Information Diffusion Modelling to Counter Semantic Attacks in Online Social Networks

THESIS

Submitted in partial fulfilment

of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

K P KRISHNA KUMAR

ID. No. 2012PHXF503H

Under the supervision of

Dr. G. Geethakumari

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

2015

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CERTIFICATE

This is to certify that the thesis entitled **Information Diffusion Modelling to Counter Semantic attacks in Online Social Networks** and submitted by **K P Krishna Kumar** ID No **2012PHXF503H** for award of Ph.D. of the Institute embodies the original work done by him under my supervision.

Signature of the Supervisor

Name in capital letters **DR. G. GEETHAKUMARI**

Designation Asst Professor, Dept. of CSIS

Date:

Acknowledgements

I would like to thank Dr. G. Geethakumari for all her suggestions and constant support during this research. Her valuable guidance and encouragement throughout the period were critical factors which contributed towards completion of the work. Through her untiring efforts, she helped me to critically analyse the problems in a systematic manner and consider innovative approaches to evolve practical solutions.

I would also like to thank Dr. Tathagata Ray and Dr. Suchetana Chakraborty, members of my doctoral advisory committee for their constant review and invaluable suggestions in steering the work. I would also like to express my gratitude to other members of the faculty in the Department of Computer Science and Information Systems Dr. Chittaranjan Hota, Dr. Gururaj, Dr. Bhanu Murthy, Dr. Aruna Malapati, Mr. KCS Murti, Mr. Digambar Powar, Mr. Abhishek Thakur and Mr. Rakesh Prasanna for all their suggestions and encouragement during various presentations and whenever I interacted with them.

My sincere gratitude to my fellow researcher Agrima Srivastava for our numerous discussions and brain storming sessions. These sessions helped me to analyse the problem from different perspectives to provide critical insights. I would like to thank each of the other researchers in the department Pawan, Meera, Jagan, Muthu, Prateek, Anita, Neha, Satya, Kiran for all the wonderful time we shared during our work.

I would also like to thank Military College of Electronics and Mechanical Engineering (MCEME), Secunderabad and its Department of Computer Systems for all the back end support they offered to me during the period.

Finally, my sincere acknowledgement of the sacrifices and support made by each member of my family during this period. They were my pillars of strength, always understanding and encouraging me. Without their support, this work would never have been completed.

BITS Pilani, Hyderabad Campus K P Krishna Kumar

Abstract

Semantic attacks are considered as the next wave of cyber security attacks. In this research work, we explored covert type of semantic attacks in Online Social Networks (OSN) in the form of large scale spread of misinformation, propaganda and disinformation - three counterfeits of information. Monitoring of OSNs, detection and prevention of semantic attacks require practical and effective solutions in near real time. Our proposed model has integrated the principles of Behavioural Sciences and Computer Science to formulate computationally efficient algorithms which uses the social computing properties of users of such networks .

Propagation of information is studied as a process of adoption using principles of Cognitivie Psychology and as a process of diffusion using principles of Sociology and evolutionary game theory. The essence of information diffusion modelling to detect diffusion of less credible information is the ability of users of OSNs to determine credibility of information. Our unique contribution is in the use of established theoretical framework of Psychometric analysis of users and their responses to different types of information to prove the social computing properties of users of OSNs to detect less credible information. We used Latent Trait Theory (LTT) for classification of the quality of information propagation in the communities in OSNs and the trust relationships between the users in them.

Having used LTT to prove the social computing properties of users of OSNs, we developed a *Behaviour trust model* based on the same principles to monitor communities spreading possible false information and detect sources of misinformation. The proposed metrics quantified the *social capital* of communities in terms of *entropy* in the communities and *gini coefficient* of degree distribution in the repropagation graph. Further we improved our algorithm by measuring the difference in the weighted and the unweighted eigen centrality ranks of sources of information to decide their credibility. Our main recommendation is in the establishment of an OSN reputation system for users of OSNs which could prevent spread of misinformation. Our algorithms provide a framework for the implementation of such a system at the clients' end to enable users to make informed decisions while repropagating messages.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Semantic attacks are the third wave of cyber security attacks. Social computing systems forming part of the Deep Web have proved to be an ideal platform for launching such attacks. ***Monitoring*** *of social networks,* ***Detection*** *and* ***Prevention*** *of semantic attacks, and* ***Counter measures*** *against them require practical and effective solutions in near real time. Do the social computing platforms offer solutions for the problem?*

Information operations are actions taken to affect adversary's information and information systems while defending our own systems. They are the means of attaining information superiority over the adversaries. Ensuring information security is one of the principal means of gaining information ascendancy. The domains of information superiority are not limited to military operations alone. Information superiority is of paramount importance in government, financial institutions, private businesses, hospitals and such organizations dealing with data. Hence, information security and in broader terms information assurance has great significance for all types of information operations. Information security aims to potect information and information systems from mainly three kinds of threats, viz unauthorized disclosure of information, unauthorized modification of information and unauthorized withholding of information. The triad of Confidentiality, Integrity and Availability (CIA) form the three goals of information security. When these three goals are achieved, information is considered secure.

The 21st century is considered as the 'Information age'. The advances in Information and Communication Technologies (ICT) have revolutionized the way we interact and the manner in which we receive information. The ability to stay updated, irrespective of location has changed the behavior patterns of people. There is exposure to an overload of information, which makes it quite a difficult proposition to validate their sources all the time. The security of any information system is a combination of means to ensure security of all links in the chain consisting of people, processes and technology. While processes and technology can be automated and can have strong technical solutions, it has been increasingly seen, that people are the weakest links in any information system domain.

Libicki in [1][2] has differentiated between three forms of attacks in cyber network operations - physical, syntactic and the semantic attacks. Physical attacks are aimed at the various hardware components forming part of the system. Attacks that have technology as the main aim are called syntactic attacks. The logical and syntactic properties of the technologies involved are affected by the execution of syntactic attacks. No human operators are involved. Semantic attacks are meant to change the information content or the meaning of information. This is done by the changing the human ability to analyse and perceive the information being received. Semantic attacks or cognitive hacking then refers to changing the human perceptions and their corresponding behaviours. Scheneir defines semantic attacks as those that "target the way, we humans attach meaning to content" [3]. The solution to cognitive attacks cannot always be completely technical in nature. Scheneir in [3][4] underlines with great efficacy this crucial point in the field of cyber security, pointing out that people who think to solve the problems of security through technology alone have understood neither the problem nor the technology.

1.1 Semantic Attacks

Semantic attacks are characterized by their ability to affect the human mind and the way in which information is interpreted by it. Physical and syntactic attacks have been researched extensively. The research required for semantic attacks is now receiving more attention. Computers and network infrastructure forming part of information systems remain in the technical domain and their security challenges have sound technical solutions. The weakest link in the chain of information security is the human interface. Human interactions are susceptible to semantic attacks. The types of semantic attacks possible in information systems along with the inter-disciplinary nature of the field makes them a very important topic of research.

The concept of semantic attacks is expected to play an important role in cyber security in the coming years. Schiener has described semantic attacks as the third wave of network attacks [5]. The humans form an important chain in the information operations and hence tackling security problem at the psychological domain along with the information security domain is more meaningful especially when semantic attacks are being investigated. The majority of attacks in future would be of the semantic type, which attack the human computer interface and the effects of which are not so visible as in the physical or the syntactic attacks. The strength of a chain is in the weakest link. In the use of computer mediated communication, the humans have proved to be the weakest link and the human computer interface the most susceptible to attacks [6].

This research is in the area of security challenges of information diffusion in online social networks (OSN). Different types of semantic attacks are possible in social computing systems like recommender systems, reputation systems and OSNs. Semantic attacks in reputation systems and recommender systems have been analysed extensively. However

analysis of similar attacks in OSNs have not been done. We would model information diffusion in OSNs with the aim to detect semantic attacks and their subsequent analysis so as to launch effective counter measures in time.

1.1.1 Social Computing Systems

Social computing systems are characterised by the employment of human beings as a part of computing system to achieve the desired objectives. In such systems, collaborative filter mechanisms use the known preferences of a group of users to make recommendations or predictions of the unknown preferences for other users [7]. These could be in the form of recommender systems [8][9] which are based on ratings of users to recommend new products to other similar users, reputation systems [8] [10] which are based on the rating of trust and reputation given by a group of users. Such collaborative filter systems have been in existence since the evolution of web and the exponential growth of e-commerce and financial transactions in it. The efficacy of the systems have been proved and they are increasingly being used. However, social computing systems are subjected to manipulation. Malicious users could use lies, rumours and propaganda to promote a person in reputation systems or downplay the popularity of an item in recommender systems. The real challenge of social computing systems is in protecting them from such semantic attacks which affect the manner in which information is interpreted by normal users of the system.

OSNs are characterised by interactions in the form of exchange of information like texts, video files, URL links etc. The spread of information in OSNs is a result of cognitive decision making processes of individuals and their interactions with others in society. The ability of OSNs to manipulate perceptions of large sections of society has been proved a number of times in recent years [11][12].

OSNs have emerged as important tools for communication and e-commerce. Deception in social media, which includes spread of misinformation, false profiles, spreading spams, viruses and so on is common. Using social engineering methods to extract confidential information to break into computer networks is of concern more for the cognitive domain than the technology domain. The ability of various systems to detect and counter such attacks is an important aspect of cyber security.

OSNs are also social computing systems. They have become an important source of information of all kinds, including personal, political, financial, health, governmental, religious and entertainment. They have lately been subjected to different kinds of semantic attacks aimed at manipulation of information contents and consequent changes in the behaviour of users. We intend to study semantic attacks in OSNs with special reference to the flow of information in them. While direct financial gains are the motivation for semantic attacks in other social computing systems, similar attacks in OSNs could have greater repercussions in the real world. Information as a currency for interactions is also required to be protected from manipulation.

OSNs are ideal platforms for very easy and cost effective means of transmission of

information. Equally effective and dangerous is the misuse of this media to create negative propaganda. The close knit communities like professional or personal links would cause such false propaganda to spread through the network with great speed. Frequently, the media is being misused to spread misinformation, deception and propaganda. While, most of the times the effects are limited to a small portion of online users, a concerted effort by a set of attackers have caused far reaching consequences in the recent attacks. Communal tensions prevailing in the society could be effectively used by attackers to launch semantic attacks to create fear and insecurity in the minds of affected people.

1.1.2 Taxonomy of Semantic Attacks

In [13], the authors have given a broad taxonomy of the various cognitive hacking methods used. Cognitive hackers manipulate users' perceptions and rely on their changed actions to carry out the attack. They have outlined the two forms of semantic attacks - *overt* and *covert attacks*. The overt type of semantic attacks include phishing, web defacing and spams. They affect the behaviour of individuals. But because of the overt nature of such attacks, the results are visible and counter measures can be taken against them effectively. It is the covert attacks, which are more dangerous as their detection itself is a challenge. OSNs are increasingly being used to deliberately spread false information. In these cases, though no computer or network infrastructure is broken into, the effects of such attacks are very severe. The ability to counter such attacks and take active measures for their early detection are big challenges. The dynamic nature of the web and the necessity to include multiple dimensions like computers, networking, psychology, deception and cognitive domains have posed challenges to the research community to propose suitable solutions. Figure 1.1 shows the detailed taxonomy of semantic attacks, their types and possible counter measures [5] [13].

Semantic attacks are different from the other two forms of cyber attacks. They attack the human computer interface and the effects of such attacks are not so visible as physical or the syntactic attacks. Detection of deception in the web has become increasingly difficult with the advent of Web 2.0. The lack of automated tools to detect and counter deception has been a major problem in ensuring information security in the Internet. A fully automated solution to counter various threats of social engineering, cognitive hacks and the various covert and overt attacks is not available. Quite possibly, automated tools augmented with human intervention may be the key to counter such attacks on information systems.

The types of covert semantic attacks include misinformation, unauthorized modification, deception and denial, pretexting and social engineering. The semantic web is touted as Web 3.0. Deliberate spread of false information like the pump and dump schemes of shares traded in the stock market is an example of covert semantic attack. Unauthorized modification of web sites, which changes the meaning of contents, like increasing the popularity of certain web sites, books or even shares of companies affect the decision making processes of people utilising such information systems. The use of social media to spread

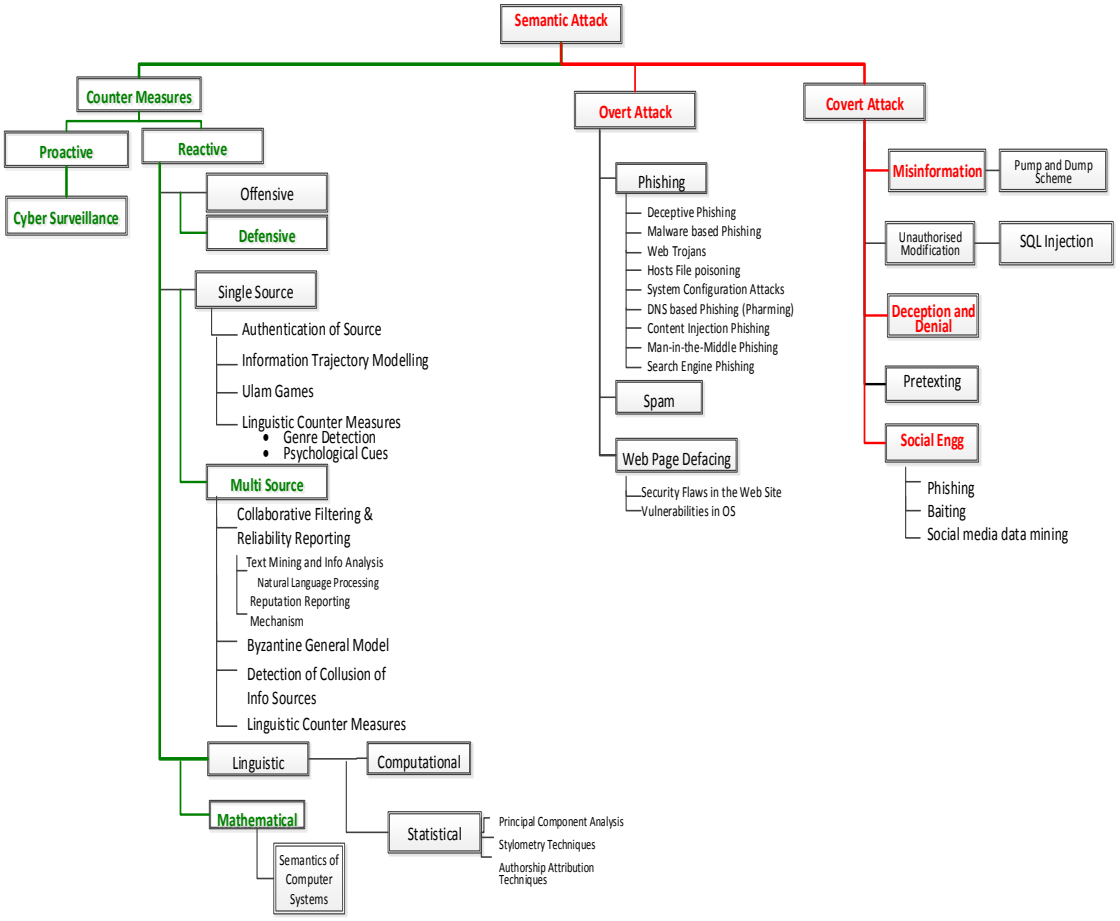


Figure 1.1: Taxonomy of semantic attacks and their countermeasures.

misinformation as well as launch semantic attacks and gain unauthorized access to computer networks by obtaining login details have been documented by many researchers in the field. We would consider covert semantic attacks in OSNs in the form of large scale diffusion of three counterfeits of information - misinformation, propaganda and disinformation [14].

Effective counter measures against semantic attacks include proactive and reactive measures. The actions taken in the absence of any declared hostilities do not fall under the realm of *Information warfare*. Nevertheless, the over dependence on the Internet to carry out running and maintenance of various physical infrastructure related activities like airport, trains, transport of goods, posts and e-commerce transactions have resulted in exposing the economic infrastructure of a country to cyber attacks. The proactive counter measures include carrying out effective cyber surveillance and detecting such attacks at the earliest. The counter measures to such attacks are required to be launched in time before the attacks could cause a change in users' behavior.

Various measures have been proposed to verify credibility of information. Depending

on the number of sources available for verification of information, single source or multi source collaborative filtering systems have been proposed. Other types of methods used are linguistic, statistical or mathematical techniques. Social media create a web of interactions which are being misused. Ability to use the same web, to create a web of trust would be ideal.

Information security is discussed from a psychological standpoint in [15]. Three main psychologically relevant dimensions such as cognitive hacking, hackers profiling and human errors along with four distinct levels of psychological approaches to information security are described. The humans form an important chain in the information operations and hence tackling the security problem at the psychological domain along with the information security domain is more meaningful especially when semantic attacks are being investigated. Four levels of psychological relevance approach to security has been made to include human errors approach, human factors approach, cognitive approach and the psychology of security approach. The role of psychology in enhancing cyber security has again recently been highlighted in [16]. The authors highlight the importance of identifying patterns of criminal and malicious activities through observing deviations from normative behaviour. There has been limited attempts to formulate solutions combining inputs from Psychology and Computer Science.

Use of deception in network security has been researched well. The dependence on Internet has added another dimension to it. A study of online deception has been done in [17], where the authors have adopted a game theory approach to conclude that deceivers may use different strategies to avoid detection and adjust their strategies dynamically. Statistical language models have been used to identify dependency of words in text without explicit feature extraction, to detect online deception [18].

Social network data is Big Data. Monitoring of OSNs, detection and prevention of semantic attacks and counter measures against them require practical and effective solutions in near real time. The present systems use content based and propagation based methods to carry out semantic analysis of contents of messages for their accurate classification. Use of statistical and linguistic modelling tools to detect deception in computer mediated communication have been done. Though algorithms using such machine learning techniques are accurate, they are considered of limited value for any real time monitoring system.

1.1.3 Gaps in Existing Research

Syntactic and semantic attacks have been researched extensively. But the coverage of semantic attacks is mostly done from a technology point of view alone and not from the human psychology and cognitive point of view. The need to analyse the human-computer interface to arrive at possible solutions of semantic attacks would require an analysis taking into account both the psychology and technology domains. The use of deception in our normal communication has been well researched. However, in computer mediated communication, the role of deception has just emerged as a new field of study. A Scientometric analysis [19]

has shown that the research papers describing the analysis of semantic attacks have rarely used the terminology of deception theories available in psychology domain.

Game theory has seen an increase in interest in modelling security attacks. The ability to predict optimum solutions for modelling real world problems using game theory has been established. The use of game theory in network security problems has shown promising results. In terms of uncertainties and vastness of the problems related to semantic attacks in cyber space, the use of game theory has been limited. There is lot of scope for applying game theory in modelling the attacker-defender scheme in semantic attacks. Their use in analysing the cognitive attacks in cyber space could lead to novel solutions to the problem.

Cyber surveillance has emerged as a requirement in today's scenario against semantic attacks. The use of cyber space to effect cognitive attacks has already been demonstrated. OSN data is Big Data and the present techniques do not offer solutions for their near real time monitoring to detect such attacks. The existing solutions based on machine learning and empirical data mining are computationally intensive to be used effectively for early detection of spread of false information. Prevention of semantic attacks requires early detection and timely launch of counter measures. In view of this, a comprehensive solution to carry out cyber surveillance and detect cyber deception in the initial stages itself has to be developed. Any counter measures which are required to be deployed, have to be done before manifestation of the effects of such misinformation. Though many algorithms have been proposed, no comprehensive solutions are described in the literature to thwart the type of cyber-attacks where social media were used to manipulative the spread of multimedia files targeting certain sections of the users [12].

Social networking sites, e-commerce portals with their specific business applications based on recommender systems and reputation systems are few examples of social computing systems. Semantic attacks in reputation systems and recommender systems have been researched extensively. Collaborative filter algorithms which make use of interpretation of data by users of the system to provide recommendations or reputation scores exist for recommender systems and reputation systems [7][20]. A systematic analysis for the use of similar algorithms has not been done on OSNs and neither solutions proposed which use such methodologies. The classification of semantic attacks in OSNs has also not been done. The open nature of OSNs is more likely to allow manipulation of information more easily than in other systems and hence the attacks need not be as technically sophisticated as in those systems. However, it is important to understand the type of attacks possible in OSNs to propose countermeasures against them.

The use of social computing properties of users of OSNs to determine credibility of information has been suggested by many authors. However, no effective algorithms have been proposed so far. Studies to prove the existence of such social computing properties of users of OSNs for detection of diffusion of less credible information has also not been attempted. The role of studies in Psychology in systems involving human beings has been

brought out. Effective use of research work in the field needs to be integrated with algorithms from Computer Science to propose solutions for semantic attacks in OSNs. Limited efforts have been made in this direction.

1.2 Objectives of the Research

The four major objectives of the research are as under:-

1. Design and implement effective algorithms for cyber surveillance to monitor specific activities in OSNs.
2. Design and implement novel and effective algorithm(s) to detect deception and counter the spread of misinformation in OSNs.
3. Design and develop a framework for establishment of trust relationships between users on the web including social media so as to prevent semantic attacks.
4. Develop a method to measure the effectiveness of counter measures launched against semantic attacks.

1.3 Scope and Problem Definition

In this research work, we have explored covert type of semantic attacks. They manipulate the way information is interpreted by human minds. We considered semantic attacks involving large scale diffusion of three counterfeits of information - misinformation, propaganda and disinformation. The scope of the proposed system would be to detect large scale spread of all the three counterfeits of information in OSNs. The extent of spread of misinformation and other less credible information could vary in OSNs. While it would be difficult to detect and limit the spread of all such information, we aim to detect information which has the potential to spread throughout the population.

Semantic attacks in the form of large scale spread of deliberate false information in OSNs have severe consequences. The challenges in the field include developing scalable solutions for their early detection, effective monitoring and targeted counter measures. The proposed system would make use of the social computing properties of users in OSNs for early detection of misinformation. Further, the system should also incorporate findings from the fields of Psychology and Computer Science to propose effective solutions to the problem. Use of statistical and linguistic modelling tools to detect deception in computer mediated communication have been done. However, the requirement is to analyse the problem from a strategic view to evolve a comprehensive solution which would consolidate all the efforts in the field.

1.3.1 Attack Model

The semantic attack discussed previously [12] can be visualized using Figure 1.2. This is the depiction of how a target population was affected by misinformation over a period of time. The target population was a section of OSN users against whom the attacks are targeted. Multimedia files were used to spread false propaganda amongst the social network users [12]. The targeted attacks found support amongst less than 15% of the population and remained confined to this set for some duration. After a certain time, the acceptance of the misinformation grew quite high and the percentage of the target population 'infected' with the misinformation displayed an exponential growth covering over 80% percent of the population in a short duration of time. The nature of spread of misinformation took the form of semantic attack which resulted in changes in behavior of persons and in these specific cases caused violence and destruction of human lives and property. Though many other media of communication were used, OSNs with their potential to spread multimedia files played a very major role in affecting the behavior of population. Monitoring of OSNs to understand the nature of such attacks and predicting them so as to undertake effective counter measures have become a necessity. As is evident from Figure 1.2, timely and effective detection of such attacks is very important. The period of attack could be divided into two parts as shown in Figure 1.2.

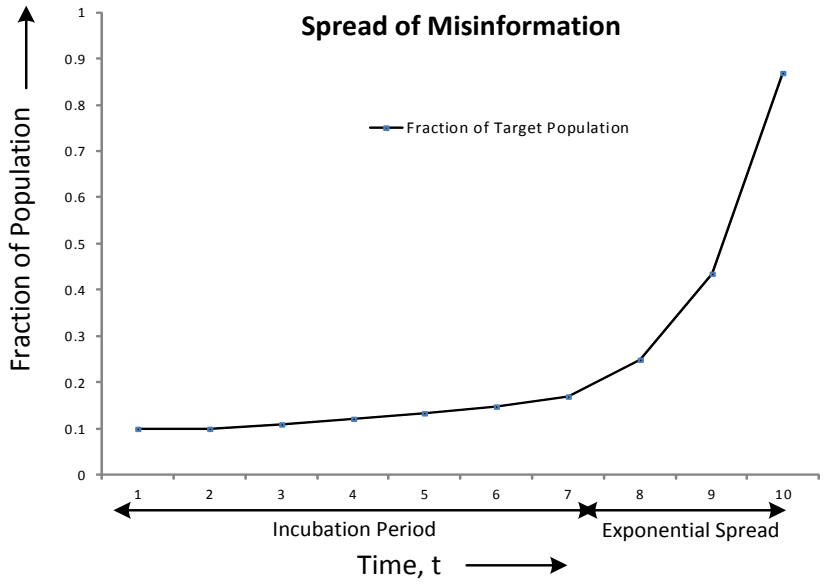


Figure 1.2: Spread of misinformation with time.

Incubation period

This is the period of time when misinformation was introduced into the target population and the semantic attack launched. During this period the acceptance of the misinformation is less and a small part of the population is affected. The spread is limited and the infected population remains the same for a long period of time. The messages are being passed on and efforts are being made by the attackers to influence the target users.

Exponential spread

This is the period of rapid increase in the percentage of infected population. A very high percentage of the target population is affected within a short span of time indicating almost an exponential spread of misinformation. The exponential spread could be triggered by the continued efforts of the attackers, or an external event which increases the acceptability of these messages.

1.4 Contributions of the Thesis

Any misinformation, once spread widely in a community, cannot easily be retracted. Preventing the spread of misinformation is a more effective method to combat misinformation, than its subsequent retraction after it has affected the population. We have proposed algorithms for early detection of semantic attacks in the form of large scale spread of misinformation, disinformation and propaganda. Our proposed model integrates principles of Behavioural Sciences and Computer Science. We prove the social computing properties of the users of OSNs and use the same for detection of diffusion of less credible information. The work proposes suitable metrics and a framework for implementation of an OSN reputation system which would enable users to make informed decisions while repropagating messages.

1.5 Outline of the Thesis

An extensive survey of the latest research from the fields of Cognitive Psychology, social network analysis and Computer Science in the area of information credibility and its diffusion is given in the next chapter. The processes of adoption and diffusion of information are studied more effectively using an *Integrated model* combining inputs from the fields of Cognitive Psychology, Sociology and Computer Science. We propose the model in Chapter 3. The proposed 'Integrated model' of analysis of semantic attacks has brought out solutions for the four objectives of the research work. The model has four stages. The first stage is the *Process of adoption* of information by individual users of OSNs, which is modeled using principles of Cognitive Psychology in Chapter 3. The second stage of *Process of Diffusion* is modeled in Chapter 4 based on work done in the fields of Sociology and evolutionary

game theory. This chapter explores the process of adoption by several users in the network and also proposes a taxonomy for different types of semantic attacks possible in OSNs. The algorithms proposed in the chapter would form the basis of a cyber surveillance system, which is the first research objective.

The proposed model is analysed in the next stage called *Process of analysis* which is described in Chapter 5. We make use of works done in the fields of evolutionary graph theory, psychometric analysis and behavioural trust for analysing information diffusion. Solutions to the second research objective which is to detect deception in the form of spread of disinformation is proposed using iterative k-core algorithm and analysis of isothermal sub graphs and bi-level graphs in the form of communities. The counter measures against the spread of misinformation in the form of behaviour trust model and trusted communities are also described in this chapter, which is the third research objective. The effectiveness of social computing properties of users to detect misinformation and identify its sources have been done using the established theoretical framework of psychometric analysis using latent trait theory. This is the fourth research objective. We propose a user centric information diffusion model in the last stage of the Integrated model called *Process of Modelling* in Chapter 6. The proposed 3-dimensional model is a variation of the standard independent cascade model for modelling information diffusion while catering for deliberate efforts made to spread any information. We have used common data sets related to different real world events obtained from 'Twitter' to explain our methodology. We present our results and analysis about the spread of propaganda during elections in Chapter 7. The recommendations of our work in the form of a framework for an OSN reputation system is also described. We give our concluding remarks in Chapter 8 along with the future scope of work.

1.6 Summary

Online deception in the form of spread of disinformation and distortion of contents of messages have increased in OSNs. We need to uncover deception so as to enable use of social media with confidence, develop measures to detect different types of disinformation so as to minimize the impact of spread of false information. Semantic attacks in the form of large scale spread of deliberate false information in OSNs have severe consequences. Handling spread of disinformation would involve effective monitoring, early detection, assessment of spread, prediction, tracking of sources of disinformation and targeted counter measures. In this chapter, we have described the problem and outlined our work in proposing effective solutions for them.

Chapter 2

Background and Related Work

"If you think technology can solve your security problems, then you don't understand the problems and you don't understand the technology". Bruce Schneier

2.1 General

The advent of Web 2.0 technologies has resulted in an exponential growth of OSNs in the last decade. Apart from the intended use as a tool to share information of interest between individuals or groups, OSNs are being increasingly used for the purposes of e-commerce like marketing, brand building and reputation systems, viral marketing, product launches and updates, customer feedback etc. They are also used for other services like real time news updates, job recruitment, health care services, technology trends, politics and financial updates to name a few. In all these services, the common denominator is the use of OSNs as a media for easy and quick dissemination of information to reach a wide audience in a cost effective manner.

This is the information age and quality of information is the metric which defines successful completion of all transactions. Even modern day warfare is centred on Network Centric Operations to ensure that right information is available to the right person at the right time. Psychological operations and use of propaganda have been tools of information warfare since olden times. The availability of OSNs has added a new dimension to be considered. The importance of dissemination of accurate information is thus equally important for armed forces as well as society at large. It is in this context that integrity of information being disseminated through OSNs becomes so critical.

OSNs have become an important source of information for a large number of people in recent years. As usage of social networks increased, the abuse of the media to spread disinformation and misinformation also increased many fold. The spread of information or misinformation in OSNs is context specific and studies have revealed topics such as health, politics, finances and technology trends are prime sources of misinformation and disinformation in different contexts which include business, government and everyday life [21]. The

information diffusion models do not take into consideration the type of information while modelling the diffusion process. The information diffusion in social networks due to misinformation or disinformation could follow different patterns of propagation and could be the result of an orchestrated campaign to mimic widespread information diffusion behaviour. The lack of accountability and verifiability allows the users an excellent opportunity to spread specific ideas through the network.

The detection of misinformation in large volumes of data is a challenging task. Methods using machine learning and Natural Language Processing (NLP) techniques exist to automate the process to some extent [Section 2.3 refers]. However, because of semantic nature of the contents, accuracy of automated methods is limited and quite often require manual intervention. The amount of data generated in OSNs is so huge as to make the task computationally expensive to be done in real time. In this research work, we have integrated principles of Behavioural Sciences with Computer Science, to propose computationally efficient solutions to be achieved in near real time.

A number of OSNs have emerged as the popular choice of interaction with people all over the world. Facebook, LinkedIn, MySpace, Twitter, Diggs are a few of them. While Facebook and MySpace are more for maintaining personal interactions across the world, LinkedIn is a network based on professional specialisations and expertise. Diggs and Twitter are for sharing general news and other information with each other, much like news sites dishing out views and real time news.

Twitter has emerged as one of the most popular micro-blogging sites. It enables users to send and receive messages, called 'tweets' in Twitter, which are limited to 140 characters. Twitter acts as a platform for social networking and news dissemination. Twitter enables propagation of news in real time. The ability to post tweets from mobile devices like smart phones, tablets and even by SMS has resulted in Twitter becoming the source of information for many users. Users can also retweet news items which they have received very easily to their followers, thus enabling quick information dissemination. These capabilities also make Twitter a platform for spreading misinformation easily.

2.2 Information, Misinformation, Propaganda and Disinformation

Human beings have an innate desire to share information with others. Sharing of information is the fundamental concept of OSNs. The desire to be the first to disseminate new information to the social network friends may result in the spread of inaccurate information. ***Misinformation*** *is false or inaccurate or misleading information* [22]. *The sender of misinformation believes the information to be true and has no intention to deceive the receiver.* ***Disinformation*** *is deliberately false information. The intention of sender is to deceive recipients.* It is spread deliberately with the aim to manipulate the behaviour of the

population. Accuracy of information is one of the important measures of quality of information. Honest mistake in the spread of inaccurate information is *misinformation*, whereas when the intention is to deceive the recipient, it is *disinformation* [21] [23]. While information, misinformation and disinformation are all informative in nature, only disinformation is deliberatively deceptive information and misinformation is misleading or false information. A disinformation campaign once launched in a social network may be further spread unintentionally as misinformation by normal users. Such disinformation campaigns can affect a large section of population. With the advent of OSNs, the time taken to achieve such large scale spread of inaccurate and misleading information has come down. The delay in taking steps to counter these misinformation cascades has invariably resulted in such campaigns changing the behaviour of population at a scale which was previously not possible.

2.2.1 Distinguishing Features

It is essential to understand the related concepts of information, misinformation, propaganda and disinformation. The distinction between information, misinformation and disinformation is difficult to be made [22]. The three concepts are related to truth, and to arrive at a universal acceptance of a single definition is almost impossible. The definition of information is clear by its very nature to the users. But what needs to be defined is the different forms it can take. We are more interested in the usage of social networks to spread specific kind of information to alter the behaviour or attitude of people. In the cyber space, manipulation of information so as to affect the semantic nature of information and the way in which it is interpreted by users is often called *semantic attacks*.

While evaluating information found on the Internet, authors [14] describe information and three of its lookalikes or counterfeits - propaganda, misinformation and disinformation. The formal definitions of the four terms given below have been obtained from Oxford English Dictionary, 3rd ed., 2007.

Information. Information is knowledge communicated concerning some particular fact, subject, or event; that of which one is apprised or told; intelligence, news.

Propaganda. Propaganda is the systematic dissemination of information, esp. in a biased or misleading way, in order to promote a political cause or point of view. Also: information disseminated in this way; the means or media by which such ideas are disseminated.

Misinformation. (a) The act of misinforming or condition of being misinformed. (b) Erroneous or incorrect information.

Disinformation. The dissemination of deliberately false information, especially when supplied by a government or its agent to a foreign power or to the media, with the intention of influencing the policies or opinions of those who receive it; false information so supplied.

