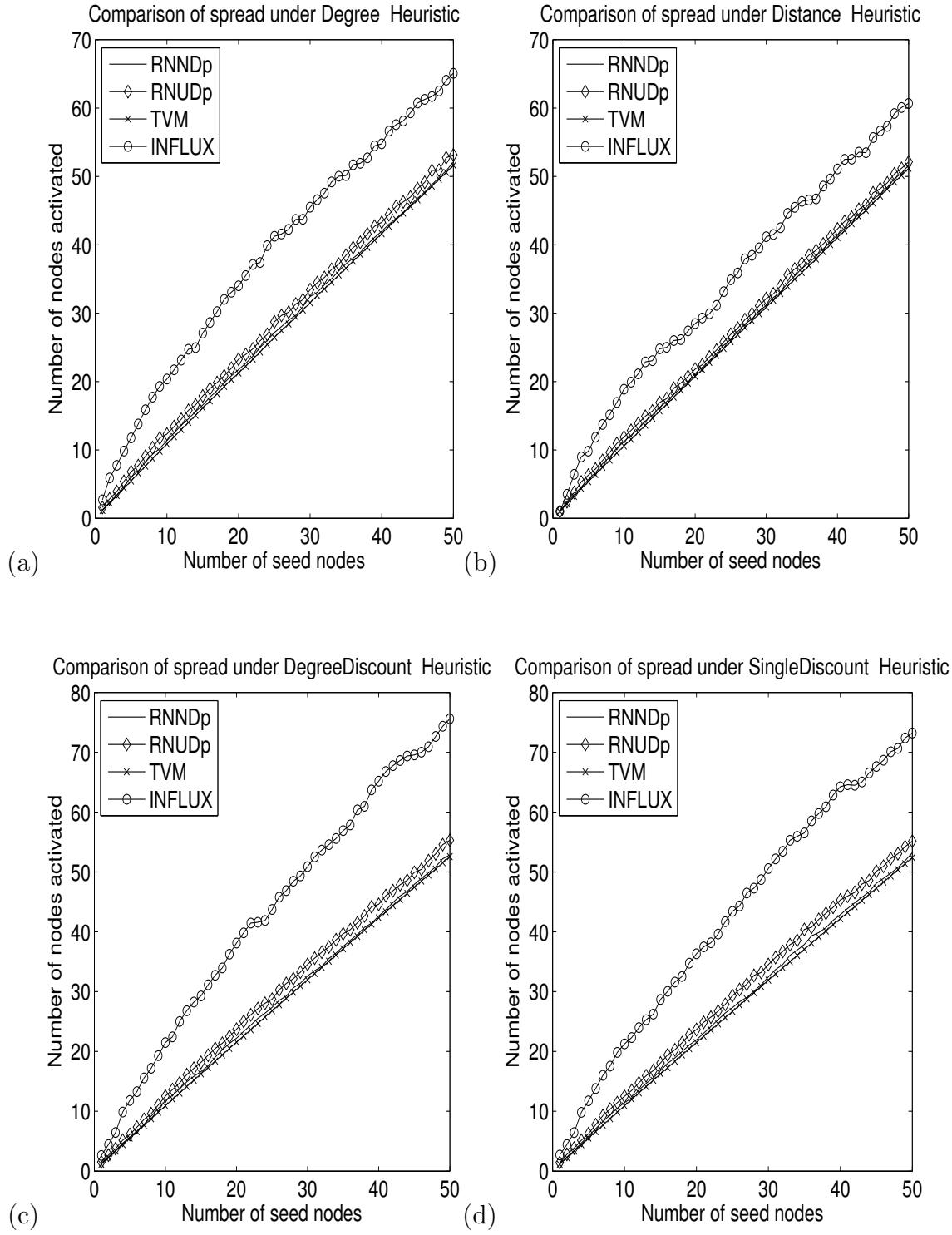


Figure 7.5: Comparison of Influx-IC to other models in HEP for (a)Degree (b) Distance (c) DegreeDiscount and (d) SingleDiscount heuristics

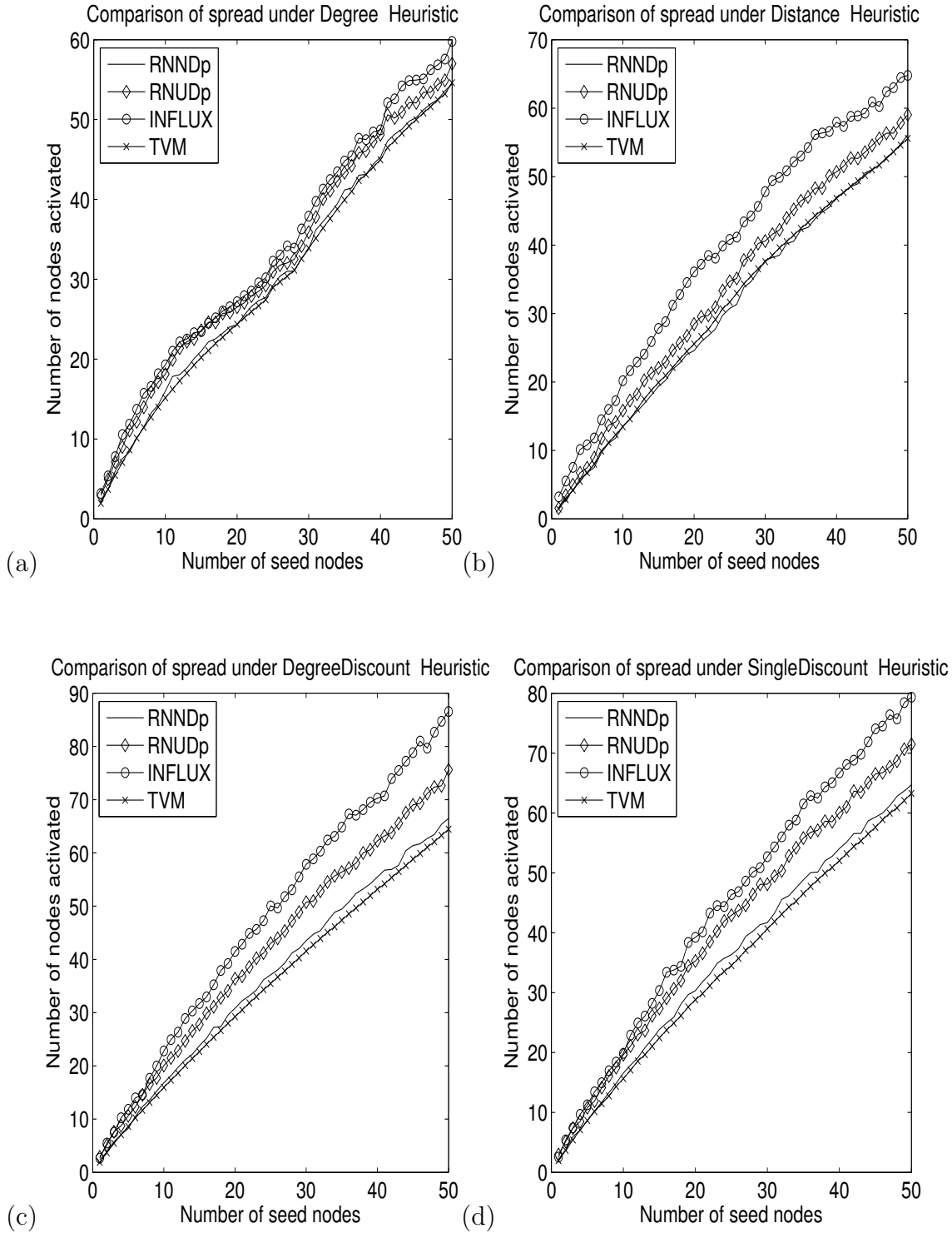


Figure 7.6: Comparison of Influx-IC to other models in PHY for (a)Degree (b) Distance (c) DegreeDiscount and (d) SingleDiscount heuristics

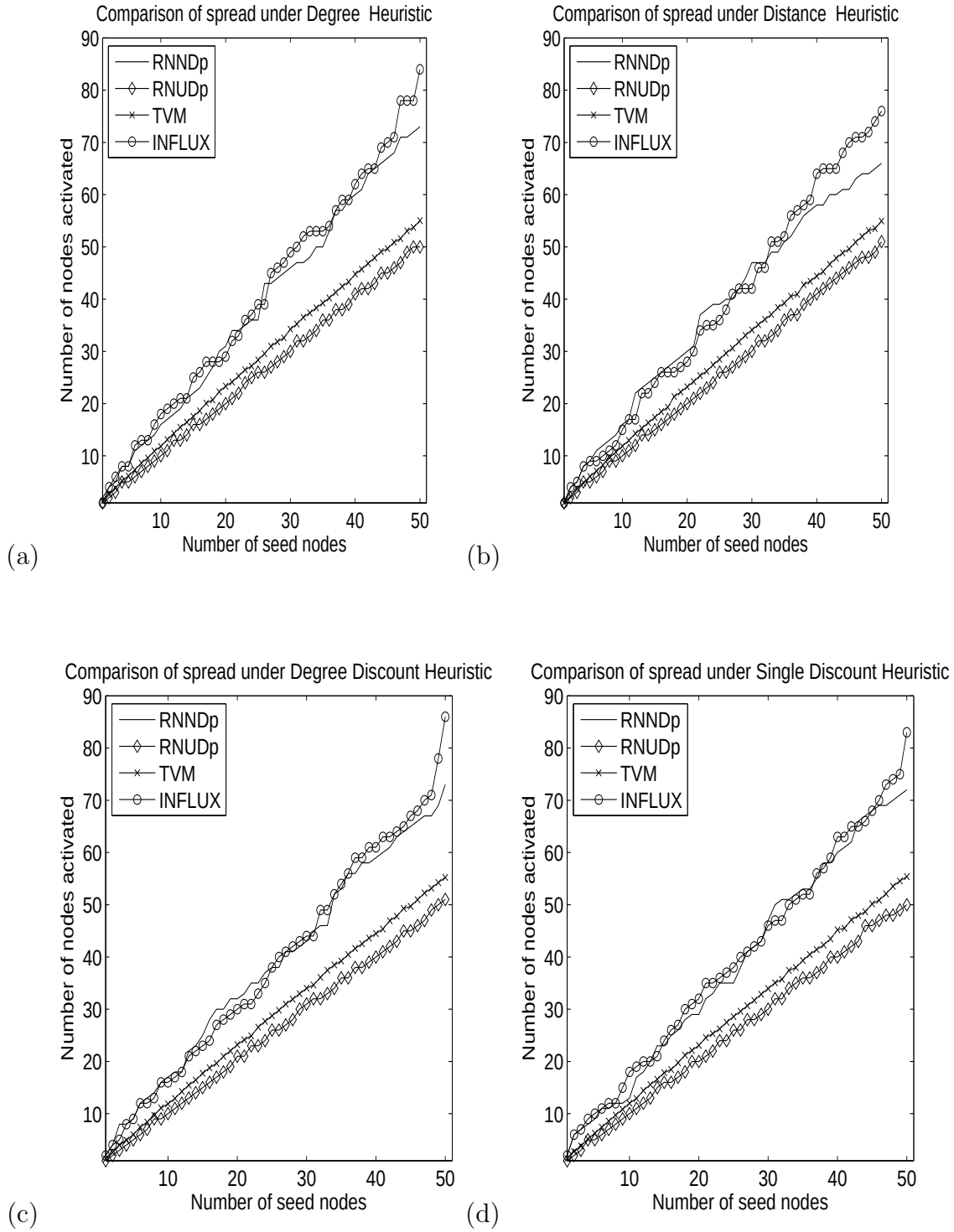


Figure 7.7: Comparison of Influx-IC to other models in Wikivote for (a)Degree (b) Distance (c) DegreeDiscount and (d) SingleDiscount heuristics

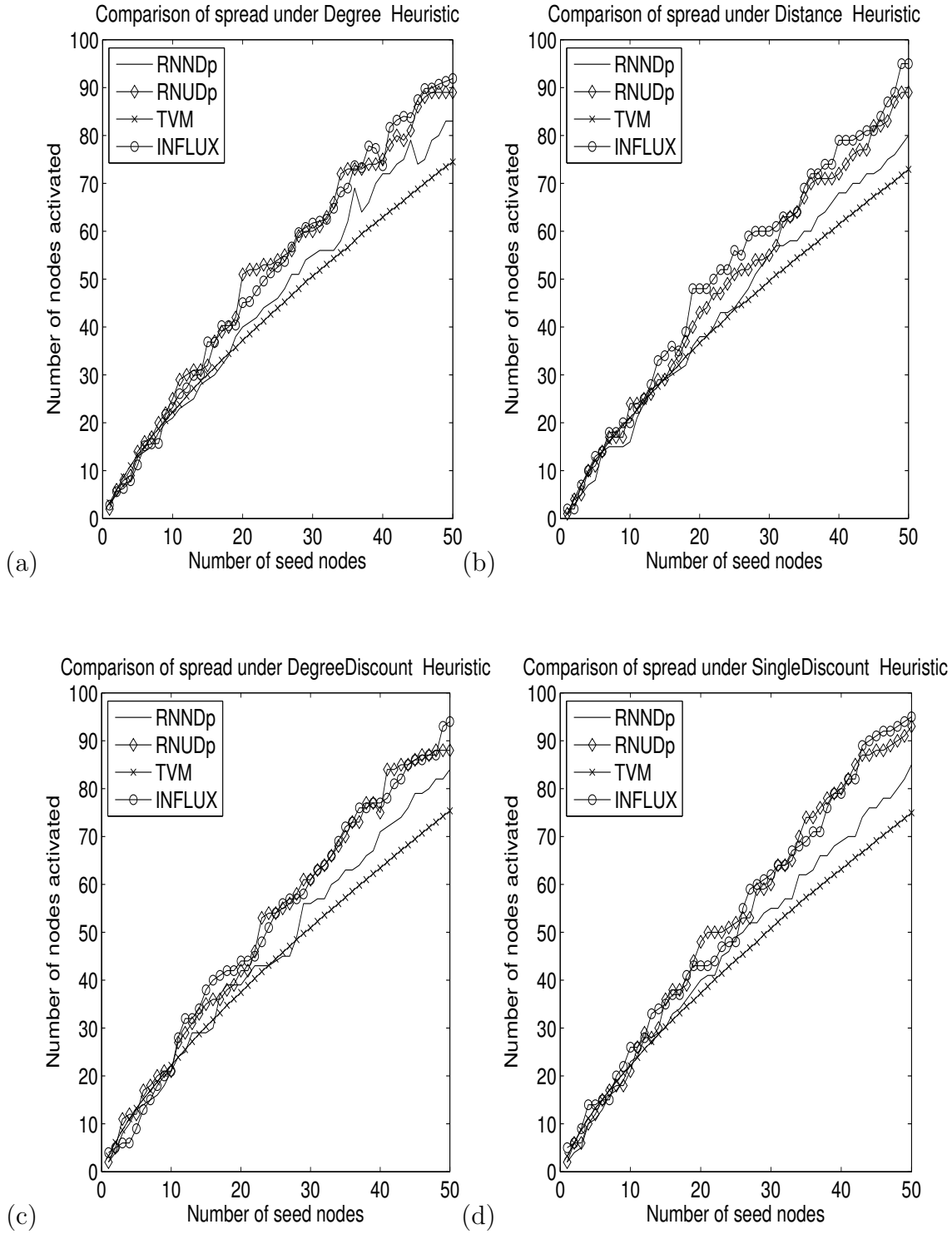


Figure 7.8: Comparison of Influx-IC to other models in Youtube for (a)Degree (b) Distance (c) DegreeDiscount and (d) SingleDiscount heuristics