The definitions have small differences and the most important fact is they involve the propagation of false information with the intention and capability to mislead at least some of the recipients. Misinformation is false information generated and spread without intention

to mislead people. Disinformation on the other hand, is deliberate falsehood generated and spread with the specific intention to mislead people [21]. As per authors, misinformation and disinformation are closely linked to information literacy, especially in terms of how they are diffused and shared as well as how people use cues to credibility and cues to deception to make judgements. While the intention of disinformation may be to deceive people, most of the persons involved in the chain of propagation may be propagating the information believing it to be true [23]. Propaganda is also concerned with spreading information which is biased. In terms of credibility, we classify propaganda, misinformation and disinformation as less credible than true information. True information would be more readily accepted by the population as compared to its lookalikes and similarly, information readily accepted by large members of the population is more likely to be credible. The advent of OSNs has made the speed of propagation of information faster, created large number of sources of information, produced huge amount of information in short duration of time and with almost no accountability about the accuracy of data. The term 'Big Data' is often associated with data in OSNs. Finding credible information after sifting out different forms of false information in OSNs has become a very challenging computational task.

2.2.2 Conceptual Explanation

The concept of information, misinformation and disinformation have been differentiated with respect to five important features by Karlova et al [21]. They are truth, accuracy, completeness, currency and deceptiveness. The authors have also given a social diffusion model of information, misinformation and disinformation as products of social processes illustrating the way they are formed and disseminated in social networks. The model suggests that people use cues to credibility and cues to deception to make judgements while participating in the information diffusion process.

Accuracy of information is one of the important measures of quality of information. In [23], authors have outlined the main features of disinformation.

- Disinformation is often the product of a carefully planned and technically sophisticated deceit process.
- Disinformation may not come directly from the source that intends to deceive.
- Disinformation is often written or verbal communication to include doctored photographs, fake videos etc.
- Disinformation could be distributed very widely or targeted at specific people or organizations.
- The intended targets are often a person or a group of people.

In order to classify as disinformation, it is not necessary that the disinformation has to come directly from the source of disinformation [23]. In the chain of dissemination of information, most of the people could actually be transmitting misleading information (hence misinformation), though only one of the intermediaries may believe that the information is actually misleading (hence disinformation). This is especially true for OSNs where the chain of propagation could be long and quite a few people are involved in the process.

2.2.3 How OSNs aid Spreading Misinformation

OSNs with its freedom of expression, lack of filtering mechanisms like reviewing and editing available in traditional publishing business coupled with high degree of lack of accountability have become an important media for spread of misinformation. Information asymmetry in OSNs play a big role in the spread of misinformation. Social networks spread information without traditional filters like editing. The advent of Web 2.0 has resulted in greater citizen journalism resulting in increase in the speed of dissemination of information using multiple online social media like social networks, blogs, emails, photo and video sharing platforms, bulletin boards etc. Summarily, the propagation of different versions of information, namely misinformation, disinformation and propaganda involves the spread of false or inaccurate information through information diffusion process involving users of social networks where all the users may not be aware of the falsehood in the information. We have used the term misinformation to denote any type of false information spreading in social networks.

The acceptance of misinformation or misleading information by the people depends on their prior beliefs and opinions [24]. People believe things which support their prior thoughts without questioning them. The same is also supported by research in Cognitive Psychology [25]. The authors have brought out that preexisting political, religious or social views make people accept information without verification if it conforms to their beliefs. Countering such ideological and personal beliefs is indeed very difficult. Another important finding was that countering misinformation may lead to amplifying the beliefs and reenforcing them.

Political astroturfing in the form of propagation of memes in Twitter was studied by the Truthy team [26] [27]. Investigating political election campaigns in US in the year 2010, the research group uncovered a number of accounts sending out duplicate messages and also retweeting messages from the same few accounts in a closely connected network. In another case, ten different accounts were used to send out thousands of posts, many of them duplicates slightly altered to avoid detection as spam. With URL shorteners available, messages containing links could be altered to give different shortened links to the same source and hence escaping the spam filters.

Use of rumours as part of a deliberate propaganda strategy to manage a population for the purposes of consumerism and war was highlighted in [28]. The authors talk about ‘rumour bombs’ which encourages rumour as a privileged communication strategy with

great efficacy to manage beliefs of population. They are characterised by rapid diffusion through electronic media in the society. Similar sentiments were also advocated in [29], where authors consider rumours as possible 'narrative landmines'. The two components of narration: what is told and how it is told are equally important. Although facts are important, truth becomes more about pre-existing and prevailing understandings of a person.

Decision making out of ignorance is often based on heuristics and the level of confidence on the decision is also low, making correction easier. Such decisions are often correct and are generally not catastrophic. False beliefs based on misinformation are held strongly and often result in greater support for a cause. Such beliefs are also very contagious and the person makes efforts to spread them to others. The persistence of misinformation in the society is dangerous and require analysis for their prevention and early detection [25] [30].

The effect of spread of misinformation have been studied extensively using methods available in Behavioural Sciences. An evaluation of all aspects of misinformation and its correction can be found in [25]. The authors have brought out the reasons for intentional and unintentional dissemination of false information and also explored the cognitive factors involved in spread of misinformation at the individual level. Internet has caused *Fractionation of Information Landscape* [31]. The phenomenon of *selective exposure* of people to information sources supporting their views has become very prevalent . This has resulted in formation of *Cyber-ghettos* or *echo chambers*, where links in social networks will follow like minded people with same views [32].

Information of an event as it unfolds like casuality figures in a natural calamity, is seldom accurate initially and the figures get updated or changed over a period of time. Such spread of misinformation is often considered benign though media is considered as one of the most important sources of misinformation. The other important sources of misinformation are governments and politicians, vested interests and rumours and works of fiction [25].

Cognitive hacking and its various counter measures are described in [33]. In this, the authors describe cognitive hack as the one which changes users' perceptions and their corresponding behaviours. The time available between posting of misinformation on the web and the corresponding change in users' behavior due to this misinformation is crucial. As covert attacks have proved, detection of such attacks is a challenge followed by providing effective measures to prevent cognitive hacking.

In [34], authors have enumerated a number of possible instances of misinformation in the Internet. They include incomplete, out-of-date and biased information, pranks, contradictions, improperly translated data, software incompatibilities, unauthorized revisions, factual errors and scholarly misconduct. However, with the advent of Web 2.0 the list has grown many times and social media is described as one of the biggest sources of information including misinformation. Internet acts as a post modern Pandora's box - releasing many different arguments for information which are not easily dismissible [35].

In [36], authors have developed a system called 'Seriously Rapid Source Review'(SRSR)

to be used by professional journalists for filtering and assessing the verity of sources in social media. The system uses a number of filtering and information cues such as content based features, aggregate information as well as location information to verify sources of information. The system would have limited use for ordinary users to judge the trustworthiness of sources. As per authors in [37], perceived credibility of sources and cognitive elaboration play important part in assessing the trustworthiness of sources. Cognitive elaboration involves active participation in information processing which involves activities like discussion of the content etc. The work clearly demonstrates the importance of credibility of sources and recency of updates for verification of information.

2.2.4 Examples of Misinformation Cascades in OSNs

There are quite a few examples when spread of misinformation in OSNs had caused large scale panic, fear and changes in behaviour of the population. The spread of misinformation of swine flu fever in Twitter in the year 2009 had caused widespread panic and fear [11]. In the year 2012, social networks were used to create panic and insecurity amongst certain communities in India for their safety, that caused a mass exodus of over thousands of people from major cities in the country to their villages [12]. The attackers made effective use of the communal tensions prevailing to launch an attack which created fear and insecurity in the minds of the affected people for their lives. Morphed video images were used to falsely depict the atrocities committed against certain communities. The whole attack was spread over a period of two months and the effect was devastating in terms of loss of lives and money. There are numerous smaller incidents which have taken only minutes to spread like wildfire creating economic chaos and crashing of stock markets at times. On 23 April 2013, hacked Associate Press Twitter account gave a false tweet that "Explosions in White House and President Obama has been injured" [38]. Though the tweet was retracted within 10 minutes, the damage to stock markets and repercussions world wide had been huge. The false identification of the 'Boston bomber' in social media [39] after bomb explosions in the Boston marathon in 2013 is another example of spread of false information with severe consequences for the victim. More recently, Twitter was used by an Indian from Bangalore to disseminate propaganda hailing Islamic state jihadists of ISIS(Islamic State of Iraq and Syria) [40]. This account was considered the most influential Islamic State Twitter account till blocked in December 2014.

Information diffusion in OSNs like Twitter, Facebook, LinkedIn etc have been studied extensively. Calamities including earth quakes and other important events like the 2011 Egyptian revolution were extensively covered in Twitter [41] [42]. It may even be appropriate to say that news no longer breaks, it tweets [43]. The factors such as indegree, page rank, retweets, mentions, influence trees govern the influence of news items in OSNs [44]. The information propagation and influence mechanism in large scale Twitter networks could be studied using the 'retweet' feature [45]. Supervised learning techniques to detect suspicious

memes in microblog platforms like Twitter have been discussed in [27]. All of these techniques use machine learning algorithms or heuristics to detect misinformation contents in the posts made in the OSN.

2.2.5 Countering Spread of Misinformation

Misinformation is easily another version of information. Countering spread of misinformation is not an easy task. The simple technique of labelling the other side as wrong is ineffective. Education of people against misinformation is necessary but not sufficient for combating misinformation. An analysis of counter measures proposed and modeled in literature against the spread of misinformation in OSNs are at times not in consonance with the effectiveness of measures suggested in Cognitive Psychology. Theoretical framework for limiting the viral propagation of misinformation has been proposed in [46] [47]. The authors have proposed a model for identifying the most influential nodes whose decontamination with good information would prevent the spread of misinformation. The solution to the problem of limiting the spread of misinformation by starting a counter campaign using k influential nodes, called the *eventual influence limitation* problem has been proposed in [48]. The influence limitation problem has also been studied in the presence of missing information. In both the papers, the basic assumption is that when an infected node is presented with correct information, it would become decontaminated. In [49], the authors have proposed ranking based and optimization-based algorithms for identifying the top k most suspected sources of misinformation in a time bound manner.

Studies in Psychology have proved that removing misinformation from infected persons is not easy [25]. The best solution to the spread of misinformation is early detection of misinformation and launch of directed and effective counter campaigns.

The strategies proposed in [25] for effective counter measures include:-

- Providing credible alternative explanation to the misinformation.
- Repeated retractions to reduce the effect of misinformation without repeating the misinformation.
- Explicit warnings before mentioning the misinformation so as to prevent the misinformation from getting reinforced.
- Counter measures be suitably biased towards affirmation of the world view of the receiver.
- Simple and brief retractions which are cognitively more attractive than the corresponding misinformation.

Any misinformation, once spread widely in a community, cannot easily be retracted. This has been suitably demonstrated during July 2012, when mass exodus of thousands of

people took place in India [50]. Preventing the spread of misinformation is a more effective means of combating misinformation, than its subsequent retraction after it has affected the population. The main aim of this work is to propose algorithms for early detection of spread of misinformation and propaganda, and propose timely measures to counter them.

2.2.6 Semantic Attacks in Social Computing Systems

Social computing systems like recommender systems and reputation systems are being increasingly used in the web. A number of algorithms are used in such systems. A detailed survey of various techniques used in recommender systems can be found in [20] [9]. The collaborative filter techniques used in recommender systems to include user based systems, item based systems and hybrid systems have been described in [7]. As per the authors, the collaborative filter systems provide much better recommendations than content based systems and are more scalable when the volume of data is very large. Manipulation of recommender systems in the form of *shilling attacks* and their possible counter measures are described in [51] [52]. A study of various types of shilling attacks on collaborative filter systems like bandwagon attack, average attack, sampling attack, random attack etc., would reveal that the manipulation of the ratings are done to either improve the ratings of the target items - called *push* attacks, or pull down the ratings of items - called *nuke* attacks. In all these cases, for the attack to become successful, the attackers inject false profiles into the systems in sufficient numbers to achieve their aim. These attacks prove that collusion of users could result in altering the correct behavior of the systems, if not detected.

Reputation systems are the backbone of many of the successful e-commerce platforms and P2P networks. Their ability to provide accurate ratings of trust have enabled users to interact with each other. Though as a collaborative filter mechanism, reputation systems provide accurate ratings of the participants, they are also subjected to different kinds of attacks by collusion between the participants. The attacks on reputation systems in the form of *sybil attacks* have been studied extensively [10][53][54]. The various classes of attacks on reputation systems to include self promoting attacks, white washing attacks, slandering attacks, orchestrated attacks and denial of service attacks are caused by collusion of users and creation of false profiles as in sybil attacks. Manipulation of the collaborative filter mechanism to change the ratings and reputation of users by collusion of attackers is one of the most important challenges in the use of such systems.

OSNs are also social computing systems. The use of collaborative filter mechanisms to predict trends in OSNs has been recently studied. The role of underlying core-periphery structure and community structure in diffusion of information in social networks was brought out in [55]. The authors have proposed algorithms for early warning analysis of large-scale protests and other such events in social networks. The collective intelligence of the web in deciding the quality of web pages by means of PageRank algorithm has been demonstrated by Google [56]. Google's PageRank algorithm uses information which is external to the web pages themselves. The back links proposed by them are a sort of peer review. Hence,

collective evaluation of quality of an item is bound to produce better results than expert evaluation in most of the social computing systems. In our work, we use a similar idea to estimate the quality of messages made in social networks based on a peer review process of ranking the repropagation links. The information diffusion processes in OSNs have been extensively studied. However limited efforts have been made to quantify the quality of contents in OSNs and detect semantic attacks in them for the spread of misinformation and disinformation through collusion of users. We studied the methodologies employed in other social computing systems to systematically evaluate their applicability to OSNs.

2.2.7 Information Diffusion Models

The flow of information in OSNs has been studied using a number of diffusion models. A study of different information diffusion models is necessary to analyse propagation of less credible information also. The simplest and the most widely used models to study flow of information in OSNs are the Independent Cascade Models (ICM) and Linear Threshold Models (LTM). Initially investigated for modelling interacting particle systems in [57] [58] and further for marketing in [59], ICM models the flow of information through the weighted directed edges of the social network graph as a set of probability functions. The LTM was initially proposed by Granovetter [60] and Schelling [61] and extended by Watts [62]. The models are further explained in [63]. This model is based on the assumption that the information, especially, misinformation is readily accepted by the nodes if it is acceptable to a certain proportion of its neighbours. A threshold value may be defined which would give a measure of resistance to the adoption of the information by the node and enables its further spread to other neighbors who have not been affected. Voter model proposes the adoption of information by a node when one of its randomly chosen neighbors has accepted the same [64]. Information Epidemiology models like Susceptible-Infected-Susceptible (SIS) model [65], Susceptible-Infected-Recovered (SIR) model [66] and further Susceptible-Infected-Recovered-Susceptible (SIRS) models have also been used to describe information diffusion processes based on the way infectious diseases spread in real world population.

Two popular heuristic algorithms which intuitively measure the influential capabilities of nodes in social networks based on the properties of the underlying graph structure are the Maximum Degree Heuristic (MDH) and the High Clustering Coefficient Heuristic (HCCH) algorithms [63] [67] [68]. The use of voter model to study the *Influence maximisation problem* is found in [69]. Cost Effective Lazy Forward (CELF) algorithm proposed another way to look at the same problem [70]. Here the aim is to identify the set of nodes where we would like to place our sensors to detect the spread of (mis)information in a social network. The influence of blogs in information diffusion and identifying the top-k nodes problem were done in [66] [71]. URL citations were used to study information epidemics in [71]. In [72], the use of PageRank computation was done to identify influential blogs. The problem of minimizing the spread of misinformation by blocking critical links has been studied in [73] [74] [75] [76].

There are several variations of the base models, especially for competing campaigns across the network like the Multi-Campaign Independent Cascade Model (MCICM) and Campaign-Oblivious Independent Cascade Model (COICM) [48]. This was one of the first attempts to study social networks with the aim to limit the spread of misinformation in them. In [77], the authors have augmented the ICM to capture the existence of multiple competing campaigns in the network.

Another method in which the problem can be modeled is by giving a user centric perspective to the process of information diffusion. From the users' perspective, the social networks are one of the parameters which would affect their decision making processes. There are inherent characteristics of each user which could be taken into account as well as the information the users have obtained from other global sources. In [78], while studying adoption of innovation, people have been categorized into five types, namely *innovators, early adopters, early majority, late majority and the laggards.* The categorization is done on time domain as to when a user adopts an innovation or an idea with respect to the rest of the population. The overall adoption follows a gaussian distribution with mean and standard deviation values deciding the categorization of the population. It has been estimated that a user in any of the five categories mentioned earlier, would exhibit the same behaviour irrespective of the innovation or the idea. The behaviour of the users are not only dictated by their immediate neighbors, but also influenced by the global adoption. The use of user centric models to study diffusion of information and the applicability of theories introduced in the diffusion of innovations by social scientists for studying diffusion of information have been described in [79]. The local models do not fully capture behaviour of users of OSNs. The users obtain information from multiple sources and hence models taking into account global signals or behaviour of the general population and the local signals are better at studying diffusion of information. The Gaussian Logit Curve Model (GLCM) models user behaviour with respect to the global population and captures the innovativeness of the users based on their actions. As per the theory of diffusion of innovations, the categorical behaviour is an innate property of a user. An innovator will act as innovator, no matter what the innovation and a laggard will always be late in adopting the same. The diffusion process described in [78] was given a mathematical model by Bass in [80].

Game theoretic modelling gives an utility point of view to the nodes for adopting information. In [81], the authors describe a mixed logical and game theoretic framework for modelling decision making under the potential of deception for online communities. Using cooperative game theory, an altogether different approach to both the top k nodes and λ coverage problem has been proposed in [68]. The use of game theoretical models for analysis of semantic attacks provides great insights into their understanding and evaluating counter measures. The semantic nature of attacks where human interactions are involved, are captured effectively using game theoretic models. Evolutionary graph theory studies the ability of a mutant gene to overtake a finite structured population [82]. Evolutionary graph theory provides the structure for interactions between agents playing games which are located at

the vertices of a graph. The interactions between agents are the games played which can be described using a payoff matrix. The effect of population structure as compared to well mixed population can be taken into account by using evolutionary graph theory. Much of game theoretic framework on evolutionary graph theory is done using simulation [82].

The use of evolutionary game theory to study diffusion of information has been attempted by number of authors [83]. As per authors, the use of machine learning methods is based on the assumption that training set is statistically consistent with the test set and rely on the fact that corresponding social network structure remains the same in future also, which may not be valid for social networks which are very dynamic in nature. The machine learning methods further ignore the decision making process of users and is based on a number of factors which include their beliefs and interactions with others. Such decision making processes and interactions can be modeled using game theory. The authors have considered users with new information as a mutant and information diffusion was considered as spreading of the mutant gene. The mutual influence process is quite similar to evolutionary process. The authors have analyzed the dynamics of information diffusion process over different types of networks to include complete networks and scale free networks. In evolutionary game theory, the emphasis on game formulation is on the concept of population and the dynamics of strategy updates of the whole population. The replicator dynamics illustrates the dynamic process of updating the strategies of the whole population. The replicator is the player who can reproduce his/her strategy under some specific rules of selection and mutation. The replicators with higher payoffs can reproduce at higher rates.

The two-player game strategies proposed are forwarding the information or not forwarding the information. The authors further use evolutionary graph theory to study games played on structured population as in OSNs. In this, the strategies are updated using predefined update rules. When information is released by a user it can diffuse over the network or it can stop suddenly. This is decided by the user's neighbours' actions as well as the actions of user's neighbours' neighbours' forwarding actions.

A survey of the latest methods of studying information diffusion in OSNs is given in [84]. The authors propose a taxonomy for information diffusion based on work done on detecting topics of diffusion, modelling information diffusion processes and identifying influential spreaders. The interdependence of information diffusion processes as competing or cooperating is important when studying spread of information in OSNs [85]. Further, studies have validated the *complex contagion principle* that stipulates that repeated exposures to an idea are crucial when the idea is controversial or contentious [86].

The selection of a particular model would depend on the type of application being considered. In order to understand the process of information spread in social networks, we constructed two taxonomies of different diffusion models in Fig. 2.1. In the first taxonomy, the models have been classified on their ability to depict multiple possibly competing information spreads in the network. Most of the models depict single information spread only. However, when multiple information campaigns have to be modeled, like depicting

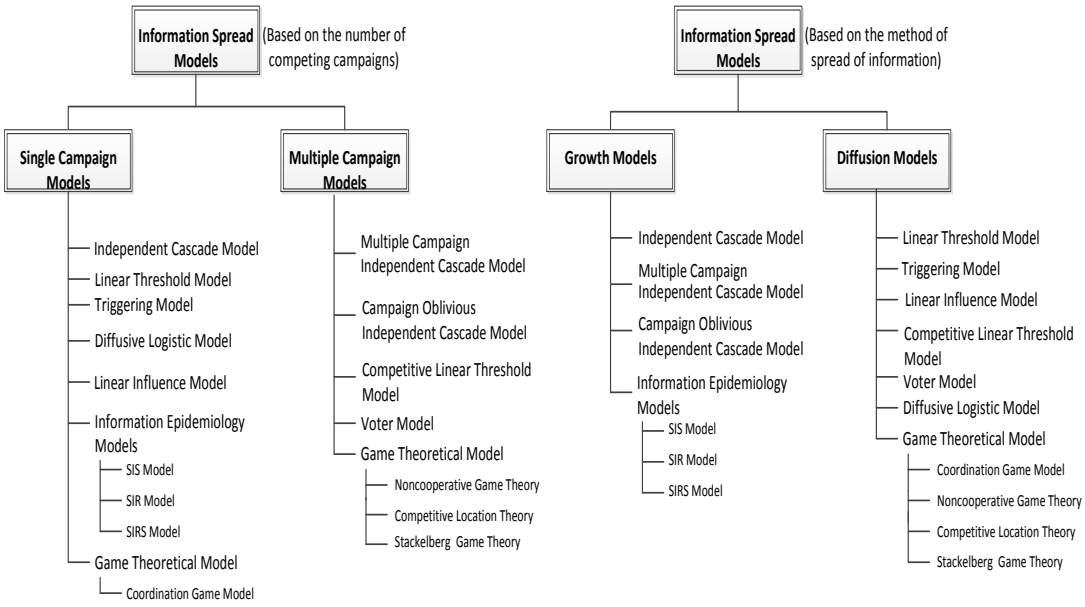


Figure 2.1: Taxonomy of information diffusion models.

counter campaigns against the spread of misinformation or when two competing advertising campaigns are launched in social networks, the response of the individual nodes to these campaigns would need to be analysed. Another way of classifying the information spread is using the growth and diffusion processes. Both are used to capture the behaviour of nodes with respect to its neighbors. Growth models are used when the information is spread to new nodes at increasing distance from the source. In this manner, it is similar to a walk through the network using depth-first search. In the diffusion models, the spread of information takes place possibly to the nodes at the same distance from the source. The spread of this kind is similar to the breadth first search of the graph. The growth process is a pure probabilistic model of information spread, whereas the diffusion process is the method by which a node gets activated if a large number of its neighbors are activated. We now give a brief description of different information diffusion processes.

The information diffusion models allow us to analyse and understand the process of information flow in OSNs. Information diffusion in OSNs like Twitter, Facebook, LinkedIn etc., have been studied extensively. However, the models seldom take into account the semantics of information. The flow of information would be different when the information is perceived as non credible by most of the users. Different OSNs provide different methods to analyse information flow in them. We carried out a specific study of one of the most popular micro-blogging site 'Twitter' so as to analyse the features used by authors to measure credibility of information flow in it. The results would enable us to detect and measure flow of non credible information. Though our study relates to Twitter, we would ideally search for a metric which is available across all OSNs and not limited to Twitter alone.

2.3 Credibility Analysis of OSNs - a Twitter Case Study

Twitter was established in 2006 and has emerged as one of the most popular micro-blogging sites. It permits users to post messages in the form of 'tweets' which are of maximum 140 characters in length. Twitter enables propagation of news in real time. The ability to posts information in the form of tweets from mobile devices like smart phones, tablets and even by SMS has resulted in Twitter becoming the source of information for many users. These capabilities also make Twitter a platform for spreading misinformation easily. Twitter has a 'retweet' feature using which users can repropagate messages they have received to their followers easily. Twitter permits users to 'follow' others to receive their tweets. The tweets of a person is sent to all his or her followers. Directed messages and replies can also be given using 'mentions' in tweets. The public nature of Twitter, ability to broadcast information to all followers, extensive use for personal communication, brand building and other e-commerce activities made us select Twitter for this case study.

2.3.1 Twitter as a Social Filter

The credibility of tweets propagated through Twitter has been analyzed in [87]. The authors have used automated methods to assess the credibility of tweets related to trending topics. The features used by the authors include retweets, texts of the posts and links to external sources. The authors used supervised learning techniques to build a classifier to estimate the credibility of tweets. Four types of features were used to classify tweets: message based features, user based features, topic based features and propagation based features. Use of message features included length of the message, positive or negative sentiments, presence of question marks or exclamation marks, and also the use of hashtags and retweets. User based features included number of followers and followees, number of past tweets etc. Topic based features were derived from user based and message based features to include fraction of tweets that contained hashtags, URLs and positive and negative sentiments. The propagation based features included the depth of the retweet tree and number of initial tweets of a topic. Best results of automatic classification of tweets were achieved using J48 learning algorithm. Sentiment features were found to be very relevant for predicting the credibility of tweets. The fraction of tweets with negative tweets was found to be more credible as well as tweets with greater number of retweets. The ability of the Twitter community to act as social filter of credible information has been clearly brought out in the paper. Credible users with large number of followers and followees along with large tweet activity have better reputation score and tend to propagate credible news. While validating the best features to be used for automatic determination of credibility of tweets, the propagation based features were ranked the best. The text and author based features alone are not sufficient to determine the credibility of tweets. The credibility of tweets increases when propagated through authors who have a higher reputation score, having written a large number of tweets before, originate at a single or few users in the network and have many retweets.

2.3.2 Twitter during Critical Events

Reliability of Twitter under extreme circumstances was also investigated in [88]. The analysis of tweets related to earthquake in Chile in 2010 has revealed that the propagation of rumours in Twitter is different from spread of credible news as rumours tend to be questioned more. The authors selected confirmed news and rumours manually from the set of tweets after the earthquake to analyse patterns of diffusion of information in the form of retweets in the network. The use of Twitter as a collaborative filter mechanism has been proved with the help of this study. Further, the authors have verified the validity of the use of aggregate analysis of tweets to detect rumours.

Credibility of tweets during high impact events was studied in [89]. The authors used source based and content based features to indirectly measure the credibility of tweets and their sources. Content based features in the tweets like number of words, special symbols, hashtags, pronouns, URLs and meta data like retweets were used. Source based features like number of followers, number of followees and age were used to measure the credibility of a user. The features were analysed for credibility using RankSVM [90] and Relevance feedback algorithms [91]. The limitation of their work is the requirement to establish ground truth using human annotation.

2.3.3 Spread of Rumours and Influence in Twitter

The spread of rumours in micro blogs was investigated in [92]. In particular, the authors investigated the spread of rumours in Twitter to detect misinformation and disinformation in them. The authors have proposed a framework using statistical modelling to identify tweets which are likely to be rumours from a given set of general tweets. They used content based, network based and microblog-specific memes for correctly identifying rumours. Content based features like lexical patterns, part-of-speech patterns, features corresponding to unigrams and digrams for each representation were used for classification of tweets. The authors used these techniques for rumour retrieval i.e., identifying tweets spreading misinformation. The belief classification of users to identify users who believe in the misinformation was done using the retweet network topology. The importance of retweet network topology has been clearly brought out in the paper. The authors have also used Twitter specific features like hashtags and URLs.

The measure of influence as given by retweet networks offer an ideal mechanism to study large scale information diffusion in Twitter [45]. The degree of influence of nodes measured by calculating the number of followers and number of retweets showed different results with little correlation between the two. The relationship between indegree, retweets and mentions as measures of influence have been further analysed in [93]. The authors have supported the claim that the users having large number of followers are not necessarily influential in terms of retweets and mentions. Influential users have significant influence across a number of topics. Influence in terms of retweets is gained only after concerted

efforts. Surveys have also shown that users are poor judges of truthfulness based on contents alone and are influenced by the user name, user image and message topic when making credibility assessments [94]. The ability to locate sources of rumour as well as classifying information as rumour was done in [95]. A number of monitor nodes were injected into the network to report on the data which could then be used to detect rumours. The work deals with identifying rumours in the absence of provenance of information. Their work provides helpful insights in terms of use of collaborative filter mechanisms to detect rumours and the fact that rumours are often initiated by a small group of connected individuals. True information is often associated with large number of sources which are often unconnected. The use of anti-rumour agents to combat rumours in the form of agents embedded in the network has also been demonstrated in [96]. The authors contend that rumours cannot be fought with authoritarian methods. Rumours can be combated with the use of messages passed from trusted friends which act as anti rumours [97].

2.3.4 Orchestrated Semantic Attacks in Twitter

Detection of suspicious memes in microblog platforms like Twitter using supervised learning techniques has been done in [26] [27] . The authors have used supervised learning techniques based on network topology, sentiment analysis and crowd-sourced annotations. The authors have discussed the role of Twitter in *political astroturf* campaigns. These are campaigns disguised as popular large scale grassroots behaviour, but actually carried out by a single person or organization. As per the authors, orchestrating a distributed attack by spreading a particular meme to a large population beyond the social network can be done by a motivated user. The paper discusses methods to automatically identify and track such orchestrated and deceptive efforts in Twitter to mimic the organic spread of information. The authors have described *Truthy*, a web service to track political memes in Twitter to detect astroturfing and other misinformation campaigns. The importance of the use of retweets to study information diffusion in Twitter has been highlighted by the authors. Network analysis of the diffusion of memes followed by sentiment analysis were used by the system to detect coordinated efforts to spread memes. The importance of detection of the spread of memes at an early stage itself before they spread and become indistinguishable from the real ones was also highlighted in the paper.

Being in the first page of the search results of any search engine is often regarded as an indicator of popularity and reputation. Search engines have introduced real time search results from social networking sites like Twitter, blogs and news web sites to appear in their first pages. A concentrated effort to spread misinformation as in political astroturf campaigns could have far reaching consequences if such search results are displayed prominently by search engines like Google. While studying the role of Twitter in the spread of misinformation in political campaigns, Mustafaraj et al have concluded that one is likely to retweet a message coming from an original sender with whom one agrees [98]. Similarly

repeating the same message multiple times indicates an effort to motivate others in the community to accept the message. The authors described an attack named *Twitter-bomb* where the attackers targeted users interested in a spam topic and send messages to them, relying on them to spread the messages further. The authors have highlighted the ability of automated scripts to exploit the open architecture of social networks such as Twitter and reach a very wide audience. Measuring hourly rate of generation of tweets seems to be a meaningful way of identifying spam accounts.

2.4 Summary

Detection of deception in the web has become one of the major challenges of Web 2.0. Lack of automated tools to detect and counter deception has been a major problem in ensuring information security in the Internet. Fully automated solutions to counter the threats of social engineering, cognitive hacks and different covert and overt attacks are not available. The methods to detect diffusion of less credible information in OSNs described in this chapter do not offer effective solutions to prevent their spread. Early detection of efforts to spread false information would be the key to prevent large scale misinformation cascades. Scalable algorithms to limit computationally expensive semantic analysis of contents of messages are required so as to enable launch of counter measures in an effective time frame.

Study of behaviour of human beings are better analysed using Behavioural Sciences and algorithms from Computer Science. The importance of Psychology in cyber security have been brought by different authors [16]. An integrated approach involving studies from Cognitive Psychology and social network analysis algorithms in Computer Science has greater potential in understanding and predicting user behaviour and developing scalable solutions for countering semantic attacks in OSNs.

The flow of information in OSNs has been modeled in a number of ways. In general, social networks are modeled as directed weighted graphs with individuals forming the nodes and interactions between them forming the edges. The direction and weights of the edges indicate the type and strength of interactions between individuals or any such qualifying parameter of relationships between them. Though modelling of social networks as graphs has become the norm, what has not been standardized is the way information diffusion is depicted in the graphs. Many such models have been proposed, each catering for a certain type of spread of information flow. There is no acceptable common model to explain the flow of misinformation and possible semantic attacks in OSNs. The study of information diffusion models has been published in [Pub4]. The Twitter case study of credibility analysis of OSNs has been published in [Pub2].

Chapter 3

Integrated Model for Study and Analysis of Spread of False Information

"There are a terrible lot of lies going about the world, and the worst of it is that half of them are true". Winston Churchill

3.1 Introduction

The spread of information in the Internet is a result of cognitive decision making processes of individuals and their interactions with others in society. The ability of OSNs to manipulate perceptions of large sections of society has been proved a number of times in recent years. Misinformation and disinformation are two forms of information. The two forms differ in the intent of their sources. Misinformation is often termed as 'accidental falsehood' as source has no intention to deceive, whereas disinformation is 'deliberate falsehood'. The intention is to manipulate human perceptions by a small set of users resulting in semantic attacks [21][22].

The importance of combining researches in the fields of Psychology and Computer Science to arrive at solutions for cyber security was highlighted in [16]. As per authors, studies in Psychology can help identify patterns of criminal and malicious activities through observing deviation from normative behaviour. Semantic security is emerging as an important part of cyber security. We have adapted principles of Cognitive Psychology and algorithms from the field of social network analysis to study deliberate propagation of biased news, propaganda and disinformation in OSNs.

The information propagation process in OSNs is similar to diffusion of innovations in a society [78][79]. Rogers in [78] highlights two different processes - Process of adoption and process of diffusion of innovations. Process of adoption is an individual process, which involves a series of stages one undergoes from first hearing about an item - an innovation or a news item- to finally endorsing it. Process of diffusion is a group phenomenon or collective behaviour which encompasses the adoption process of several individuals over a period of time. The analysis of information propagation in this manner would help us to understand the decision making processes of users and their manifestation in OSNs.

3.2 Integrated Model

The processes of adoption and diffusion of information are studied more effectively using an Integrated model combining inputs from fields of Cognitive Psychology, Sociology and Computer Science. The stages of proposed Integrated model with the fields of study are outlined in Figure 3.1.

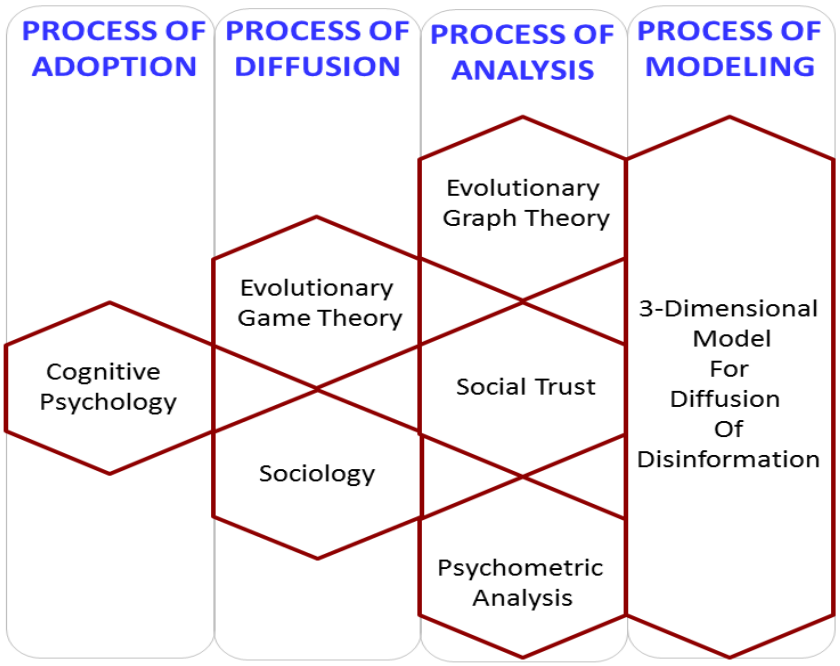


Figure 3.1: Integrated model for study and analysis of spread of new information.

The four stages of the model are elaborated below.

1. *Process of Adoption.* The adoption of a news item by an individual is a cognitive process weighing a number of parameters to decide on its credibility. We used research in the field of Cognitive Psychology to analyse behaviour of users in OSNs.

2. *Process of Diffusion.* The collective behaviour of users of OSNs are analysed using inputs from Sociology and researches in the field of evolutionary game theory.

3. *Process of Analysis.* The output from the previous stages are analysed using evolutionary graph theory, Psychometric analysis of information diffusion and Behavioural Trust model. The analysis should help us to segregate sources propagating false information and other users involved in the process.

4. *Process of Modelling.* In this stage we propose a 3-dimensional model for depicting the process of diffusion of disinformation in OSNs.

3.3 Process of Adoption of Information

The adoption of information by an individual is a cognitive process which we analyse using methods in Psychology. We evaluate the cognitive decision making processes of individuals to understand factors which influence them. Further, we study research work done in information propagation in OSNs to identify suitable metrics to analyse them. We classify people based on the process of adoption of information and propose an initial model to show their different transitions.

3.3.1 Cognitive Evaluation of Information

The effects of spread of misinformation using methods of Psychological sciences have been studied extensively and an evaluation of all aspects of misinformation and its correction can be found in [25]. The authors have brought out the reasons for intentional and unintentional dissemination of false information and also explored the cognitive factors involved in spread of misinformation at individual level. As per the authors of [25], spread of misinformation is a result of the cognitive process of adoption of information by receivers based on their assessment of the truth value of information. The users employ their cognitive powers to make decisions regarding the truthfulness of information they access. This decision by the receiver is based on a set of parameters which can be characterised by asking four relevant questions. These questions are given below and illustrated in Figure 3.2.

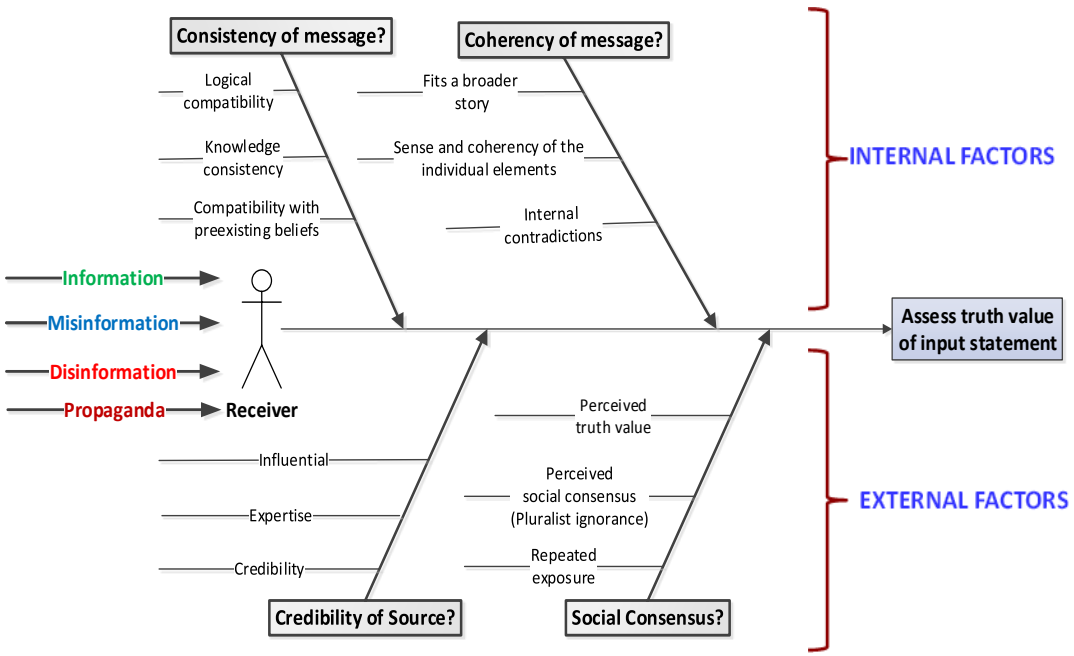


Figure 3.2: Cognitive process of assessing cues to misinformation or deception - 4Cs.

Cognitive Consistency. Messages are accepted easily when they are compatible with the recipient's preexisting beliefs. Misinformation which is compatible with other knowledge increases its probability of acceptance and decreases its chances of correction. Users receiving information in OSNs are more likely to accept them irrespective of other factors if such information conforms to what they already believe as true.

Cognitive Coherency. Misinformation which has been organised without internal contradictions and in agreement with common assumptions about human behaviour and motivation is readily accepted. Stories which fit well without inconsistencies are easier for cognitive processing and stories which are easier to process are more readily accepted and more resistant to correction. The text of messages in OSNs and the clarity with which ideas are conveyed in them would have an impact on their acceptability.

Credibility of Source. The acceptance of a message is directly proportional to the credibility of source. When the answers from previous cognitive reasoning do not provide a conclusive answer, recipients assess the information based on the perceived credibility and expertise of the communicator. OSNs provide a platform for receiving messages from multiple sources whose direct measurement of credibility would not always be easy.

Social Consensus. Familiarity of a belief due to continued exposure to it creates a social consensus which is considered as a 'secondary reality test'. Social-consensus information is very powerful in communities existing in social networks. The existence of 'echo chambers' in social media has already been proved. These echo chambers create social consensus by repetition which results in *pluralistic ignorance.*

The factors of *cognitive consistency* and *cognitive coherency* are intrinsic to individuals and a reflection of their personal ideologies, values and beliefs. This would remain the same when decisions are made while using OSNs or other wise. They could be defined as *Internal factors*. While they are important, *credibility of source* and *social consensus* assume greater significance in OSNs. They are external to an individual. The perceived credibility of source and general acceptability of information resulting in social consensus could be termed as *External factors*.

Direct measurement of cognitive decision making is difficult. But they can be evaluated based on indirect measures of reactions of users to messages. With this in mind, we took the popular micro-blogging site 'Twitter' as a case study and analysed literature to determine the most important features which could help us measure adoption of information by individual users in Section 2.3 of Chapter 2. Though our work relates to Twitter we would ideally search for a measure which is available across all OSNs and not limited to Twitter alone.

3.3.2 Analysis of Measuring Credibility of Tweets

A summary of analysis of the literature on information propagation in Twitter along with metrics used to detect credibility of tweets is given in Table 3.1. The present efforts to detect the spread of misinformation in OSNs can be broadly classified based on *consistency of message, coherency of message, credibility of source* and *social consensus* factors.

Table 3.1: Comparison of metrics for measuring credibility of tweets.

Criteria	**Metrics**	**Authors**	**Accuracy**	**Usefulness for fast detection**	**Remarks**
Consistency of message	Questions, affirms, denial, NLP techniques, retweets, mentions	[87], [89], [27], [26], [98], [93]	Retweets are better than mentions. Others are accurate.	Computationally intensive. Requires ground truth.	Content analysis required. Metrics are an indirect measure.
Coherency of message	Questions, affirms, denial, No of words, pronouns, hashtags, URLs, exclamation marks, negative and positive sentiments, NLP techniques	[88], [87], [89], [92], [27], [26]	Decision tree algorithms with a combination of various factors are accurate.	Computationally intensive. Requires ground truth.	Content analysis required. Metrics are an indirect measure.
Credibility of Source	Tweets, retweets, mentions, indegree, user name, image, followers, followees, age	[87], [89], [27], [26], [98], [45], [93]	Retweets are more accurate.	Yes	Direct measurement possible.
Social Consensus	Retweets	[27], [26]	Good	Yes	Direct measurement possible.

Differences in reactions of users to tweets based on their acceptability have been brought by all the authors. The authors have used a combination of features based on sources of information, contents of tweets, network and propagation based features for assessing credibility of tweets. The analysis has highlighted the following aspects.

- Automated means of detecting credibility of tweets are accurate, but computationally intensive and manual inputs are required.
- Retweets form a unique mechanism available in Twitter for studying information propagation and segregating misinformation. Repropagation feature is available in some form in most of the OSNs.
- Direct measurement of credibility of sources and social consensus parameters are possible. The other parameters permit indirect measurement.
- Analysing the information propagation using models in Computer Science and concepts of Cognitive Psychology would provide efficient solutions for detection and countering the spread of misinformation.

3.4 Modelling Acceptance of Information using Cognitive Psychology

The acceptance of messages by individual users is a cognitive process. We categorize users based on the factors influencing them in their decision making processes.

Profiling users

We are interested in studying the spread of false information in the network. We have already seen the four aspects, which can be called *4Cs - Internal factors* of Cognitive Consistency of message, Cognitive Coherency of message, and *External factors* of Credibility of source and Social Consensus about the message.

It is reasonable to estimate that users would use some or all of these factors to decide on the credibility of news items. The factors would not have equal influence in the decision of users to accept information as credible. The *Coherency* and *Consistency* values are based on cognitive factors internal to users. They are difficult to estimate but are the first tools to be used by users to evaluate any information presented to them. Information which appeals to their internal values are likely to be accepted easily. The acceptance could be in the form of agreement to the news contents or total disagreement of the news item. In both cases, users are convinced of the credibility of the news item or lack of it rather strongly. Along with perceived *Credibility* of source, these factors would be used by users to form their initial decision.

We would like to divide these people under two categories - **For** and **Against** the news item. Obviously, the classification is dynamic and would vary from one topic to another and may remain the same across different news items of the same topic. While a small proportion of people fall in these categories, majority would base their decisions taking also into account factors of *Social Consensus*. We want to categorize them as **Neutral**, as their decisions are influenced by external factors mainly and they themselves do not have strong convictions about the news item - *For* or *Against* it. We assume most of the people would fall in this category especially when the information is not depicting extreme views. The intensity of *Neutral* behaviour is not same across all people and would differ in the amount of additional information required by them to migrate to other two types of profiles. The *Neutral* types can also stay the same after considering all factors. The division of people into these types is similar to division of people as members, ex-members and potential members in [99] when profiling people for spread of social movements. Classification of people has also been done for purposes like adoption of innovations [78]. The proposed categorisation of people relates to the manner of acceptance of new information which could also be less credible information.

- **For.** People with strong support for the credibility of the news item. Decision is based on internal cognitive factors and relying mainly on the consistency and coherency

aspects of the news item.

- **Against.** People who are strongly against the news item and do not consider it credible. Decision is again based on the internal cognitive factors and relying mainly on the consistency and coherency aspects of the news item. These could also be people who know the information to be false.

- **Neutral.** People who mainly depend on others for confirmation of credibility of the news item. They are either not concerned about the news item or are not convinced themselves about the credibility of the news item. They would evolve to either the *For* type or the *Against* type or remain as *Neutral*, depending on the inputs they receive about the social consensus factor from the environment.

The process of adoption of information using Cognitive Psychology based on user profiles described above as well as *internal* and *external* factors influencing the decision making processes of users is depicted using a transition diagram in Figure 3.3.

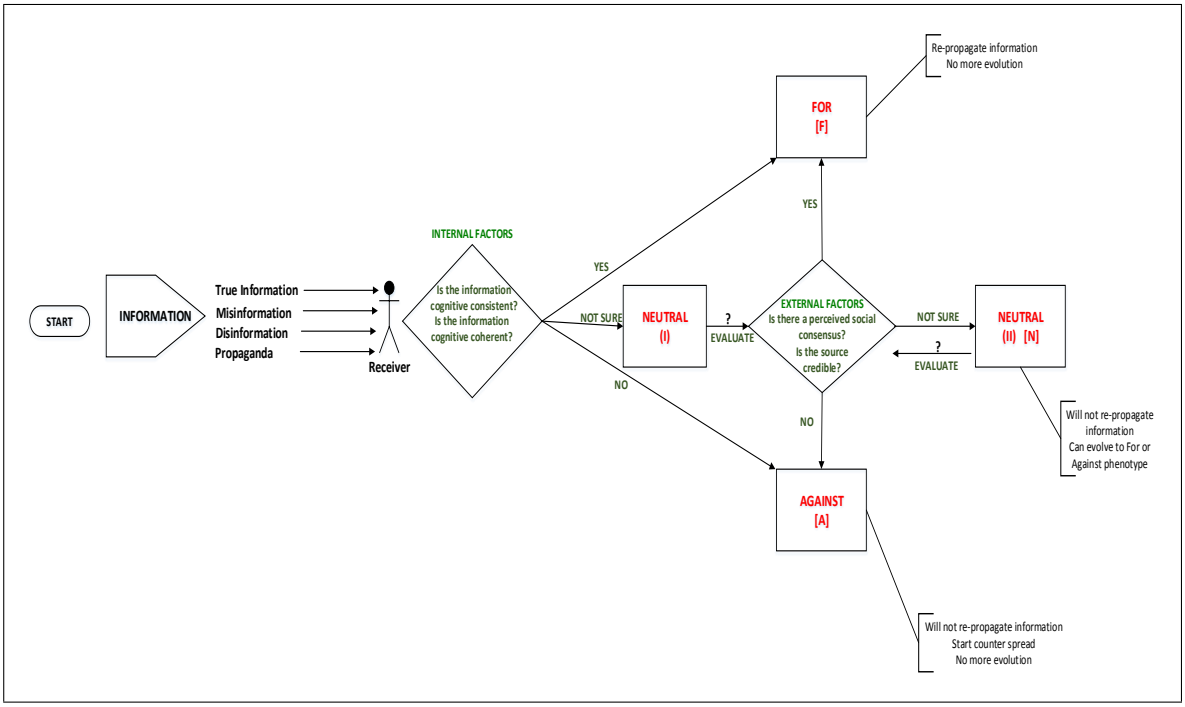


Figure 3.3: Modelling adoption of information using Cognitive Psychology.

The model depicts the decision making processes of individuals based on their inherent profiles as *For*, *Against* or *Neutral* about the received information. While internal factors enable the *For* and *Against* types to decide about accepting the information, the *Neutral* types wait for more information from the social network. Subsequent exposure to similar messages and the perceived social consensus about them enable users to accept the information or reject it. They may also continue to remain *Neutral*.

3.5 Summary

In this chapter we analysed the decision making processes of individuals to accept information they receive in OSNs. While some users would be intrinsically *For* or *Against* accepting the news item, a large number of users remain *Neutral* and accept or reject the same due to perceived social consensus. We also developed a model to describe this decision making process of individuals based on inputs from cognitive factors affecting the decision making of individuals. The importance of looking at sources of information and the perceived social consensus of their messages have been brought out. We would now study the process of diffusion of information in OSNs based on studies in Social sciences and evolutionary game theory and extend the proposed model to include the same in the next chapter. The use of Cognitive Psychology to understand and model decision making processes has been published in [Pub2].

Chapter 4

Process of Diffusion of False Information

"A lie can travel halfway round the world while the truth is still putting on its shoes." Mark Twain

4.1 Introduction

Diffusion of information in OSNs is a result of adoption of information by several individuals over a period of time. The process of diffusion could be studied from the perspective of Social sciences to understand how large number of people adopt a certain behaviour. The decision to accept information due to social interactions is modeled using evolutionary game theory. As different types of individuals interact over social networks, social influences would result in changes in behaviour of people. Evolutionary game theory provides an ideal framework to model resultant behaviour in a social network graph with users as vertices of the graph and edges as interactions. We aim to extend the model developed in the previous chapter to show the process of diffusion of false information.

4.2 Analysing Diffusion of Information using Sociology

The analysis of factors have revealed that credibility of sources of information and social consensus are important to study acceptance of information by users. We studied social consensus using repropagation as a metric of credibility. Grouping together the repropagated messages of each source would enable us to understand the level of acceptance of the messages of the sources. We modeled the interactions in OSN as a repropagation graph and measured variations in acceptance of messages from a source using the metric of gini coefficient. In order to study the diffusion of new information and their acceptance by users, we used data obtained from Twitter for different contexts. Further analysis of them would require evaluating the contents of messages grouped by their sources.

4.2.1 Data Sets

We obtained the required data from Twitter. The classification of information as genuine information or misinformation is with respect to the context in which they are studied. The data from Twitter pertains to different contexts as defined in Table 4.1. We assume that the messages could be segregated using keywords provided by subject matter experts to group together messages related to same events. We used Twitter API to collect the tweets. The spreadsheet tool TAGS v5 used for collection of tweets using the Search API was provided by Martin Hawskey [100]. Only the 'Higgs' anonymized data set was obtained from a public source as indicated. We would use these data sets throughout our work to validate the proposed methodology. In addition, we would carried out simulation on synthetic data sets also which are used extensively in other similar works.

- **Egypt.** We investigated the spread of news related to the political unrest and massive protests in Egypt during the period from 13 Aug 2013 to 23 Sep 2013. The tweets were collected using the keyword 'egypt'.

- **Syria.** We tracked the events of use of chemical agents in Syria and all news related to it using the keyword 'syria'. The tweets were collected over the period between 25 Aug 2013 and 21 Sep 2013.

- **Bodhgaya.** The spread of information about terrorist attacks on 7 Jul 2013 at 'Bodhgaya' temple in India was tracked for a period of nineteen days from 07 Jul 2013 to 25 Jul 2013. The tweets were collected using the keyword 'bodhgaya'.

- **MyJihad.** We tracked a particular hashtag 'MyJihad' which we observed had contents which were controversial and the frequency of tweets were quite high. The tweets were collected over a period of eight days between 20 Jul 2013 and 27 Jul 2013.

- **Telangana.** The spread of politically sensitive information in India over the demand for a separate state of Telangana was studied using the keyword 'telangana'. The tweets were collected over a period of eight days between 23 Jul 2013 and 30 Jul 2013 prior to the government decision being announced.

- **Andhra.** There was wide spread stir against the bifurcation of the state of Andhra Pradesh in India after the decision was announced. We tracked the movement using the keyword 'andhra' and 'telangana' for the period from 30 Sep 2013 to 09 Oct 2013.

- **Phailin.** The coast of Odisha and Andhra Pradesh were hit by a severe cyclone 'Phailin' between 10-11 Oct 2013. We tracked the event in Twitter using the keyword 'phailin' for a period from 10 Oct 2013 to 13 Oct 2013.

- **Higgs.** The data set of anonymized tweets pertains to messages in Twitter before, during and after the announcement of the discovery of a new particle with the features

Table 4.1: Details of Twitter data sets.

Data set	Users	Tweets	Sources	Retweets	Period	Type
Egypt	27532	141682	10850	51723	13 Aug 2013 to 23 Sep 2013	Civil Movement
Syria	25415	104867	11452	44671	25 Aug 2013 to 21 Sep 2013	Political
Bodhgaya	4573	8457	660	4230	07 Jul 2013 to 25 Jul 2013	Terrorism
MyJihad	1166	5925	140	3232	20 Jul 2013 to 27 Jul 2013	Religious
Telangana	2671	6787	464	2177	23 Jul 2013 to 30 Jul 2013	Political
Andhra	3255	25463	1385	9064	30 Sep 2013 to 09 Oct 2013	Political
Phailin	4408	16190	1567	7408	10 Oct 2013 to 13 Oct 2013	Natural Calamity
Higgs	425008	14855875	42176	423198	01 Jul 2013 to 07 Jul 2013	Scientific

of Higgs boson on 4th July 2012. The messages posted in Twitter about this discovery pertain to the period between 1 July and 7 July 2012 [101]. The data set is publicly available at http://snap.stanford.edu/data/higgs-twitter.html.

We used the 'retweet' mechanism in Twitter to study the patterns of spread of messages of each source. Users propagate any news item they receive by retweeting it to their followers. This could be considered as a positive affirmation by the retweeting node about the acceptance of the message. It also enables us to monitor the spread of a particular news item in the network.

4.2.2 Establishment of Ground Truth

We used human annotation to classify the tweets and their sources. We concentrated only on tweets which have been retweeted, as we were interested only in misinformation which has the potential to spread in the network. We classified each of the retweeted tweets as possible misinformation or genuine information. Each of the original source was also classified as a potential source of misinformation even if one of his/her tweets were retweeted and classified as misinformation. We repeated the procedure with all the data sets. Finally, we had a list of sources which are considered spreading misinformation as well as a list of messages.

4.2.3 Analysis of Sources

Having classified the data sets, we performed a detailed analysis of all the sources. We plotted graphs for each source of tweets - both misinformation and genuine information. We used retweets as a combined measure of acceptance of credibility of sources as well as general acceptability of the messages. For each source we plotted the cumulative percentage of tweets made by the source against the cumulative percentage of users who have retweeted them. We considered total number of users as those who have retweeted at least

a single tweet of the source. Figure 4.1 is for four sample sources whose messages have been classified as normal users and Figure 4.2 are for a sample set of four users classified as 'misinformers'.

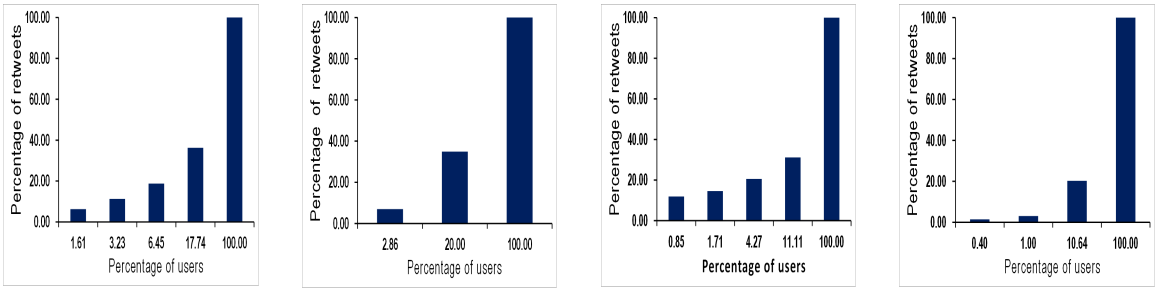


Figure 4.1: Distribution of retweets of four different sources of genuine information amongst users retweeting their tweets.

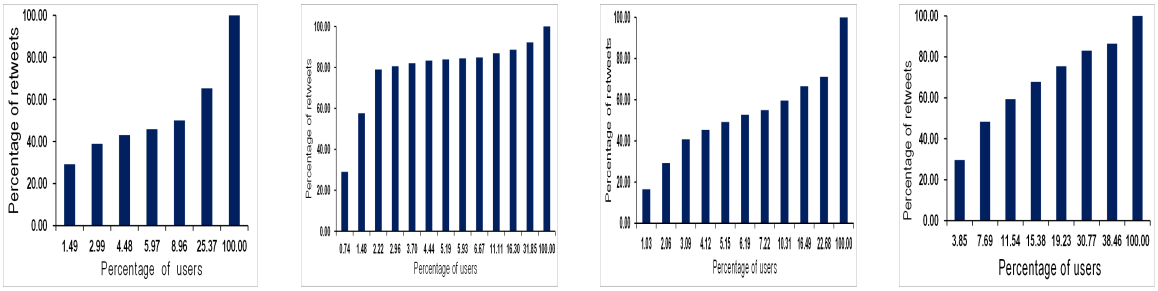


Figure 4.2: Distribution of retweets of four misinforming sources amongst users retweeting their tweets.

The difference in types of acceptance levels of the two types of sources is clear from Figure 4.1 and Figure 4.2. In the case of misinforming sources, there seems to be a smaller set of users who are retweeting most of the tweets of the source. This points towards the fact that while the source has generated large number of tweets, only a small fraction of the users have found it credible enough to retweet them. Most of the users did not find all the tweets of these sources worth repropagating. For source spreading genuine information, most of their tweets were acceptable to large proportion of the retweeting nodes. The disparity in retweeting between the retweeting nodes is much sharper in the case of sources who misinform as compared to the sources who do not. This also points towards the fact that there could be collusion of users who retweet each others tweets in order to ensure their spread in the network. This in turn would result in greater communication links between the colluding nodes as compared to their communication with others. The presence of such colluding nodes and disparity in their retweet behaviour can be detected using suitable network algorithms like core and community detection algorithms in graphs.

4.2.4 Results and Discussion

The difference in levels of acceptance of different sources were measured by drawing a repropagation graph or retweet graph and calculating the gini coefficients of the distribution of retweets.

Repropagation graph

We constructed a new graph called repropagation graph, which is a bipartite graph with two types of nodes - user nodes and message nodes (tweets). Directed edges are made from the tweets to the source user node and also from the retweeters to the tweet message nodes. We call the users retweeting others' tweets as retweeters. User nodes are the social network users who participate in the information propagation process and the message nodes are the actual messages or tweets in this case. An example of a section of the graph is given in Figure 4.3. In the figure, Node1 to Node7 are the user nodes and RT1 to RT3 are the message nodes. When a source creates a news item in the OSN, a directed edge is made from the message node to the source node. When another user node repropagates the message, a directed edge is created from the user node to the message node. Thus, every message node would have one outgoing link and one or more number of incoming links, if the message has been repropagated. The in-degree of each user node would give the number of messages generated by the user node and in-degree of each message node would give the number of user nodes involved in the repropagation of the message.

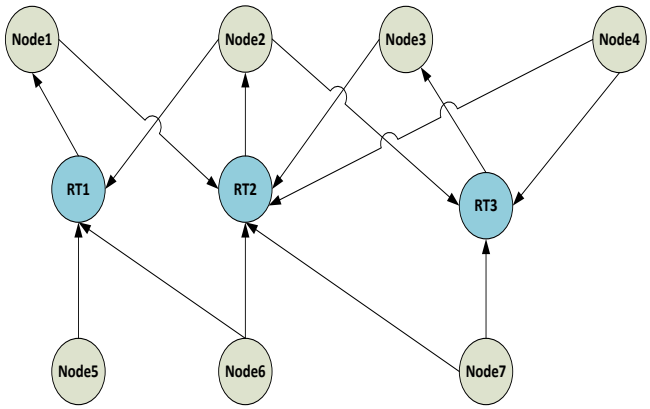


Figure 4.3: Repropagation graph of spread of information in OSNs.

The retweet graph enables us to view the messages as well as their sources together. The representation of the complete retweets in this form has the following implications.

- The complete propagation of tweets in the network could be identified. The graph would depict as to who 'infected' whom and the tweets involved.
- The graph would enable the use of standard PageRank algorithm [56] to calculate the general acceptance level of the user nodes and the message nodes. The rating of

the source nodes would depend on the rating of their tweets which in turn would be decided by the rating of the retweeter nodes.

- The disparity between the retweeters of a particular source node could be calculated using *gini coefficient* as would be explained subsequently.
- The possibility of collusion between users in propagating misinformation in the network by retweeting each other's messages could be identified using community detection and core identification algorithms. Frequent retweet behaviour between the nodes would result in formation of cycles of length 4 and the nodes would fall in the same community or core of the retweet graph.

A part of the repropagation graph from the 'egypt' data set is shown in Figure 4.4. An exploded view showing the two different types of nodes and edges between them in the 'bodhgaya' data set is shown in Figure 4.5. We used Gephi software for all visualisations of graphs in this work [102].

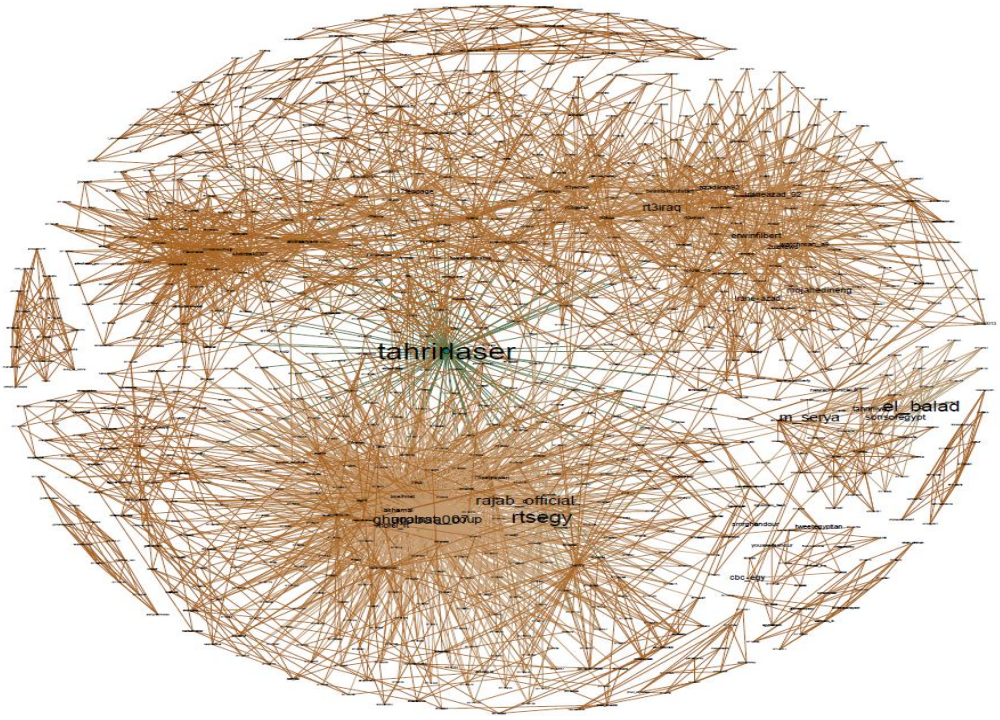


Figure 4.4: Repropagation graph of a part of the Egypt data set.

Gini coefficient as a measure of disparity

Gini index is a measure of statistical dispersion developed by the Italian statistician and sociologist Corrado Gini. It is a measure of inequality of a distribution. It is a ratio with values between 0 and 1. The numerator is the area between the Lorenz curve of the distribution and the uniform distribution line. The denominator is the area under the uniform distribution

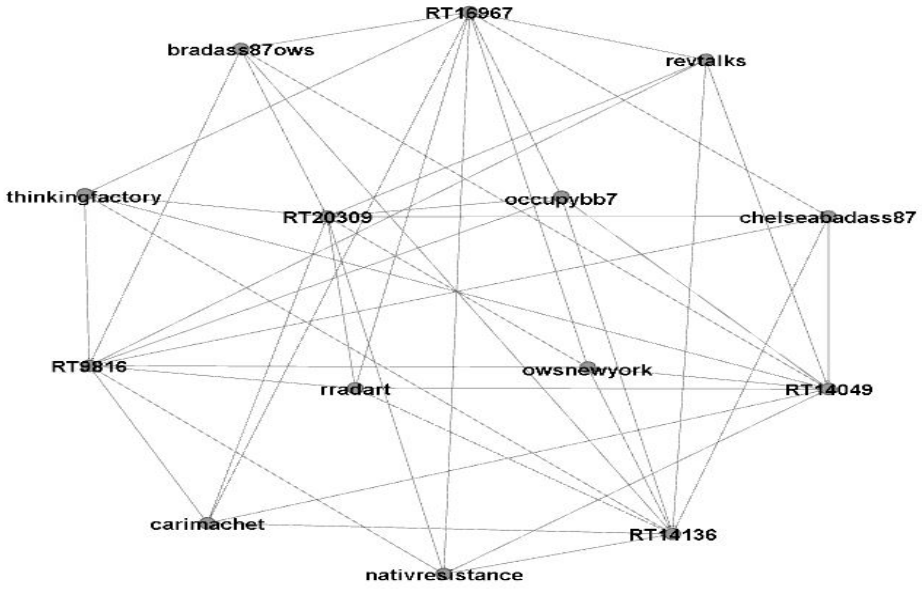


Figure 4.5: Detailed view of a section of the repropagation graph of Egypt data set showing the two types of nodes and edges between them. Nodes starting with RT are the message nodes and others are the user nodes.

line. Gini coefficient is often used as a measure of disparity in the distribution of any measurable quantity in a target population. Often used to measure the distribution of incomes in a target population, we measured how the retweets are distributed amongst the retweeters of a particular source node. We calculated the gini coefficient, G using the equation given below. Let X_k be the cumulative proportion of the population variable, for k=0,...n and X_0 = 0 and X_n = 1 and let Y_k be the cumulative proportion of the re-propagated news items, for k = 0,..,n and Y_0 = 0 and Y_n = 1. If X_k and Y_k are indexed such that $X_{k-1} < X_k$ and $Y_{k-1} < Y_k$, the gini coefficient, G is given by

$$G = 1 - \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k + Y_{k-1}) \tag{4.2.1}$$

The gini coefficients calculated for the four normal users shown in Figure 4.1 and the misinforming users in Figure 4.2 are given in Figure 4.6. The gini coefficients vary between 0 and 1. Perfect equality of distribution is denoted by gini coefficient of 0 and perfect inequality by 1. The actual values lie in between with normal users having a value closer to 0. A high value of gini coefficient for a source node would indicate greater difference in acceptance levels of its tweets and consequent reduction in credibility. The gini coefficients of misinforming sources had higher values as compared to the normal users and indicated their suitability for classification of sources.