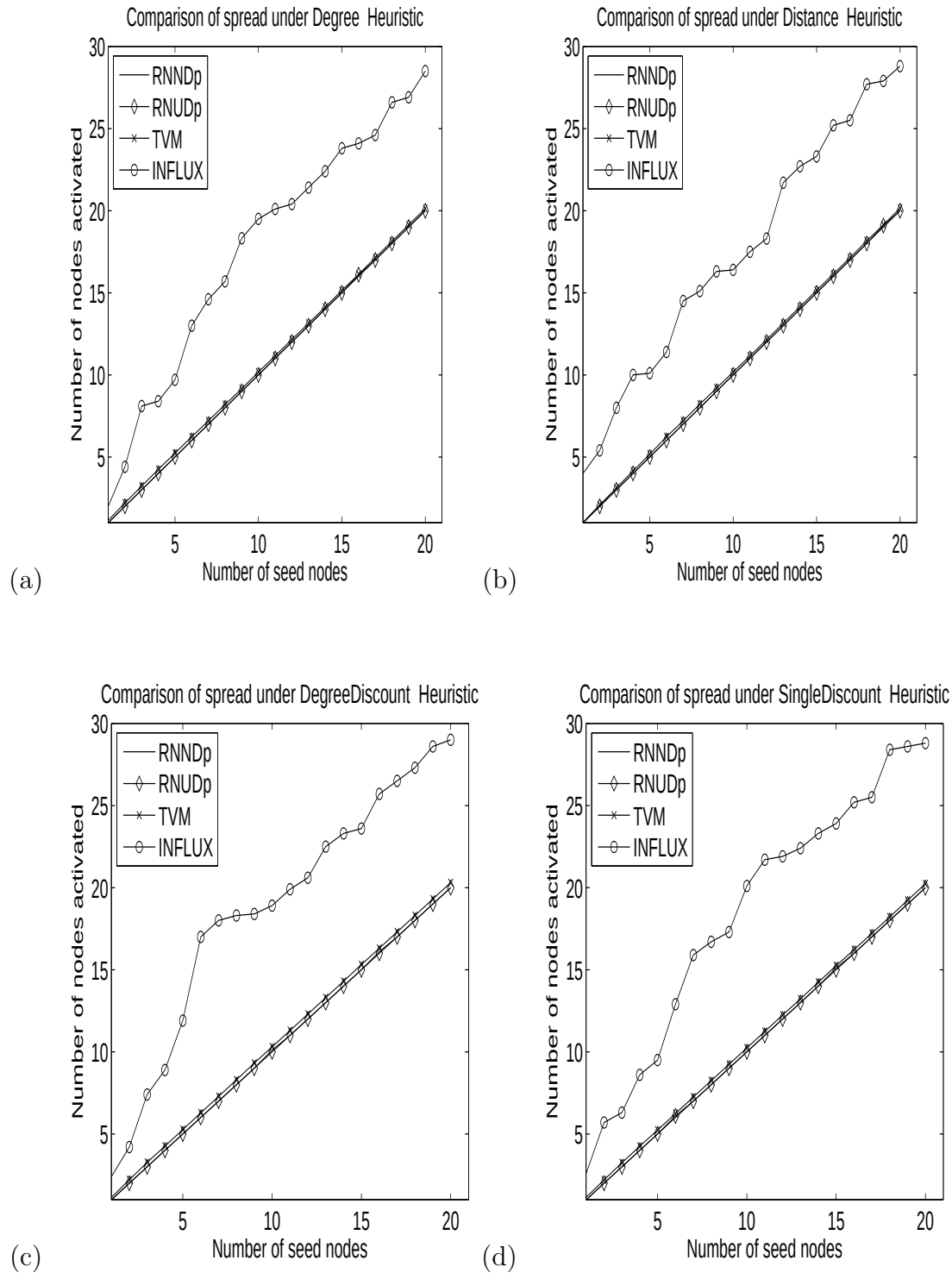


Figure 7.9: Comparison of Influx-IC to other models in Infectious for (a)Degree (b) Distance (c) DegreeDiscount and (d) SingleDiscount heuristics

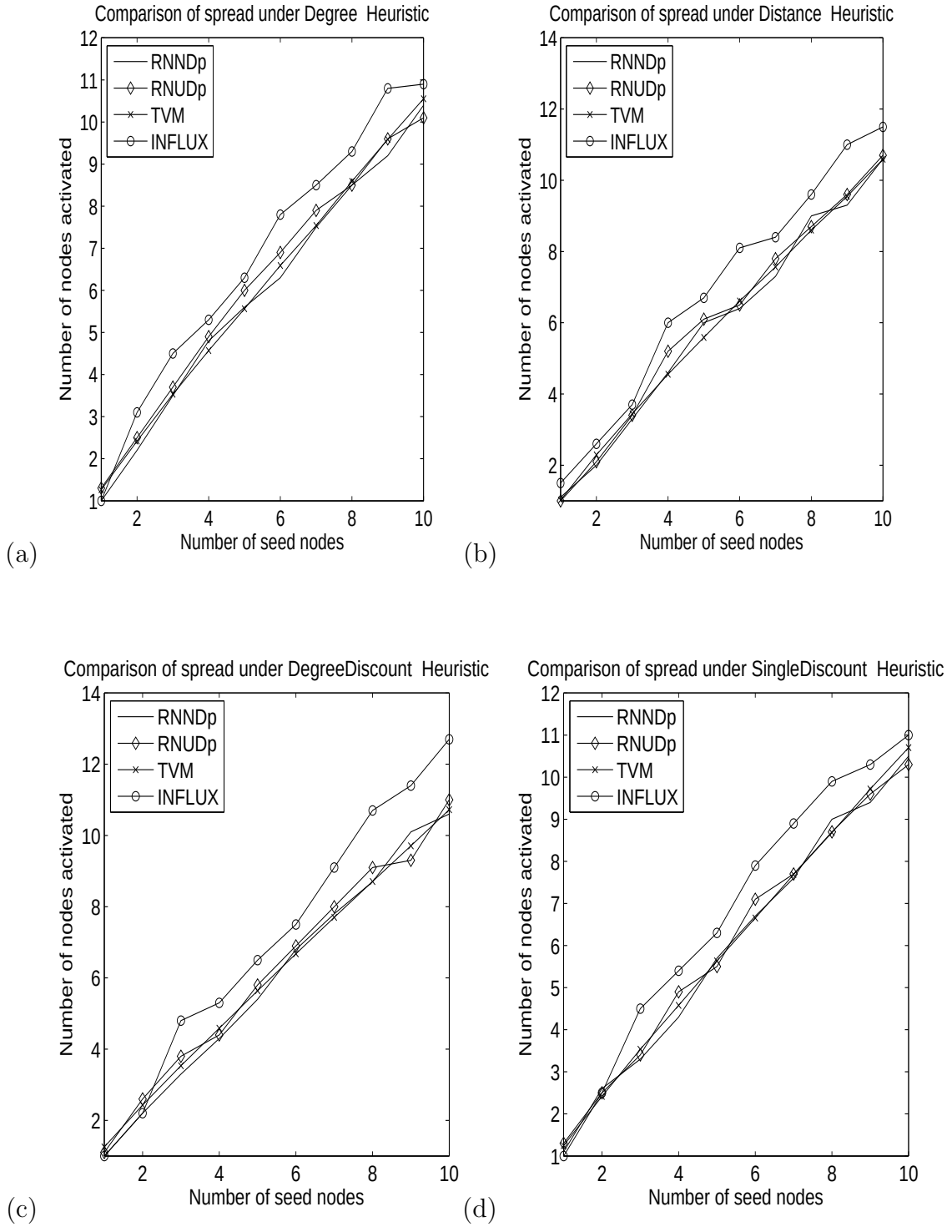


Figure 7.10: Comparison of Influx-IC to other models in Twitter for (a)Degree (b) Distance (c) DegreeDiscount and (d) SingleDiscount heuristics

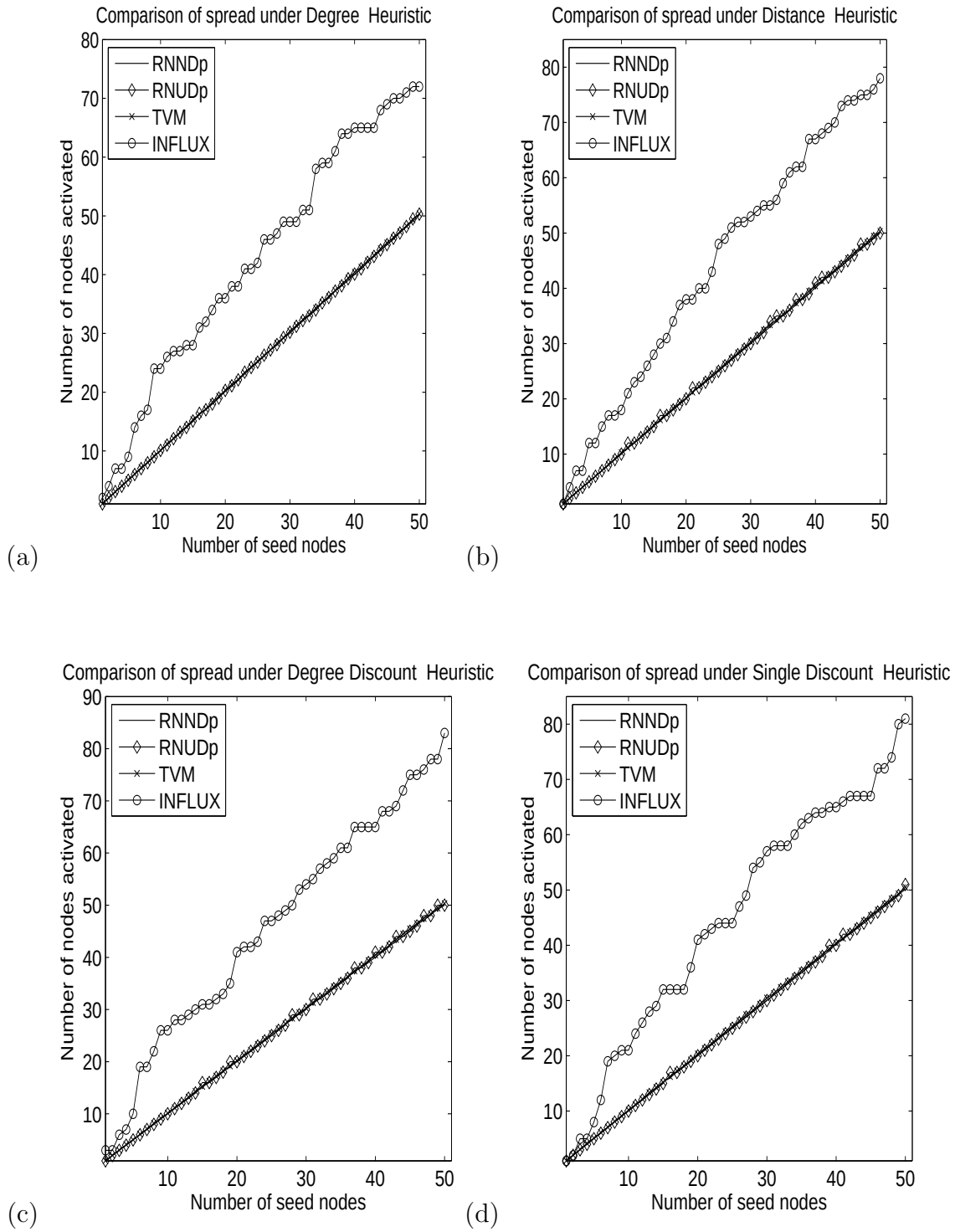


Figure 7.11: Comparison of Influx-IC to other models in Email for (a)Degree (b) Distance (c) DegreeDiscount and (d) SingleDiscount heuristics

Table 7.6: Twitter- Influx-IC Performance gain(in %)

-	Degree	Distance	SingleDiscount	DegreeDiscount
RNNDp	4.6	7.8	4.5	16.5
RNUDp	7.3	6.95	6.36	13.38
TVM	3.7	7.9	2.7	15.6

Table 7.7: Email- Influx-IC Performance gain(in %)

-	Degree	Distance	SingleDiscount	DegreeDiscount
RNNDp	30.4	35.64	38.148	39.75
RNUDp	30.1	35.89	37.03	39.6
TVM	30.05	35.5	37.67	39.39

There is an increase of up to 40.6% information spread when *Influx* is employed (as in Table 7.3) as compared to RNNDp, RNUDp and TVM, for each of the chosen standard heuristics. It is clear that the *Influx* approach yields better outcome since it reflects the interactive nature of the users. When the social network has highly interactive users, the influence they exert on each other is also high. Therefore, high influence has resulted in vast spread of information. In contrast, the existing approach of using the TVM which includes a constant value to represent influence, although the network had high levels of interactions. Due to this, TVM failed to reflect the dynamism of spread.

7.5.2 Efficacy of Influx-IC in dynamic scenario

In this section the effectiveness of the new approach in the case of dynamic changing interaction intensities among users is investigated. Most of the work predicting information spread, often considers the social network to be static. Therefore, the changes occurring in the network is not reflected in the outcome. However, in real world, users vary their interaction intensities among friends. At some point of time two connected users may be interacting more often and at some later point of time there may be few interactions. When the number of interaction are few over a certain time space, the influence also decreases. Also, when interactions increase, the influence increases. This scenario is shown in Figure 7.12. The time is divided into three slots T1, T2 and T3. In each of these time slots, the interactions of few of the connected pair of users vary. The impact of this scenario is investigated in this section.

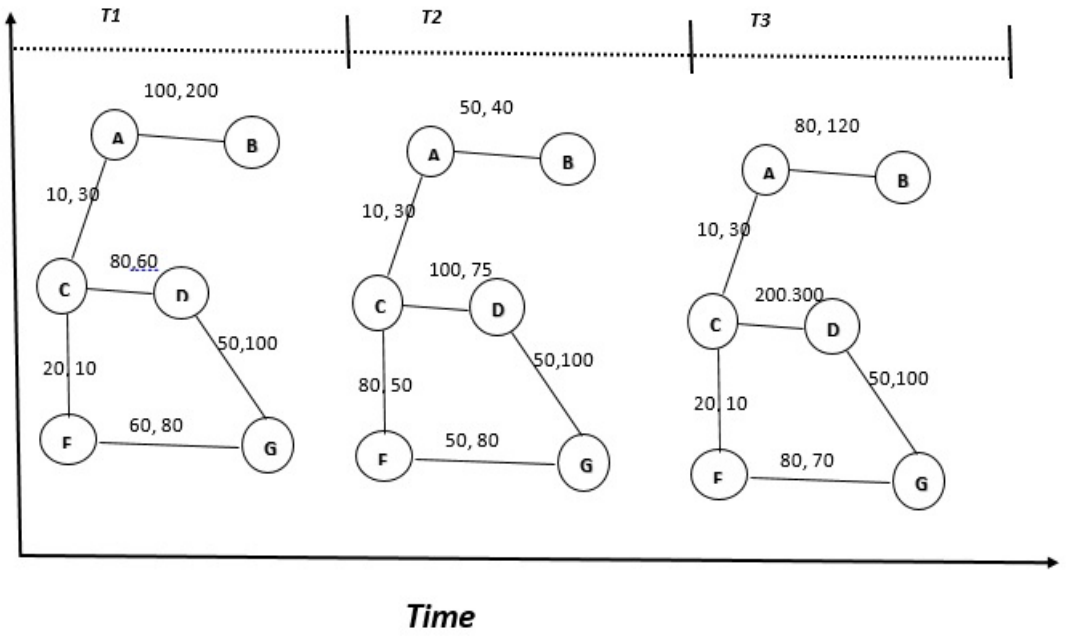
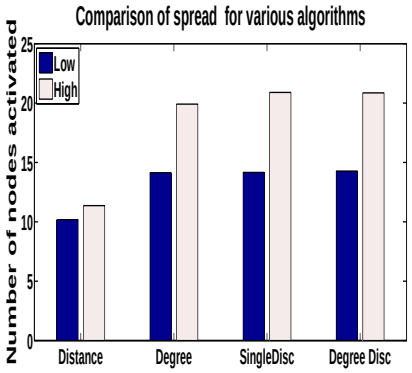


Figure 7.12: Changing interaction rates among the contacts

7.5.2.1 Results and discussions

To this end, the HEP and PHY datasets are analyzed for different activity logs representing small and large number of interactions among the users. The *Influx-IC* model runs 1000 iterations and 10 initial seeds are chosen. The results are as shown in figure 7.13.

(a)



(b)

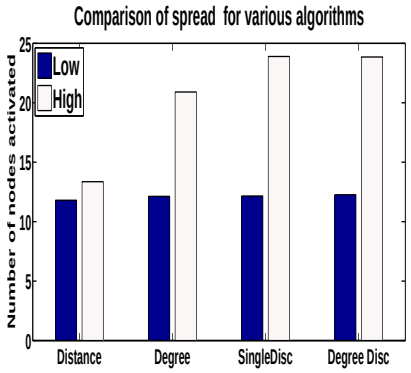


Figure 7.13: Efficacy of Influx-IC in (a)HEP and(b) PHY dataset

It is conclusive from the results that the increased interactions among the connected users, help in the wider spread of information in the network. The implication of using the Influx approach in predicting the spread in social networks is direct. When a constant uniform probability of influence is used, the standard algorithms such as degree, distance, degreediscount, singlediscount etc., fetches the same outcome irrespective of the changes within the social network. Hence, these algorithms show no response to the interaction dynamics among the users in the network. Thus, for these algorithms results are unaffected by the increase or the decrease in the interactions of the users in the network. The worst effect of such an approach would be to have an expected result at hand, even in the case when there is no interactions among the users. However the *Influx-IC* model, on the contrary, reflects the changes within the social network. Thus the new model predicts information spread better.

7.6 Summary

In this chapter, the solution to the effective information spread is presented by estimating the user influence. For this purpose, a novel approach named *Influx* is designed. The role of a realistic value of user influence is emphasized in the research work. The results show that using novel *Influx-IC* model based on the Influx approach, predicts higher information spread when compared to other models such as: TVM, RNNDp-IC and RNUDp-IC. The *Influx-IC* model also responds to the changes within the social network such as increased or decreased activities between the pair of users. In the dynamically changing scenario of addition and deduction of friends as well as the interactions among them within the social network, the traditional approaches and models are silent, where as the *Influx-IC* responds to these changes.

The next chapter discusses a new centrality approach to evaluate users. Further, two new heuristics are proposed to fetch top influential users.

Chapter 8

Finding the top influential users

Truth is ever to be found in the simplicity, and not in the multiplicity and confusion of things. -Isaac Newton

This chapter addresses the problem of finding the influential users in the social networks. A novel centrality measure is introduced here. This is further used to rank users and fetch top influential users. This chapter presents the methodology and the outcomes to substantiate the new approach.

8.1 Background

The runtime concerns with the Greedy approach, paved the way for various alternatives such as: CELF (Leskovec et al., 2007b), CELF++ (Goyal et al., 2011b), Mix Greedy and New Greedy (Chen et al., 2009), which aimed to reduce the execution time. However, several improvements to the original Greedy approach that has been proposed for influence maximization, which were developed to reduce runtime, have not yet been proved efficient. Therefore, there has been a need to look for heuristics to tackle the efficiency issue in influence maximization. In this context, various heuristics have been proposed to reduce the run time of influence maximization and to fetch optimal seed set. Out of these heuristics, degree centrality heuristic is proved to be efficient and close to optimal solution (Chen et al., 2010a).

The degree concept fetches users with the highest degree (a.k.a. contacts), with the belief that, such users will trigger a vast outbreak of information.

However, in real world, a user will interact only with a small percentage of his/her contacts, raising suspicion on the viability of degree heuristic. Also, various variations of degree heuristic raise similar concerns.

8.2 Problem description

The aim of the work, discussed in this section, is to fetch the set of top influential users in the social network such that they are capable of influencing their peers to adopt and spread the information further. This is formulated as problem 8.2.1

Problem 8.2.1. *Given a directed weighted social graph $G(V, E)$, a constant k, fetch the seed set I, where $|I| = k$, that maximizes the spread $\sigma(I)$ in the social network.*

8.3 Proposed methodology

In this section a new centrality approach namely the *Outdegree Rank* is proposed. Further, two new heuristics to fetch the top influential users are discussed.

8.3.1 Outdegree rank centrality

To make the degree concept viable in real world, the *Outdegree Rank* centrality is proposed for evaluating users in the social networks. Unlike existing works, *Outdegree Rank* heuristic considers the user attribute i.e., the interaction count, for fetching top influential users. Before getting into the details of *Outdegree Rank*, the following two terms are mentioned again.

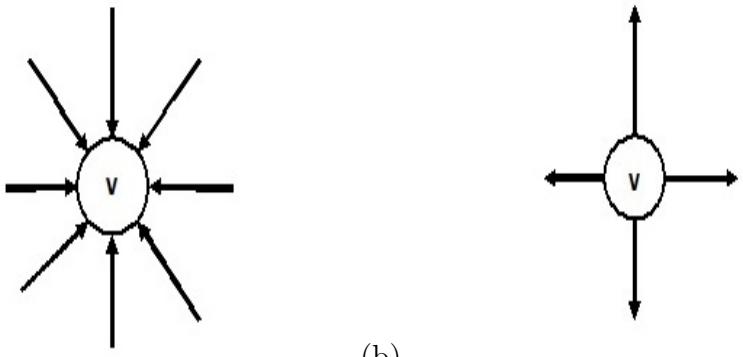
Contact degree: In a social network graph, for a node v, its contact degree is referred to as the number of edges incident on it and is denoted as **Cd(v)**.

Interaction degree: In an interaction graph, for a node v, its interaction degree is the number of edges incident on it and is denoted as **Id(v)**.

A user may have a large number of contacts (Cd(v)), but may have interactions with only few of them. In such a scenario, we have $Id(v) \ll Cd(v)$.

Therefore, in this case, the degree centrality approach which uses *Cd(v)* to fetch influential users, may not fetch optimal seed set.

The *Outdegree* centrality is developed on the interaction graph of a social network. The *Id(v)* of a node in the graph is further specified in terms of *Indegree* and *Outdegree* centrality. The *indegree(v)* represents the popularity index and *outdegree(v)* represents the participation index. The concept of indegree is explored in the popular PageRank heuristic for finding popular web pages (Page et al., 1999) and identifying key users in social networks (Heidemann et al., 2010). Figure 8.1 shows the *indegree(v)* and *outdegree(v)* of node *v*.



(a) (b)

Figure 8.1: (a) Indegree and (b) Outdegree of a node v in the interaction graph

The various combinations of indegree and outdegree with comparison to Cd(v) is considered as a cartesian product of

$$\{Indegree(v), Outdegree(v)\} \times \{\ll, \rightarrow\} \times \{Cd(v)\}.$$

At this point, the four cases with reference to the *Outdegree* and *Indegree* of node v are as follows:

i. $Outdegree(v) \ll Cd(v) \Rightarrow node(v)$ is less interactive.
ii. $Outdegree(v) \rightarrow Cd(v) \Rightarrow node(v)$ is highly interactive.
iii. $Indegree(v) \rightarrow Cd(v) \Rightarrow node(v)$ is highly popular.
iv. $Outdegree(v) \rightarrow Cd(v)$ & $Indegree(v) \rightarrow Cd(v) \Rightarrow node(v)$ is highly interactive as well as popular.

In the research work, the *Outdegree(v)* centrality is further explored to find solution to influence maximization.

8.3.2 Outdegree Rank heuristic for fetching influential users

Among the cases discussed above, case(ii) is further discussed. For any node v, if *Outdegree(v)* $\rightarrow$ *Cd(v)*, it shows that node v maintains communications and interaction with almost all of his/her contacts. Such a node has more potential to spread information in the network, when compared to other nodes. Based on this surmise, the *Outdegree(v)* is used to rank the nodes in the network. This is the *Outdegree Rank (OR)* heuristic. Algorithm 4 details steps to fetch the seed set I.

Input: Directed graph $G_I(V_I, E_I)$ and k

Output: Seed set I of k users

1 Initialize $I = \phi$;

2 Compute the outdegree out_v for each vertex $v \in V_I$

3 Repeat k times steps 4 and 5

4 $u = argmax_v\{out_v | v \in V_I - I\}$

5 I= $I \cup u$

6 Output I

Algorithm 4: Outdegree Rank to fetch seed set I

The input to algorithm 4 is the interaction graph, $G_I(V_I, E_I)$ and a constant k. In algorithm 4, step 2 computes the outdegree of each node v. Steps 4 and 5, picks k nodes with highest outdegree and adds them to the set I.

Complexity: When the number of nodes in G_I is $\bar{m}$, step 2 takes $O(\bar{m})$. This is followed by the k times execution of steps 4 and 5. Thus, the complexity of algorithm 4 is $O(\bar{m} + k)$.

8.3.3 Applying the discount concept on OutDegree Rank

The discount concept proposed by Chen et al. (2009) on the degree centrality is also applicable to *Outdegree Rank* centrality. Let v be a neighbor node of u. If u is already in the seed set, then while considering the inclusion of node v into seed set on its *Outdegree Rank*, the edge $e(u, v)$ towards its degree should not be counted. Thus v's degree is discounted by the count of all its neighbors which are already in the seed set. With reference to the case presented by Chen et al. (2009), for a vertex v with η_v neighbors already in the seed set, $v's$ degree is discounted as in Eq(8.3.1). This heuristic is named as *Outdegree Rank Discount*.

$$2\eta_v - (Cd_v - \eta_v)\eta_v p \tag{8.3.1}$$

where Cd_v is the degree of the node v.

Algorithm 5 details steps in Outdegree Rank Discount heuristic. The input to Algorithm 5 are size of seed set k, the contributor graph i.e. $G_c(V_c, E_c)$ and the interaction graph of the same, represented as $G_I(V_I, E_I)$. Step 2 computes the outdegree i.e. out_v, from the interaction graph and degree of the node. The top k influential nodes are fetched from steps 5 to 10. In steps 8 and 9, the degree count of selected influential nodes are discounted as in Eq 8.3.1.

Input: $G_c(V_c, E_c)$, $G_I(V_I, E_I)$ and k

Output: Seed set I of k users

1 Initialize $I = \phi$;

2 For each vertex $v \in V_I$, compute the outdegree out_v, degree Cd_v and initialize $t_v = 0$

3 Repeat k times steps 4 to 6

4 $u = argmax_v\{out_v | v \in V_c - S\}$

5 I= $I \cup u$

6 Repeat steps 7 and 8 for each neighbour v of u and $v \in V_c - I$

7 $t_v = t_v + 1$

8 $Cd_v = Cd_v - 2t_v - (Cd_v - t_v)t_v p)$

9 Output I

Algorithm 5: Outdegree Rank Discount to fetch seed set I

Complexity of algorithm 5:The step 2 computes the outdegree and degree of each node. For an interaction graph of $\bar{m}$ nodes, it takes $O(\bar{m})$. The top k influential nodes are fetched from steps 4 to 9. In steps 7 and 8, the degree count of selected influential nodes are discounted as in Eq(8.3.1). Thus, the complexity of algorithm 5 is $O(\bar{m} + k.\bar{m})$.

Unlike the Degree heuristic, in which the $Cd(v)$ is almost static, the $Outdegree(v)$ is frequently varying according to the changes in the interaction rate of the user. In this way, the OutDegree Rank reflects the dynamic changes occurring in the social network. Thus, Outdegree Rank is a viable solution in the real world.

8.4 Results and analyses

In this section the performance gain of the *Outdegree Rank* and the *Outdegree Rank Discount* heuristics in the *Influx-IC* model are highlighted. The performance of these heuristics are evaluated for two scenarios. In the first scenario the new heuristics are evaluated in *Influx-IC* model and other heuristics are evaluated

in the original setup in IC model. In the second scenario all the heuristics are employed in the *Influx-IC* model. The highestdegree (a.k.a degree), distance, singlediscount and degreediscount heuristics are used to fetch the influential users. These are used to predict the information diffusion and their outcomes compared to the proposed approaches.

8.4.1 Performance of the new heuristics

This section discusses the first scenario where on one hand the standard heuristics, i.e., highestDegree, distance, degreediscount and singlediscount, are employed in the independent cascade model which uses an uniform influence value to estimate information diffusion. On the other hand, the Outdegree Rank as well as, the Outdegree Rank Discount heuristics are employed in the *Influx-IC* model which uses estimated influence to predict the information diffusion. Therefore, the heuristics are designated as Outdegree Rank with Influence Estimate (*ORIE*) and to Outdegree Rank with Influence Estimate- Discount (*ORIE-Discount*). The outcome of this first case is shown in figure 8.2 to figure 8.8. The performance gain of ORIE and ORIE-Discount, when compared to other state-of-the-art approaches, is shown in Table 8.1.

Table 8.1: ORIE and ORIE Discount Performance gain(in %)

	HighestDegree		SingleDiscount		DegreeDiscount		Distance	
Compared to →	ORIE	ORIE Disc	ORIE	ORIE Disc	ORIE	ORIE Disc	ORIE	ORIE Disc
HEP	40.73	40.78	39.47	39.87	39.2	39.6	40.8	41.25
PHY	32.6	35.37	21.9	25.13	20.5	23.71	31.47	34.27
YouTube	31.7	32.36	31.2	31.9	30.8	31.5	33.1	33.72
Infectious	56.3	57	56.1	56.86	55.9	56.7	56.4	57.14
Twitter	22.2	24.1	20.7	23	20.5	22.87	22	24.5
Email	40.04	41.41	39.9	41.3	39.6	40.9	40.02	41.37
Wikivote	33.68	29.5	33.2	28.9	33.46	29.2	33.8	29.6