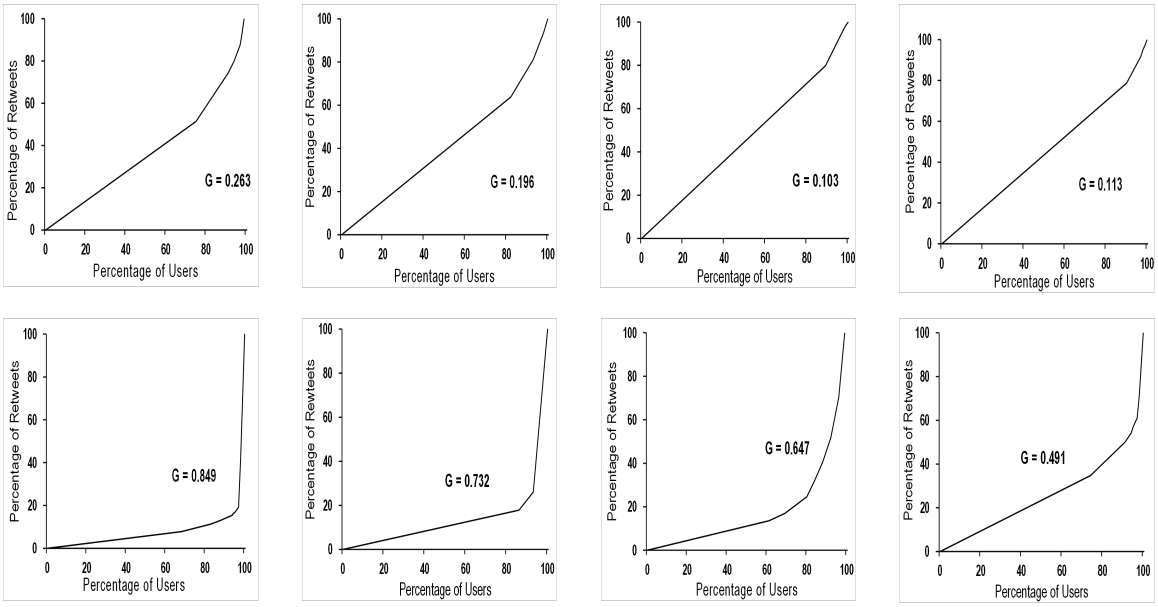


Figure 4.6: Gini coefficients for four sample users spreading genuine information (top set of figures) and for four sample users spreading disinformation (bottom set of figures).

Construction of a binary classifier

The classification of sources as credible or non-credible is based on a binary classifier. The binary classifier bases its decision on the threshold value of gini coefficient for the sources in the data set. This value is normally above 0.4. However, the exact value needs to be determined by evaluating the sources which have higher gini values. Following the heavy tailed distribution, these numbers are small. When we decrease the threshold values, false positives would increase. Similarly, an increase in threshold value results in some misinforming sources being left out. We arrive at a threshold figure by starting with the maximum gini value in the graph and slowly reducing them till we continue to get sources of misinformation.

Detecting sources who collude with each other could be done using patterns of communication. Even if one non credible source is identified, the others could be detected by using standard community detection algorithms based on modularity [103] in the retweet graph. The whole methodology is outlined in Figure 4.7. We present the algorithm of the binary classifier for detection of colluding nodes in Algorithm 1.

Validation of results

We validated the proposed methodology based on its ability to detect all the colluding nodes, i.e, True positives and the computation requirements for the same. The distribution of gini coefficients of degree distribution of all retweeting nodes of the source nodes in all the data sets being studied is given in Figure 4.8. The graph shows a heavy tailed distribution with

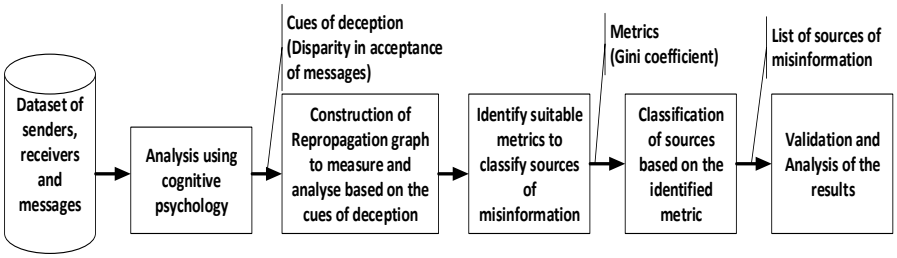


Figure 4.7: Analysing the process of diffusion using Sociology in conjunction with Cognitive Psychology.

Algorithm 1 Methodology for detection of colluding nodes to spread misinformation

Input: Details of users and messages (tweets) involved in the spread of information for specific 'context'
Preprocessing Step: repropagationgraph ← re-propagation bi-partite graph with user nodes and message nodes
sourcelist ← List of sources of messages in the repropagation graph
ginithreshold ← Threshold value of gini coefficient for disinforming sources (usually 0.4)
k ← length(sourcelist)
for (i = 1 → k) **do**
    sourcesubgraph[i] ← Subgraph of messages (tweets) and users retweeting them for sourcelist[i]
    Gini[i] ← Gini coefficient of distribution of degree of all user nodes only in sourcesubgraph[i]
end for
for (i = k → 1) **do**
    if (Gini[i] > ginithreshold) **then**
        potentialdisinformers[i] ← sourcelist[i]
    end if
end for
for (i = 1 → length(potentialdisinformers)) **do**
    colludingnodes[i] ← Interacting nodes with potentialdisinformers[i] in the same community using modularity based community detection algorithm
end for
Output: *potentialdisinformers:* List of potential disinformers in the data set.
Output: *colludingnodes:* List of colluding nodes for each of the potentialdisinformers.

few source nodes having higher gini coefficients. This also is intuitive in that the social network is mainly used by genuine users to spread information rather than misinformation.

The number of source nodes with gini coefficients above the threshold value (taken as 0.4 in our data sets) varied from 0.5% to 1.6% of the total users nodes. All sources associated with spread of misinformation had high gini coefficients, while all nodes with high gini coefficients were not spreading misinformation. The methodology could segregate potential sources of disinformation to less than 2% of the source nodes. The identification of colluding nodes using community detection algorithm [103] with these sources nodes enabled detection of all misinforming user nodes. We are interested in estimating the gini coefficients of communities only initially. The worst case complexity of the algorithm is proportional to the largest number of nodes in a community, which would always be much smaller than the total number of user nodes.

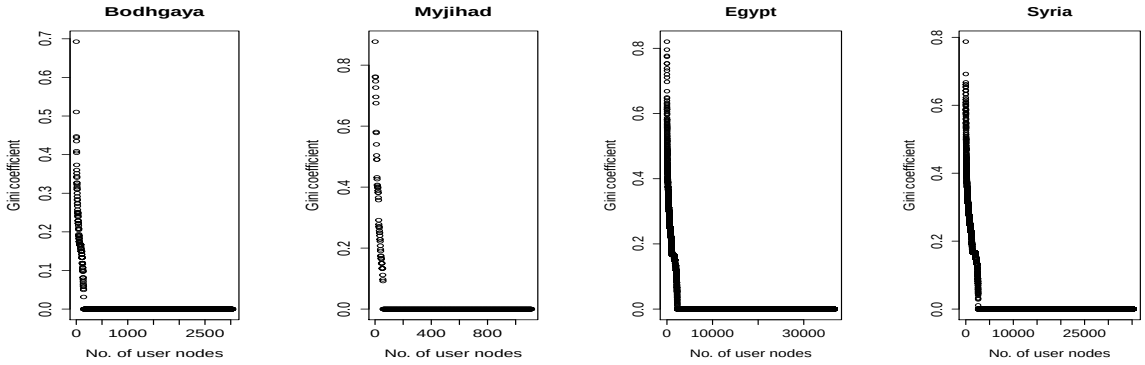


Figure 4.8: Distribution of gini coefficients in data sets.

Detection of communities of information propagation

In order to examine the propagation of news items as measured by the retweets in the graph, we used community detection algorithms based on modularity [103]. The distribution of communities for two data sets - Egypt and Syria are shown in Figure 4.9. As we see in the figure, the propagation of news items is marked by large number of isolated communities with very few communities having connections with others. So identification of communities where the misinformation may be spreading would indicate whether the spread is limited to a single community or it has spread across communities. The number of user nodes in the communities would also indicate the users who are 'infected' with the news items or 'susceptible' to them.

Based on principles of Cognitive Psychology and using gini coefficient as a measure of acceptability of a source, the methodology adopted by us identified the sources of misinformation rather than the messages. The efforts to detect the sources are more robust and scalable for implementation in large social media monitoring systems. Collaborative filter algorithms based on repropagation of messages should precede content based algorithms for effective cyber surveillance of spread of large scale misinformation. Having semantically analysed the messages in the data sets we classified efforts made to deliberately spread messages to influence other users which we call as semantic attacks in the next section.

4.3 Semantic Attacks in OSNs

The deliberate spread of false information is *disinformation*, which differs from *misinformation* only in terms of intent of the user. A coordinated group of individuals, effectively using the underlying concerns of users of social networks, can cause semantic attacks. The deliberate spread of such false information would result in wider spread amongst the users than otherwise possible.

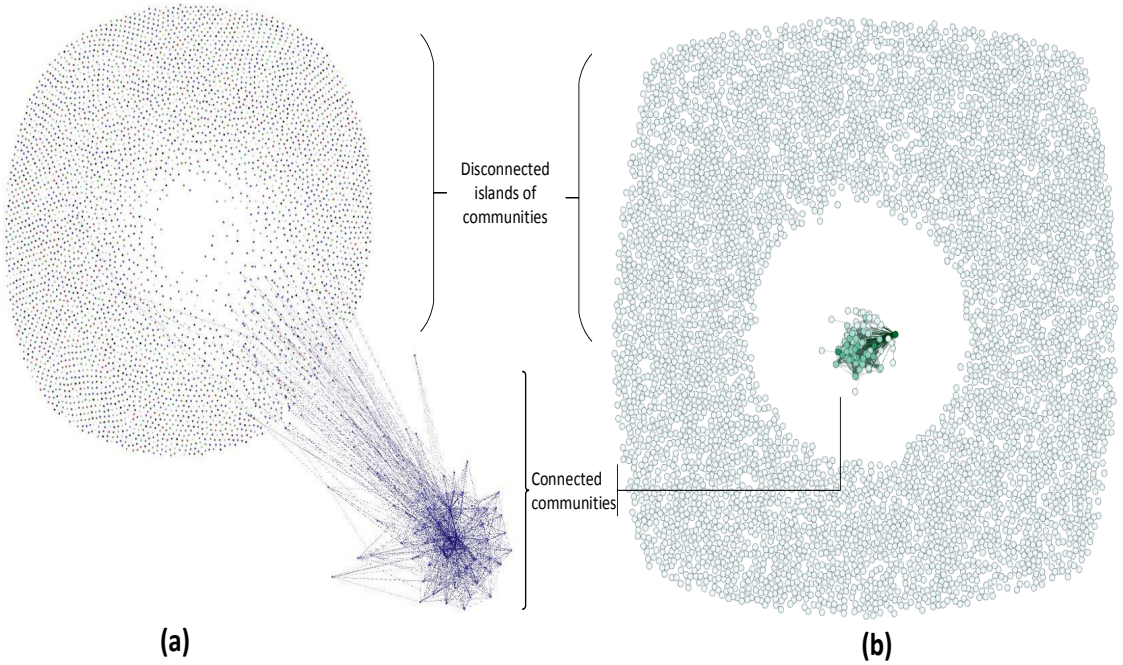


Figure 4.9: Distribution of communities in the (a) Egypt and (b) Syria data sets.

4.3.1 Classification of Semantic Attacks

We carried out in depth analysis of the data sets using automated network analysis algorithms and human annotation of data, with the aim of understanding the types of semantic attacks possible in them. The data sets studied represent a wide domain and include events which attracted lots of comments in Twitter. On analysis, we found that around 20% of messages in all the data sets relate to information which have no correlation with the event being discussed and the information in them could be classified as false information. Around 30% of information propagation in the data sets related to the event being discussed were making unsubstantiated claims and speculations. There are different types of attacks possible in OSNs. We consider semantic attacks which manipulate spread of information either by introducing a new set of users or creating multiple pseudo user profiles. Similar to the aim of such attacks in other social computing systems, semantic attacks in OSNs are aimed at manipulation of users' behaviour in a manner desired by the attackers. Understanding the intentions of sources of messages are important in classifying such attacks as given below.

1. **Disinformation attacks**. We define *Disinformation attacks* as the spread of disinformation, propaganda and lies by a source knowing fully well that the information is false or misleading. The intention can be assessed by the efforts made in propagation of the news item and possibly the relevance of information to the event being considered. In the analysis of data sets we found disinformation being propagated in two forms - Sybil attacks and Shill attacks.

- *Sybil attacks*. Sybil attacks are launched in social computing systems like recommender systems and reputation systems when a malicious user creates multiple 'pseudonymous' identities and use their combined power either to give false recommendations about items or false ratings in reputation systems [53]. When a malicious user of OSN creates multiple identities to ensure the spread of a news item, the intention of the source is deliberate spread of false or biased information. Such attacks to influence users of OSNs could be termed as Sybil attacks. In the data sets, we saw that mostly such news items have little relevance to the event being discussed.
- *Shill attacks*. Shill attacks in recommender systems are launched by a group of users who enter the system and give false opinions about items with the intention to mislead other users. These users are called *shills* and their opinions are used to either *nuke* the ratings of target items or *push* ratings of target items [104]. The injection of false profiles into the recommender system would result in modifying the ratings and consequent recommendations to the buyers. As in recommender systems, the users of OSNs act as shills, collude with each other to cause the spread of a desired news item. The coalition of malicious users could cause spread of false information. The ease with which false profiles can be created in OSNs make shill attacks an effective method to spread disinformation in OSNs.

2. **Misinformation attacks**. This attack involves the actual spread of misinformation related to the topic. Incomplete information and reporting stories as they develop by news agencies are examples of spread of misinformation. Even without any malicious intent, the credibility and acceptability of sources could result in large scale spread of false information. This could be in the form of unsubstantiated news, speculation or biased information about the subject. In the chain of spread of disinformation, most of the users would not have any intention to spread false information and hence would be spreading misinformation. Other examples of spread of misinformation were seen in all data sets [1].

4.3.2 Proposed Taxonomy of Semantic Attacks in OSNs

OSNs are susceptible to spread of disinformation and misinformation. Using detailed analysis of different data sets, we propose a taxonomy for semantic attacks in OSNs based on intention of source of messages and methodologies adopted for spread of information. The proposed taxonomy is given in Figure 4.10. We give details of each in the subsequent sections.

[1] https://twitter.com/naveentirthani/status/382706087452884993/photo/1

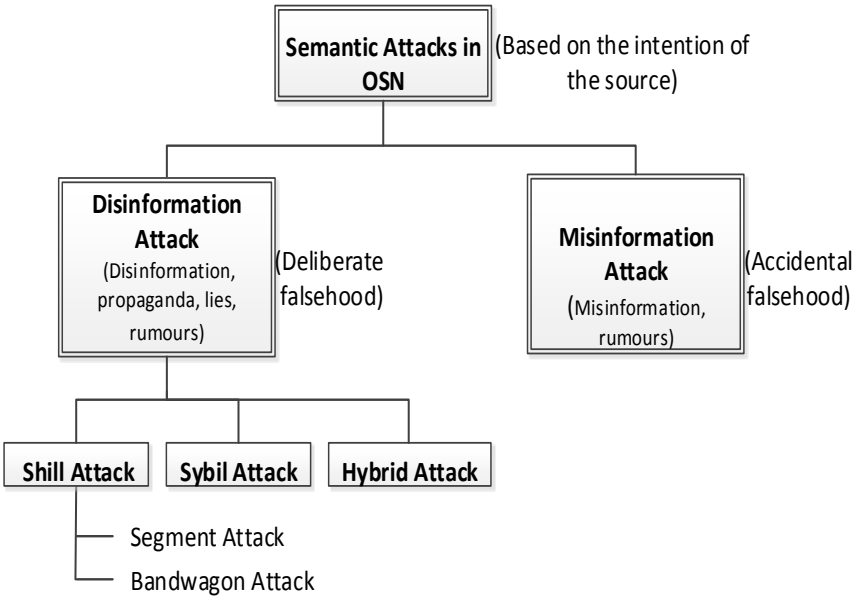


Figure 4.10: Taxonomy of semantic attacks in OSNs.

4.3.3 Sybil Attacks

In recommender systems, sybil attacks involve creation of multiple profiles by the same user to nuke or push up the recommendations of target items. Multiple profiles can be created in OSNs with the aim to propagate a target news item in the network. With URL shorteners available, the messages may seem different, but they would be pointing to the same web page. Detecting these false messages using content analysis of all the tweets would be computationally expensive.

The users carrying out sybil attacks would have greater communication links between them as they would be engaged in propagating similar messages. This would result in the formation of cores in the retweet graph. The detection of cores in a graph can be done using *k*-core decomposition algorithm [105]. *k*-core of a graph is a sub graph where all nodes have a degree *k* or more. We applied k-core decomposition algorithm to the day wise distribution of tweets in the Andhra data set. Figure 4.11 shows the inner most cores of Day 4 and Day 5 of the period of collection. The appearance of a set of core users increased the coreness of the Andhra data set from 5 to 9 on these days, indicating the extreme level of mutual communication between the new set of users. A visual analysis would also show that the names of all the users start with 'aum' and the retweets shown in the figure point towards a single URL, which was misinformation [2].

[2]http://pic.twitter.com/mLXgltMW36

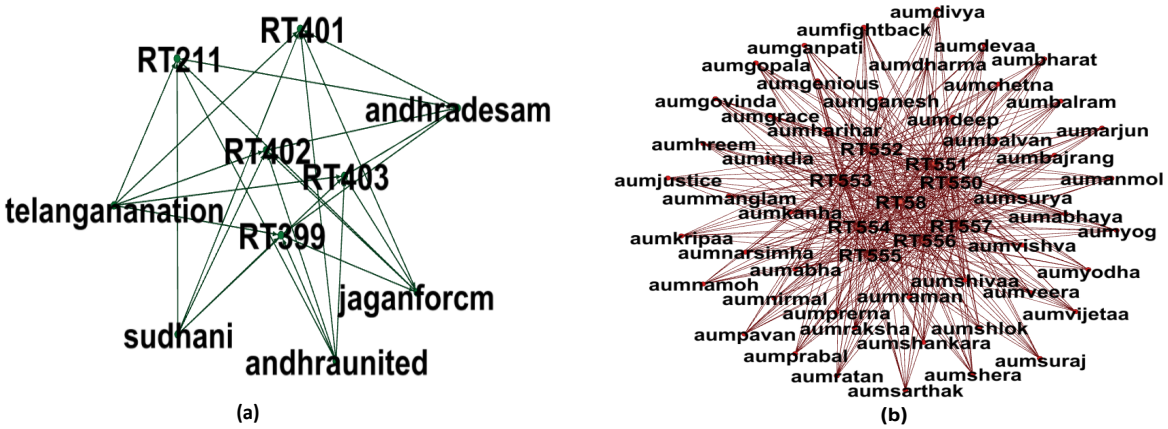


Figure 4.11: View of inner most core of the retweet graph of Andhra data set on (a) Day 4 and (b) Day 5 of the period of collection.

4.3.4 Shill Attacks

Shill attacks in recommender systems are carried out with the aim of increasing the ratings of target items [9]. The different types of shill attacks could include Random attack, Average attack [106], Bandwagon or Popular attack [107], Probe attack, Segment attack, Love/hate attack [108], Sampling attack, Perfect knowledge attack [109] etc. The ease of creation of user profiles and lack of sophisticated attack resistant systems make OSNs vulnerable to shill attacks. Manipulation of preferential propagation of news items in OSNs could be achieved by a limited subset of the attack types in recommender systems. We describe the more prevalent types in OSNs below.

Bandwagon Attack

Bandwagon or Popular attacks in recommender systems are carried out by creating profiles with high ratings to well known popular items and highest possible rating to the target item. The injected profiles can easily push predictions of the target items. The attack is easy to implement and are effective [107] [110]. In the *reverse bandwagon attack*, the aim is to nuke the products by giving low ratings to the least popular items and the lowest possible rating to the target item [108][51]. Bandwagon attacks in OSNs are carried out by clubbing false information along with messages and news items which are popular and of great interest to a large section of population. In this type of attack, news items are propagated by making certain keywords in the message similar to the ones trending at that point in time. In Twitter, *hashtags* are used to tag news pertaining to a single event or entity to form a common thread. The target news item which needs to be propagated in the network is referred in the message using a shortened URL and the rest of the characters permissible in the tweet are used to include all popular hashtags. The aim would be to get bracketed with all the top

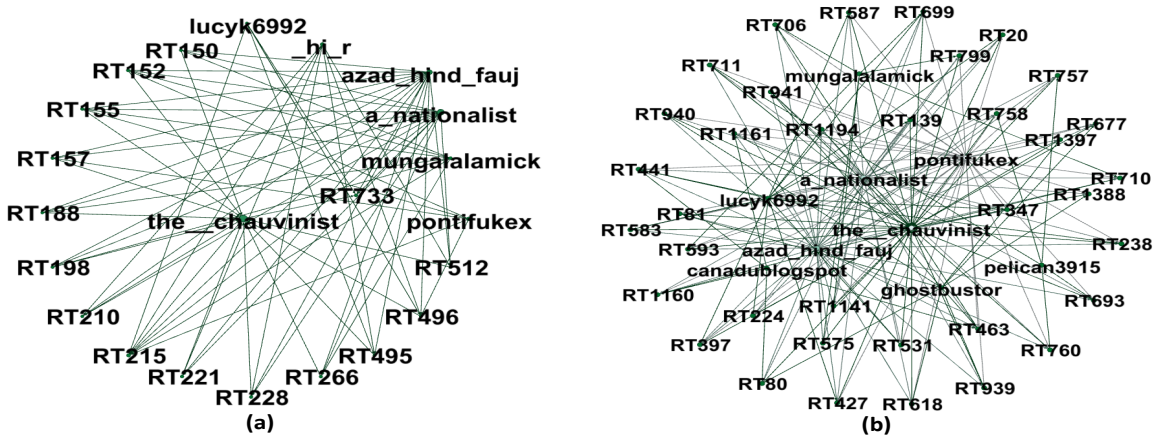


Figure 4.12: The tweets in the core of the (a) Bodhgaya and (b) MyJihad data sets are examples of Segment attacks.

trending topics and thus make it visible to a wider population and ensure its propagation. The promoted news item normally would have no similarity with the other items. The example of sybil attack shown in Figure 4.11 is also a type of Bandwagon attack [3].

Segment Attack

Segment attacks in recommender systems are targeted at a specific group of users who are likely to buy a product. In this, high ratings are inserted for those items which the users in the segment would like and low ratings for others. The similarity between the segment and injected profiles is high and the target item also gets more recommendations [51] [111].

In OSNs, this type of attack is targeted at a specific group or entity. Twitter sees lots of activity during crisis events like earthquakes, acts of terrorism etc. Often the handling of the event by the administration or acts by a community are commented upon by lots of users in social networks. Targeted attacks to change the perception for or against a section or community are often witnessed. Core analysis of the Bodhgaya and MyJihad data sets as shown in Figure 4.12 are examples of these type of attacks. The message nodes begin with 'RT' in their names. The users and the messages in both the data sets were similar in nature. Efforts were noticed to spread false information against a particular community to incite perceptions against them in the wake of bomb blasts at Bodhgaya [4].

4.3.5 Hybrid Attack

This combines other types of methodologies to carry out semantic attacks. In this, the attackers make use of Bandwagon and Segment attack methodologies coupled with Sybil

[3] https://twitter.com/aumsatya/status/381027667350269952/photo/1
[4] http://twitpic.com/aa0wg1

attacks to augment the authenticity of their messages and ensure more effective spread.

We have used terminology from other social computing systems to describe semantic attacks in OSNs. The classification enables us to identify different types of semantic attacks and help analyse intention of sources in spreading information. This would form the basis for a framework for early detection of deliberate spread of false information in an acceptable time frame. In the next section, the interactions between users to spread information is modeled as a game between the sender and the receiver and the diffusion process is further studied using evolutionary game theory.

4.4 Understanding Diffusion of Information using Evolutionary Game Theory

There are different types of information diffusion models used to describe information diffusion in networks as described in Section 2.2.7. We are interested in information which has the potential to spread and overtake the complete population or at least major part of the population. The formation of information cascades is the result of decisions of a series of users to forward information that they receive in the network. The users' decision to forward messages would depend on four different factors, which determine the credibility of information. The study of spread of innovations and the classification of users based on their time of acceptance of innovation have been done [78]. The applicability of the same to study information diffusion in OSNs has been proposed in [79]. The user based modelling is what we also attempt using evolutionary games played by users in a structured population.

The decision making processes of human beings are not always strictly rational. Traditional game theory is based on rationality of people. In the case of misinformation or disinformation, the decision taken in situations of uncertainty, may not reflect rational behaviour. The dynamics of games played is also not reflected in traditional game theory. The extensive form of games tries to bring in dynamicity to some extent. But the lack of perfect information and large number of information sets, make it very difficult to model games of even medium complexity using traditional game theory. Traditional game theory also fails when the behaviour of a player changes based on what he learns from seeing others play. Further, the extensive form of the game gives strategies only at decision points corresponding to selections made prior to the game play. Evolutionary game theory naturally brings in the dynamic element and hence is considered more appropriate to model human behaviour while studying information cascades.

Evolutionary game theory enables application of mathematical theory of games to biological contexts. This is based on the concept that *frequency dependent fitness* introduces a strategic aspect to evolution. Evolutionary game theory has interested social scientists because the 'evolution' need not be biological, but it can be cultural also, which refers to the change in beliefs and norms over time [112]. We used evolutionary game theory for the

decision making process due to its simplicity as well as its ability to capture the learning process of human beings. The three things which favour evolutionary game theory are - the ability to model non-rational behaviour, dynamic nature of decision making and ability to model cultural evolution.

Libicki says that prior beliefs or opinions inherent in people affect how they interpret information. When people believe something to be true, they are more likely to believe something that supports their inherent beliefs [24]. This may lead people to believe misinformation. The individuals have their own beliefs and exhibit their behaviour based on them, which are inherent in them. However, when they are exposed to a different behaviour they may adopt the new one, provided it conforms to their inherent beliefs. They are also influenced by the opinion of those around them. Once, the behaviour is changed, they continue to exhibit the new behaviour. It is in this context, we propose to use evolutionary game theoretical model to explain the spread of new information to include misinformation also in social networks.

Evolutionary graph theory studies the ability of a mutant gene to overtake a finite structured population [82]. Information diffusion in structured population has been studied using evolutionary game theory in [83]. The authors have proposed two player game strategies of forwarding information or not forwarding the information. The authors have considered new information as a 'mutant' and the process of information diffusion as spreading of mutant gene. The diffusion of information is modeled as evolutionary games with strategies of players that are updated using predefined update rules.

4.4.1 Modelling New Information as Mutants

Evolutionary game theory was developed for explaining biological phenomenon on the basis of mathematical theory of games. It was developed to explain the spread of mutants in a target population. The application was based on the fact that the type of biological agents impart a strategic aspect to evolution. The evolutionary framework could explain how certain mutants are able to invade a whole population though starting with relatively small numbers. Using this model, we treat the new information as a mutant and explain its spread in the population. The nodes forming the social network are hard wired to behave in a particular manner, rather than being rational. When a mutant invades the target population, the spread of the mutant is based on the relative success it achieves in making inroads into the population.

Evolution in the context of the game being described here, refers to the cultural evolution. This evolution captures the change in belief and acceptance of a new norm when exposed to a 'mutant'. The 'mutant' in this case is the new information to which the population is subjected to. The population would adopt a strategy which would increase its fitness. The fitness, we would refer to is its reproductive ability. For the model, it would indicate the decision of each member to adopt the 'mutant' strategy, thereby increasing the population of the 'mutant', or continue with the existing strategy of remaining unaffected.

The adoption could be measured in terms of decision to forward or not forward information. In the latter case, the strategy to remain unaffected, would indicate that the population of users with 'mutant' strategy and others have not changed after the interaction.

The pairwise interaction between the mutants and the existing population is modeled as a two player game. The payoffs to the agents are frequency dependent on the number of other agents or mutants they are interacting with. This would result in change in the type of strategies adopted by the nodes in the population.

4.4.2 Evolutionary Stable Strategy and Evolutionary Dynamics

There are primarily two approaches to study evolutionary game theory. The first approach uses Evolutionary Stable Strategy (ESS) as the principal tool of analysis. The second model studies the process by which the frequency of strategies change in the population and properties of evolutionary dynamics within that model. The first is more 'static' in nature leading to ESS without looking at the underlying processes leading to evolution and hence changes in behaviour or strategies of population. The second approach looks at modelling the population dynamics taking into account the standard stability concepts used in the analysis of dynamical systems.

The players modify their strategies over time based on replicator dynamics. The game is played repeatedly. For the replicator dynamics, the state of the population is denoted by a vector s_i, where s_i denotes the frequency of the strategy *i* in the population. The replicator dynamics allows us to model how the distribution of strategies changes over time. Replicator dynamics assuming pairwise interaction equally likely between any two individuals in the population would be different when the same is not true. In OSNs, this assumption would not hold good. In such cases correlated interactions can occur. The correlation present in the interactions would determine the eventual outcome. It is also assumed that the initial conditions of the state of the population is such that all conditions are equally likely. As the amount of correlation increases, the evolutionary stable strategy is bound to evolve faster. Hence deciding on the replicator dynamics accurately is important to model the evolutionary process.

In order to apply the 'utility' value in the traditional game, we need to define 'fitness' value of the population, an objective function which should reflect the cultural evolutionary fitness of people. We assume strategies with higher payoff will spread through the population - by learning, copying, inheriting or infecting [113]. The payoffs depend on the frequencies of the strategies in the population. The frequencies will change according to the payoffs. This would yield a feedback loop. This dynamics is the object of evolutionary game theory. The feedback depends on the population structure, on the underlying game and on the way strategies spread. The strategies spread through social interactions which are determined by the underlying graph structure, which restricts interactions to neighbouring nodes only. The evolution of strategies in this manner is studied using concepts of evolutionary graph theory.

4.4.3 The Information Spread Model

The dynamics of evolution in a finite well mixed population of N individuals is studied in terms of *Moran Process* [114]. The stochastic process defines the reproductive mechanism, where at each time step, one individual is randomly chosen to reproduce and another one is chosen to die and be replaced by a duplicate of the first one. In this manner, the total population is maintained the same. All the resident members are generally considered having the same *fitness* level of 1. A mutant with a relative fitness level *r* is introduced into the population. *Fixation Probability* is defined as the probability that the mutant will overtake the complete population. Evolutionary graph theory allows us to study the population dependent success of game theoretic strategies in a structured population. We assume that the population remains constant during the period of spread of new information.

4.4.4 Evolutionary Graph Theory

Evolutionary graph

The dynamics of population growth is described using replicator equations when the population is well mixed. However, most of the time, as in OSNs, the interactions are not random. The structure of the population is modeled as a directed, weighted graph called an *evolutionary graph*. The study of evolutionary graph theory was initiated by Lieberman et al [114]. The dynamics of population growth when modeled in this way is quite different from a well mixed population. The application of game theoretical aspects to evolutionary graph theory has given new insights into cooperation in human behaviour [82]. Evolutionary graph theory studies the ability of mutant gene to completely affect a finite structured population. The reproduction process is modeled as a stochastic process. Using evolutionary graph theory, we model the spread of new information as a mutant gene in OSNs which would enable us to use the theory of population dynamics to answer important questions of its spread in terms of time and size.

The users of the social network graph are the vertices of an *evolutionary graph*. For two vertices v_i and v_j, the directed, weighted edge from v_i to v_j has a weight w_{ij}. In an evolutionary graph, w_{ij} represents the probability with which if v_i is selected to reproduce then it replaces v_j. For any given v_i, $\Sigma_j w_{ij} = 1$. In an evolutionary graph the nodes are replaced based on their fitness values. *Neutral drift* is the process of reproduction in an evolutionary graph, in the special case where the mutant introduced has the same fitness *r* as the resident population, i.e, $r = 1$. If a mutant is introduced which has a greater fitness value than the resident population, then the probability of its spread is higher. The *Fixation probability* is an important aspect to be considered.

Fixation probability

Fixation Probability is defined as the probability that a mutant gene can overtake a complete population of N individuals and is defined for *Moran process* as ρ_1, where

$$\rho_1 = \frac{1 - 1/r}{1 - 1/r^N}$$

A population with lower fixation probability is more resistant to invasion by a mutant and can be considered more evolutionary stable [82]. The result has been proved for a wide variety of evolutionary graphs where $\forall v_i$, $\Sigma_j w_{ij} = \Sigma_j w_{ji}$ in [114]. When evolutionary graph theory is applied to game theory, the evolutionary fitness f_i of an individual v_i is related to their game theoretic payoff (P) by the equation [82].

$$f_i = 1 - w + w.P$$

The parameter 'w' is used to relate the fitness and payoff obtained from games. If w = 1, the fitness and payoff obtained are the same. If w = 0, it is neutral drift as the payoff has no role to play. We calculate the information spread in two phases using each of the values in our model. The other values of *w* could be used to vary the weightage of either of the two parameters.

Update rules

The evolutionary game model should also specify the replicator dynamics or as in the case of evolutionary graphs the update rules. The update rules specify how the strategies spread in the population. There are broadly three kinds of pre defined update rules possible which decide the evolutionary strategies of the population. They are - Birth-Death (BD), Death-Birth (DB) and the Link Dynamics (LD) update rules [82]. They are briefly described below.

- *Birth-death (BD) update rule*. A vertex v_i is selected for reproduction based on its fitness and then a neighbor v_j is selected at random for replacement. v_j is replaced by a duplicate of v_i.
- *Death-birth (DB) update rule*. A vertex v_i is selected randomly to die. A neighbor v_j is selected based on its fitness values. A duplicate of v_j replaces v_i.
- *Imitation (IM) update rule*. A vertex v_i is chosen for updating its strategy. The vertex v_i adopts a strategy of one of its neighbors or remains with its own strategy based on its fitness values.

The three update rules are based on 'fertility selection'. The payoffs affect the reproductive success of the nodes. The cultural reproduction takes place based on the above rules. The strategies or beliefs held by the individuals change to conform to their surroundings. We

consider that the strategies themselves do not mutate. But we investigate the circumstances under which a resident population is affected when exposed to mutants having different strategies.