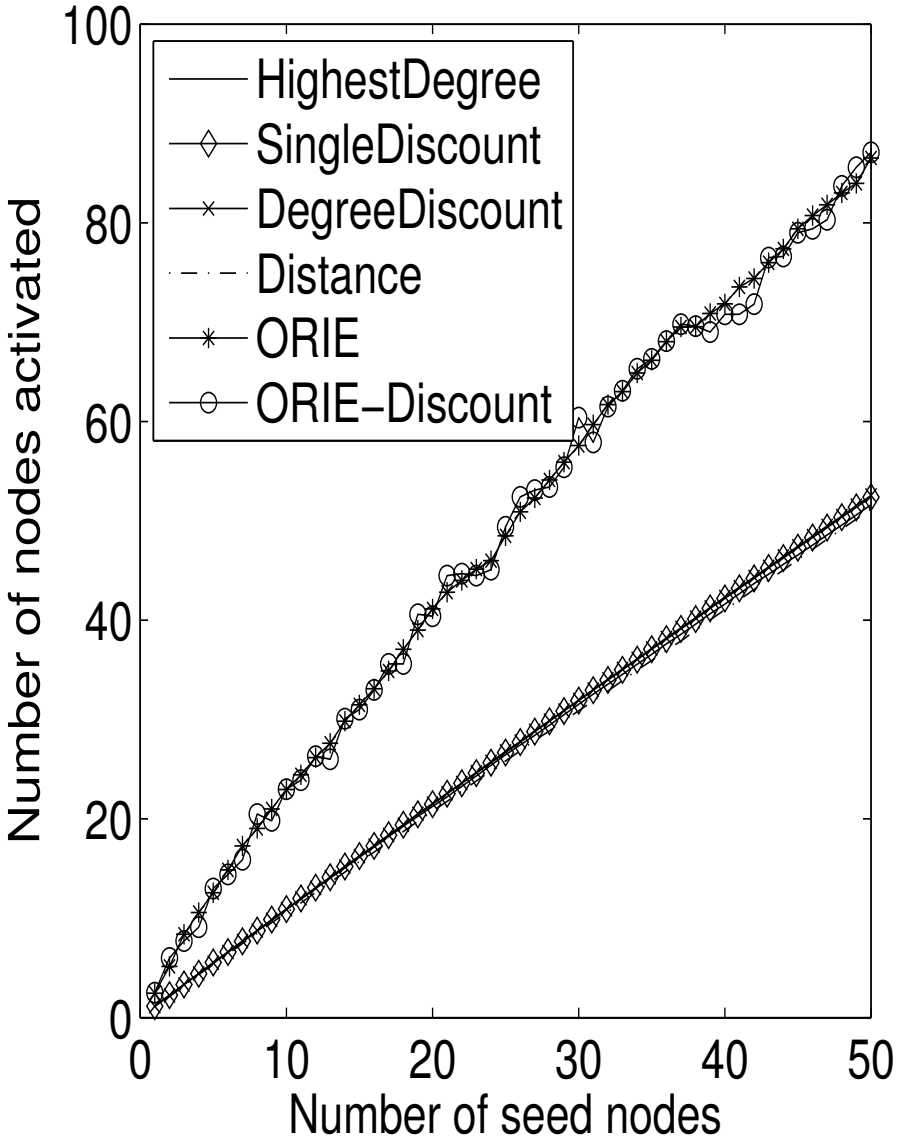


Figure 8.2: Performance of ORIE and ORIE-Discount in HEP dataset

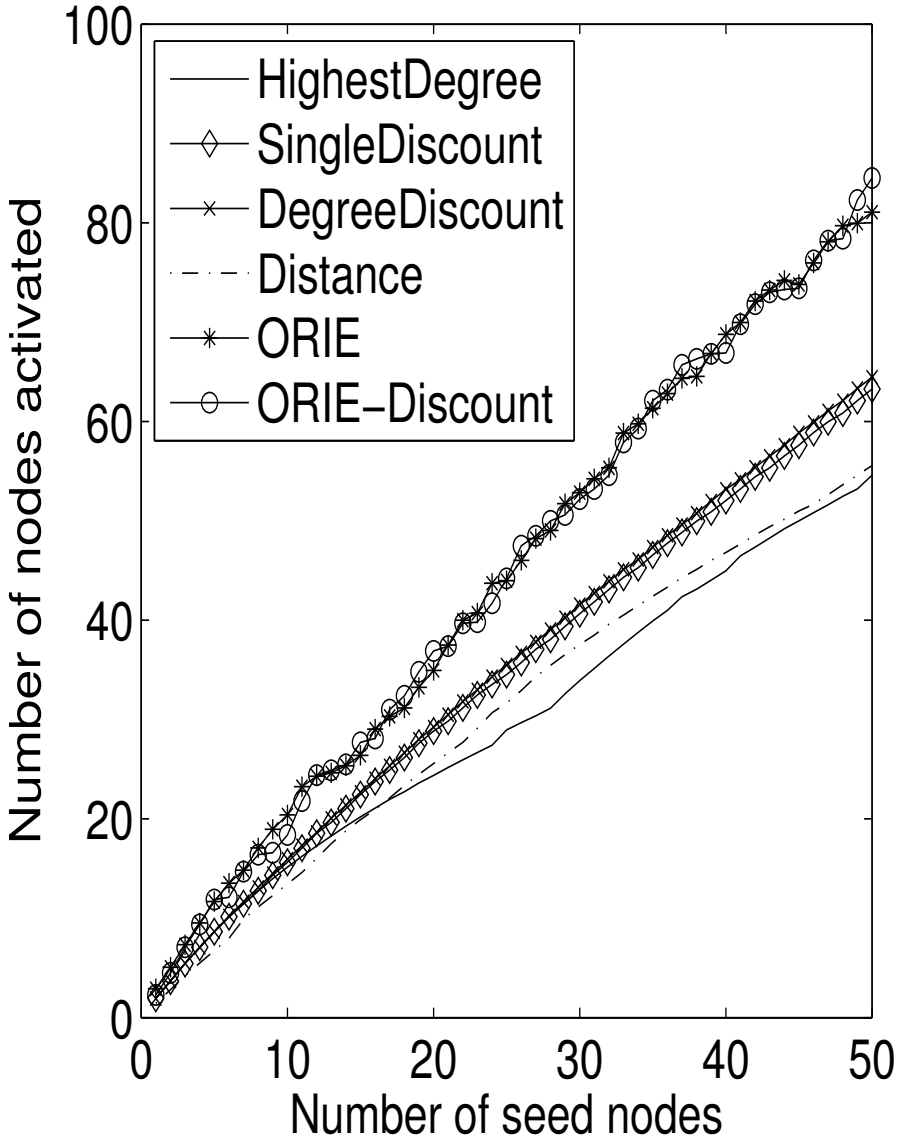


Figure 8.3: Performance of ORIE and ORIE-Discount in PHY dataset

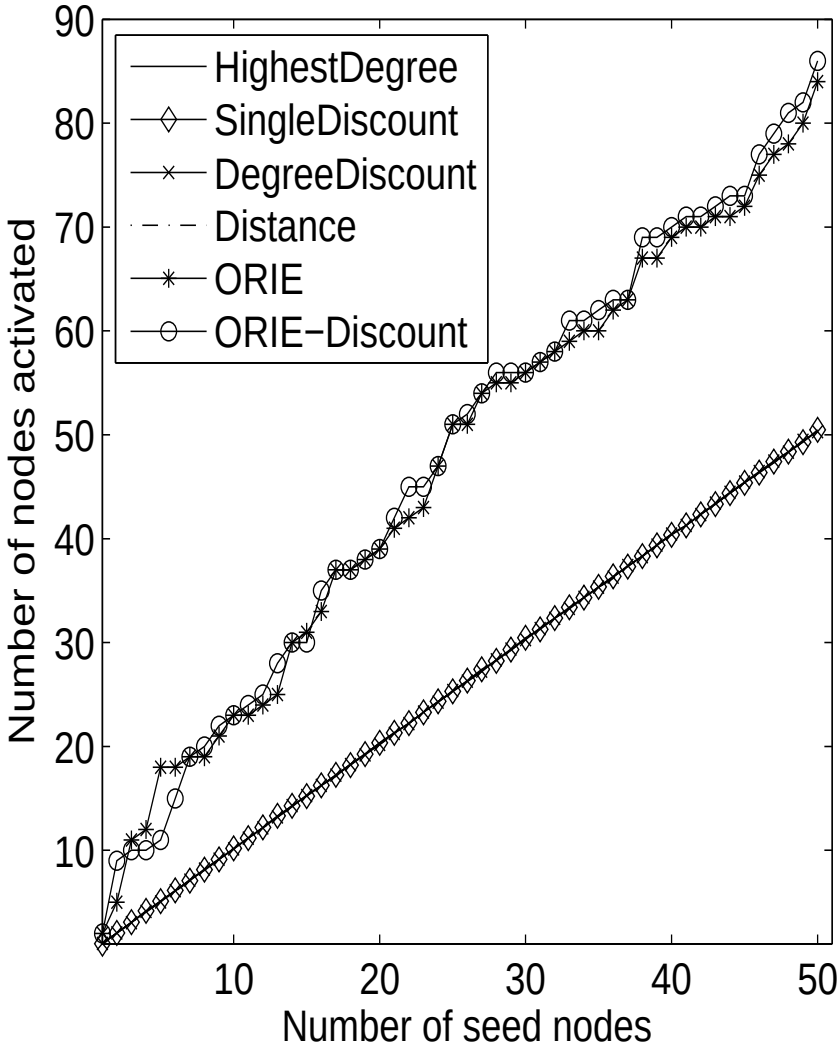


Figure 8.4: Performance of ORIE and ORIE-Discount in Email dataset

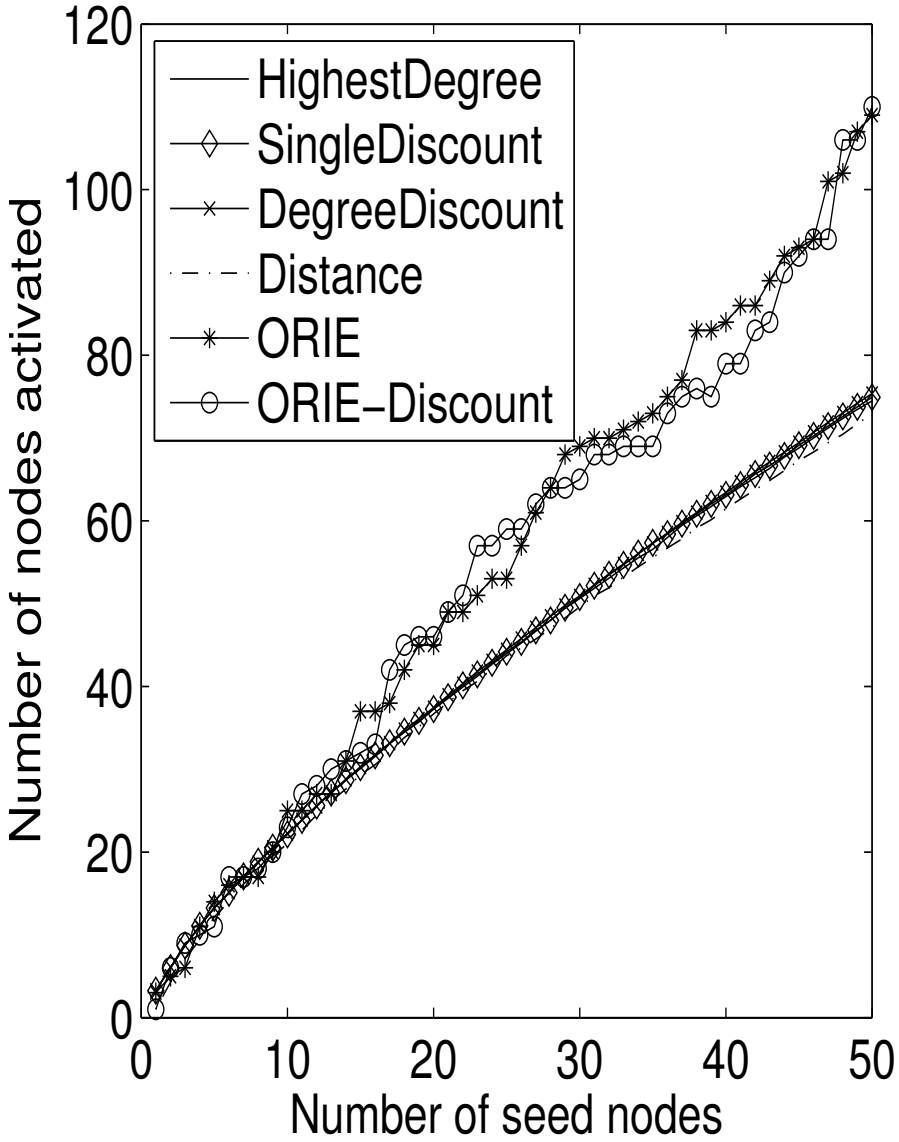


Figure 8.5: Performance of ORIE and ORIE-Discount in YouTube dataset

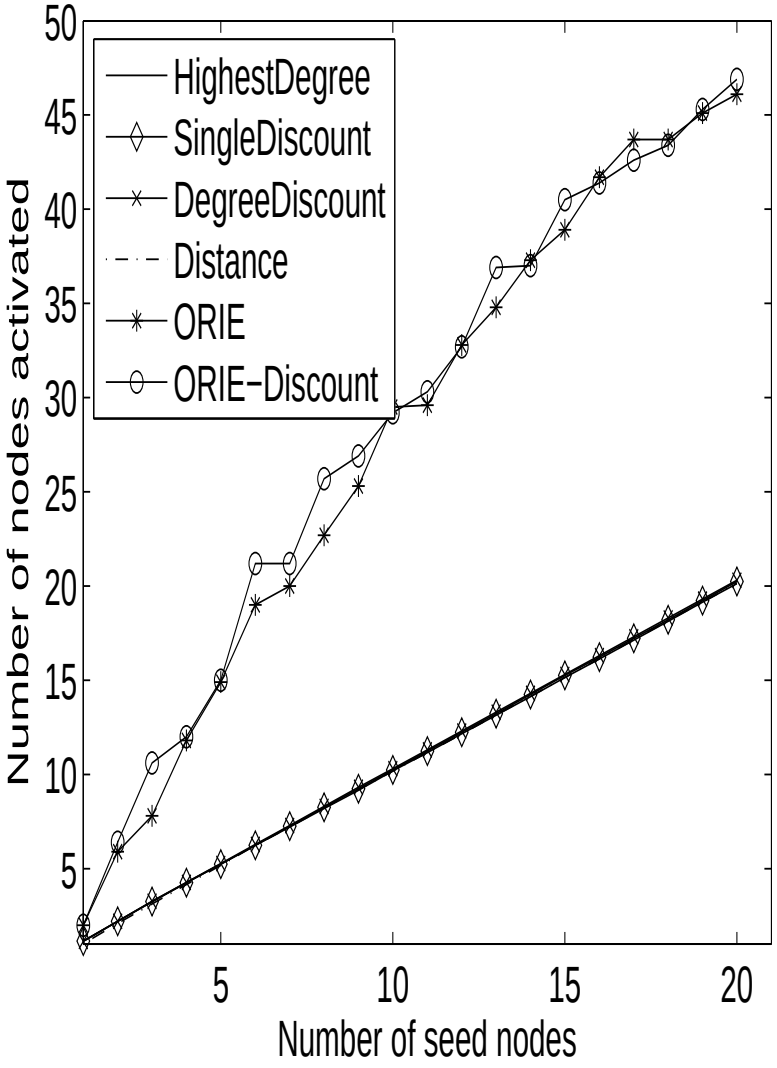


Figure 8.6: Performance of ORIE and ORIE-Discount in Infectious dataset

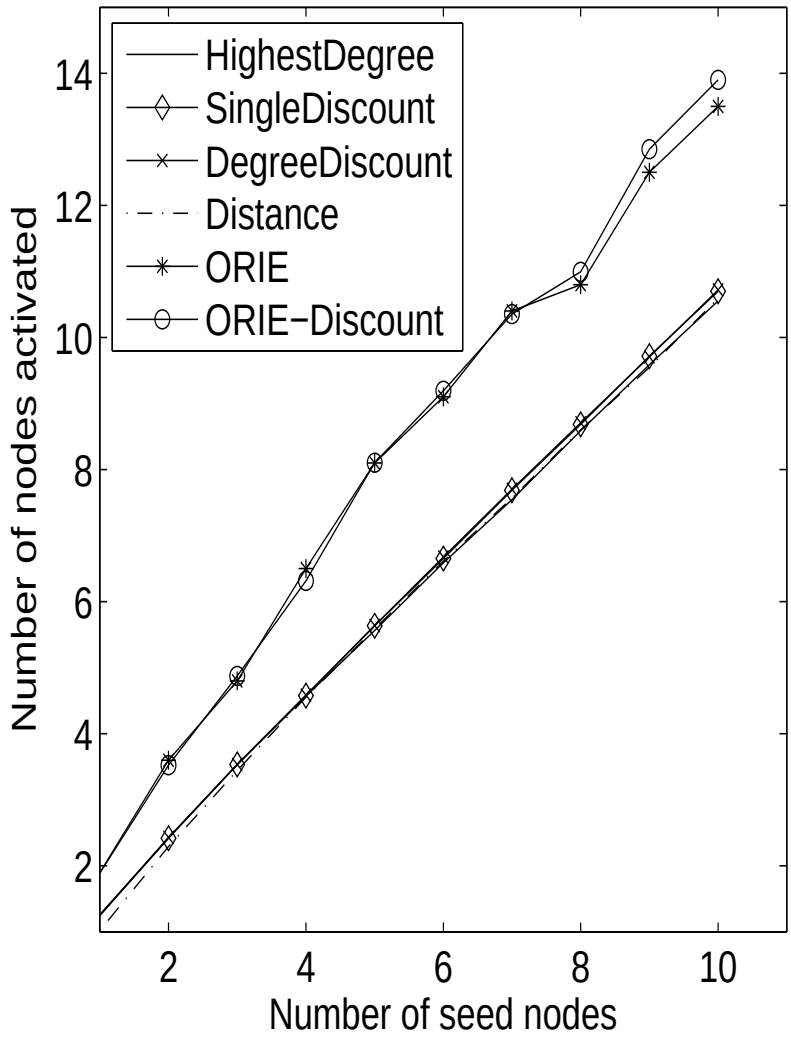


Figure 8.7: Performance of ORIE and ORIE-Discount in Twitter dataset

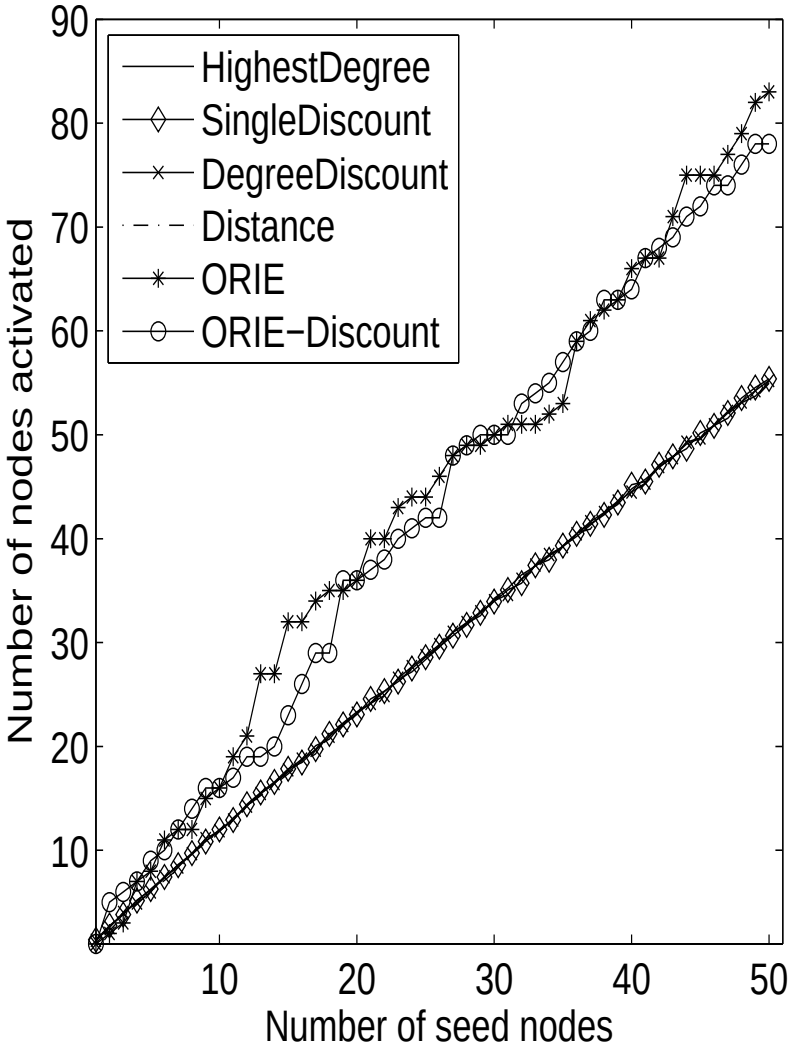


Figure 8.8: Performance of ORIE and ORIE-Discount in Wikivote dataset

With ORIE and ORIE-Discount, there is a gain in the prediction of the spread of information in the range of 20.5% - 57%, when compared to other state-of-the-art approaches. Specifically, in HEP there is up to 41.25% increase in information diffusion.In PHY, there is 35% increase, in YouTube up to 33%, in Infectious 57%, in Twitter up to 24%, in Email up to 41% and in Wikivote up to 33% increase in information spread when compared to the proposed approach is seen. This gain in the spread is possible, since the *Influx-IC* model is based on the approach that estimated user influence from the user interaction. This outcome is primarily for the reason that both, the heuristics, as well as, the influence estimates approach are designed on the interaction rate of the users. The use of the contributor graph ensured that only highly interactive users were available for the application. These highly interactive users resulted in higher influence among them, which finally led to the vast outcome of spread. Conclusively, the ORIE and ORIE-Discount outperforms standard approaches that are based on the degree centrality.

8.4.2 State of Art heuristics in Influx-IC

In the latter case, Influx is employed to standard heuristics. To distinguish these heuristics from their original setup, they are named as HighestDegree-IE, DegreeDiscount-IE, SingleDiscount-IE, Distance-IE. The predicted information spread under these heuristics are compared to ORIE and ORIE Discount heuristics. The outcomes are as shown in Figure 8.9 to Figure 8.15.