4.4.5 Modelling Spread of Information

Cooperator defector game

The game of cooperator and defector is a well studied model in game theory. The spread of new information can be modeled as a problem of cooperation between the agents. The agents can either cooperate or defect. In this case, *cooperate* means accepting the mutant information and further forwarding it and *defect* means not accepting the information. In a well mixed population the defectors win and the mutant new information becomes extinct, especially when the credibility of the new information is low. This game represents a Prisoner's dilemma between the two players. However, in an OSN graph, where the players engage in an iterated Prisoner's dilemma with all their neighbors, the cooperators can win due to the principle of *network reciprocity* [115]. The payoff matrix for the game is shown at Table 4.2.

Table 4.2: Payoff matrix for cooperator defector game.

	cooperate	**defect**
cooperate	b-c	-c
defect	b	0

Here *b* is called the benefit of the act of cooperation and *c* is the cost of such an act. In this game, the cooperator incurs a cost of *c* for each of its neighbors and earns a benefit of *b* for each node which accepts and become a cooperator. A *cooperator*, who has accepted the new information (mutant gene), when connected to *p* cooperators and *q* defectors gets the following payoff:

$$Payoff = b \,.\, p - c \,.\, (p + q)$$

Benefit to Cost ratio

The benefit to cost ratio determines whether a node would accept the new information if a certain proportion of its neighbours has accepted the same. Let us consider the effect of b/c ratio for each of the update rules. ρ_C and ρ_D are the fixation probabilities of the cooperators and defectors respectively,

- For BD update rule, for any value of *b* and *c*, the fixation probability $\rho_C < \rho_D$ and $\rho_C < 1/N$. The cooperators are never favoured and they become extinct [116].

- For DB update rule, the fixation probability $\rho_C > \rho_D$ and $\rho_C > 1/N$, if $B/C >$ k, where k is the mean degree, B is the total Benefit and C is the total Cost.
- For IM update rule, the fixation probability $\rho_C > \rho_D$ and $\rho_C > 1/N$, if $B/C >$ k + 2, where k is the mean degree.

The BD update rule never favours the cooperator. As per this update rule, the cooperators would become extinct and defectors would prevail. Hence new information would not spread in the population. For the mutant to invade the resident population, we use either DB update rule or IM update rule. The equations give a necessary condition for a mutant to invade a resident population, but not a sufficient condition. In a cooperator defector game, a node would adopt a strategy which would give it maximum benefit. Hence, the B/C ratio for each vertex would give an indication of the likelihood of spread of new information. For any vertex the following holds good for calculation of B/C ratio. If the number of cooperators are *p* and the number of defectors are *q*, the total benefit to cost ratio for the node is given below:

$$Benefit/Cost(B/C) = [(p\ .\ b)]/[(p+q)\ .\ c]$$

Structure of population

Based on the perceived values of b/c ratio, we can categorize the population into three types.

- *For type*. For them the value of b/c is very high, say MAX, with the result that irrespective of the ratio of p/(p+q), the total Benefit to Cost ratio (B/C) would be always greater than 1 and they would always forward the new information. Here b/c $>>$ p/((p+q).
- *Against type*. The value of b/c is 0. They would not forward the new information under any circumstances, as probably they understood it as non credible information or misinformation or they were not active in the network during the time of spread. Here the ratio p/(p+q) is irrelevant.
- *Neutral type*. The value of b/c ratio varies between the maximum value, MAX and 0. Hence the Neutral types would accept new information only when the ratio of p/(p+q) is favourable and makes the product p/(p+q) * b/c > 1. In other words, the ratio p/(p+q) would play a decisive role in mutation of the *Neutral* type to the *For* type. They might mutate to *Against* type also.

The model requires specifying the structure of population in terms of the proportion of *For, Against and Neutral* types. This is a valid assumption as some knowledge about the vulnerability of the population is important to study the effect of semantic attacks on them. This information could be based on studies conducted elsewhere or the reaction of the population to similar attacks in the past.

4.4.6 Information Propagation

The initial structure of the population is specified in terms of *For, Against and Neutral* types. We model the spread of new information in this population by introducing a *mutant* with a certain fitness level *r*. The value of *r* would vary for different types of information. We use two predefined methods by which the mutant can spread through the network based on the work done in evolutionary graph theory [82].

Neutral Drift. During the neutral drift phase, the new information spreads between nodes of equal fitness, i.e, the *For* types. If the ratio of *For* types is large in the population, there is a possibility of large number of people getting infected. The number of infected population at the end would be a function of the network structure, initial *For* types in the population and the percentage of initial infected population. This stage is similar to acceptance of information by innovators and early adopters who base their decisions of adoption on their own beliefs.

IM Update. This is the phase of spread by frequency dependent selection of strategies. The players adopt a strategy based on the response of others in the population. This is the phase of *fertility selection*, where payoffs of the game affect the reproductive success of players. As more and more *For* types get infected, the product of ratio of cooperators in the population and the b/c ratio would become greater than 1 for any node and the mutation of the *Neutral* type to *For* type would happen as per IM update rule. This would further increase the number of *For* types. This is the phase where mutation happens and using the IM update rule the vertices exhibiting less evolutionary fit strategy, like the *Neutral* type will adopt the more fit *For* type strategy or possibly the *Against* type strategy. The fitness is defined in terms of *reproductive fitness*, indicating the ability to contribute towards subsequent spread and if the total infected population increases more than a defined critical mass, the rate of spread would become exponential and affect maximum population. This phase signifies the spread of new information due to the inherent nature of the individuals to conform to the opinions and beliefs held by the society. A node will adopt a different strategy if it gets a favourable B/C ratio.

Another possible mutation is when the *Against* types play a more active role. This would happen when the new information from the *For* types is against a particular entity, a community, a person or any such belief which invites a reaction from the *Against* types. In a way, this works in the lines of Newton's law of every action eliciting an equal and opposite reaction. The mutation of *Neutral* types to *Against* types can also occur based on the b/c ratio of adopting a more fitter strategy. The social computing phenomenon of the OSNs is seen here which has reacted in a manner to counter the spread of misinformation. For a large proportion of new credible information, we may not have this type in the population.

The above process can be expressed in terms of a Payoff matrix given in Table 4.3. The payoff matrix gives reproductive fitness of the strategies for every pair wise interactions in spreading new information. The value in each cell gives the availability of the player displaying the strategy for the spread of new information in the next time interval.

Table 4.3: Payoff matrix for evolutionary games.

		Player2 : Node2		
		For	**Neutral**	**Against**
Player1: Node1	**For**	1,1	1,f_1	1,-1
	Neutral	f_1,1	0,0	f_2,-1
	Against	-1,1	-1,f_2	-1,-1

The value of f_1 is defined in Equations 4.4.1 and 4.4.2. If *m* is the actual proportion of *For* types amongst the total number of neighbors, i.e, the actual p/(p+q) ratio defined earlier, the fitness value *f1* of the payoff matrix will change depending on the value of *m* and the ratio b/c.

$$f_1 = 0, \text{ if } m \text{ . b/c} \leq 1 \tag{4.4.1}$$

$$f_1 = 1, \text{ otherwise} \tag{4.4.2}$$

Similarly for f_2, the value is defined in Equations 4.4.3 and 4.4.4. The defectors, *q* would consist of both *Against* types and *Neutral* types. If *n* is the actual proportion of *Against* types amongst the total number of neighbors, i.e, the ratio of s/(p+q), where s is the number of *Against* types, the fitness value f_2 of the payoff matrix will change depending on the value of *n* and the ratio b/c. We indicate the mutation by assigning a value of -1 for *Against* types, as they will definitely not available in the next time interval for spreading new information. In fact, they could play a role in limiting spread of misinformation.

$$f_2 = 0, \text{ if } n \text{ . b/c} \leq 1 \tag{4.4.3}$$

$$f_2 = -1, \text{ otherwise} \tag{4.4.4}$$

The overall extent of spread of new information depends on initial state of *For* types in the population and number of *Neutral* types who have mutated to *For* type. The structure of the network graph has also an important role to play. In this model of spread of new information, we assume that the *For* and the *Against* types do not mutate at all. This is a fair assumption, as views held by them are quite strong and changes in them for the duration of spread is less. We also assume that users maintain their strategies and behave as per their phenotypes in subsequent interactions in the OSN graph.

4.5 Integrated Model

The initial model proposed in the previous chapter using Cognitive Psychology could be extended to include the diffusion of new information using Sociology and evolutionary game theory. Modelling diffusion of new information using evolutionary game theory is depicted in Figure 4.13. Updating of strategies using evolutionary graph theory is depicted

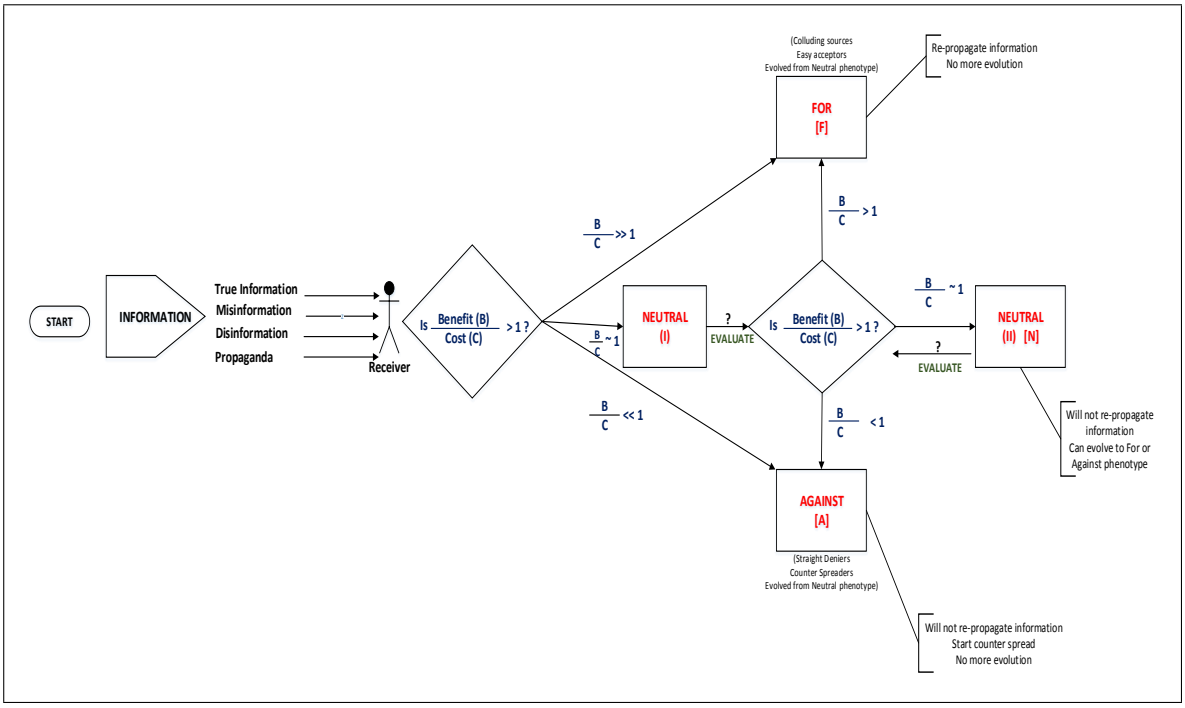


Figure 4.13: Modelling diffusion of misinformation using evolutionary game theory.

in Figure 4.14. We combine the above models with the Cognitive theory model proposed in Figure 3.3 to give an Integrated model for spread of new information in Figure 4.15.

The Integrated model brings out the fact that the individual adoption process of information by an individual is reflected as a group phenomenon when people interact in OSNs. The changes in behaviour is dependent on the type of neighbours and hence we have frequency dependent updating of strategies in evolutionary game theory. The nodes are not subjected to homogenous mixing and hence mutants can survive and infect a section of the population. The presence of different strategy profiles and the quantum of initial percentage of population required to cause new information to spread has been studied and analysed in the next chapter.

The proposed model has support from other works in literature where spread of rumours have been modeled. In [117], the authors have proposed a SEIZ model (Susceptible, Infected, Exposed, Skeptic) to model rumours and news diffusion in Twitter data sets. In the Susceptible(S) state, the individuals have not yet heard of the news item. Individuals in Susceptible(S) state get Infected(I) when they accept the message they receive and stay infected. In the Exposed(E) state, the individuals take some time before they accept the information and get infected. Skeptic(Z) state represents an individual who has heard about the message, but chooses not to repropagate it. The corresponding states for the above model in our Integrated model is represented in Figure 4.15. The users in 'S' state transition to the 'For' [I] state or 'Against' [Z] state. They may also chose to remain 'Neutral' [E] state and further transition to 'I' state or 'Z' state. Our contention is that the individuals

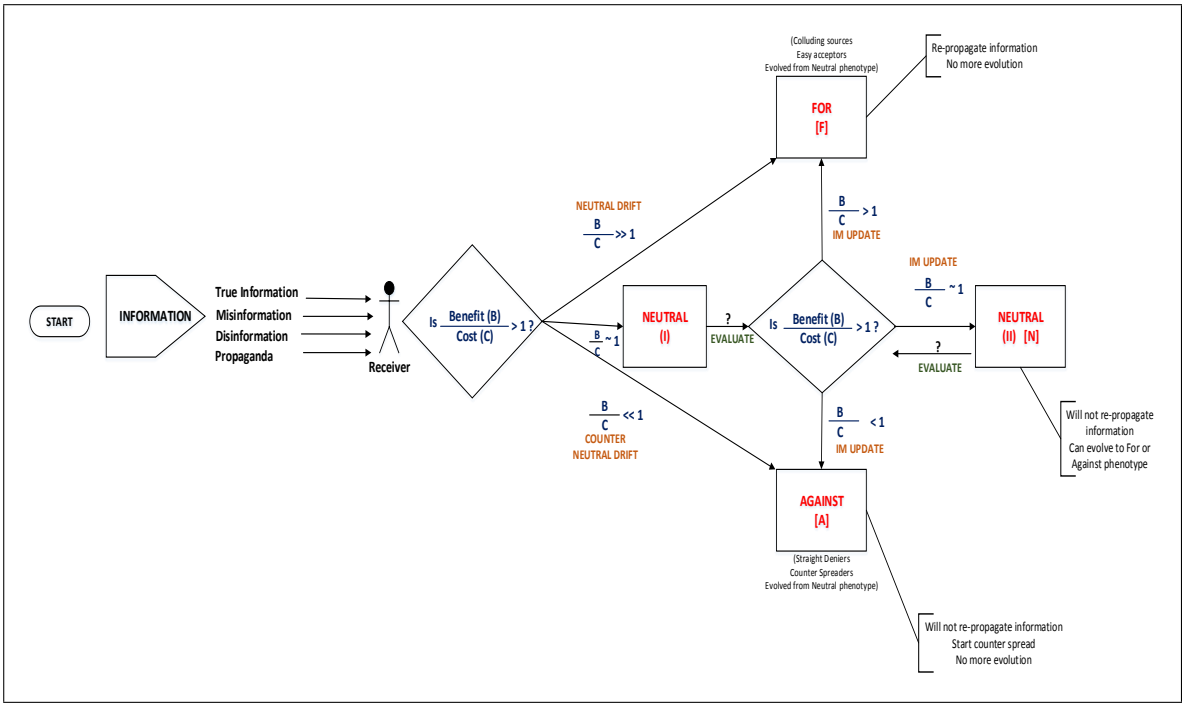


Figure 4.14: Updating strategies using evolutionary graph theory.

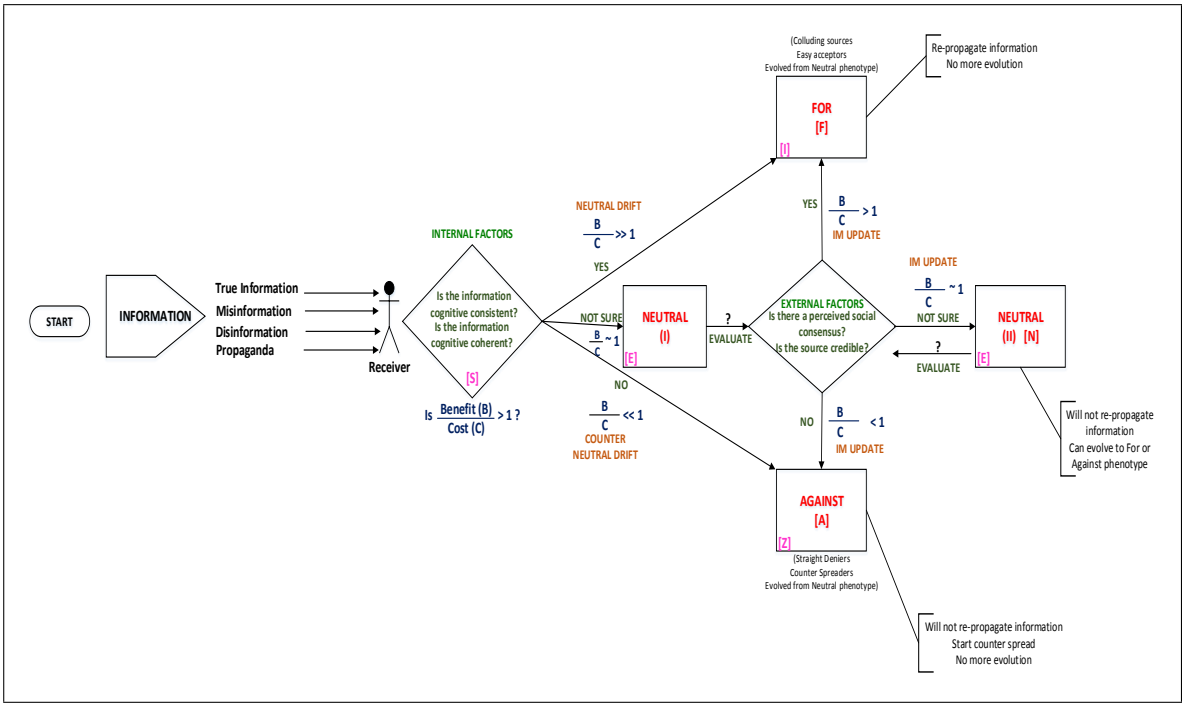


Figure 4.15: Integrated model depicting diffusion of new information using principles of Cognitive Psychology, Sociology and evolutionary game theory

in ‘Z’ compartment may react against stories, if they believe them to be completely false. This could be in the form of propagation of new messages countering the news items they receive. The study of propaganda during elections has shown that political parties and their leaders are often quick enough to rebut stories propagated by opponents.

4.6 Summary

In this chapter we have analysed diffusion of information and the differential acceptance of it based on its credibility. We used gini coefficient to measure this and detect sources of misinformation. The intention of sources and methodology adopted to deliberately propagate certain kind of information in OSNs were used to develop a taxonomy for semantic attacks in OSNs. Further, the interactions between sender and receiver of information were modeled as an evolutionary game. Combining inputs from Behavioural Sciences and Computer Science to analyse propagation of new information in OSNs, we have proposed an Integrated model of information diffusion. The Integrated model depicts their convergence for analysis of spread of new information including misinformation in OSNs. The profiling of users has similar basis and explanation in all the methods.

The results of analysing diffusion of information using Cognitive Psychology have been published in [Pub2] and [Pub6]. The taxonomy of semantic attacks has been published in [Pub3][Pub7]. The use of evolutionary game theory to model information propagation has been published in [Pub8]. In the next chapter, we would use data from real world data sets and simulation to carry out analysis and validation of our model.

Chapter 5

Process of Analysis of Semantic Attacks

"A rumor is a social cancer: it is difficult to contain and it rots the brains of the masses. However, the real danger is that so many people find rumors enjoyable. That part causes the infection. And in such cases when a rumor is only partially made of truth, it is difficult to pinpoint exactly where the information may have gone wrong. It is passed on and on until some brave soul questions its validity; that brave soul refuses to bite the apple and let the apple eat him." Criss Jami

5.1 Introduction

In this chapter we have carried out the analysis of different data sets based on the proposed models. We used theoretical concepts from evolutionary graph theory to analyse evolutionary games played by users. We carried out Psychometric analysis of users in the communities formed in the bi-level evolutionary graphs to prove their social computing properties to detect non credible information and development of trust between users. We hhave developed a Behavioural trust model to segregate communities to be monitored and also evaluate credibility of sources. The results are validated using synthetic data sets as well as real world data sets obtained from Twitter and Facebook.

5.2 Integrated Model

The convergence of profiling of users based on Cognitive Psychology and evolutionary game theory for studying the spread of new information is depicted using an Integrated model in Figure 5.1. The profiling of population along with decision boxes representing the modelling based on Cognitive Psychology and the profiling based on evolutionary game theory is shown together. The intersection of the two models and further their representation in the *Replacement evolutionary graph* enables using graph analysis algorithms to detect sources of information whose messages could cause information cascades in the network.

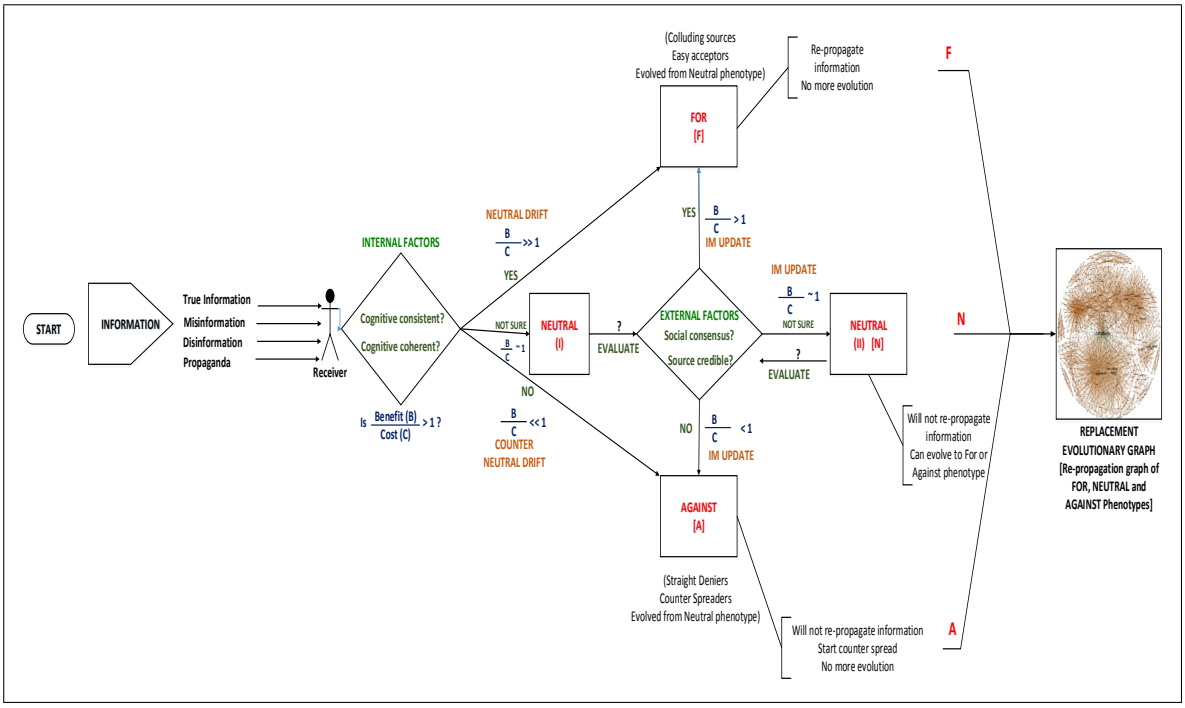


Figure 5.1: Integrated model of profiling population and its representation in replacement evolutionary graph.

5.2.1 Replicator Equations for Games on Evolutionary Graphs

Analytical results have been obtained for replicator equations of evolutionary dynamics for regular and other different types of random graphs including scale free graphs under neutral drift and evolution of cooperation when players are involved in playing Prisoners' dilemma [116]. There is a lack of exact analytical results for deriving replicator equations and fixation probabilities for random, heterogenous graphs. However, the basic principles for evolution for cooperation, which in our case is the spread of new information, would remain the same. An excellent review of use of game theoretical aspects of evolutionary graph theory is given in [82]. We would outline the results for regular graphs and analyse their importance in applying to OSN graphs. We have used extensive simulation to validate the models in different real world and synthetic data sets.

In a cooperator-defector game the payoffs for different strategies are given by Table 4.2. Consider an Interaction graph, H with degree h, Replacement graph, G with degree g, with an Overlap graph, L between the two with degree l. The individuals play games with all their h neighbours in the Interaction graph H and their strategies are updated in Replacement graph, G. These interactions decide the total payoff for each player. Let us denote the payoff of each player by P. The fitness, $F = 1 - w + w.P$ [82].

The value of w determines the relative contribution of the game to fitness. If $w = 0$, all players have equal fitness and the game does not affect the fitness as would happen in case of neutral drift. When $w = 1$, it is strong selection. Payoff due to the game is equal to fitness.

Generally, analysis has been carried out in case of weak selection, when w $<<$ 1.

In case of Prisoner's dilemma when studying interactions between cooperators and defectors, the cooperators would benefit only when benefit, b $>$ cost, c. For IM update, the condition for the same is b/c $>$ h(g+2)/l. If ρ_C is the fixation probability of cooperators and ρ_D is the fixation probability for defectors, for the mutant corporators to spread, the condition required is $\rho_C > \rho_D$ [116].

The equations which would define the replicator dynamics of spread of misinformation would be different in a well-mixed population as compared to a population structured in the form of graphs. Let us define the replicator equations for a well mixed population. In a finite size population of **N** individuals, we consider a game with *n* strategies. Let the payoff matrix be given by **A.** a_{ij} is an element in the matrix **A** and it denotes the payoff for strategy *i* when played against strategy *j*.

If f_i is the global frequency of strategy *i*, then $\sum_{i=1}^{n} f_i = 1$. f_i denotes the frequency with which strategy *i* plays against strategy *j* in a well-mixed population. The average payoff for a person playing strategy *i* is given by $\sum_{j=1}^{n} a_{ij} f_j$ = $\mathbf{e}_i$.**Af** [116]. Here $\mathbf{e}_i$ is the *i*-th unit column vector, **f**=(f_1,...,f_n). The change in proportion of population playing strategy *i* is given by the replicator equation for a well mixed population as

$$\dot{f}_i = f_i[\mathbf{e}_i.\mathbf{Af} - \mathbf{f}.\mathbf{Af}] \tag{5.2.1}$$

$$\mathbf{f}.\mathbf{Af} = \sum_{i=1}^{n} (\mathbf{e}_i.\mathbf{Af}) f_i = \sum_{i,j=1}^{n} a_{ij} f_i f_j \tag{5.2.2}$$

The Equation 5.2.2 gives the average payoff of the population. The relation indicates that the change in f_i is the product of f_i and the difference between the average payoff of a person playing strategy *i* and the average payoff of the population. This also leads to the conclusion that strategies which are fitter tend to increase their proportion in the population at the cost of others.

The structuring of population in the form of graphs would change the probability of strategy *i* meeting strategy *j* not equal to f_i. The amended replicator equation is given by Equation 5.2.3 [116].

$$\dot{f}_i = f_i[\mathbf{e}_i.\mathbf{(A + B)f} - \mathbf{f}.\mathbf{(A + B)f}] \tag{5.2.3}$$

The matrix **B** = (b_{ij}) takes into account the effect of local interactions. The transformation of the payoff matrix would only affect the payoffs of strategies of different types. Matrix **B** is derived from **A** and satisfies **f**.**Bf** =0 [116]. The replicator equation given at Equation 5.2.3 would yield the same condition as b/c $>$ h(g+2)/l, for cooperators to succeed in a structured population. The effect of network reciprocity works in favour of the cooperators and depending on the fitness of the mutant false information, the information could spread to a large section of the population. Cooperators working in clusters would be able to gain a higher payoff than defectors and can expand in networks.

5.2.2 Experimental Results

Data sets

In this section, we use results from simulation to model the proposed information diffusion process. We also validate the importance of presence of clusters of sources of information in the network to spread less credible information. We validated the proposed methodology on three different data sets - both real world and synthetic data sets. We considered the following types of data sets.

- **Synthetic data set.** We used Preferential attachment model proposed by Barbarasi and Albert [118] to generate a scale free network of 5000 nodes.
- **Social network data set.** Facebook like network of an online community with 1899 nodes representing students of University of California and edges represent the messages sent between them during the period from April to October 2004 [119].
- **Facebook data set.** This data set consists of 'friends lists' from Facebook [120]. The data set consists of 4039 nodes and 88234 edges.

Experimental Setup

The simulation was carried out to validate the proposed model of diffusion of new information in OSNs. In particular we were interested in the coordination required between user nodes to spread less credible information. For this, we introduced clusters of users of sizes from one agent to multiple agents constituting around 0.5% of the total nodes to spread new information. This would constitute a probable set of colluding users in the real world OSNs. The coordinated effort by the user nodes would indicate spread of deliberate efforts to spread new information. We wanted to study the diffusion process by which less credible information could disseminate to a sizeable proportion of the population. We divided the population into *For*, *Neutral* and *Against* phenotypes. The *For* proportion was varied from 0.1 to 0.5 of the total users in the population. Out of the remaining, *Neutral* proportion was kept at 0.9 and *Against* phenotypes at 0.1. The simulation of diffusion of new information was carried out in two stages.

Neutral Drift. This is the initial stage as described in Section 4.4.6. The initial adopters of the new information accept the information because of their internal belief in the same. These are the nodes which would adopt the information even if one of its neighbors has been infected. These are the 'early adopters' of new information. These are the *For* types in the population which are neighbors of the original spreaders of new information. We evaluated the spread of new information with and without Neutral drift to understand the role played by the colluding sources and the initial adopters in the final size of infected population.

IM Update. The users update their strategies based on their phenotypes and IM update rule. The IM update rule as given in the payoff matrix at Table 4.3 was applied. The IM update was applied after the Neutral drift. The Benefit to Cost ratios were varied as multiples

of K, where K = k + 2. Here k is the degree of the node and the rule permits the node to retain its phenotype without mutation. As the infection spreads more and more *Neutral* types would get mutated to *For* type.The experiments were repeated with early adopters in the form of Neutral drift and without them. In each case the clusters of initial sources of disinformation were varied from one agent to 0.5% of the total number of nodes.

We carried out extensive simulation on all the three data sets. The simulation setup is shown in Figure 5.2. We used Netlogo [121] to carry out our simulation. Netlogo is an agent based simulation platform which provides adequate capability to model diffusion of new information and frequency dependent updating of strategies.

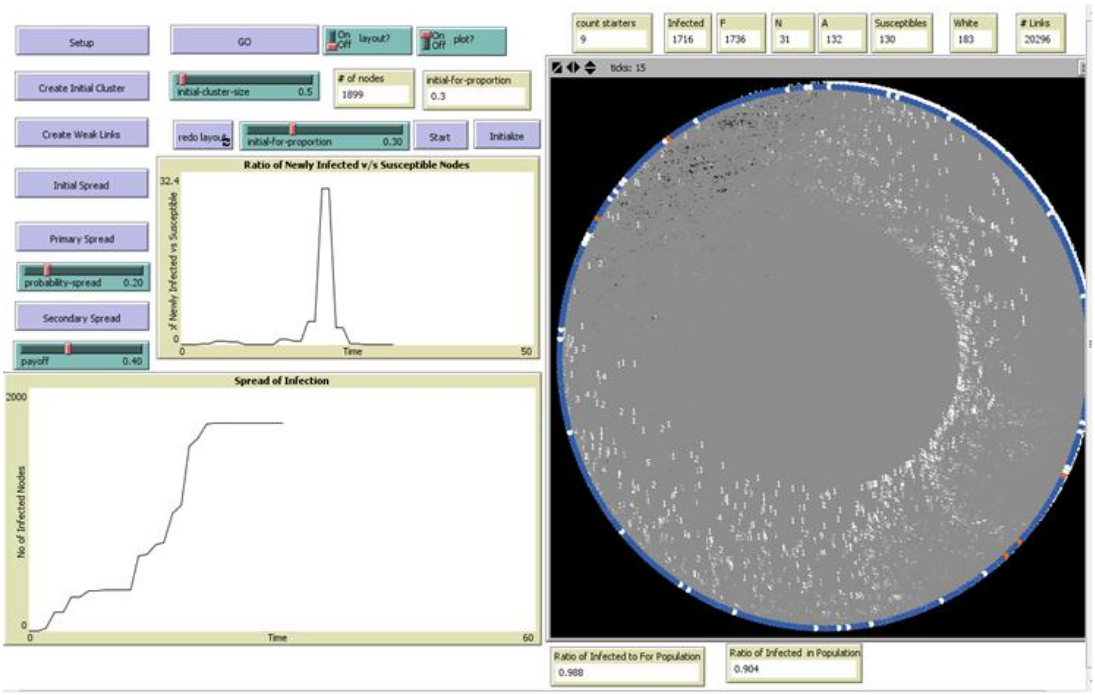


Figure 5.2: Simulation setup in Netlogo used to study diffusion of misinformation.

5.2.3 Results and Analysis

The information diffusion process was studied to understand the proportion of *For* types and *Benefit to Cost+* (B/C) ratio required for information cascades to occur. Less credible information would have less B/C ratio for most of the population and the initial *For* types would also be less. Under these circumstances, we would like to evaluate the effect of clustering of sources and Neutral drift to cause information cascades.

The results of the final infected percentage of users for the Facebook like social network data set is given at Figure 5.3, for Synthetic scale free network is given at Figure 5.4 and for Facebook network data set is given at Figure 5.5.