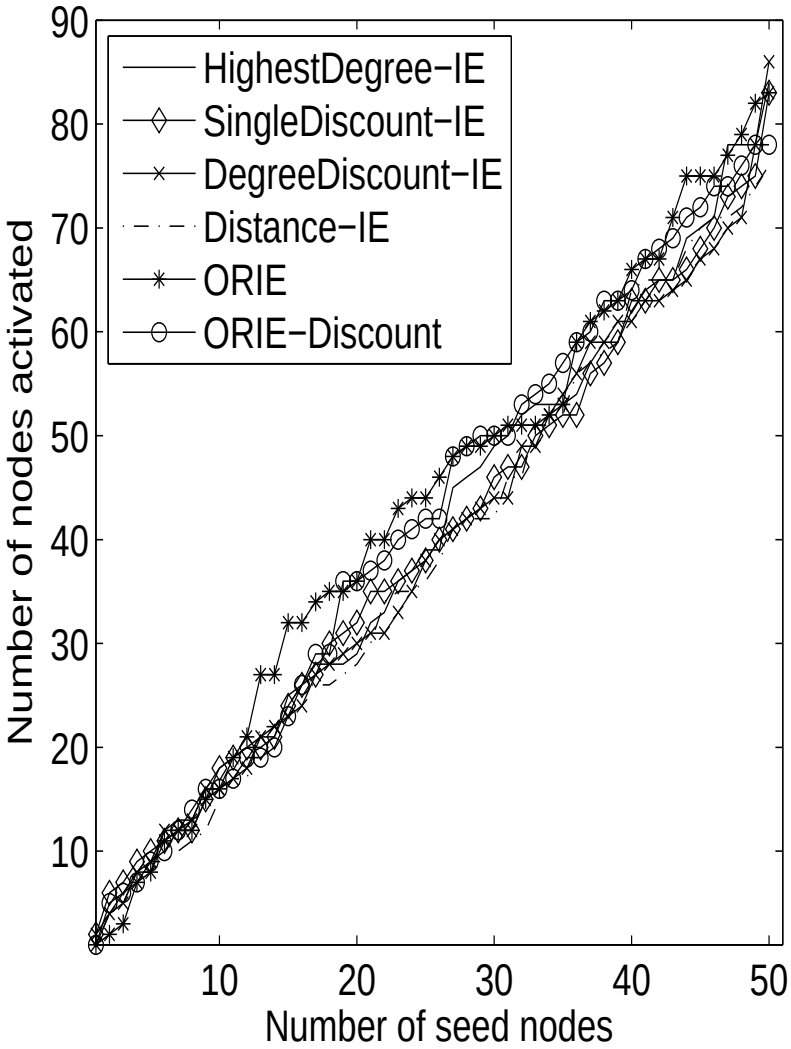


Figure 8.9: Influx employed to various heuristics on Wikivote dataset

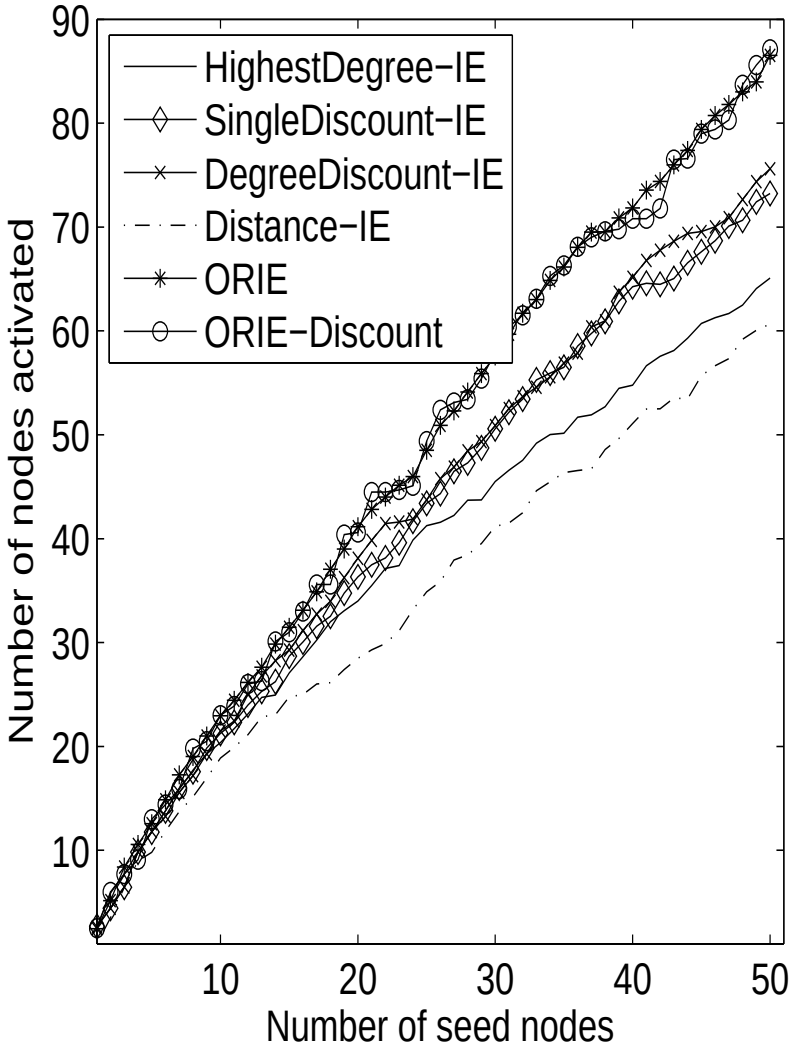


Figure 8.10: Influx employed to various heuristics on HEP dataset

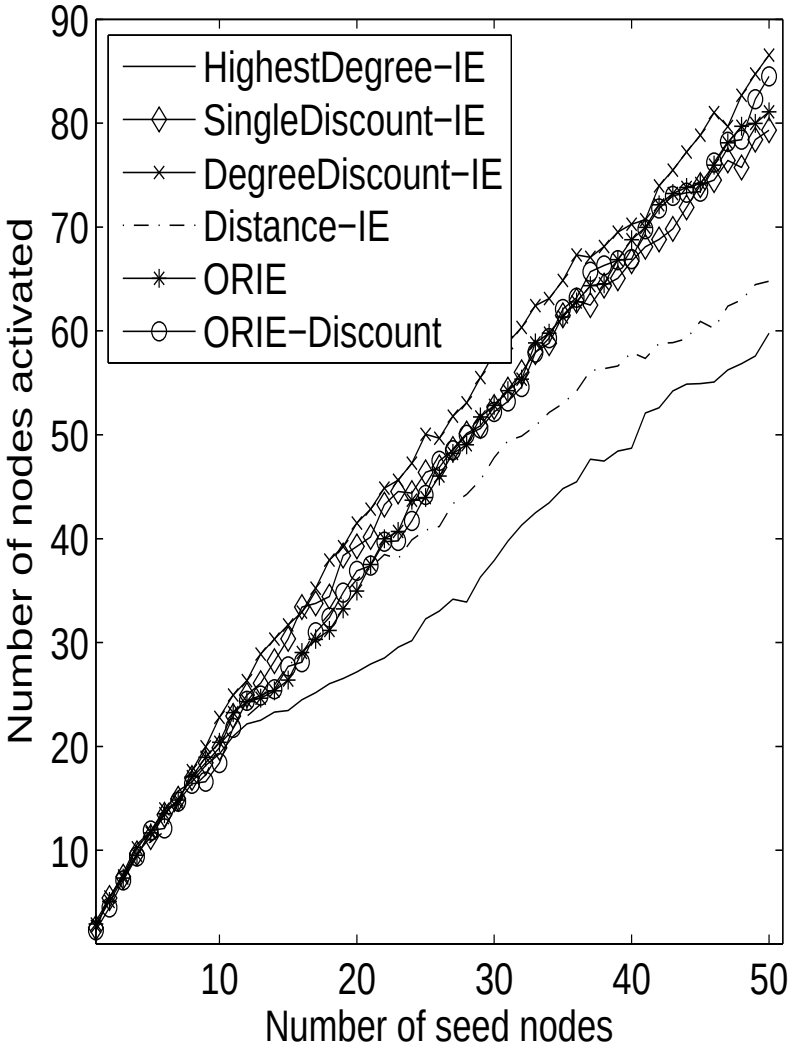


Figure 8.11: Influx employed to various heuristics on PHY dataset

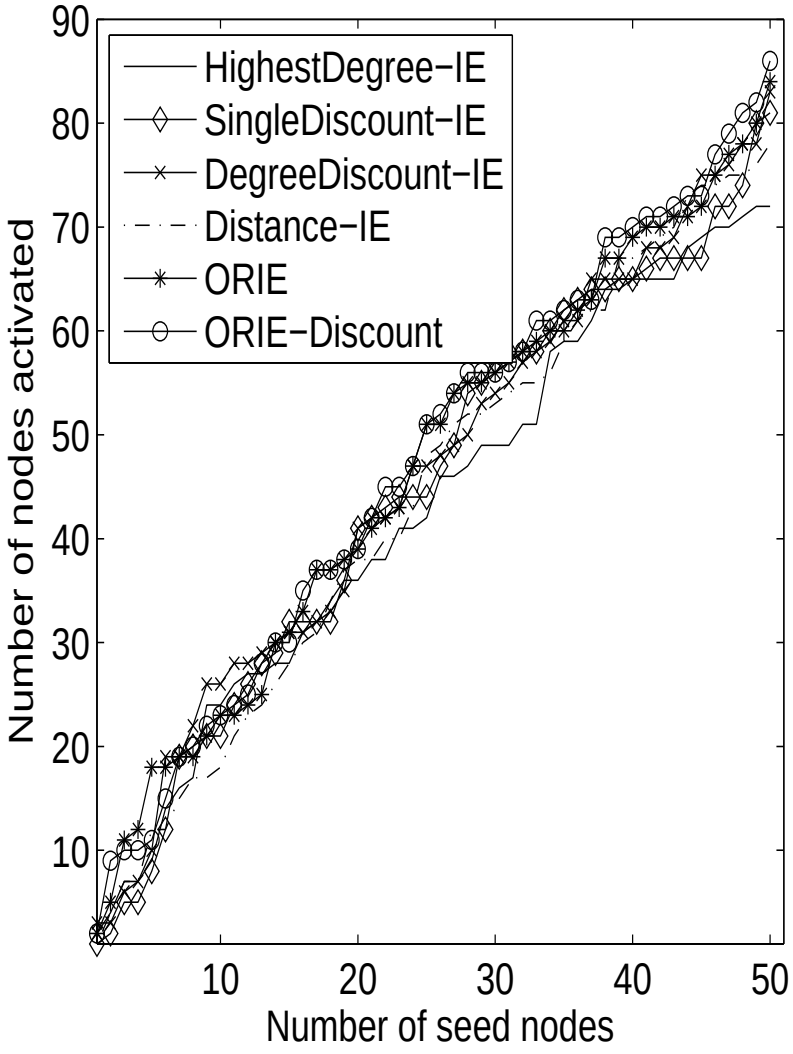


Figure 8.12: Influx employed to various heuristics on Email dataset

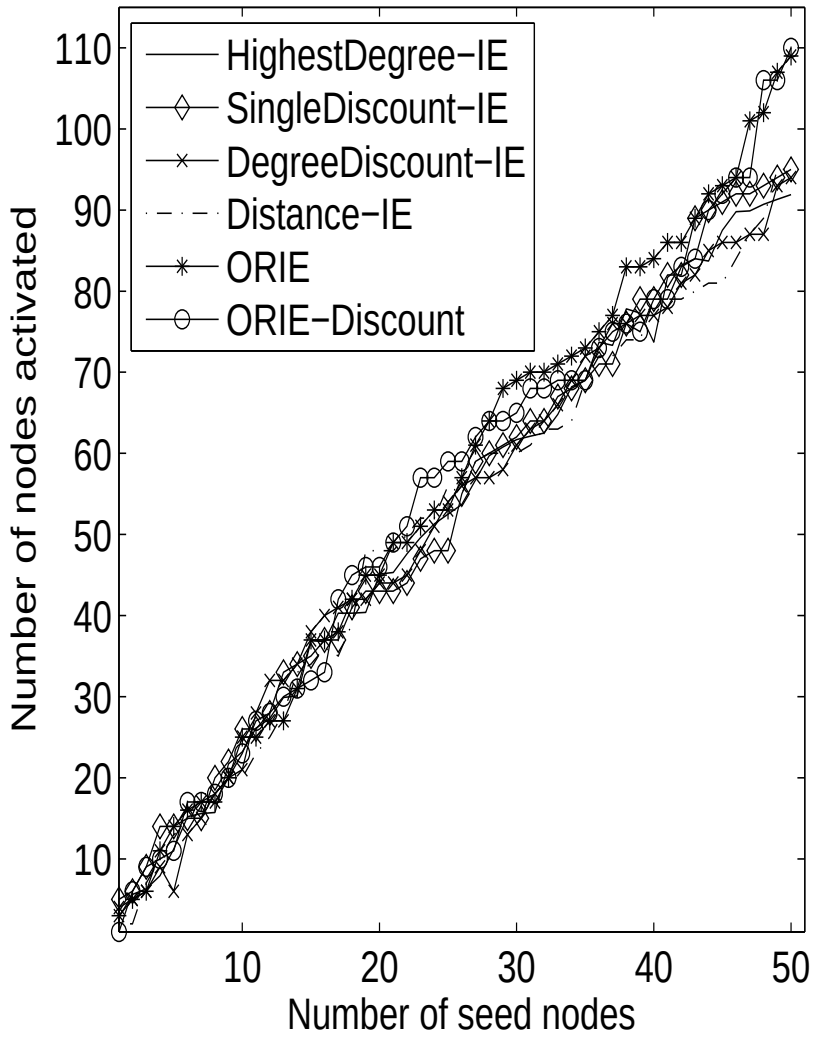


Figure 8.13: Influx employed to various heuristics on YouTube dataset

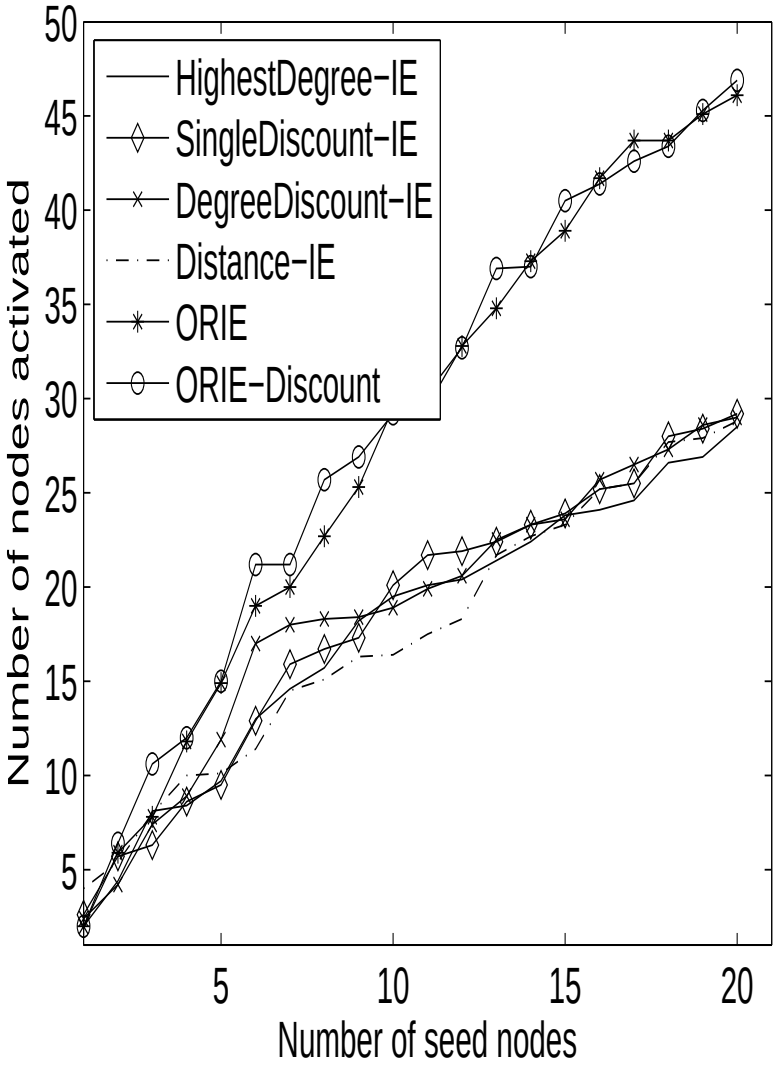


Figure 8.14: Influx employed to various heuristics on Infectious dataset

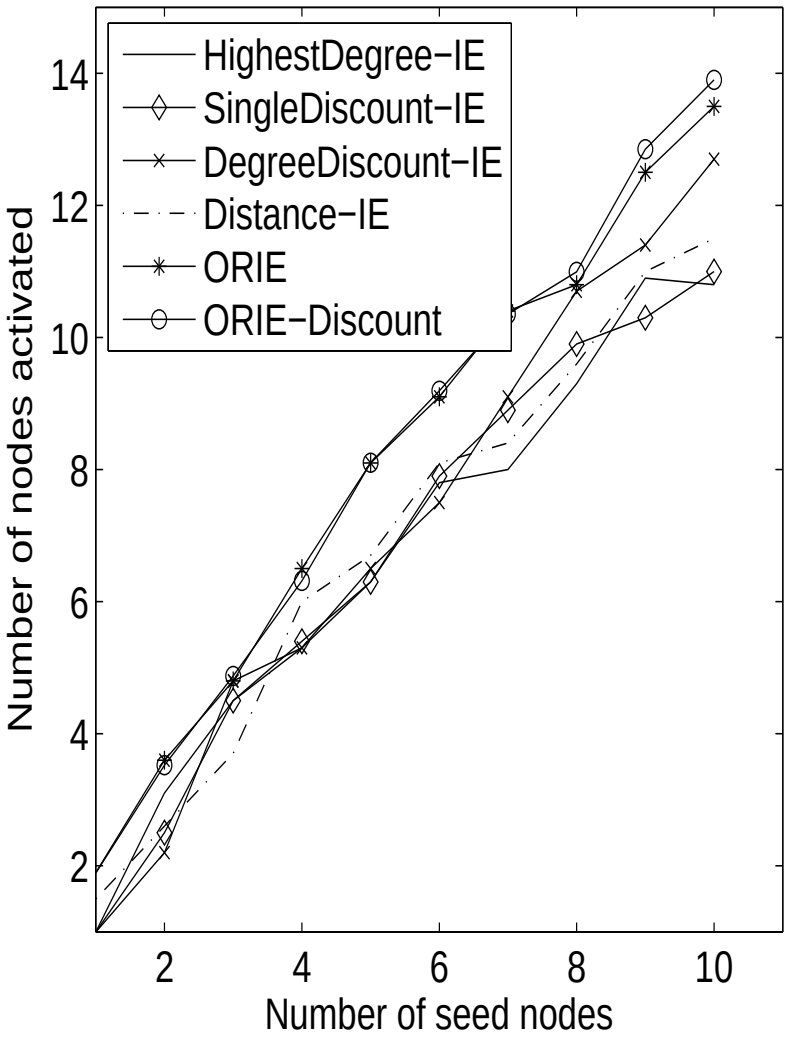


Figure 8.15: Influx employed to various heuristics on Twitter dataset

The performance gain of ORIE and ORIE-Discount when compared to the chosen heuristics is shown in Table 8.2.

Table 8.2: Standard heuristics in Influx-IC model

	HighestDegree-IE		SingleDiscount-IE		DegreeDiscount-IE		Distance-IE	
Compared to →	ORIE	ORIE Disc	ORIE	ORIE Disc	ORIE	ORIE Disc	ORIE	ORIE Disc
HEP	24.4	25.62	14.9	16.33	12	13.6	29.5	30.68
PHY	26	29.25	2.5	6.13	-1.7	1.12	21	23.37
YouTube	15.6	16.45	12.8	13.63	13.76	14.5	12.8	13.63
Infectious	38.17	39.23	36.65	37.73	37.09	38.16	37.5	38.6
Twitter	20	22.3	18.5	21.4	6	8.6	14.8	17.2
Email	14.2	16.2	3.57	5.8	1.19	3.48	7.14	9.3
Wikivote	-1.2	-7	1	-6	-3	-10	8	2.5

There is an increase in information diffusion of 30% in HEP, up to 29% in PHY, up to16% in YouTube, up to 38.6% in Infectious, up to 22% in Twitter, up to 16% in Email and upto 8% in Wikivote. The information spread did not see gain in Wikivote dataset for degree, degree discount and single discount heuristics. However, the ORIE and the ORIE-Discount heuristics performed better than the Distance heuristic. Moreover, for other datasets, there is a gain in the range of 6%- 38.6% in the spread of information when ORIE and ORIE-Discount is employed. The degree centrality fetched the seed set, that only have high degree but are not interactive in the network. As a result, the influence they exerted on their peers was not high to propagate information. The outcome of Distance-IE heuristic emphasizes that, users who are in close proximity need not influence each other. The SingleDiscount-IE and DegreeDiscount-IE heuristics fetches seeds close to ORIE and ORIE-Discount, due to the fact that already those nodes induced in the seed set are discounted from the degree value. The result of PHY dataset shows a reduction in the spread. This is the case where the users with high degree maintain high levels of

interaction. Such a scenario in real world networks is rare, as observed by Jakob (2012). The outcome emphasizes that even in a scenario when standard heuristics employ *Influx* method, the proposed OR and OR Discount heuristics perform better. Thus, assuredly, Outdegree Rank and Outdegree Rank Discount are effective viable solutions and are contributions towards influence maximization.

8.5 Summary

In this chapter, two heuristics for the influence maximization problem are proposed. The aim of proposing the Outdegree Rank heuristic is to emphasize the role of users in information spread. The degree centrality, which is a popular metric in picking influential users will fail; adhering to the fact that users do not interact with all of their contacts. In such a scenario, the *Outdegree Rank* metric provides a realistic solution. This heuristic is extended to include the discount concept to design another heuristic named as Outdegree Rank Discount. Both these heuristics are employed in the *Influx-IC* model. The outcome shows that the new heuristics are able to spread information to a major portion of the network when compared to other approaches.

In this research work, the influence maximization problem is solved in three stages. In the first stage the network is pruned to pick probable candidates of information diffusion. In the second stage influence among the users is estimated with the aim that when the third stage fetches the influential users, it can be evaluated with a realistic influence value. Thus, in the third stage influential users are picked. At this stage, the diffusion model has information propagators and realistic influence value to predict the diffusion on the social networks that is initiated by the influential users. The solution proposed in this research is an amalgamation of three aspects; i.e. network pruning, estimating user influence and influential users, that efficiently predicted the diffusion.

Chapter 9

Conclusion and future work

I may not have gone where I intended to go, but I think I have ended up where I intended to be. -Douglas Adams.