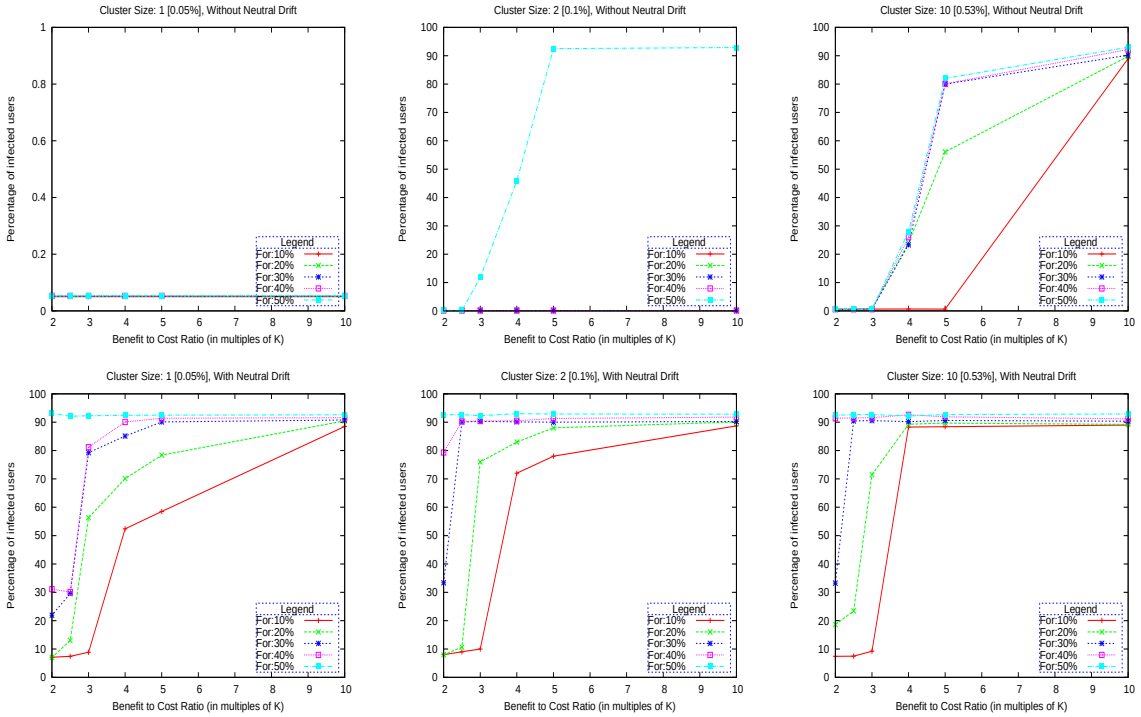


Figure 5.3: Diffusion of information in a real social network data set with 1899 nodes. Benefit to Cost Ratio is in multiple of K, where K = k+2, and k is the degree of the node.

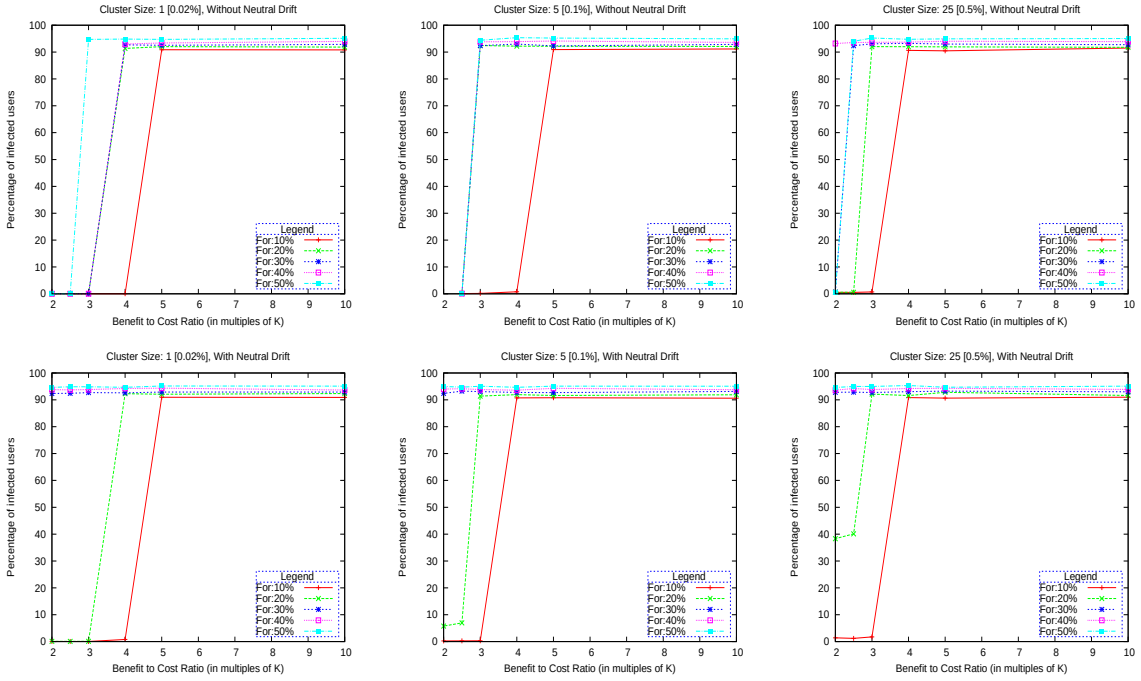


Figure 5.4: Diffusion of information in a synthetic social network data set with 5000 nodes constructed using preferential attachment model. Benefit to Cost Ratio is in multiple of K, where K = k+2, and k is the degree of the node.

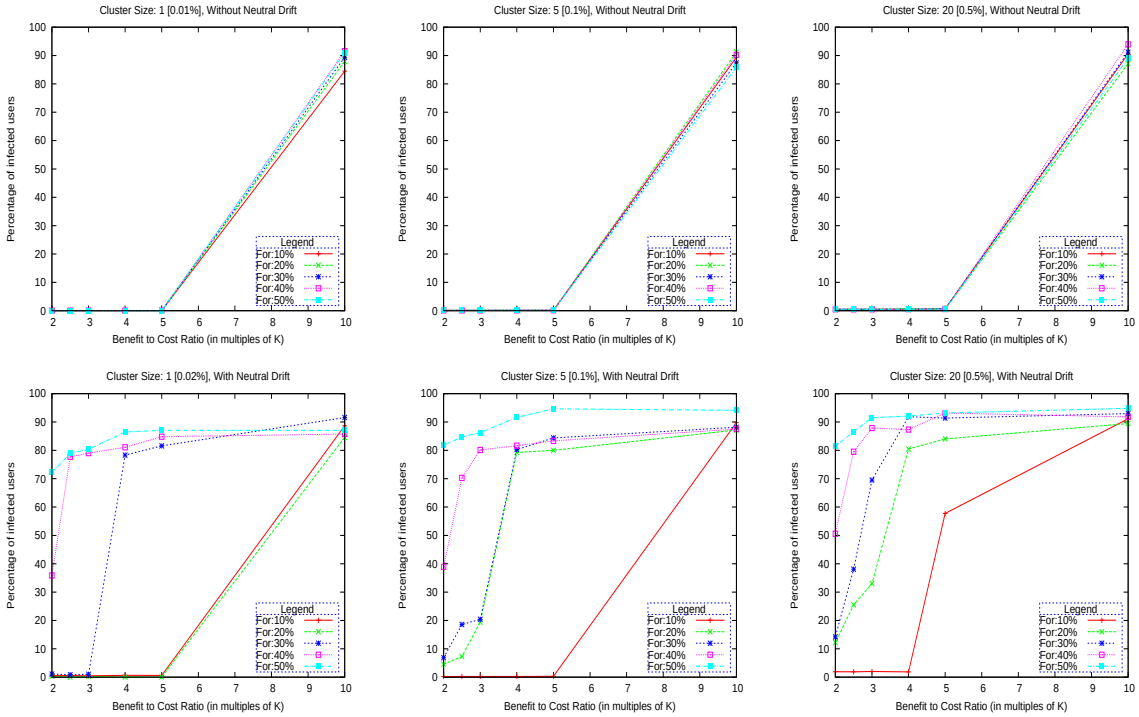


Figure 5.5: Diffusion of information in Facebook social network data set with 4039 nodes. Benefit to Cost Ratio is in multiple of K, where K = k+2, and k is the degree of the node.

The top set of figures of each data set shows the spread without initial adopters in the form of Neutral drift. The bottom set of figures are for spread of infection in the population with Neutral drift. The size of the cluster of sources of new information was varied from one in the left most figure to around 0.5% in the right most figure in each set. The extend of spread with and without coordination of the users to spread new information as well as the requirement of initial adopters of less credible information are very clear from the graphs, when the initial *For* types is between 10 to 20% and the B/C ratio for adopting the information is very less between 2K to 3K. When the B/C ratio is low and *For* types are also less, information cascades, which results in majority of the population accepting the information, are most likely to occur due to collusion between sources and the presence of initial adopters of information. When the users coordinate to spread new information, the spread is much larger than when they do not. This is important to overcome the threshold levels of utility for each user defined by the *Benefit to Cost* ratio. The spread of new information becomes limited if the number of initial adopters of the information is limited.

We varied the proportion of *For* types in the population. On other values of *For* population and B/C ratio, that is, when the initial *For* types are big enough and the benefit of accepting the information is also high, the effect of Neutral drift and clustering of users are not pronounced. Information cascades in such cases can occur without them also. However,

the higher proportion of initial *For* types would indicate credible information than otherwise. Hence information which is acceptable to a greater proportion of the population can spread without early adopters or clustering of sources. This is also intuitive, as the higher levels of acceptance would probably indicate true information and their spread would require no coordination or collusion of users. The clustering of users and early adopters are required when the initial proportion of *For* types are limited between 10 to 20% of the population, suggesting possible false information. The coordination of users to spread information would be detectable in the replacement evolutionary graph for the type of information being propagated. The greater coordination between sources and early adopters in the spread of new information would result in the formation of clusters of users who have adopted the new information and could be detected using network analysis as has been demonstrated in our experiments on real world data sets in Section 5.3.

We then investigated the role played by the *Against* phenotypes for countering the spread of misinformation. In the previous case, we had not assigned any proactive role to the *Against* phenotypes. However, to be more realistic to the social computing nature of user nodes, the reactions of *Against* types need to be captured. The counter diffusion against new information would be applied in the IM update phase as per payoff matrix given at Table 4.3. The updating of strategy would involve the transformation of *Neutral* types into *Against* type also when the Benefit to Cost ratio is favourable to this change. The results obtained in all the data sets shows a reduction in the final infected population. Moreover, the presence of such *Against* types would indicate differing views about the information being propagated and competing information campaigns in the network. This would again point towards greater probability of false information. The verification of the presence of such counter spread against misinformation is validated in our experiments on real world data sets.

A sample S-shaped curve obtained from our scale free network based on preferential attachment data set is shown in Figure 5.6 for B/C ratio of 2K and initial *For* types around 10%. The S-shaped curve is very similar to the one proposed by Rogers et al in [78] for adoption of new ideas by a society. The authors suggested the presence of different categories of population when new ideas are introduced in the society in the form of *Early adopters, Early majority, Late Majority* and *Laggards*. The classification of population into *For, Neutral* and *Against* types and their mutation follow similar classification of population for adoption of new ideas or inventions. The simulation results obtained verify that modelling the spread of new information which spreads and affects a large section of the population would also follow the same classification reflected by updating of strategies of different phenotypes in the population. Here also, a critical mass of initial adopters of information between 10 to 25% of the population can cause information cascades resulting in a majority of the population accepting the information- both credible and possibly false information.

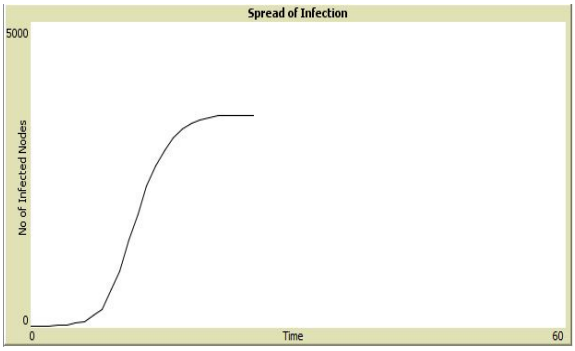


Figure 5.6: S-shaped curve formed in the diffusion of information in Facebook network data set when the B/C ratio was 2K and initial *For* types were around 10%.

5.3 Analysis of Evolutionary Replacement Graphs

The proposed information diffusion model should help us detect clustering of users to spread different types of information. We would analyse diffusion of information based on two concepts described in evolutionary graph theory. The concept of *isothermal graphs* would help us to understand the necessary and sufficient conditions for a mutant gene to spread in the whole population. *Bi-level evolutionary graphs* are used to study spread of multiple mutants.

5.3.1 Isothermal Graphs

Consider a graph G=(V,E), where V is the set of vertices and E is the set of edges. W=[w_{ij}] specifies the adjacency matrix of the graph, where for vertices v_i and v_j, the quantity w_{ij} specifies the weight of the directed edge from v_i to v_j. w_{ij} represents the probability that if v_i is selected for reproduction it replaces v_j with a probability given by w_{ij}. For any given vertex v_i, $\sum_j w_{ij}$=1. Isothermal graphs are evolutionary graphs which satisfy the condition $\sum_j w_{ij} = \sum_j w_{ji}$ [114].

Theorem 1. *(Isothermal Theorem)* [114]. An evolutionary graph is isothermal iff the fixation probability of a randomly placed mutant is ρ_1.

Here, ρ_1 is the fixation probability of Moran process in a well-mixed population of N individuals and is defined by Equation 5.3.1 [82].

$$\rho_1 = \frac{1 - 1/r}{1 - 1/r^N} \tag{5.3.1}$$

For those graphs which are not isothermal, the fixation probability depends on the fitness of the mutant as well as the structure of the graph [82]. A graph or a subgraph which satisfies the property is likely to help in the spread of mutants. The existence of near isothermal property between certain vertices would indicate possible clustering of players adopting

mutant strategy. The variation of degree of the vertices from the isothermal property could be measured in order to identify the existence of clusters in real world graphs as well as estimate their probability of invading the population. We have used iterative *k*-core decomposition algorithm on the graph to isolate such vertices in our experiments on real world OSN data sets.

5.3.2 Bi-level Evolutionary Graphs

Bi-level graphs are defined as evolutionary graphs representing relationships between communities [122] [123]. The vertices in the bi-level graph represent communities. Each of the vertices of the bi-level graph, generally denoted by **B**, is itself an evolutionary graph of individuals. The bi-level graph is an ideal way of studying the spread of mutants, especially when one of the vertices could be isothermal and can be labelled as leader and the others are followers. If the number of leaders increases and they spread the new information including misinformation, the fixation probability of the whole graph increases. The bi-level structure also helps in the study of multiple mutants in the graph.

5.3.3 Construction of Bi-level Graphs with Isothermal Vertices

The necessary condition found using the theoretical concepts of evolutionary graph theory was the presence of isothermal vertices for the mutants to invade the population. The presence of complete isothermal vertices would not be realistic in the real world evolutionary graphs. However there could be a large set of isothermal vertices pointing towards greater probability for the mutants to spread in the population. Detection of these isothermal vertices is important. As the mutant information spreads, more and more *Neutral* types would adopt the fitter *For* strategy of forwarding the new information.

We would describe the methodology used to construct bi-level graphs where each vertex would be an evolutionary graph consisting of isothermal vertices. The Replacement evolutionary graph would be constructed as a repropagation graph as was described earlier in Figure 4.3. We are making use of repropagation by users as a measure of acceptance of the source for the context being studied. We capture the assortativity of the network by grouping together interacting users with the same degree. In order to group the users who are strongly connected than the rest of the network we use *k*-core decomposition algorithm [105]. A *k*-core is a subgraph in which all nodes have at least *k* neighbour nodes. Since there could be multiple such cores in the graph we would obtain multiple clusters of users with their messages which satisfy the isothermal property. In order to detect all such clusters in the network, we use *k*-core decomposition algorithm iteratively. We identify the inner most core of the network with its user nodes and message nodes. We separate them and delete them. Any edges emanating from them to the rest of the graph are also removed. We continue the process till we reach the outer most cores and there would be nodes which would become completely disconnected. The iterative k-core algorithm is given in Algorithm 2.

Algorithm 2 Iterative k-core decomposition algorithm for detection of potential sources of disinformation

Input: Interactiongraph Social Network graph of the section of the users for the 'context' being evaluated
Input: Interaction details Details of users and messages(tweets) involved in the spread of information
Preprocessing Step:*Replacementgraph* repropagationgraph ← repropagation bi-partite graph with user nodes and message nodes
inputgraph ← repropagationgraph
corelist ← List()
while (inputgraph != NULL) **do**
    maxcore ← Max(Value of coreness) of nodes in the inputgraph using k-core decomposition algorithm
    maxcorelist ← User nodes and message nodes with (corevalue == maxcore)
    corelist[maxcore] ← corelist[maxcore] + maxcorelist
    inputgraph ← inputgraph – subgraph of nodes in corelist[maxcore]
end while
Innercorelist ← Subset of Isothermal subgraphs forming the inner cores in the Replacement graph
Output: *PotentialSources*: List of all sources of messages in the *Innercorelist*

At the end we would have segregated different cores of the graph with vertices in each core having the same degree and hence satisfying the isothermal property. The whole graph could now visualised as shown in Figure 5.7. The inner cores would consist of leaders propagating different messages and the outer cores would have the followers. The inner cores are shown as sets A, B and C with core values 5, 4 and 3 respectively. These sets could be collapsed into three vertices A, B and C to form bi-level evolutionary graph. Ideally there would have been edges emanating from all the vertices inside the collapsed vertices to other vertices in the graph. The absence of such isothermal property would result in the information propagated by the source vertices not being adopted by all the other vertices. However the probability of the messages in the inner cores being propagated by a large section of the population is maximum as compared with the rest of the messages. Once the messages are separated, the sources of messages could be easily identified. The number of sources segregated for the next stage of analysis would be limited a small percentage of the total number of sources.

5.3.4 Experiments on Real World Data Sets

In this section we would evaluate our methodology for analysis of spread of information in OSNs. We used the same data sets obtained from Twitter described in the previous chapters for our experiments. The results obtained after applying our algorithm are shown in Figure 5.8. The vertices identified in the inner cores are less than 5% of the total number of nodes. For each of the separated cores we identified the messages involved and then the sources of these messages. Only for the Higgs data set, where we had the weighted edges between the sources and retweeters, we used the weights to identify isothermal vertices.

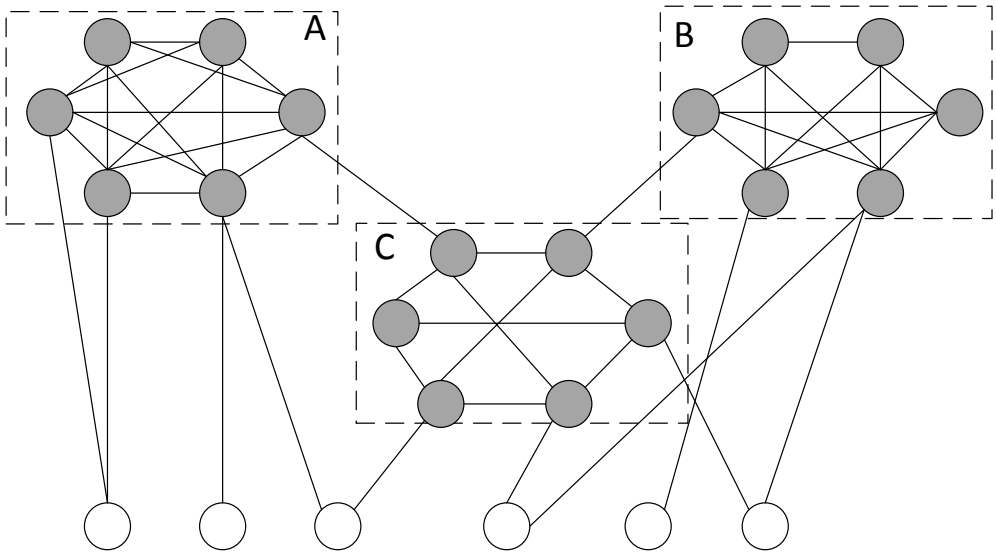


Figure 5.7: Structure of Bi-level evolutionary graphs with leaders (set A, set B and set C) forming isothermal evolutionary graphs and others as followers. The vertices in the sets A, B and C could be collapsed to three vertices A, B and C respectively.

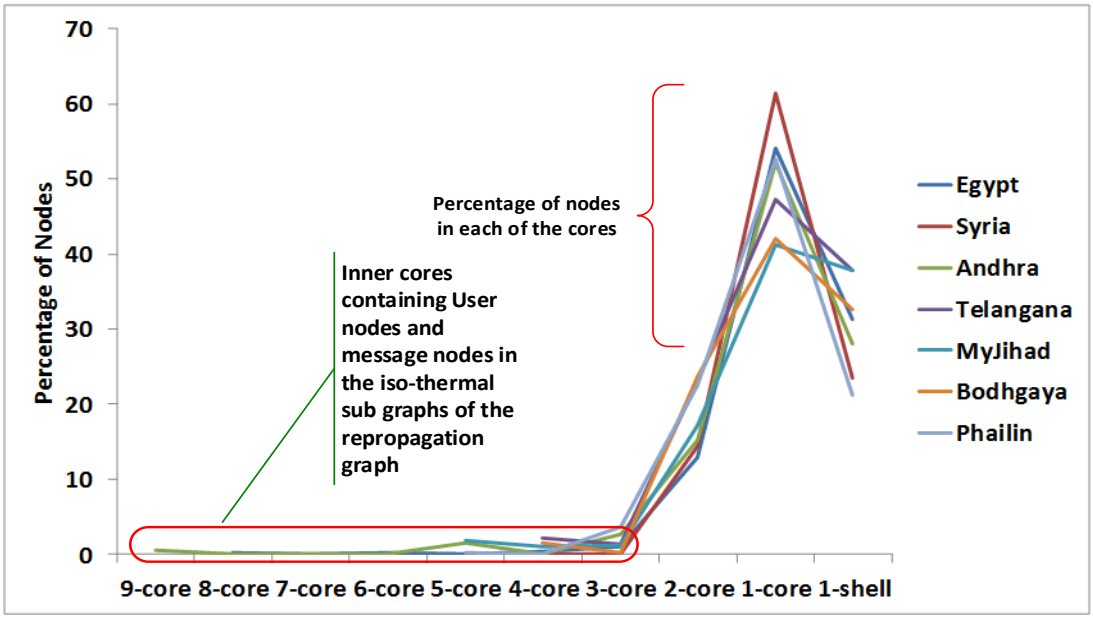


Figure 5.8: The segregation of inner most cores using iterative k-core decomposition algorithm.

We identified the most frequently repropagated messages in the data set and separated the sources of these messages as shown in Figure 5.9. We used measures of precision and

recall to answer two questions. The first one was, how many of the segregated sources in the inner cores were also the sources whose messages were frequently repropagated? The second question was how many of the sources whose messages are most frequently repropagated could be identified in the inner cores? The first question relates to 'recall' and the second one relates to 'precision'. The precision and recall figures are given for each of the data sets in Table 5.1. The figures of recall were always 100%, indicating the ability of the algorithm to separate all the sources who are more frequently repropagated, as we move from the inner most cores towards the outside. But as the figures of precision indicate all the segregated sources using the algorithm have not been frequently repropagated. The percentages of source nodes which have been segregated are also shown. In that manner, the results show the percentages of source nodes which have been segregated to achieve a recall figure of 100%. The presence of sources in the inner cores is a necessary condition but not a sufficient one for identification of frequently repropagated sources. The algorithm has enabled us to critically reduce the number of sources to be monitored for checking the quality of messages being propagated. The formation of such inner cores happen quite early and hence could be used to separate them prior to any rigorous semantic analysis of the messages.

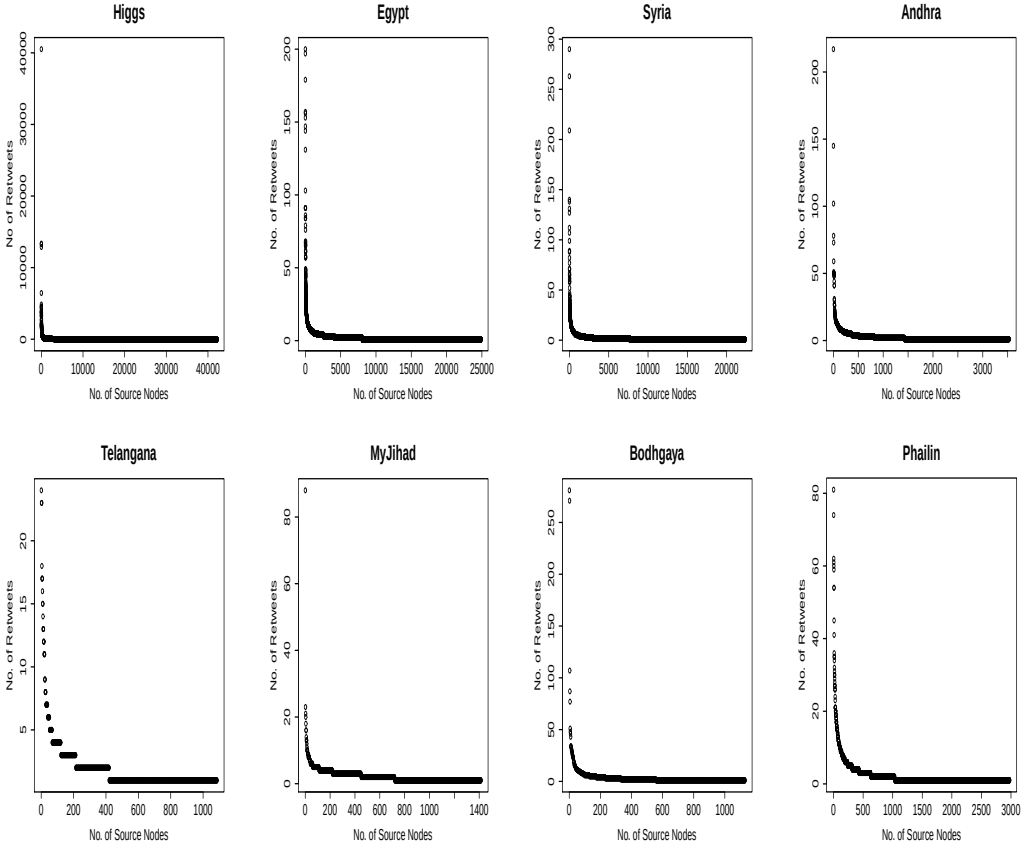


Figure 5.9: Frequency of retweets in data sets.

Table 5.1: Analysis of results.

	Recall	Precision	Percentage of Nodes
Higgs	100%	98%	2.37%
Egypt	100%	67%	1.27%
Syria	100%	58%	2.27%
Andhra	100%	70.7%	5.4%
Telangana	100%	85.71%	4.5%
MyJihad	100%	95.4%	15.7%
Bodhgaya	100%	91.5%	4.39%
Phailin	100%	81.9%	9.89%

The existence of isothermal property or near isothermal property in real world graphs is a clear indication of support for the cooperators. This is in support of our earlier conclusion that the spread of mutants is clearly beneficial if the mutants form clusters in the evolutionary graph. It would also lead from this discussion that, in order to detect spread of any type of information including misinformation which has the potential to reach a substantive proportion of the population, it is enough if we concentrate on detecting the quality of information being spread in such clusters. Further study of these clusters could be using standard community detection algorithms as explained in the next section.

5.3.5 Bi-level Graphs as Communities based on Modularity

While monitoring the spread of messages by single isolated nodes would be expensive, the formation of bi-level graphs provided a better alternative. The spread of messages could be studied by looking at a different type of bi-level evolutionary graphs. The bi-level graphs or communities in the repropagation graph were detected using modularity based community detection algorithm proposed by Blondel et al in [103]. The algorithm would identify user and message nodes which have greater communication links between them. This would ensure similar users identified in the isothermal graphs would form part of the same community along with others involved in repropagation of messages following similar strategies. The analysis of community formation in terms of bi-level graphs would enable the study of spread of different strategies and multiple mutants in the graph. The proposed model would help in modelling the spread of different types of information in OSNs accurately than any other generic model. The subsequent analysis and methodology followed in our work concentrate on the analysis of communities formed in the repropagation graphs. The evolutionary bi-level graphs obtained for various data sets are shown in Figure 5.10.

Each of the node in the graph is a community consisting of an evolutionary graph of user nodes and message nodes. The broad structure of graphs of communities remained the same for the repropagation graphs of all the data sets. There were large number of disconnected communities which had no edges to other communities. The few communities

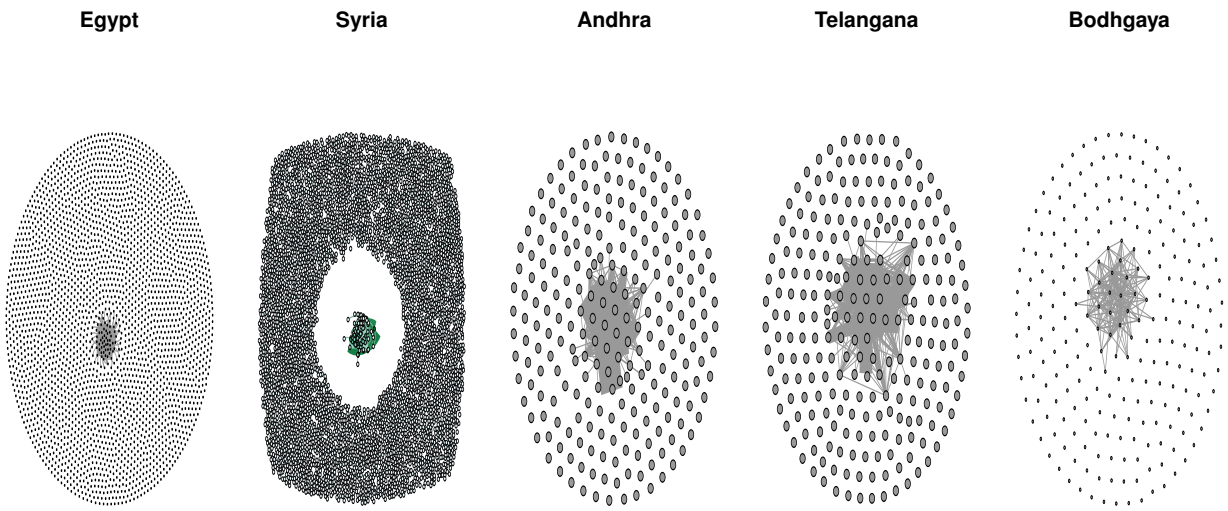


Figure 5.10: Communities in data sets as bi-level graphs.

which had edges to other communities represented probabilities of their evolution. The large disconnectedness reflects the preferential methodology followed by users in accepting and forwarding information which they agree with and hence similar to the formation of 'echo-chambers' or 'cyber-ghettos' discussed earlier.

The theoretical concepts behind the proposed information diffusion model helped us to analyse propagation of information in OSNs. The two types of bi-level evolutionary graphs proposed in this section would form the basis of our work. While the detection of sources whose messages are more likely to cause information cascades using bi-level graphs with isothermal vertices was found effective, further analysis would now be done on bi-level evolutionary graphs based on modularity measures.

The ability of users of OSNs to determine the credibility of information is the essence of using information diffusion modelling to detect and counter semantic attacks. We would like to clearly prove this property of OSN users and the effectiveness of monitoring information diffusion processes in communities of OSNs. Towards this we used the established framework of Latent trait theory to measure the ability of users of OSNs to detect misinformation.

5.4 A Psychometric Analysis of Diffusion of Information using Latent Trait Theory

In this section we have used Psychometric analysis based on Latent Trait Theory [LTT] [124] to study quality of information propagation in communities in OSNs. More importantly, we would like to definitely establish the social computing properties of users of OSNs and quantify their capabilities to determine quality of information. The collective intelligence of users of OSNs could be used to determine credibility of information. They could also be used to study trust relations between users in these communities.

The latent trait of ability of users could be used as a measure of social computing to distinguish between true information and misinformation in the network. Using repropagation features available in these networks as an affirmation of credibility of information, we constructed a dichotomous item response matrix which is evaluated using different models in LTT. This enabled us to detect presence of misinformation and also evaluate trust of users in the sources of information. Affirmation of credibility of sources would lead to trust between users and sources of information. Quantification of the trust relationships could be used to construct a polytomous matrix. The matrices are analysed using polytomous latent theory models to evaluate the types of trust and evaluate credibility of sources of information.

OSNs are also effective social computing systems. The users of such networks could be used to detect spread of false information. Latent trait theory is also called *Item response theory* and is used extensively in the educational domain to measure ability of students as a latent trait based on their responses to different items or questions as well as estimate the quality of question papers in measuring the ability of students. The process involves matching the ability of students to the difficulty levels of questions. We used the latent trait of ability of users in OSNs to determine credibility of information in the form of news items and messages.

5.4.1 Latent Trait Theory for Evaluation of Credibility of Information

Social networks have been defined as social computing platforms to assess quality of information propagating in them. Rumours and false information were questioned more by users than credible information [87]. The inherent ability of social network users to identify non-credible information is a function of information literacy of the user. Information literacy is defined as a set of abilities requiring individuals to recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information [125]. Evaluating the credibility of information is a critical aspect of information literacy. The ability of a user and the parameters of news items (henceforth called items) would determine the probability that the user would consider the items to be credible. The manifestation of credibility of items would be in the form of repropagation of such items.

Latent trait theory provides an ideal framework for the study of an underlying latent

trait and manifestation of the corresponding response. The theory provides for estimation of item parameters and ability of users in social networks [124]. Unlike educational usage, where the intention is to measure the ability of students accurately, we would like to use the ability of social network users to estimate the credibility of items. Estimation of item parameters and population parameters are iterative processes and we would use different models in LTT to accurately model acceptance of information in OSNs.

The basic assumptions that we make while we develop the latent theory model is that each individual possesses some amount of the underlying trait of ability in the context of information literacy. At each of the ability levels, an individual uses his or her ability to assess the credibility of information and decides to repropagate it. We would also assume that ability of individuals increase monotonically as we move from one end of the scale to the other. The probability of repropagating a typical item would depend on the difficulty in assessing the credibility of item and the ability levels of individuals.

Let the ability of users of OSN to assess credibility of items be θ. Let P(θ) be the probability that an individual with ability θ would repropagate an item. As the ability increases, the probability of correctly assessing the credibility of items also increases. Let 'b' be the difficulty of assessing the credibility of an item. We would assume that as the difficulty parameters of items increase, greater ability of users would be required to correctly assess credibility of items. If θ_1 and θ_2 are two ability levels such that $\theta_1 < \theta_2$, then users with ability of θ_2 would have a higher probability of correctly assessing the credibility of an item as compared to users with ability level of θ_1. The plot of correct response to an item and the ability of users is described using an item characteristic curve (ICC) as shown in Figure 5.11. Each item would have a separate characteristic curve.

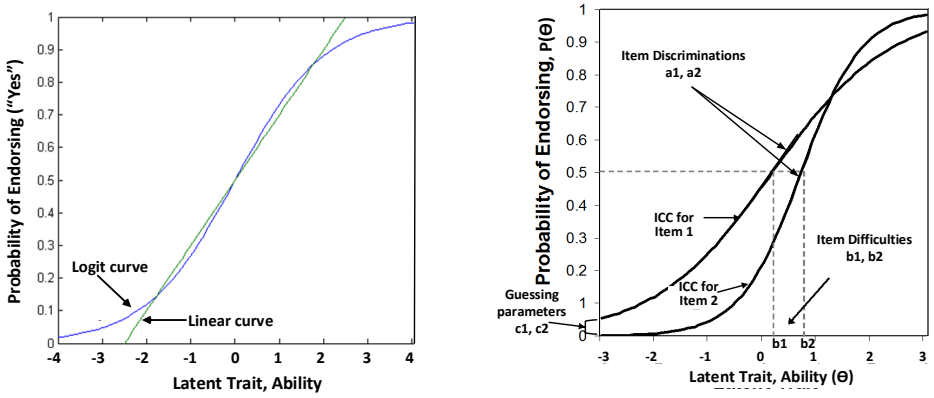


Figure 5.11: (a) Non-linear regression of probability of endorsing and latent trait, ability (b) Item characteristics curves for two items with different item parameters.

Each items has certain parameters. Let 'a' be the discrimination parameter which determines how well items can differentiate between individuals having ability above the item location and below the item location on the ability/difficulty scale. Steeper the curve, more

would be the discrimination parameter and the items would be able to discriminate better. As the slope decreases, the discrimination reduces and the probability of correct response at higher and lower levels of ability would be the same. Let 'c' be the guessing parameter, which takes into account the probability that the user could not evaluate the credibility of items and repropagated them nevertheless. This is similar to guessing the answer by examinees in case of multiple choice questions.

The difficulty of items and abilities of users are plotted together on the same scale on the X-axis called *conjoint scaling*. The probability of endorsing the item is plotted on the Y-axis. We would assume a binary response of 1 for repropagating and 0 for not repropagating an item. This type of responses are called dichotomous responses in LTT. We would assume a logit curve to plot the probabilities of endorsing an item rather than a linear curve as shown in Figure 5.11(a). A sample plot of ICC to explain different item parameters is given in Figure 5.11(b).

5.4.2 Item Response Matrix for Dichotomous Responses

The use of LTT model for evaluating information propagation in OSNs would require segregation of items in terms of messages and users who propagate them. Given a social network graph G, where G= (V,E), V is the set of user nodes in G and E is the set of edges between them. Let M be the set of messages propagating in the graph G, which are required to be analysed. For determining the credibility, we need to group messages based on context. We assume that messages could be separated based on keywords provided by the subject matter experts.

The analysis of messages in M using LTT would require us to separate out messages which have been repropagated at least once. In an item response matrix with items as columns and users as rows, we have to remove all rows which have all 1's and all 0's as this would mean ability levels of positive infinity and negative infinity respectively. Similarly columns with all entries as 0's and 1's are also removed as they would indicate difficulty levels of either positive infinity or negative infinity. These are standard practices in modelling of LTT using computers. In real terms, we are removing items or users who would not contribute towards analysis of either the items or the users.

The propagation of items in the network could be visualised using a repropagation graph. We constructed the bi-partite repropagation graph as explained earlier in Section 4.2.4. The accuracy of the model would depend on the accuracy of data in terms of responses of users to items received by them. While repropagation could be termed as confirmation of credibility, users not repropagating an item could be seen as a result of uncertainty about the credibility of the item as well as not having exposed to it at all. In order to tackle the second scenario and remove users who probably would not have even received the item, we used community detection algorithms in a repropagation graph. Community detection algorithms based on modularity would be able to isolate users in the social network graph who have more interaction edges within the community than external to it.

Detection of communities. We used community detection algorithms to group together users propagating similar items. The nodes in a community are connected and the property of modularity would ensure that the nodes have greater similarity with other nodes inside the community than outside it. The probability of nodes inside the community to have received all items propagating in the community is very high. Hence, the absence of an edge between a user and a message node in the repropagation graph would indicate lack of acceptability of the message for the user. We used community detection algorithm proposed by Blondel et al [103].

Construction of item response matrix. An item response matrix is constructed for each community in the repropagation graph. Since the cardinality of the communities vary, we consider communities of sufficiently large sizes. The item response matrix consists of items as columns and users as rows. A repropagation of an item by a user would result in an edge in the repropagation graph and is indicated by 1 in the matrix. The absence of edges in the repropagation graph is indicated by 0 in the matrix. For each community we separated messages from the community as columns in the item response matrix. All users who have repropagated at least one of these messages are the rows in the matrix. This item response matrix is then used for further evaluation to detect the appropriate model using LTT. The outline of the proposed algorithm for construction of item response matrix is given Algorithm 3.

Algorithm 3 Construction of dichotomous item response matrix.

Input: Social Network graph G(V,E) of the section of users for the 'context' being evaluated
Input: Details of users and messages(tweets),M involved in the spread of information
Preprocessing Step: $R_g \leftarrow$ Repropagation bi-partite graph obtained from G and M
$\mathcal{S} \leftarrow$ Set of all communities based on modularity in R_g
for all x in $\mathcal{S}$ **do**
  R[x] $\leftarrow$ Set of user nodes in x
  C[x] $\leftarrow$ Set of message nodes in x
  rows $\leftarrow |R[x]|$
  columns $\leftarrow |C[x]|$
  $IRM_x \leftarrow$ Matrix of *rows* and *columns* with all elements set to 0 for community x
  for $i = 1$ to rows **do**
    for $j = 1$ to columns **do**
      if $\exists$ edge $IRM_x[i]$ to $IRM_x[j]$ in R_g **then**
        $IRM_x[i][j] \leftarrow 1$
      end if
    end for
  end for
end for
Output: Item Response Matrix, IRM for dichotomous responses

5.4.3 LTT Models based on Dichotomous Responses

Repropagation of items by a user is considered as a confirmation of the credibility of the items as well as their sources. In [126], the authors have established this fact and have proposed metrics for quantifying behavioural trust based on proportion of repropagation of

messages of a source. In this section we aim to use repropagation as a measure of acceptance of credibility of news items in the form of binary responses. This would result in a dichotomous item response matrix. In subsequent sections, we intend to calculate trust value of a source based on metrics provided in [126].

The responses of users to various messages are functions of their latent trait of ability. As per LTT, there are four models which could describe them. These are Rasch model, 1-parameter logistic model (1PL), 2-parameter logistic model (2PL) and 3-parameter logistic model (3PL) [124]. They differ in the way item parameters are considered. Our aim would be to estimate the most appropriate model which fits the observed data of responses. This would enable us to study the item parameters as a function of ability of users. More importantly, we would like to estimate the quality of news items from the observed responses. This is akin to studying the quality of question papers using IRT based on the ability of students and their responses to the questions. The estimation of item parameters would throw light on quality of information diffusion in OSNs and possibly segregate likely misinformation and disinformation contents in them.

In order to model the responses of endorsing an item as a function of the ability of a user, we used the basic Rasch model as the initial null hypothesis. In the Rasch model, only the ability of the user and difficulty of estimating the credibility of items are considered. The parameters of discrimination, *a* is kept as 1 and the guessing parameter, *c* is kept as 0. The other models are the alternate hypotheses which we compared against the null hypothesis. A brief of the different models are given below. We summarise all the models in Table 5.2.

Table 5.2: Item parameters for all latent trait models.

Model	Difficulty, *b*	Discrimination, *a*	Guessing parameter, *c*	Probability of endorsing
Rasch	variable	1	0	$P(X=1\|\theta,b)=\frac{e^{(\theta-b)}}{1+e^{(\theta-b)}}$
1PL	variable	constant,k	0	$P(X=1\|\theta,b)=\frac{e^{k(\theta-b)}}{1+e^{k(\theta-b)}}$
2PL	variable	variable	0	$P(X=1\|\theta,a,b)=\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}$
3PL	variable	variable	variable	$P(X=1\|\theta,a,b,c)=c+(1-c)\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}$

(a) Rasch Model. In this model, it is assumed that all items relate to the latent trait equally and only the difficulty varies between items. The value of discrimination parameter is kept as 1. Let X be a random variable which takes a value of 1 or 0, as we deal with dichotomous values. If X=1, we assume that the user endorses the item for its credibility and repropagates it. If X=0, the user does not endorse the credibility of the item due to different reasons. P(X=1) is the conditional probability of random variable X taking a value of 1, given parameters of θ and *b*. The Item Response Function (IRF), P(X) is given by Equation 5.4.1.

$$P(X = 1|\theta, b) = \frac{e^{(\theta-b)}}{1 + e^{(\theta-b)}} \tag{5.4.1}$$

(b) 1PL Model. The 1-parameter logistic model is similar to Rasch model except for the fact that the value of discrimination, *a* is considered to be a constant, *k* while estimating P(X). The IRF is given by Equation 5.4.2.

$$P(X = 1|\theta, b) = \frac{e^{k(\theta-b)}}{1 + e^{k(\theta-b)}} \tag{5.4.2}$$

(c) 2PL Model. In the 2-parameter logistic model, the IRFs vary in their discrimination *a* and difficulty *b* parameters. The guessing parameter is set to 0. Equation 5.4.3 gives the IRF for 2PL model.

$$P(X = 1|\theta, a, b) = \frac{e^{a(\theta-b)}}{1 + e^{a(\theta-b)}} \tag{5.4.3}$$

(d) 3PL Model. In the 3-parameter logistic model, the IRFs include a guessing parameter, *c* which denotes a non-zero probability of endorsing an item at lower levels of latent trait. The IRF for 3PL model is given by Equation 5.4.4.

$$P(X = 1|\theta, a, b, c) = c + (1 - c)\frac{e^{a(\theta-b)}}{1 + e^{a(\theta-b)}} \tag{5.4.4}$$

5.4.4 Evaluation of Selected Models

Given a set of data and a proposed mathematical model which describes the distribution of parameters in the data set, we would use maximum likelihood test that best explain the data in terms of largest probability or likelihood. The technique of maximum likelihood is used to estimate the parameters of a model and test hypotheses about those parameters. The determination of latent trait, ability θ of users based on responses of items with parameters *a, b* and *c* were done in an iterative manner. We now give the methodology adopted for a set of items to identify the appropriate model which could fit the observed responses.

The four models are nested models with the Rasch model having the most constrained parameters. We carried the likelihood ratio test (LRT) for each pair of models starting with Rasch model as the null hypothesis. If L is the likelihood function and $\hat{\Omega}_0$ and $\hat{\Omega}$ are the likelihood estimators of Ω_0 and Ω, the Likelihood Ratio Test (LRT) is given by Equation 5.4.5. The value of L(θ_i) would be given by the item response functions for each model given earlier and the combined probability distribution is given by Equation 5.4.6. Normally log likelihoods are used. Let $L(x)_f$ and $L(x)_r$ be the log likelihoods of the 'full' (least constrained) and 'reduced' (more constrained) models. Let n_f and n_r be the number of parameters in the full and reduced models. Then the log likelihood ratio test is given by Equation 5.4.7. Log Likelihood Ratio, L will be distributed as χ^2 statistic with degrees of freedom equal to $(n_f - n_r)$.

Here, we test the null hypothesis $H_0 : \Theta \in \Omega_0$ versus $H_a : \Theta \in \Omega_a$ and use λ as the test statistic. The value of λ where $0 \leq \lambda \leq 1$ and $\lambda \leq k$, is chosen in such a way that the Type-I error, α is very low as determined by the p-values. The p-values of χ^2 is the probability of randomly selecting a χ^2 from the estimated distribution of χ^2 with $(n_f - n_r)$ degrees of freedom that is greater than the χ^2 observed in the data. That is, we estimate the p-value as the probability, p($\chi^2 > \chi^2_{obs}$). p-value is regarded as a probability quantifying the strength of evidence against the null hypothesis in favour of the alternative hypothesis. Smaller the p-value, stronger is the evidence against the null hypothesis in favour of the alternative hypothesis. We use p-values < 0.01 to reject or accept the null hypothesis. We selected the best model amongst the four which showed the best fit of the observed data. The model which fits the data could vary between communities in the graph. The model has a bearing on the analysis of the type of information propagating in the communities.

$$\lambda = \frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})} = \frac{max_{\Theta \in \Omega_0} L(\Theta)}{max_{\Theta \in \Omega} L(\Theta)} \tag{5.4.5}$$

$$L(\Theta) = \prod_{i=1}^{N} L(\theta_i) \tag{5.4.6}$$

$$L = 2(L(x)_f - L(x)_r) \tag{5.4.7}$$

5.4.5 Analysis of Data using Selected Model

Once the model was selected, we plotted the Item Characteristic Curves (ICC), Item Information Curves (IIC) and Test Information Curves (TIC) for each of the item response matrices in the data set to study the types of information propagating in communities.

Item Characteristic Curve. The ICC defined in LTT is the cumulative form of the logistic functions defined in Equations 5.4.1, 5.4.2, 5.4.3 and 5.4.4. The general characteristics of the curve is in the form of S-shape where the slope of the curve changes as a function of ability level and reaches a maximum value when the ability level equals the difficulty of the item [124]. Using these curves we would get to know the ability levels of users who repropagated the news items, the difficulty and discrimination of the news items and the guessing parameters, if any, in their propagation. While positive discrimination of news items augur well about the quality of item, it would be the items with negative discrimination which would have to be separated. Positive discrimination would indicate that the probability of endorsing an item would increase with increase in the ability levels of the users. Negative discrimination would indicate the reverse, i.e, the probability of endorsing an item decreases with increase in the ability levels of users. The occurrence of negative discrimination is not usual in LTT and two reasons are cited in [124] for their occurrence.

- *Incorrect Response.* Incorrect response to an item would always have a negative discrimination index if the correct response has a positive value. This would mean,

errors in modelling the correct response of endorsing an item could result in negative discrimination index.

- *Misinforming Item.* There is something wrong with the item which may be due to misinformation contents in it, due to which higher ability users did not endorse it. Negative discrimination is a warning that the item needs further attention.

Based on the above explanation, along with the results obtained in multiple data sets, discrimination parameters could be used to classify different types of information as given in Table 5.3. The possibility of incorrect response being modeled could be verified by reversing the values in the item response matrix and observing the ICC. The nature of ICC with positive and negative discrimination is shown Figure 5.12 [124].

Table 5.3: Interpretation of discrimination parameter, *a*.

Range of discrimination parameter, *a*	Classification
$a < threshold_{low}$	Possible misinformation
$threshold_{low} \leq a < threshold_{high}$	Credible information, widely accepted
$a > threshold_{high}$	Credible information

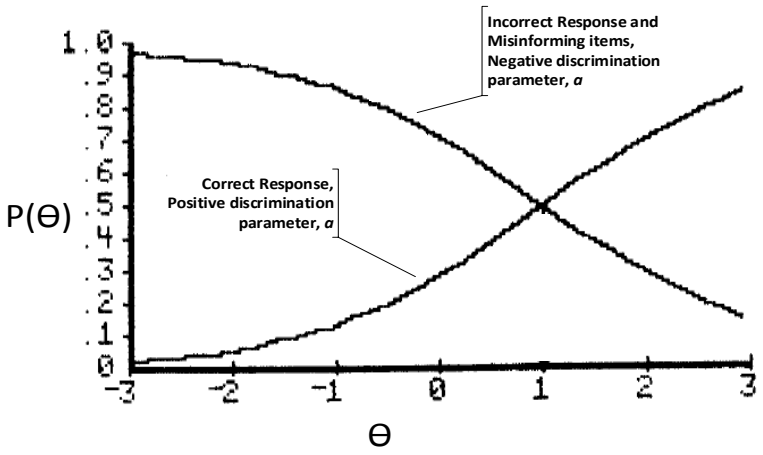


Figure 5.12: Item characteristic curves showing positive and negative discrimination parameters.

The presence of small negative discrimination index may not be alarming. However, large number of items having high values of negative discrimination index would certainly need attention. We retain items which displayed negative discrimination parameter even after the response values were reversed. A set of labels for item discrimination parameter values were proposed in [124]. We extended the same to include negative discrimination

also so as to cover the type of items encountered in different real world data sets. The proposed numerical values for $threshold_{low}$ and $threshold_{high}$ are given in Table 5.4. This occurrence in OSNs may not be surprising. While the questions in examinations are set by qualified teachers, the authors of messages in OSNs are normal users and the ubiquitous availability of social media enables them to express their thoughts without editorial filtering of any kind.

Table 5.4: Labels for item discrimination parameter values.

Label	Range of values as per [124]	Proposed Ranges for News Items	Proposed Labels
None	0	-infinity	Possible misinformation
Very Low	0.01 - 0.34	$<$ -1.7	Possible misinformation
Low	0.35 - 0.64	-1.69 - -0.65	Possible credible information
Moderate	0.65 - 1.34	-0.64 - +1.34	Credible information, widely accepted
High	1.35 - 1.69	+1.35 - +1.69	Credible information
Very High	$>$ 1.7	$>$+1.7	Credible information
Perfect	+infinity	+infinity	Credible information

Item Information Function. The Item information function (IIF) gives the estimate about the ability of a user from the information provided by the item. The amount of information depends on how closely the difficulty of the item matches the ability of the user. The IIFs for 1PL, 2PL and 3PL models are given in Table 5.5. For an item *i*, the probability of endorsing at ability θ is given by $P_i(\theta)$, and $Q_i(\theta)$ is defined as 1-$P_i(\theta)$.

Table 5.5: Information functions for all dichotomous latent trait models.

Model	Information Function
Rasch/1PL Model	$I_i(\theta, b_i) = P_i(\theta)Q_i(\theta)$
2PL Model	$I_i(\theta, a_i, b_i) = a_i^2 P_i(\theta)Q_i(\theta)$
3PL Model	$I_i(\theta, a_i, b_i, c_i) = a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta)-c_i}{1-c_i}\right]^2$

Information. Information in statistics and psychometrics is defined as the reciprocal of the precision with which a parameter could be estimated [124]. If variance of estimating the ability of a user is denoted by σ^2, the amount of information, I is given by Equation 5.4.8. If the amount of information at each ability level is plotted against ability we will get a continuous curve. Larger the information, the ability of the user could be determined with greater accuracy.

$$I = \frac{1}{\sigma^2} \tag{5.4.8}$$

Test Information Function. Test Information Function (TIF) gives the complete distribution of information across all ability levels for all items. TIF is given by the sum of Item Information Functions (IIFs) and is defined by Equation 5.4.9 [127]. The IIF and TIF are directly related to the square of discrimination parameters for both 2PL and 3PL models. The plot of TIF would give a good idea of the variation of discrimination parameters, which is important for making a decision about the quality of information spread in the network. The curve would indicate whether there is even spread across all abilities or peaks at specific abilities which could then be marked for further analysis.

$$I_j(\theta_j) = \Sigma_i I_{ij}(\theta_j, b_i) \tag{5.4.9}$$

Invariance of Item Parameters. As per LTT, the parameters of ICC, which include discrimination parameter *a*, difficulty parameter *b*, and guessing parameter *c*, are invariant across different groups of users. The determination of item parameters are independent of the distribution of ability levels of the users in the data set. This is called group invariance [124]. The item parameters are properties of the item and not of the users responding to the item.

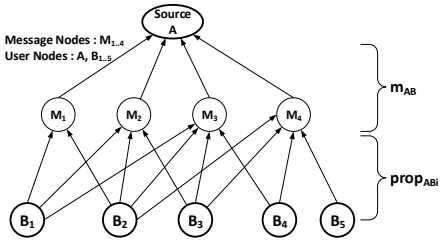
5.4.6 LTT Models based on Polytomous Responses

The use of repropagation features to determine credibility of messages was shown in the previous section using dichotomous matrices. This could further be extended to measure the credibility of sources of information in terms of behaviour trust. In this section, we explore modelling the latent trait of trust of users in sources of information in OSNs as indicated by repropagation of their messages. Behavioural trust between users are polytomous and could take a range of values from no trust to complete trust as described in [128].

Quantifiable measures of behavioural trust in OSNs have been described in [126]. Repropagation of information is an indication of credibility of information and trust in the source for the context specified by the information. Consider two users A and B as shown in Figure 5.13(a). Let m_{AB} be the number of messages sent by A to B and $prop_{AB}$ be the number of messages of A repropagated by B, and $prop_B$ be the total number of messages repropagated by B from all sources. The direct trust from B to A could be calculated by one of the methods given in Equations 5.13(b).

5.4.7 Item Response Matrix for Polytomous Responses

The construction of item response matrix is done from the repropagation graph given in Figure 5.13(a). We consider sources of information as items in the columns of the matrix and all users as rows in the matrix. For denoting trust between two users, we use equation



(a) Sample repropagation graph

(i) T(B,A) = $\frac{prop_{AB}}{prop_B}$

(ii) T(B,A) = $\frac{prop_{AB}}{m_{AB}}$

(b) Equations for behavioural trust

Figure 5.13: Estimation of behavioural trust in OSN using repropagation as a measure of trust.

given in Figure 5.13(b)(i), which gives the trust value of user B for source A, based on the proportion of messages of A out of all messages repropagated by user B. The algorithm for construction of item response matrix for polytomous trust values is given in Algorithm 4.

Algorithm 4 Construction of polytomous item response matrix.

Input: Social Network graph G(V,E) of the section of users for the 'context' being evaluated
Input: Details of users and messages(tweets),M involved in the spread of information
Preprocessing Step: $R_g \leftarrow$ Repropagation bi-partite graph obtained from G and M
$\mathcal{S} \leftarrow$ Set of all communities based on modularity in R_g
for all x in $\mathcal{S}$ **do**
  R[x] $\leftarrow$ Set of user nodes in x
  C[x] $\leftarrow$ Set of source nodes in x
  rows $\leftarrow |R[x]|$
  columns $\leftarrow |C[x]|$
  $IRM_x \leftarrow$ Matrix of *rows* and *columns* with all elements set to 0 for community x
  for $i = 1$ to rows **do**
    for $j = 1$ to columns **do**
      if $\exists$ path $IRM_x[i]$ to $IRM_x[j]$ of length=2 in R_g **then**
        $IRM_x[i][j] \leftarrow \frac{prop_{AB}}{prop_B}$
      end if
    end for
  end for
end for
Output: Item Response Matrix, IRM for polytomous responses

5.4.8 LTT Models based on Polytomous Responses

The models for application of polytomous responses differ from models for dichotomous responses. The number of responses for polytomous values are more than two. There are two main models considered for modelling trust between users and sources. They are Graded Response Model and Generalized Partial Credit Model. A brief description of various models are given below.

(a) Graded Response Model (GRM). GRM is used when the outcome categories of responses are ordered [129]. The trust scores determined by equation given in Figure 5.13(b)(i) would enable ordering of responses. Hence we use GRM as the first choice of models for polytomous item response matrices. GRM models the probability for any given response category or higher, so for any given difficulty sub model, it would be like 2PL model. Equation 5.4.10 gives $P_{ik}(\theta)$, the probability of responding in item category *k* (*k*=0,1,2,...,*m*) of item *i* at trait level, θ.

$$P_{ik}(\theta) = \frac{e^{a(\theta - b_{ik})}}{1 + e^{a(\theta - b_{ik})}} - \frac{e^{a(\theta - b_{ik+1})}}{1 + e^{a(\theta - b_{ik+1})}} \tag{5.4.10}$$

The equation is summarised as $P_{ik} = P^*_{ik} - P^*_{ik+1}$, where P^*_{ik} represents the category threshold function for category *k* of item *i*.

(b) Generalised Partial Credit Model (GPCM). GPCM is used for both ordered polytomous responses and unordered polytomous responses [130]. We used the same only when the polytomous item responses did not fit the GRM model after repeated trials. GPCM models the probability of adjacent response categories. Equation 5.4.11 gives the $P_{ik}(\theta)$, probability of responding in item category *k* (*k* = 0, 1, 2, ..., *m*) of item *i*, a_i is the item discrimination parameter, b_i is the item difficulty parameter and d_i is how far the item is located from the threshold.

$$P_{ik}(\theta) = \frac{e^{\sum_{j=0}^{k} a_i(\theta - b_i + d_j)}}{\sum_{i=0}^{m-1} e^{\sum_{j=0}^{i} a_i(\theta - b_i + d_j)}} \tag{5.4.11}$$

5.4.9 Evaluation of Selected Models

We evaluated different models similar to the methodology adopted for dichotomous models using likelihood ratio test and *p-values* < 0.01. We used the constrained and unconstrained versions of GRM to evaluate item response matrices. Constrained model is similar to the Rasch model, where the discrimination parameters were kept the same. In unconstrained model, they were varied between different items. GPCM was used to evaluate only those matrices which did not converge to a stable solution with GRM.

5.4.10 Analysis of Data using Selected Model

After selecting the correct model for each of the response matrices, we plotted Item Response Category Characteristic Curves (ICC), Item Information Curves (IIC) and Test Information Curves (TIC) for each of the item response matrices in the data sets to study trust on the sources of information in the communities.

Item Information Function. The Item Information Functions (IIF) for GRM and GPCM are given in Table 5.6. The meaning of the function is same as that of dichotomous responses. The presence of a^2, the square of the discrimination parameter in the equations is similar, and hence plotting the function would help in evaluating the quality of trust information in the communities.

Table 5.6: Information functions for all polytomous latent trait models.

Model	Information Function
GRM	$I_i(\theta) = \sum_{k=1}^{m} a_i^2 \frac{\left[P_{ik}^*(\theta)\left[1-P_{ik}^*(\theta)\right]-P_{ik+1}^*(\theta)\left[1-P_{ik+1}^*(\theta)\right]\right]}{P_{ik}(\theta)}$
GPCM	$I_i(\theta) = a_i^2 \left\{ \sum_k k_i^2 P_{ik} - \left(\sum_k k_i P_{ik} \right)^2 \right\}$

Test Information Function. The Test Information Function (TIF) is again given by the sum of the Item Information Functions (IIF). The TIFs for GRM and GPCM are given in Equation 5.4.12. The TIFs would again be used for initial analysis of communities and their segregation based on a^2 values.

$$I_j(\theta_j) = \Sigma_i I_{ij}(\theta_j, b_i) \tag{5.4.12}$$

5.4.11 Experiment Results

The credibility and trust worthiness is context specific. We used the Twitter data sets used in earlier chapters to evaluate the proposed methodology.

5.4.12 Results

We constructed retweet graph as repropagation graph and item response matrices as per the methodology given in Algorithms 3 and 4. The different communities were analysed using appropriate models. A section of the item response matrix for dichotomous responses constructed for a community in the Bodhgaya data set is given at Figure 5.14. A detailed view of ICC curves and their interpretation is given in Figure 5.15. The use of discrimination parameter, a, to classify the information is shown. Positive values of a indicates credible information. When a is approximately 0, it would mean very less discrimination between ability levels of users and hence such information would have either complete acceptance or rejection by the users, and hence not required to be considered further. The value of a less than 0 indicates possible misinformation. The figures highlight repropagated messages which fall in different categories.

	row.names	RT1012	RT133	RT136	RT150	RT152	RT155	RT156	RT157	RT164	RT170	RT176	RT188	RT190	RT195	RT198	RT203	RT210	RT212	RT215
1	_hi_r	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
2	2611usa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	a_nationalist	0	0	0	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1
4	akandpal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5	ashishvashist	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	ashoksharma_20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	azad_hind_fauj	0	0	0	1	1	1	0	1	1	1	0	1	0	0	1	0	1	0	1
8	criminalsingh	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	deepumankani7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	geostation	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
11	hariharohari	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	headboiler	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	hinduidf	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	iambuddha119	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
15	imaheshmange	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	indspk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	ivivek22	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	jayjay1950	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
19	justushavarthor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
20	lcia88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	lucyk6992	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
22	manjunathkumarr	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	mungalalamick	0	0	0	1	1	1	0	0	0	0	0	1	1	0	1	0	0	0	1
24	pkruler	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	pontifukex	0	0	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	1
26	ramachandrag	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	roshin_raj	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	rudeerasberryy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	saffronarya	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	surender_wgl	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	swaroopvaidya78	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	telfordatheist	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
33	the__chauvinist	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
34	varshasinghs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	vedicallah	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	vicharak1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 5.14: Sample Item Response Matrix for dichotomous responses in a community in the Bodhgaya data set.

A sample item response matrix for polytomous responses in a community in Bodhgaya data set is given in Figure 5.16. The results and interpretations of some of the communities in the data sets are shown in Figures 5.17 and 5.18.

Each row in Figures 5.17 and 5.18 represents sample output of analysis of information propagation in communities from a data set. Figure 5.17 is for dichotomous responses and Figure 5.18 is for polytomous responses for different data sets. We show only the results for four data sets, but similar results were obtained for other data sets also. For each set of dichotomous responses, we show item characteristics curves (ICC), item information curve (IIC) and test information curves (TIC). For each set of polytomous responses, we show one of the item response category characteristics curve, item information curve and test information curve. For determining credibility of information, the value of discrimination is used to make decisions as per ranges given in Table 5.2. The use of TIC and IIC gives a visual representation of the type of information in the community. Both figures represent information as proportional to square of discrimination parameter, a. The values of ability levels at which the information functions are maximum also indicates the levels at which discrimination parameters are also maximum. We need to concentrate only at these levels. Positive and negative discrimination values could be further assessed from the ICC. TIC and IIC enable us to carry out initial estimation of quality of information. A monotonic rising TIC, would indicate good information and would require no further analysis. As

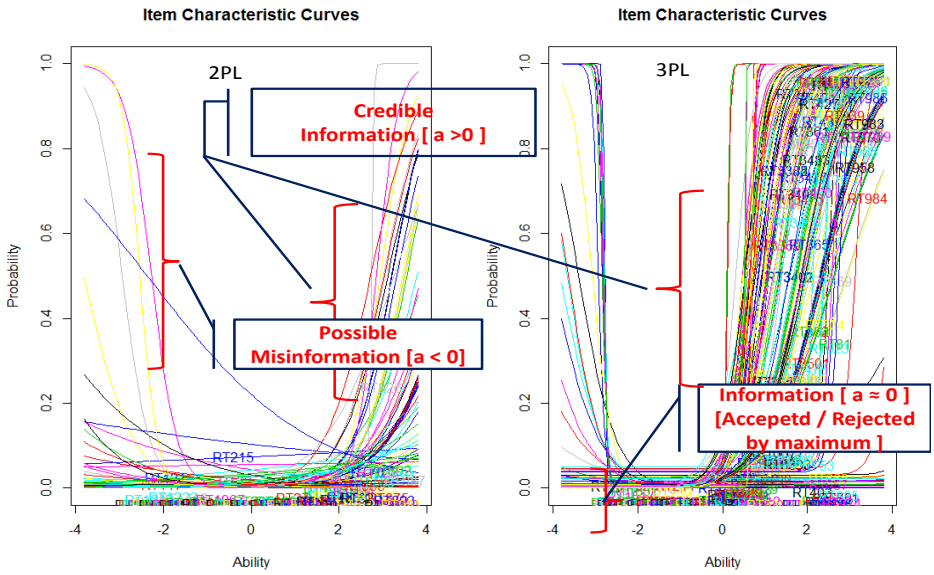


Figure 5.15: Sample item characteristic curves and their interpretation in 2PL and 3PL models.

	row.names	abhisek_yadav	ashwinisingh	atripathi01	charaknarender	harpreets_kohli	imac_too	ipallabidowerah	machbio	mahadevrdesai	naman_kapoor02
1	80o00	1	0	0	0	0	0	0	0	0	0
2	abhinav450213	0	0	0	0	0	0	0	0	0	0
3	abhiru80	0	0	0	0	0	0	0	0	0	0
4	abhisek_yadav	0	0	0	0	0	0	0	0	0	0.3
5	alok02013716	0	0	0	0	0	0	0	0	0	0
6	appujain	0	0	0	0	0	0	0	0	0	0
7	ashishiacr	0	0	0	0	0	0	0	0	0	1
8	barianilesh	0	0	0	0	0	0	0	0	0	0
9	crystalclearpra	0	0	0	0	0	0	0	0	1	0
10	digesh_123	0	0	0	0	0	0	0	0	0	1
11	dreamemerchant	0	0	0	0	0	0.0625	0	0	0.0625	0
12	drrmittal	0	0	0	0	0.1428571	0	0	0	0	0
13	globalgujju	0	0	0	0	0	0	0	0	0	1
14	globalhindi	0	0	0	0	0	0	0	0	0	0
15	harpreets_kohli	0	0	0	0	0	0	0	0	0	0
16	harshkushwaha10	0	0	0	0	0	0	0	0	0	0
17	hit3n	0	0	0	0	0	0	0	0	0	0
18	iaminsane4eva	0	0	0	0	0	0	0.5	0	0	0
19	ipallabidowerah	0	0	0	0	0	0	0	0	0	0.75
20	isachinpathak	0	0	0	0	0	0	0	0	1	0
21	ivsk1	0	0	0	0	0	0	0	0	0	0
22	kanjebro	0	0.5	0	0	0	0	0	0	0	0
23	kunalhariom	0	0	0	0	0	0	0	0	0	0
24	madhura_april	0	0	0	0	0	0	0	0	0.07692308	0
25	manishmig	0	0	0	0	0	0	0	0	0	0
26	mohandb	0	0	0	0	0	0	0	0	0	0
27	namo_4_pm	0	0	0	0	0	0	0	0	0	0
28	ninadvarkhede	0	0	0	0	0	0	0	0	0.5	0
29	pavankleo	0	0	0	0	0	0	0	0	0	0
30	prouddilliwala	0	0	0	0	0	0	0	0	0	1
31	rajmohan_k	0	0	0	0	0	0	0	0	0	0
32	ratigirl	0	0	0	0	0	0	0	0	0	0
33	ravigoklani1	0	0	0	0	0	0	0	0	0	0
34	rockdgod	0	0	0	0	0	0	0	0	0	0
35	rockrider2	0	0	0	0	0	0	0	0	0	0
36	saffronarya	0.2	0	0	0	0	0	0	0	0	0
37	sanjayvalliath	0	0	0	0	0	0	0	0	1	0
38	santoshkmr534	0	0	0	0	0	0	0	0	0	0
39	sathishswift	0	0	0	0	0	0	0	0	0	0
40	shaurya50832869	0	0	0	0	0	0	0	0	0	0
41	sk2210	0	0	0	0	0	0	0	0	0	0
42	ssachin_d	0	0	0	0	0	0	0	0	0	0
43	themanthan	0	0	0.07692308	0.07692308	0	0	0	0.07692308	0.07692308	0
44	tweeted_now	0	0	0	0	0	0	0	0	0	0

Figure 5.16: Sample item response matrix for polytomous responses in a community in the Bodhgaya data set

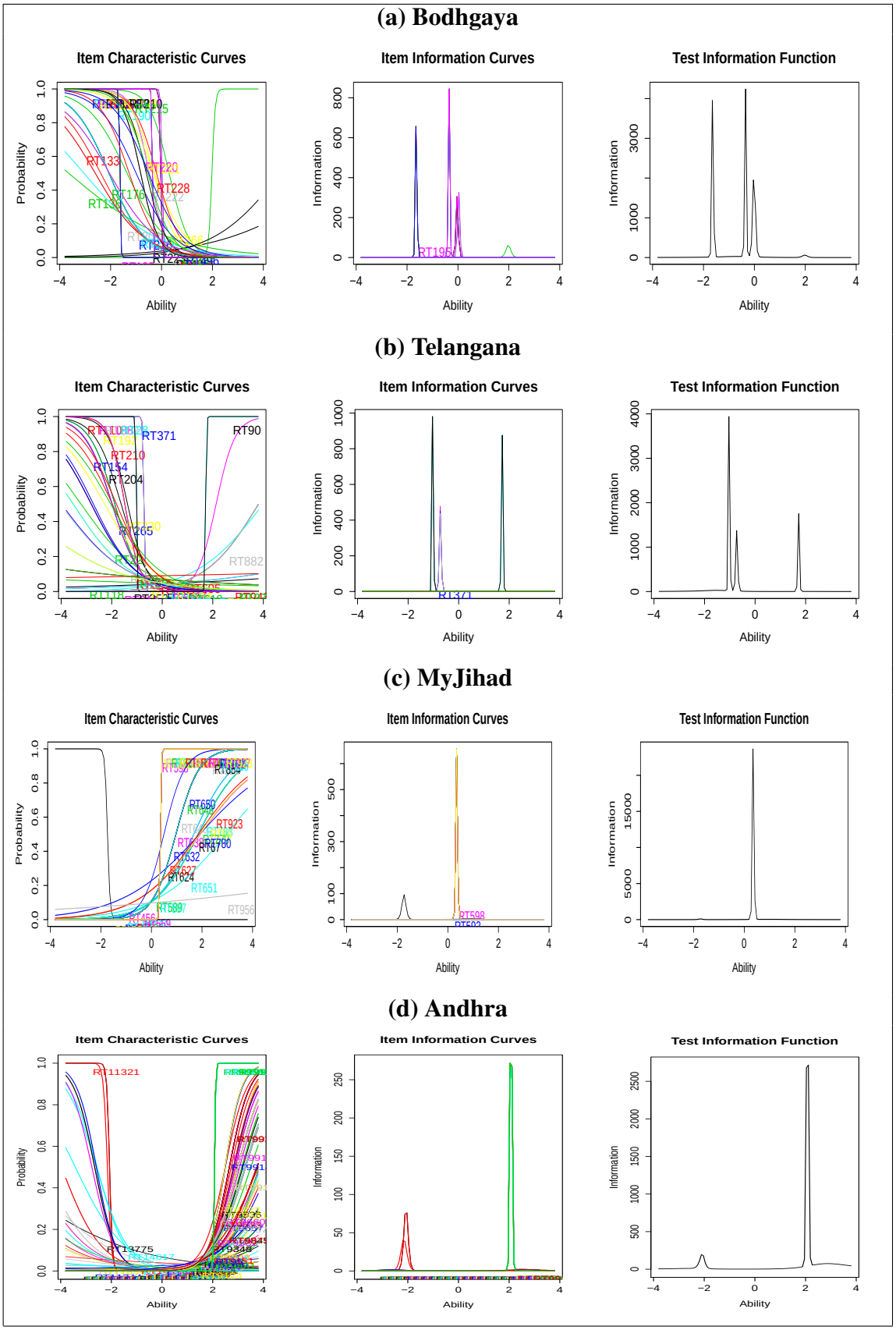


Figure 5.17: Sample item characteristic curves, item information curves and test information curves of dichotomous responses for different communities in data sets.