With the increased popularity and pervasiveness of social networks, more and more applications are tailor-made for it. Predicting close to accurate spread of information is central to viral marketing applications. Making the solution to influence maximization viable in the real world is the aim of the research work. Influence maximization and information diffusion are like two faces of the same coin, and have to be studied in each other's context.

In this research, various contributions are made towards influence maximization. The research presents a new model to depict the role of users during the information propagation process. Along with this model, the solution to influence maximization is attempted in three phases. Although, various approaches are earlier proposed to solve influence maximization problem and for better information spread, there are still a few concerns to be addressed. These issues are discussed in this research and in that aspect a holistic approach is proposed.

This research work presents a better solution to influence maximization problem in three phases; 1) find the contributors, 2) estimate influence and 3) fetch initiators. The proposed approach is more scalable when compared to the existing approaches. This work finds contributors who are most likely to spread and also adopt the information in the network, thereby meeting the desired final

outcome. Influence estimation is largely being previously unexplored and is addressed in this research. A new diffusion model named as *Influx-IC* model is also designed to predict information diffusion. This research also designs new heuristics namely the Outdegree Rank and OutDegree Rank Discount. These heuristics give a new dimension to fetch initiators by considering the activity rate of users. Overall, the amalgamation of these phases makes the solution user centric. Thus, a new dimension to influence maximization problem is explored. The simulations support the claim and provide a new milestone to the solution to influence maximization problem.

In future, the aspects of network structure, user influence and influential users, can be dealt in detail by using more user data. When the constraints on privacy of user data for applications is flexible, more user features can be used in each of the phases to provide a better working model. Yet another direction is verifying the validity of the proposed approaches on other diffusion models. The temporal dynamics of the social networks is not explored in depth, which can be dealt in the future. A scenario where campaigns of similar products but from rival enterprises are pushed in a same social networks is not studied in this research. This case of multiple competitors can also be considered as an extension to this work. Negative influence (Jung et al., 2011) can inhibit the diffusion of information. In this work positive influence is considered. The presented model can be extended to reflect negative influence. Dynamic influence maximization (Pan et al., 2017) is the newest trend. The proposed model could be extended to include the temporal dynamics for a better perspective.

References

Agarwal, G. and Kempe, D. (2008). "Modularity-maximizing graph communities via mathematical programming." *The European Physical Journal B*, 66(3), 409–418.

AlFalahi, K., Atif, Y. and Abraham, A. (2014). "Models of Influence in Online Social Networks." *International Journal of Intelligent Systems*, 29(2), 161–183.

Alon, N., Gamzu, I. and Tennenholtz, M. (2012). "Optimizing budget allocation among channels and influencers." *Proceedings of the 21st international conference on World Wide Web*. ACM, 381–388.

Arenas, A., Duch, J., Fernández, A. and Gómez, S. (2007). "Size reduction of complex networks preserving modularity." *New Journal of Physics*, 9(6), 176.

Bakshy, E., Rosenn, I., Marlow, C. and Adamic, L. (2012). "The Role of Social Networks in Information Diffusion." *Proceedings of the 21st International Conference on World Wide Web*. ACM, New York, NY, USA, 519–528.

Belák, V., Mashhadi, A., Sala, A. and Morrison, D. (2016). "Phantom cascades: The effect of hidden nodes on information diffusion." *Computer Communications*, 73, 12–21.

Bhagat, S., Goyal, A. and Lakshmanan, L. V. (2012). "Maximizing product adoption in social networks." *Proceedings of the fifth ACM international conference on Web search and data mining*. 603–612.

Borgatti, S. P., Carley, K. M. and Krackhardt, D. (2006). "On the robustness of centrality measures under conditions of imperfect data." *Social Networks*, 28(2), 124 – 136.

Breiger, R., Boorman, S. and Arabie, P. (1975). "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling." *Journal of Mathematical Psychology*, 12(3), 328–383.

Brin, S. and Page, L. (1998). "The Anatomy of a Large-scale Hypertextual Web Search Engine." *Comput. Netw. ISDN Syst.*, 30(1-7), 107–117.

Brown, P. E. and Feng, J. (2011). "Measuring user influence on twitter using modified k-shell decomposition." *Fifth international AAAI conference on weblogs and social media*. 18–23.

Burt, R. S. (1987). "Social Contagion and Innovation: Cohesion versus Structural Equivalence." *American Journal of Sociology*, 92(6), 1287–1335.

Burt, R. S. (2009). *Structural holes: The social structure of competition*. Harvard university press.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, P. K. (2010). "Measuring user influence in twitter: The million follower fallacy." *Icwsm*, 10(10-17), article no. 30.

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C. and Zhou, T. (2012a). "Identifying influential nodes in complex networks." *Physica a: Statistical mechanics and its applications*, 391(4), 1777–1787.

Chen, W. (2009). "Hep-Phy dataset." Www.microsoft.com/en-us/research/people/weic/#publications [accessed in June 2013].

Chen, W., Lin, T., Tan, Z., Zhao, M. and Zhou, X. (2016). "Robust Influence Maximization." *CoRR*, abs/1601.06551. Http://arxiv.org/abs/1601.06551.

Chen, W., Lu, W. and Zhang, N. (2012b). "Time-critical influence maximization in social networks with time-delayed diffusion process." *arXiv preprint arXiv:1204.3074*.

Chen, W., Wang, C. and Wang, Y. (2010a). "Scalable Influence Maximization for Prevalent Viral Marketing in Large-scale Social Networks." *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 1029–1038.

Chen, W., Wang, Y. and Yang, S. (2009). "Efficient Influence Maximization in Social Networks." *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 199–208.

Chen, W., Yuan, Y. and Zhang, L. (2010b). "Scalable Influence Maximization in Social Networks Under the Linear Threshold Model." *Proceedings of the 2010 IEEE International Conference on Data Mining*. IEEE Computer Society, Washington, DC, USA, 88–97.

Chen, Y.-C., Zhu, W.-Y., Peng, W.-C., Lee, W.-C. and Lee, S.-Y. (2014). "CIM: Community-based influence maximization in social networks." *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(2), article no.25.

Chun-Ping, L., Yu-Rong, L., Da-Ren, H. and Lu-Jin, Z. (2008). "A Formula of Average Path Length for Unweighted Networks." *Communications in Theoretical Physics*, 50(4), 1017.

Cicalese, F., Cordasco, G., Gargano, L., Milanič, M., Peters, J. and Vaccaro, U. (2015). "Spread of influence in weighted networks under time and budget constraints." *Theoretical Computer Science*, 586, 40–58.

Clauset, A., Shalizi, C. R. and Newman, M. E. J. (2009). "Power-Law Distributions in Empirical Data." *SIAM Rev.*, 51(4), 661–703.

Coja-Oghlan, A., Feige, U., Krivelevich, M. and Reichman, D. (2015). "Contagious sets in expanders.", 1953–1987.

Cordasco, G., Gargano, L. and Rescigno, A. A. (2015). "Influence Propagation over Large Scale Social Networks." *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 1531–1538.

Cormen, T. H., Stein, C., Rivest, R. L. and Leiserson, C. E. (2001). *Introduction to Algorithms*. Second edition. McGraw-Hill Higher Education.

Dayama, P., Karnik, A. and Narahari, Y. (2012). "Optimal incentive timing strategies for product marketing on social networks." *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 703–710.

Deyasi, K., Upadhyay, S. and Banerjee, A. (2014). "Communication on structure of biological networks." Technical Report arXiv:1401.4913. Https://cds.cern.ch/record/1644154.

Dhamal, S., J., P. K. and Narahari, Y. (2016). "Information Diffusion in Social Networks in Two Phases." *IEEE Transactions on Network Science and Engineering*, 3(4), 197–210.

Domingos, P. and Richardson, M. (2001). "Mining the Network Value of Customers." *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 57–66.

Dubois, E. and Gaffney, D. (2014). "The multiple facets of influence: identifying political influentials and opinion leaders on Twitter." *American Behavioral Scientist*, 58(10), 1260–1277.

Estrada, E. and Rodríguez-Velázquez, J. A. (2005). "Subgraph centrality in complex networks." *Phys. Rev. E*, 71, 056103, number of pages 9.

Everett, M. G. (1985). "Role similarity and complexity in social networks." *Social Networks*, 7(4), 353 – 359.

Everett, M. G. and Borgatti, S. (1988). "Calculating role similarities: An algorithm that helps determine the orbits of a graph." *Social Networks*, 10(1), 77 – 91.

Everett, M. G. and Borgatti, S. P. (1999). "The centrality of groups and classes." *The Journal of Mathematical Sociology*, 23(3), 181–201.

Everett, M. G. and Borgatti, S. P. (2005). "Extending centrality." *Models and methods in social network analysis*, 35, 57–76.

Fang, X., Hu, P. J.-H., Li, Z. L. and Tsai, W. (2013). "Predicting Adoption Probabilities in Social Networks." *Information Systems Research*, 24(1), 128–145.

Foti, N. J., Hughes, J. M. and Rockmore, D. N. (2011). "Nonparametric Sparsification of Complex Multiscale Networks." *PLoS ONE*, 6(2), 16431(2).

Freeman, L. C. (1978). "Centrality in social networks conceptual clarification." *Social networks*, 1(3), 215–239.

Fung, W. S., Hariharan, R., Harvey, N. J. A. and Panigrahi, D. (2011). "A general framework for graph sparsification." L. Fortnow and S. P. Vadhan (eds.), *STOC*. ACM, 71–80.

Ganesan, K. (2016). "Case study on ripple effects of Ice bucket challenge on social media channels." Http://www.digitalvidya.com/blog/[Online; accessed 7-September-2016].

Ganesh, A., Massoulié, L. and Towsley, D. (2005). "The effect of network topology on the spread of epidemics." *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, volume 2. IEEE, 1455–1466.

Gargano, L., Hell, P., Peters, J. G. and Vaccaro, U. (2015). "Influence diffusion in social networks under time window constraints." *Theoretical Computer Science*, 584, 53–66.

Geppert, G. (2016). "How Influence Marketing Differs from Celebrity Endorsement." https://http://www.convinceandconvert.com/digital-marketing/influence-marketing-differs-from-celebrity-endorsement/[Online; accessed 7-November-2016].

Goyal, A., Bonchi, F. and Lakshmanan, L. V. (2010). "Learning Influence Probabilities in Social Networks." *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, USA, 241–250.

Goyal, A., Bonchi, F. and Lakshmanan, L. V. S. (2011a). "A Data-based Approach to Social Influence Maximization." *Proc. VLDB Endow.*, 5(1), 73–84.

Goyal, A., Bonchi, F., Lakshmanan, L. V. S. and Venkatasubramanian, S. (2013). "On minimizing budget and time in influence propagation over social networks." *Social Network Analysis and Mining*, 3(2), 179–192.

Goyal, A., Lu, W. and Lakshmanan, L. V. (2011b). "CELF++: Optimizing the Greedy Algorithm for Influence Maximization in Social Networks." *Proceedings of the 20th International Conference Companion on World Wide Web*. ACM, New York, NY, USA, 47–48.

Goyal, A., Lu, W. and Lakshmanan, L. V. (2011c). "Simpath: An efficient algorithm for influence maximization under the linear threshold model." *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 211–220.

Granovetter, M. (1978). "Threshold models of collective behavior." *American journal of sociology*, 1420–1443.

Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A. (2004). "Information Diffusion Through Blogspace." *Proceedings of the 13th International Conference on World Wide Web*. ACM, New York, NY, USA, 491–501.

Guo, H., Viktor, H. L. and Paquet, E. (2007). "Pruning Relations for Substructure Discovery of Multi-relational Databases." *Knowledge Discovery in*

Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007, Proceedings. 462–470.

Han, M., Yan, M., Cai, Z. and Li, Y. (2016). "An exploration of broader influence maximization in timeliness networks with opportunistic selection." *Journal of Network and Computer Applications*, 63, 39–49.

Han, M., Yan, M., Cai, Z., Li, Y., Cai, X. and Yu, J. (2017). "Influence maximization by probing partial communities in dynamic online social networks." *Transactions on Emerging Telecommunications Technologies*, 28(4).

Hanneman, R. A. and Riddle, M. (2005). *Introduction to social network methods*. University of California Riverside.

He, X. and Kempe, D. (2016). "Robust Influence Maximization." *CoRR*, abs/1602.05240.

Heidemann, J., Klier, M. and Probst, F. (2010). "Identifying Key Users in Online Social Networks: A PageRank Based Approach." Association for Information Systems, 1–21.

Hinz, O., Skiera, B., Barrot, C. and Becker, J. U. (2011). "Seeding strategies for viral marketing: An empirical comparison." *Journal of Marketing*, 75(6), 55–71.

Hinz, O. and Spann, M. (2008). "The impact of information diffusion on bidding behavior in secret reserve price auctions." *Information Systems Research*, 19(3), 351–368.

ImpulseDigital (2017). "Celebrity Endorsement and Influencer marketing. Where lies the difference?" Http://www.theimpulsedigital.com/blog/celebrities-vs-influencers-marketing/ [accessed on July 2017.

Isella, L., Stehl, J., Barrat, A., Cattuto, C., Pinton, J.-F. and den Broeck, W. V. (2016). "Infectious network dataset – KONECT." Http://konect.uni-koblenz.de/networks/sociopatterns-infectious [accessed in 2016.

Iyengar, R., Van den Bulte, C. and Valente, T. W. (2011). "Opinion leadership and social contagion in new product diffusion." *Marketing Science*, 30(2), 195–212.

Jabeur, L. B., Tamine, L. and Boughanem, M. (2012). "Active microbloggers: identifying influencers, leaders and discussers in microblogging networks." *International Symposium on String Processing and Information Retrieval*. Springer, 111–117.

Jakob, N. (2012). "Participation Inequality." Http://www.nngroup.com/article/participation Inequality.html [accesed in March 2016.

Jendoubi, S., Martin, A., Liétard, L., Hadji, H. B. and Yaghlane, B. B. (2017). "Two evidential data based models for influence maximization in Twitter." *Knowledge-Based Systems*, 121, 58–70.

Jiang, J., Wilson, C., Wang, X., Sha, W., Huang, P., Dai, Y. and Zhao, B. Y. (2013). "Understanding latent interactions in online social networks." *ACM Transactions on the Web (TWEB)*, 7(4), article no. 18.

Johnson, T. (2009). "Mathematical Modeling of Diseases: Susceptible-Infected-Recovered (SIR) Model." Http://op12no2.me/stuff/tjsir.pdf [accessed in March 2016.

Jung, K., Heo, W. and Chen, W. (2011). "IRIE: A Scalable Influence Maximization Algorithm for Independent Cascade Model and Its Extensions." *CoRR*, abs/1111.4795.

Kandhway, K. and Kuri, J. (2017). "Using Node Centrality and Optimal Control to Maximize Information Diffusion in Social Networks." *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(7), 1099–1110.

Kasthurirathna, D., Harre, M. and Piraveenan, M. (2015). "Influence modelling using bounded rationality in social networks." *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 33–40.

Kempe, D., Kleinberg, J. and Tardos, E. (2003). "Maximizing the Spread of Influence Through a Social Network." *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 137–146.

Kempe, D., Kleinberg, J. and Tardos, E. (2005). "Influential Nodes in a Diffusion Model for Social Networks." *Proceedings of the 32Nd International Conference on Automata, Languages and Programming*. Springer-Verlag, Berlin, Heidelberg, 1127–1138.

Kephart, J. O. and White, S. R. (1993). "Measuring and modeling computer virus prevalence." *Research in Security and Privacy, 1993. Proceedings., 1993 IEEE Computer Society Symposium on*. IEEE, 2–15.

Kermack, W. O. and McKendrick, A. G. (1927). "A Contribution to the Mathematical Theory of Epidemics." *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 115(772), 700–721.

Kimura, M., Saito, K. and Motoda, H. (2009a). "Efficient Estimation of Influence Functions for SIS Model on Social Networks." *Proceedings of the 21st International Jont Conference on Artifical Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2046–2051.

Kimura, M., Saito, K., Nakano, R. and Motoda, H. (2009b). "Extracting influential nodes on a social network for information diffusion." *Data Mining and Knowledge Discovery*, 20(1), 70–97.

Kirby, J. and Marsden, P. (2006). *Connected Marketing: The Viral, Buzz and Word of Mouth Revolution*. Butterworth-Heinemann.

Kleinberg, J. M. (1999). "Authoritative Sources in a Hyperlinked Environment." *J. ACM*, 46(5), 604–632.

Kristina, L. (2009). "Digg 2009 dataset." Http://www.isi.edu/lerman /downloads/digg2009.html [accessed in 2015.

Kutzkov, K., Bifet, A., Bonchi, F. and Gionis, A. (2013). "STRIP: Stream Learning of Influence Probabilities." *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 275–283.

Leskovec, J., Adamic, L. A. and Huberman, B. A. (2007a). "The Dynamics of Viral Marketing." *ACM Trans. Web*, 1(1).

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J. and Glance, N. (2007b). "Cost-effective Outbreak Detection in Networks." *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 420–429.

Leskovec, J. and Krevl, A. (2014). "SNAP Datasets: Stanford Large Network Dataset Collection." http://snap.stanford.edu/data [accessed in 2015.

Lin, Z. and Bei, Y. (2014). "Graph indexing for large networks: A neighborhood tree-based approach." *Knowledge-Based Systems*, 72, 48–59.

Lisa, R. (2008). "Social influence." *The Blackwell Encyclopedia of Sociology*. Oxford Blackwell, 4426–4429.

Liu, B., Cong, G., Xu, D. and Zeng, Y. (2012). "Time constrained influence maximization in social networks." *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 439–448.

Lorrain, F. and White, H. C. (1971). "Structural equivalence of individuals in social networks." *The Journal of Mathematical Sociology*, 1(1), 49–80.

Lü, L., Zhang, Y.-C., Yeung, C. H. and Zhou, T. (2011). "Leaders in social networks, the delicious case." *PloS one*, 6(6), e21202.

Mahyar, H. (2015). "Detection of Top-K Central Nodes in Social Networks: A Compressive Sensing Approach." *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, New York, NY, USA, 902–909.

Mathioudakis, M., Bonchi, F., Castillo, C., Gionis, A. and Ukkonen, A. (2011). "Sparsification of Influence Networks." *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 529–537.

McCracken, G. (2010). "How Ford Got Social Marketing Right." https://hbr.org/2010/01/ford-recently-wrapped-the-firs/[accessed in September-2016].

Misiolek, E. and Chen, D. Z. (2006). "Two flow network simplification algorithms." *Information Processing Letters*, 97(5), 197 – 202.

Mochalova, A. and Nanopoulos, A. (2013). "On The Role Of Centrality In Information Diffusion In Social Networks." *ECIS*. 1–12.

Mohite, M. and Narahari, Y. (2011). "Incentive compatible influence maximization in social networks and application to viral marketing." *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems, 1081–1082.

Moore, C. and Newman, M. E. J. (2000). "Epidemics and percolation in small-world networks." *Phys. Rev. E*, 61, 5678–5682.

Muchnik, L., Pei, S., Parra, L. C., Reis, S. D. S., Andrade Jr, J. S., Havlin, S. and Makse, H. A. (2013). "Origins of power-law degree distribution in the heterogeneity of human activity in social networks." *Sci. Rep.*, 3.

Myerson, R. (2003). "Graphs and Cooperation in Games." B. Dutta and M. Jackson (eds.), *Networks and Groups*. Studies in Economic Design, Springer Berlin Heidelberg, 17–22.

Narayanam, R., Skibski, O., Lamba, H. and Michalak, T. P. (2014). "A Shapley Value-based Approach to Determine Gatekeepers in Social Networks with Applications." *ECAI*, volume 14. 651–656.

Newman, M. E. (2006). "Modularity and community structure in networks." *Proc Natl Acad Sci U S A*, 103(23), 8577–8582.

Nguyen, H. and Zheng, R. (2012). "Influence spread in large-scale social networks–a belief propagation approach." *Machine Learning and Knowledge Discovery in Databases*. Springer, 515–530.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999). "The PageRank citation ranking: bringing order to the web." *Standford Infolab*.

Pal, S. K., Kundu, S. and Murthy, C. (2014). "Centrality measures, upper bound, and influence maximization in large scale directed social networks." *Fundamenta Informaticae*, 130(3), 317–342.

Pan, T., Kuhnle, A., Li, X. and Thai, M. T. (2017). "Popular Topics Spread Faster: New Dimension for Influence Propagation in Online Social Networks." *arXiv preprint arXiv:1702.01844*.

Pastor-Satorras, R. and Vespignani, A. (2001). "Epidemic spreading in scale-free networks." *Physical review letters*, 86(14), 3200–3203.

Pham-Gia, T. and Hung, T. (2001). "The mean and median absolute deviations." *Mathematical and Computer Modelling*, 34(7-8), 921 – 936.

Qin, Y., Ma, J. and Gao, S. (2016). "Efficient influence maximization under TSCM: a suitable diffusion model in online social networks." *Soft Computing*, 1–12.

Quirin, A., Cordn, O., Santamara, J., Vargas-Quesada, B. and Moya-Anegn, F. (2008a). "A New Variant of the Pathfinder Algorithm to Generate Large Visual Science Maps in Cubic Time." *Inf. Process. Manage.*, 44(4), 1611–1623.

Quirin, A., Cordón, O., Guerrero-Bote, V. P., Vargas-Quesada, B. and Moya-Anegón, F. (2008b). "A quick MST-based algorithm to obtain Pathfinder networks (, n- 1)." *Journal of the American Society for Information Science and Technology*, 59(12), 1912–1924.

Raghavan, U. N., Albert, R. and Kumara, S. (2007). "Near linear time algorithm to detect community structures in large-scale networks." *Physical review E*, 76(3), 036106.

Rauch, J. E. (2007). *Missing Links, The: Formation and Decay of Economic Networks*. Russell Sage Foundation.

Robert, H. (2008). "Applicability of Graph Metrics when Analyzing Online Social Networks." volume 1. 215 – 239.

Rodrigue, J.-P., Comtois, C. and Slack, B. (2009). *The geography of transport systems*. Second edition. Routledge, New York.

Rodriguez, M. G. and Schölkopf, B. (2012). "Influence maximization in continuous time diffusion networks." *arXiv preprint arXiv:1205.1682*.

Rogers Everett, M. (1995). *Diffusion of innovations*, volume 12.

Romero, D. M., Galuba, W., Asur, S. and Huberman, B. A. (2011). "Influence and Passivity in Social Media." *Proceedings of the 20th International Conference Companion on World Wide Web*. ACM, New York, NY, USA, 113–114.

Rudat, A. and Buder, J. (2015). "Making retweeting social: the influence of content and context information on sharing news in twitter." *Computers in Human Behavior*, 46, 75–84.

Saito, Kimura, M., Ohara, K. and Motoda, H. (2010). "Efficient Estimation of Cumulative Influence for Multiple Activation Information Diffusion Model with Continuous Time Delay." *Trends in Artificial Intelligence, PRICAI 2010 Proceedings*. 244–255.

Saito, K., Kimura, M., Ohara, K. and Motoda, H. (2012). "Efficient discovery of influential nodes for SIS models in social networks." *Knowledge and Information Systems*, 30(3), 613–635.

Saito, K., Kimura, M., Ohara, K. and Motoda, H. (2013). *Which Targets to Contact First to Maximize Influence over Social Network*, chapter Social Computing, Behavioral-Cultural Modeling and Prediction: 6th International Conference, SBP 2013, Washington, DC, USA, April 2-5, 2013. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg, 359–367.

Saito, K., Kimura, M., Ohara, K. and Motoda, H. (2016). "Super mediator A new centrality measure of node importance for information diffusion over social network." *Information Sciences*, 329, 985 – 1000. Special issue on Discovery Science.

Saito, K., Nakano, R. and Kimura, M. (2008). "Prediction of Information Diffusion Probabilities for Independent Cascade Model." *Proceedings of the 12th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Part III*. Springer-Verlag, Berlin, Heidelberg, 67–75.

Serrano, M. A., Bog, M. and Vespignani, A. (2009). "Extracting the multiscale backbone of complex weighted networks." *Proceedings of the National Academy of Sciences*, 106(16), 6483–6488.

Shang, J., Zhou, S., Li, X., Liu, L. and Wu, H. (2017). "CoFIM: A community-based framework for influence maximization on large-scale networks." *Knowledge-Based Systems*, 117, 88–100.

Shapiro, M. and Delgado-Eckert, E. (2012). "Finding the probability of infection in an SIR network is NP-Hard." *Mathematical biosciences*, 240(2), 77–84.

Shapley, L. S. (1953). "A value for n-person games." *Contributions to the Theory of Games*, 2(28), 307–317.

Shawndra Hill, C. V., Foster Provost (2007). "Learning and Inference in Massive Social Networks." *International Workshop on Mining and Learning with Graphs*.

Singh, L. (2005). "Pruning social networks using structural properties and descriptive attributes." *Fifth IEEE International Conference on Data Mining (ICDM'05)*. 4.

Statista (2016). "Number of social media users worldwide from 2010 to 2020." http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/[Online; accessed 7-October-2016].

Subbian, K., Aggarwal, C. and Srivastava, J. (2016). "Mining Influencers Using Information Flows in Social Streams." *ACM Trans. Knowl. Discov. Data*, 10(3), 1–28.

Sung, J., Moon, S. and Lee, J.-G. (2013). "The influence in Twitter: Are they really influenced." *Behavior and Social Computing*, 8178, 95–105.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor. Isbn: 0385721706.

Tanbeer, S. K., Leung, C. K.-S. and Cameron, J. J. (2012). "DIFSoN: Discovering influential friends from social networks." *CASoN-2012*. IEEE, 120–125.

Tang, J., Musolesi, M., Mascolo, C. and Latora, V. (2009). "Temporal distance metrics for social network analysis." *Proceedings of the 2nd ACM workshop on Online social networks*. ACM, 31–36.

Tang, Y., Shi, Y. and Xiao, X. (2015). "Influence maximization in near-linear time: A martingale approach." *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, 1539–1554.

Tarjan, R. E. (1972). "Depth-First Search and Linear Graph Algorithms." *SIAM J. Comput.*, 1(2), 146–160.

Teng, Y.-W., Tai, C.-H., Yu, P. S. and Chen, M.-S. (2015). "Modeling and Utilizing Dynamic Influence Strength for Personalized Promotion." *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 57–64.

Treagus, P. (2014). "The Dark Knight: A Case Study of Viral Marketing." http://philtreagus.com/the-dark-knight-a-case-study-of-viral-marketing/[Online; accessed 7-September-2016].

Wang, C., Tang, J., Sun, J. and Han, J. (2011). "Dynamic Social Influence Analysis Through Time-Dependent Factor Graphs." *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*. IEEE Computer Society, Washington, DC, USA, 239–246.

Wang, W., Liu, Q.-H., Cai, S.-M., Tang, M., Braunstein, L. A. and Stanley, H. E. (2016). "Suppressing disease spreading by using information diffusion on multiplex networks." *arXiv preprint arXiv:1604.01644*.

Wang, X.-G. (2016). "A new algorithm for the influence maximization problem in dynamic networks or traffic sensor networks." *Multimedia Tools and Applications*, 1–12.

Wang, Y., Cong, G., Song, G. and Xie, K. (2010). "Community-based greedy algorithm for mining top-k influential nodes in mobile social networks." *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1039–1048.

Wang, Y., Fan, Q., Li, Y. and Tan, K.-L. (2017). "Real-Time Influence Maximization on Dynamic Social Streams." *arXiv preprint arXiv:1702.01586*.

Wang, Z., Qian, Z. and Lu, S. (2013). "A Probability Based Algorithm for Influence Maximization in Social Networks." *Proceedings of the 5th Asia-Pacific Symposium on Internetware*. 1–7.

Weng, J., Lim, E.-P., Jiang, J. and He, Q. (2010). "TwitterRank: Finding Topic-sensitive Influential Twitterers." *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, USA, 261–270.

White, S. and Smyth, P. (2003). "Algorithms for Estimating Relative Importance in Networks." *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 266–275.

Wilder, B., Yadav, A., Immorlica, N., Rice, E. and Tambe, M. (2017). "Uncharted but not uninfluenced:Influence maximization with uncertain

network." *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1305–1313.

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. and Zhao, B. Y. (2009). "User Interactions in Social Networks and Their Implications." *Proceedings of the 4th ACM European Conference on Computer Systems*. ACM, New York, NY, USA, 205–218.

Wilson, C., Sala, A., Puttaswamy, K. P. N. and Zhao, B. Y. (2012). "Beyond Social Graphs: User Interactions in Online Social Networks and Their Implications." *ACM Trans. Web*, 6(4), 1–31.

Xiang, R., Neville, J. and Rogati, M. (2010). "Modeling Relationship Strength in Online Social Networks." *Proceedings of the 19th International Conference on World Wide Web*. ACM, New York, NY, USA, 981–990.

Yadav, A., Chan, H., Xin Jiang, A., Xu, H., Rice, E. and Tambe, M. (2016). "Using Social Networks to Aid Homeless Shelters: Dynamic Influence Maximization under Uncertainty." *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 740–748.

Yadav, A., Wilder, B., Petering, R., Rice, E. and Tambe, M. (2017). "Influence Maximization in the Field: The Arduous Journey from Emerging to Deployed Application." *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*.

Yang, J. and Leskovec, J. (2010). "Modeling information diffusion in implicit networks." *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 599–608.

Yang, W.-S., Dia, J.-B., Cheng, H.-C. and Lin, H.-T. (2006). "Mining social networks for targeted advertising." *System Sciences, 2006. HICSS'06.*

Proceedings of the 39th Annual Hawaii International Conference on, volume 6. IEEE, Washington, DC, USA, 137a–137a.

Zafarani, R. and Liu, H. (2009). "Social Computing Data Repository at ASU." Http://socialcomputing.asu.edu.

Zhang, H., Mishra, S., Thai, M., Wu, J. and Wang, Y. (2014). "Recent advances in information diffusion and influence maximization in complex social networks." *Opportunistic Mobile Social Networks*, 37.

Zhang, X., Zhu, J., Wang, Q. and Zhao, H. (2013). "Identifying influential nodes in complex networks with community structure." *Knowledge-Based Systems*, 42, 74–84.

Zhao, X., Yuan, J., Li, G., Chen, X. and Li, Z. (2012). "Relationship Strength Estimation for Online Social Networks with the Study on Facebook." *Neurocomput.*, 95, 89–97.

Zhou, F., Mahler, S. and Toivonen, H. (2010). "Network Simplification with Minimal Loss of Connectivity." *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. 659–668.

Zhou, F., Mahler, S. and Toivonen, H. (2012a). "Review of bisonet abstraction techniques." *Bisociative Knowledge Discovery*. Springer, 166–178.

Zhou, F., Mahler, S. and Toivonen, H. (2012b). "Simplification of Networks by Edge Pruning." M. R. Berthold (ed.), *Bisociative Knowledge Discovery*, *Lecture Notes in Computer Science*, volume 7250. Springer, 179–198.

Zhu, H., Huang, C., Lu, R. and Li, H. (2016). "Modelling information dissemination under privacy concerns in social media." *Physica A: Statistical Mechanics and its Applications*, 449, 53 – 63.

Zhuang, H., Sun, Y., Tang, J., Zhang, J. and Sun, X. (2013). "Influence maximization in dynamic social networks." *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 1313–1318.

List of publications

Journals

1 Sumith N, Annappa B and Swapan Bhattacharya. Social network pruning to build optimal social network, Elsevier Knowledge Based Systems (SCI journal). Vol. 117, pp 101-110.

2 Sumith N, Annappa B and Swapan Bhattacharya. A holistic approach to influence maximization in social networks: STORIE. Elsevier Applied Soft Computing. First revision submitted.

3 Sumith N, Annappa B and Swapan Bhattacharya. Influence Maximization in Large Social Networks: Heuristics, Models and Parameters. Elsevier Future generation computer systems. Under review.

Conferences

1 Sumith N, Annappa B and Swapan Bhattacharya. Data Reduction By Removal Of Lurkers In OSN. Proceedings of IEEE TENCON 2013, Xi'an, China, 22-25th, October, 2013(Scopus indexed).

2 Sumith N, Annappa B and Swapan Bhattacharya. RnSIR: A New Model for Information Spread in Online Social Networks, IEEE TENCON 2016, Singapore, Nov, 22-25, 2016(Scopus indexed).

Book chapter

1 Sumith N, Annappa B and Swapan Bhattacharya. A Holistic Approach to Influence Maximization, Book Chapter. Springer Hybrid intelligence for social networks. Accepted for publication in Oct, 2017.

Brief Bio-Data

Sumith N.
Research Scholar
Department of Computer Science and Engineering
National Institute of Technology Karnataka Surathkal
P.O.Srinivasanagar
Mangaluru 575025
Phone: 00824 – 2473044
Email: s.nireshwalya@gmail.com

Qualification

1 The research scholar has earned her Master degree M.Tech in Computer Science and Engineering from N.M.A.M.I.T, Nitte, Karkala in the year 2010.

2 She has earned her B.Tech in Information Science and Engineering from P.A.College of Engineering, Mangaluru in 2004.

Work experience

She has worked in the below mentioned colleges which are affiliated to Vivesvaraya Technological University, a collegiate state university in Karnataka State, India.

1 Srinivas Institute of Technology, Mangaluru from July 2011 till December 2012. Worked as Asst. professor in the department of Computer Science and Engineering. Handling theory and practical session for graduate computer science engineering program and post graduate M.Tech program.

2 ShreeDevi Institute of Technology, Mangaluru from February 2010 till July 2010.

Worked as Lecturer in the department of Computer Science and engineering. Handling theory and practical session for graduate computer science engineering program

3 St.Joseph Engineering college, Mangaluru from August 2005 till August 2008. Worked as Lecturer in the department of Computer Science and engineering. Handling theory and practical session for graduate computer science engineering program .

4 Vivekananda College of Engineering and technology, Puttur from August 2004 till July 2005. Worked as Lecturer in the department of Computer Science and engineering. Handling theory and practical session for graduate computer science engineering program

Publications prior research

1 Co authored, A study on different approach to manage identity, International conference on computing and control engineering, held at Dr.MGR University, Chennai on April 12-13, 2012.

2 Co authored, A survey on different background detection algorithms, National conference XPLORE-09 organised by IEEE-DSCE Students branch held at DSCE on 07th November 2009

3 Co authored, Implementation of motion detection system, in National conference XPLORE-09 organised by IEEE-DSCE Students branch held at DSCE on 7, November, 2009

4 Co authored, An approach to hidden node problem, presented in National conference on Recent Trends in Wireless Communication System, held at SDM college, Mangaluru , on 20, March, 2009.

5 Co authored, A dynamic extensible authentication protocol for device authentication in transport layer, International Journal for Computer Science

and Information Technologies(IJCSIT), volume 3(2),2012-3488-3492.

6 Co author, Application Layer security issues and its solutions, International Journal of Computer Science Engineering and Technology(IJCSET), volume 2, issue 6, June 2012

Awards and Achievements

1 Bagged a gold medal for highest scorer in M.tech CSE for the year 2008-2010.

2 Received Dr. K M Hebbar award as top scorer in M.Tech CSE for year 2008-2009.

3 Certification in Privacy and Security in Online Social Media by NPTEL, October, 2016.