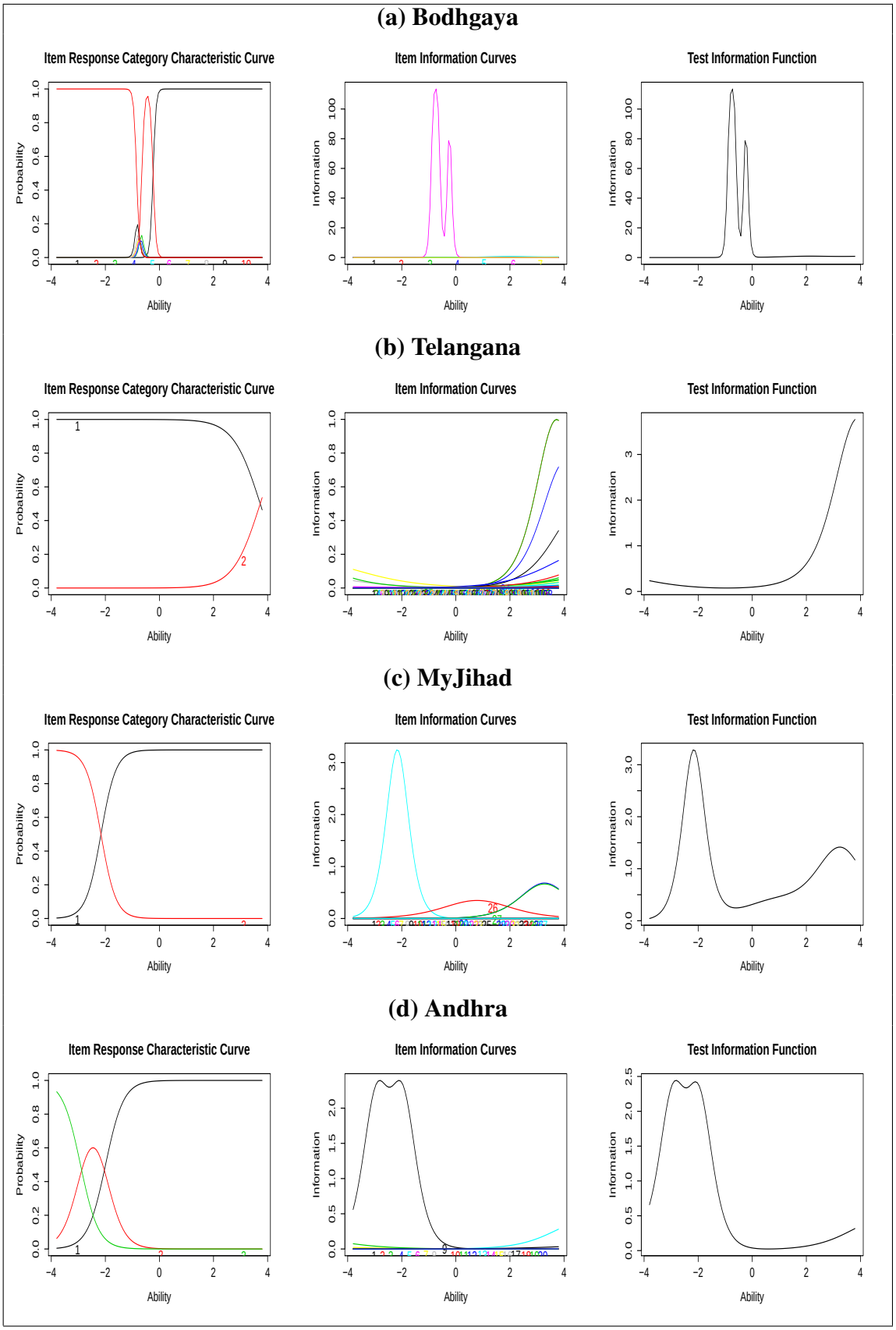


Figure 5.18: Sample Item response category characteristic curves, item information curves and test information curves for polytomous responses for different communities in data sets.

social media is mostly used by legitimate users, we expect most of the communities to be segregated in this manner. This would enable us to concentrate on those communities in which there is a greater possibility of flow of misinformation.

The spikes in IICs and TICs would occur for items with both negative and positive discrimination parameters because of the presence of a^2 term in their functions. The ICCs of the first two rows of dichotomous responses in Figure 5.17 indicate the presence of misinforming items. The ICCs of these two rows indicate sizeable presence of negative and near zero discrimination levels in the items. The figures in third row has a single item which could be classified as misinforming. The value of negative discrimination in the ICC of the fourth row is below the threshold to be classified as misinforming. Analysis of TICs would indicate spikes in the curve when ability levels are < 0 for only the dichotomous responses in the first two rows. Though not improbable, negative discrimination ICCs were not observed when ability levels are > 0. This would mean, when TIC and also IIC with information spikes at positive ability levels would not have negative discrimination and hence items would not be classified as misinforming items. The absence of such spikes in the IIC and TIC of fourth row figures of dichotomous matrices are clear. The presence of misinforming items in the communities of first two rows and items contributing towards it could be identified by these figures. The absence of misinforming items in the communities depicted in the fourth row and a single misinforming item in the third row are also evident from the figures.

The presence of spikes in polytomous matrices in Figure 5.18 are also analysed in a similar manner. Spikes in TIC are present in the polytomous matrices of first, third and fourth rows. The monotonic increase in TIC of the second row of polytomous responses is the ideal case of presence of trusted sources in the community without any misinforming sources. Spikes in TICs and IICs at negative ability levels could be matched with Item Response Category curves (IRC) and the presence of negative discrimination in them. IRC curve is for each category of item and for each row, we have shown only one IRC curve of the community which shows negative discrimination parameter. Similar to dichotomous responses, there were no information spikes at positive ability levels for items with negative discrimination.

The accuracy of the classification was verified in the data sets using human annotation. The figure of 'precision', which refers to the amount of actual false / unverified contents in the segregated messages was very high at around 95%. The identification of possible sources of misinformation is very important as further monitoring of OSNs would involve monitoring the activities of suspected sources of misinformation.

5.4.13 Time Complexity of the Algorithms

The proposed methodology could provide an effective social media monitoring framework only if the time complexity of the algorithm is less. The running time of the algorithm would

depend on the size of the data sets and would be proportional to the number of nodes. However, the algorithm enables parallelization of computation. After identification of different communities in the data sets, analysis of communities for quality of information propagation can be taken up independently. While each of the communities can be processed in parallel, the number of communities which require analysis could be reduced. We need to isolate communities with sizeable amount of interactions and number of nodes. We would use entropy as defined subsequently in Equation 5.5.3 as a measure to analyse information contents in the bipartite graphs. The distribution of entropy in different communities form a tailed distribution as given in Section 5.5.5. The analysis of communities could be limited to communities with higher entropy.

5.4.14 Visualisation of Trusted Communities

The use of LTT models to evaluate behaviour trust in terms of polytomous item response matrices of communities gives us a novel method of visualisation of trust in these communities. The TIF and IIF curves indicate distribution of credibility of information as well as trust in their sources. Even spread of information indicates greater trust and uneven distribution would mean formation of clusters of users in the community with no universal acceptance of credibility. Measure of trust is also an indication of social capital of these networks and we feel our work enables quantifying social capital of the networks and their visualisation in an effective manner. We investigate the trust in greater detail in the next section.

5.5 Trust

We would use the results proved in the previous section to further develop a behaviour trust model to determine credibility of sources as well as formation of trust communities. Trust is a subject which has been explored from different perspectives in OSNs. The underlying social structure of OSNs enables development of unique trust relationships. We bring out the necessity of developing a trust based reputation system for OSNs not only to rate the sources for providing accurate and timely information but also to detect malicious users forming clusters to spread false information.

5.5.1 Behavioural Trust Model

The calculation of trust in social networks is based on the social relations specified by the social structure of the network. Trust based reputation systems use local trust values rather than global values. This is important, as the social relations and the interactions reveal the degree of knowledge and hence the trust an individual places on people around him. The local values are more relevant and appropriate for such calculations.

The behavioural aspects of trust between people and information disseminated in OSNs are specified by the interactions in terms of comments, repropagations and directed messages. These aspects form an important part of social capital of a network, which is a measure of value of the network. Credibility of information shared through interactions increases trust and hence the social capital of OSNs.

5.5.2 Social Capital of OSNs

The social capital of an OSN is the collective value of the network. There are many definitions for social capital. We used the definition given in [131]. The authors have defined social capital as a collective resource that facilitates cooperation at the level of small groups. As per the authors, it exists not with an individual but through relationships between actors and is based on density of interactions. It is these interactions which would define the social capital of a network and the formation of communities of like minded persons. The number of interactions and the nature of interactions determine the social capital of the OSN at the network level as well as at the community level. The social capital is thus a function of *quantum of interactions* and *quality of interactions*.

The development of social capital is dependent on social trust. Social trust aids the development of social capital, the growth of which again improves trust in the OSN. The formation of *trusted communities* is important for development of social trust. Measuring social trust in the communities formed in OSNs in terms of information diffusion would also help us to understand the credibility of news propagating in them. Evaluating provenance of information is the first step towards measuring credibility. Important aspects of source of data and how the data is manipulated in OSNs form part of provenance of data.

Provenance of data. The meta information of data could be used to initially segregate possible non-credible information. We evaluated provenance of data - *where-provenance* and *how-provenance* by evaluating the source of the message and the community to which the source belongs. Further we look at how the data is manipulated in the communities. Deliberate efforts to spread information by collusion of users is a clear indication of loss of credibility. We used the work done by Adali et al [126] to enable us to assign local trust values to sources of information in OSNs. In their paper, authors have used conversation and repropagation aspects of communication to assign trust values to the information from a source. We analysed trust in communities as in the bi-level evolutionary graphs constructed earlier.

Repropagation graph. Trust is context specific. We segregated messages based on keywords specified by subject matter experts to define the context. Repropagation is an effective estimate of trust by the receiver on the source and the information propagated by it. We describe our method of evaluating the local trust values of information diffusion in OSNs. We constructed repropagation graph as discussed in previous chapters and as illustrated in Figure 5.13(a). We call the repropagation graph also as the trust graph. The interactions in terms of repropagation of news items is provided in all social networks.

Considering repropagations as assertions of trust and treating repropagation graphs as trust graphs in a social network, we propose metrics to measure the quantum of interactions as a function of number of nodes in the network and the number of edges in them. A message which has never been repropagated could be considered as being voted by users of the network as not worth being repropagated.

Consider two users A and B. Let m_{AB} be the number of messages sent by A to B and $prop_{AB}$ be the number of messages of A repropagated by B, and $prop_B$ be the total number of messages repropagated by B from all sources. The direct trust from B to A could be calculated by one of the methods given in Figure 5.13(b) [126] and reiterated below. We now define metrics to evaluate social capital of communities based on trust, namely *Entropy* and *Gini coefficient*.

$$T(B, A) = \frac{prop_{AB}}{prop_B} \tag{5.5.1}$$

$$T(B, A) = \frac{prop_{AB}}{m_{AB}} \tag{5.5.2}$$

Entropy. The density of interactions in the repropagation graph would indicate the quantum of information flow. We use *Entropy* function of the distribution of weighted degree of nodes in a repropagation graph to measure uncertainty of information. We define Entropy as under:-

$$H(d) = -plogp - (1-p)log(1-p) \tag{5.5.3}$$

Here H(d) refers to the entropy of the distribution of nodes and their weighted degrees in the network. p is the proportion of nodes and 1-p is the proportion of weighted edges. Higher the value of Entropy, more would be the interactions between nodes. The increase in quantum of interactions could be due to increased participation by more number of nodes or greater interactions between particular sets of users.

Gini Coefficient. We had defined gini coefficient earlier in Section 4.2.4. Here we use the same metric to measure credibility of sources based on behaviour trust. The quality of information flow is indicated by the credibility of sources. We evaluate credibility by the evenness in acceptance of sources by other users. While evaluating the trust value of source A, we use equation given in Figure 5.5.2. We would consider m_{AB} as those messages of A which have been repropagated by at least one user in the network as shown in Figure 5.13. The direct trust from user B_i to source A, T(B_i,A) is the fraction of messages of A which have been repropagated by B_i which were considered worth repropagating by the collective intelligence of the network. The distribution of trust values so assigned is the same as the distribution of the weighted edges between users in the repropagation graph. We use *Gini coefficient* (Ref Equation 4.2.1) of the weighted degree distribution of the repropagation graph to evaluate the variation in the trust values of the sources. The trust worthiness of a source is more if it is trusted by more number of users and the distribution of trust values is more uniform. The semantic analysis of the data sets revealed the efficacy of gini coefficient to quantify the differential acceptability of sources of information.

5.5.3 Monitoring System

The rate of growth of entropy in the communities can be used to monitor the information content in communities. A social network monitoring system would need to monitor only those communities with a higher growth of entropy and gini coefficient indicating possible spread of less credible information. We summarise the proposed methodology of an OSN monitoring system to prevent disinformation cascades by early detection of the deliberate spread of false information in Algorithm 5. The interpretation of output labels of the communities given by the algorithm is given in Table 5.7. We used measures of entropy and gini coefficient to form trust communities. We further evaluated the credibility of sources based on differences in their weighted and unweighted eigen centrality ranks.

Trust communities

Instead of measuring the entropy and gini coefficient for each subnetwork of every source, we used the metrics for each community. Communities in the repropagation graph are defined in terms of modularity, which indicates greater similarity with users within the community than outside. As greater interactions occur between users within the same community and hence generally more trust, we treat them as trust communities. We used community detection algorithms based on modularity measures to identify these communities [103]. In all data sets, the graph had a large number of disconnected communities and less number of connected communities. We used measures of entropy and gini coefficient to measure the information content and quality of interactions in these communities. This would form the basis of our OSN monitoring system.

Algorithm 5 Estimation of social capital of communities in repropagation graph.

Input: Social Network graph G(V,E) of the section of users for the 'context' being evaluated
Input: Details of users and messages(tweets),M involved in the spread of information
Preprocessing Step: $R_g \leftarrow$ Repropagation bi-partite graph obtained from G and M
$\mathcal{C} \leftarrow$ Set of all communities based on modularity in R_g
for all x in $\mathcal{C}$ **do**
    E[x] $\leftarrow$ entropy(degree distribution of user nodes in x)
    G[x] $\leftarrow$ gini(degree distribution of user nodes in x)
end for
$g_{thresh} \leftarrow$ threshold value of gini coefficient
$e_{thresh} \leftarrow$ threshold value of entropy
$\mathcal{K} \leftarrow$ Set of all cluster of communities based on spatial clustering
for all y in $\mathcal{K}$ **do**
    if $E[y] \geq e_{thresh}$ and $G[y] \geq g_{thresh}$ **then** $Label(y) \leftarrow (H, H)$
    end if
    if $E[y] \geq e_{thresh}$ and $G[y] < g_{thresh}$ **then** $Label(y) \leftarrow (H, L)$
    end if
    if $E[y] < e_{thresh}$ and $G[y] \geq g_{thresh}$ **then** $Label(y) \leftarrow (L, H)$
    end if
    if $E[y] < e_{thresh}$ and $G[y] < g_{thresh}$ **then** $Label(y) \leftarrow (L, L)$
    end if
end for
Output: *LabelsofCommunities*: Labels of Social Capital for all communities in the repropagation graph

Table 5.7: Interpretation of labels for communities based on values of entropy and gini coefficient.

Entropy	Gini Coefficient	Final Rating	Interpretation
High [H]	High [H]	Monitor High	Heavy spread of information, but low credibility. Spread to be monitored
High [H]	Low [L]	Good	Heavy spread of information of High Credibility
Low [L]	High [H]	Monitor Low	Low spread, Low in credibility. Spread to be monitored
Low [L]	Low [L]	Low	Low spread, credible information, less value

5.5.4 Estimation of Credibility of Sources

While the Algorithm 5 could segregate communities to be monitored, we would like to rate the sources also as per credibility. Evaluating the credibility of sources and quantifying perceived social consensus of messages are important to provide early warning to users. While gini coefficient of sources in the sub graph containing them in the repropagation graph is the first measure of credibility, we would use eigen centrality scores of source nodes to further improve our classification.

The use of left principal eigen vector in a trust network to assign global trust values in a P2P network was proposed in [132]. EigenTrust algorithm aggregates the local trust values to provide a global trust value to each user of a P2P network. In order to neutralise the higher recommendations to a source by colluding users, the ratings are normalised in EigenTrust algorithm. We propose a similar algorithm using the weighted and unweighted eigen centrality ranks of the sources to determine their credibility. The weighted centrality ranks of the sources are calculated initially, where the sources get ranks based on weighted edges from the users repropagating their messages. Then, we set these weights to unity and calculate the unweighted eigen centrality ranks. When equal weights are given to all users, efforts by a single user or a small set of users would not be sufficient to improve the rankings of the sources. The difference in the weighted and unweighted eigen centrality ranks could then be used to decide on the credibility of sources.

The methodology for rating the sources is given in Algorithm 6. The interpretation of ranking of the nodes is given in Table 5.8. The threshold for classification of the nodes is obtained using spatial clustering algorithm [133]. Note that higher the EC score, better the rankings (lower in value) of the nodes.

A source whose weighted EC rank is much higher than its unweighted EC rank is considered less credible and its credibility score is set to zero. The presence of large number of nodes whose weighted EC ranks are much higher than their Unweighted EC ranks, would also indicate the requirement of constant surveillance of the sources in the data set.

Algorithm 6 Estimation of credibility of sources.

Input: Social Network graph G(V,E) of the section of users for the 'context' being evaluated
Input: Details of users and messages M (tweets) involved in the spread of information
Preprocessing Step: $RG \leftarrow$ Repropagation graph obtained from G and M consisting of sources and users propagating their messages
Preprocessing Step: $R_g \leftarrow$ Repropagation bi-partite graph obtained from G and M consisting of sources, messages and users propagating their messages
$W_{ij} \leftarrow$ weight of the directed edge from $node_i$ to $node_j$ in RG
$W_{ij}(RG) \leftarrow$ Number of messages of $node_j$ repropagated by $node_i$ in RG
$WEC[x] \leftarrow$ Weighted Eigen Centrality Score of user nodes in V
for all edges $\in RG$ **do**
    $W_{ij}(RG) \leftarrow 1$
end for
    $UWEC[x] \leftarrow$ Unweighted Eigen Centrality Score of user nodes in V
    RankWEC $\leftarrow$ Rank all x as per WEC[x]
    RankUWEC $\leftarrow$ Rank all x as per UWEC[x]
for all x $\in$ V **do**
    CredibilityScore[x] $\leftarrow 1$
    if (RankWEC[x] - RankUWEC[x]) $> \delta$ **then**
        CredibilityScore[x] $\leftarrow 0$
    end if
end for
for all x $\in$ V **do**
    $SG_x \leftarrow$ Sub graph of R_g containing source node x, user nodes repropagating messages of x and the corresponding messages
    Gini[x] $\leftarrow$ Gini Coefficient of degree distribution of user nodes in SG_x
    if Gini[x] $> gini_{thresh}$ **then**
        CredibilityScore[x] $\leftarrow 0$
    end if
end for
    NonCredibleUsers $\leftarrow$ List of users whose (Credibility Score = zero)
    $N_{cred} \leftarrow$ number of credible users
    $N_{noncred} \leftarrow$ number of non-credible users
    $cred_{thresh} \leftarrow$ threshold credibility
if ($N_{noncredible} - N_{credible} > cred_{thresh}$) **then**
    Mark data set for close monitoring
end if
Output: Data sets containing non credible sources of information
Output: NonCredibleUsers, List of non credible sources in the data set

Table 5.8: Interpretation of labels of credibility of sources.

Weighted Rank [RankWEC]	**Unweighted Rank [RankUWEC]**	**Final Rating**	**Interpretation**
High [H]	High [H]	Good	Highly credible source
High [H]	Low [L]	Malicious	Low Credibility, Possibly malicious
Low [L]	High [H]	Good	Highly Credible, Presence of other malicious collectives in the network
Low [L]	Low [L]	Low	Low Credibility, Possibly not malicious

5.5.5 Experiment Results

Trust communities

We used the same data sets from Twitter to validate our proposal. We constructed retweet graph as the repropagation graph or the trust graph. The social capital of the communities in the graphs were estimated using algorithm given in Algorithm 5. The distribution of entropy of communities and their gini coefficients showed a tailed distribution with few communities having higher values for both metrics and most of them having very low values. Representative plots of distribution of entropy and gini coefficient for three data sets - Egypt, Syria and Higgs are given in Figure 5.19.

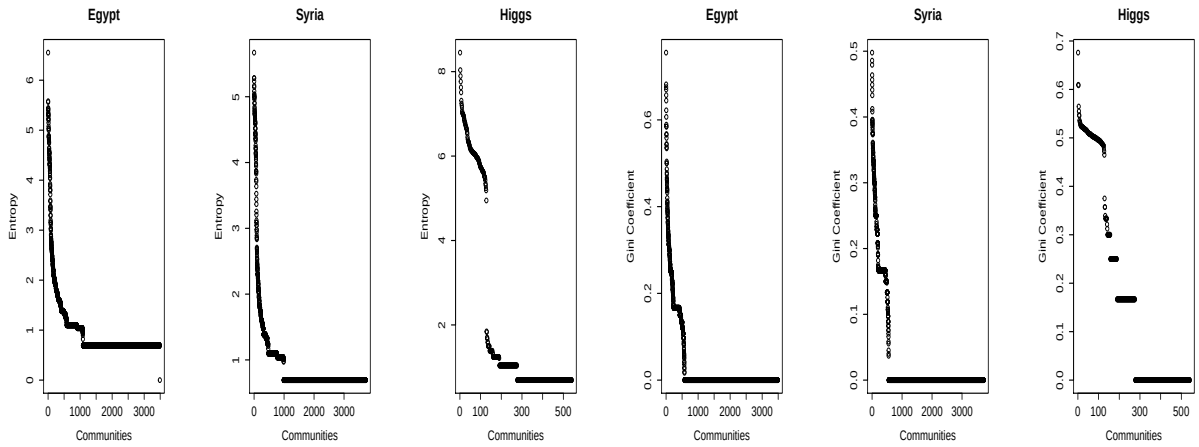


Figure 5.19: Distribution of entropy and gini coefficients of communities in three data sets.

Size of communities. We studied the distribution of user nodes and message nodes in the communities. The number of user nodes in each of the communities followed a tailed distribution. A combined figure displaying the number of user nodes and gini coefficients of all communities is given in Figure 5.20. The communities with more user number of user nodes and higher gini coefficients are marked for further monitoring.

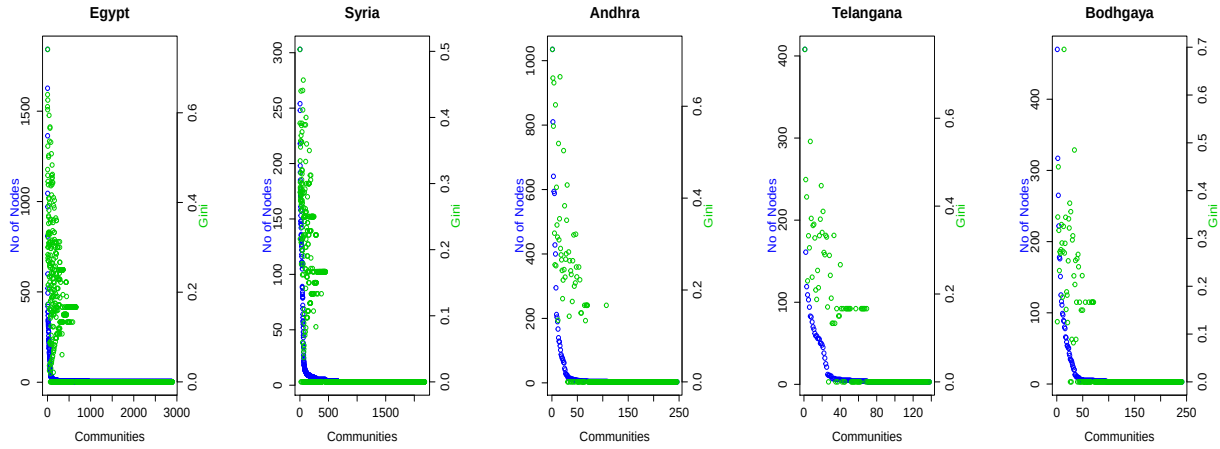


Figure 5.20: Distribution of number of nodes and gini coefficient in the communities of each of the data sets.

We studied the variation of entropy and gini coefficient across all communities. A plot of entropy values and gini coefficients of the communities was made. In order to segregate the communities based on entropy and gini values, we used spatial clustering algorithm proposed in [133], which cluster points based on medoids. This enabled us to spatially cluster communities with similar entropy and gini coefficients in the plot. The plot for the various data sets are given at Figure 5.21. The possible interpretation of the values of gini coefficient and entropy was given in Table 5.7. Higher entropy value indicates greater interactions in the community and lower gini coefficient indicates more even distribution of repropagated news items among the nodes and hence greater credibility of information.

The classification into four quadrants in each of the data sets was done based on *k*-medoid spatial clustering algorithm [133] and approximate threshold value of Gini coefficient of 0.4 observed in all data sets for credible information. The different clusters of communities are shown in the graph. The communities in clusters shown in (H,H) and (L,H) quadrants have a high probability of less credible information. The credibility of information was highest in the communities in (H,L) quadrant. The information in communities in (L,L) were not misinformation, but possessed less value. The measure of 'recall', for identifying potential misinforming communities in the (H,H) and (L,H) communities was over 95%. The 'precision' was between 65 to 70%. We propose the use of this method as the first stage of the social media monitoring system before more computationally intensive semantic analysis is carried out. The methodology adopted has proved to be very efficient in terms of computation, especially for large networks and when a near real time monitoring system is required. The number of communities in the (H,H) and the (L,H) quadrants, which requires further semantic analysis is less than 5% of the total communities in the graph.

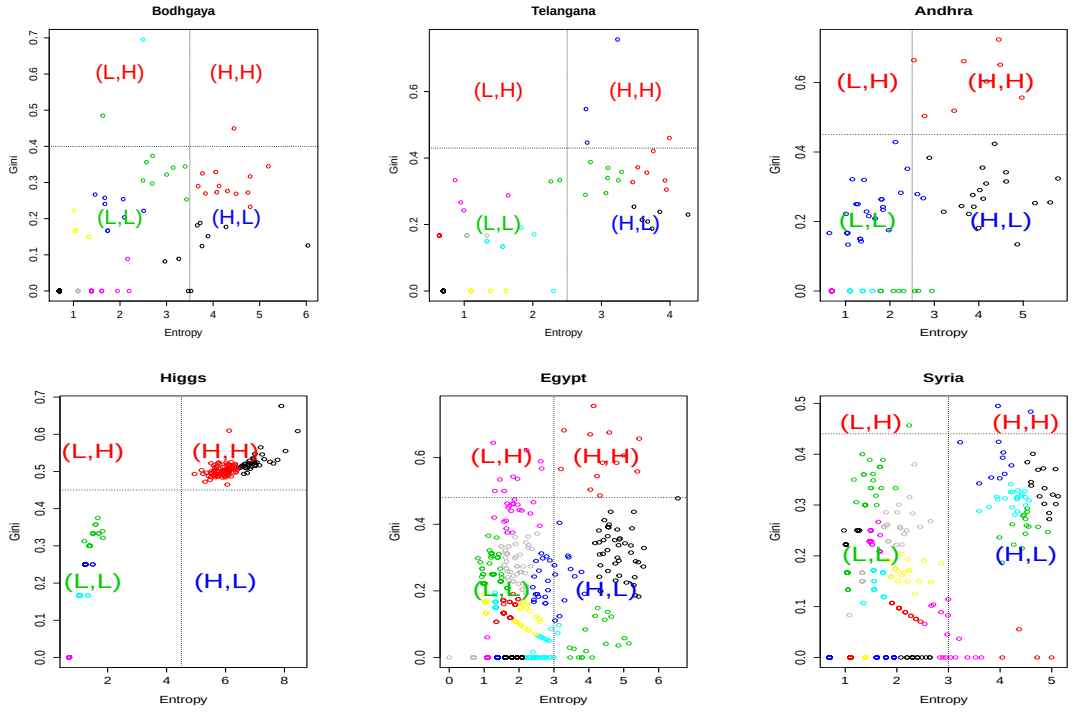


Figure 5.21: Variation of entropy and gini coefficient values of communities in the data sets.

Credibility of sources

The results obtained for the graphs of weighted and unweighted EC ranks for three data sets are shown in Figure 5.22. The colors of the nodes indicates the clusters identified by the clustering algorithm [133].

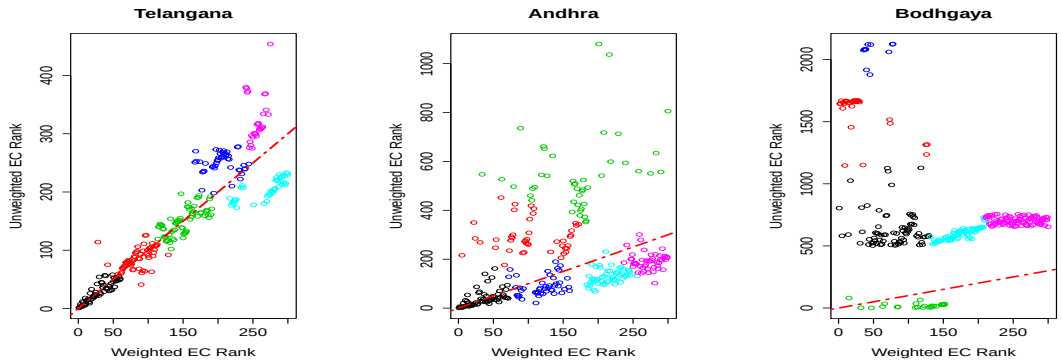


Figure 5.22: Comparison of weighted and unweighted eigen centrality rankings of nodes in the data sets.

In the figure, if both types of EC ranks were comparable, the distribution of nodes would lie on the line where the two EC ranks are equal (shown in dotted lines in the figures). The

first panel for the Telangana data set is the ideal one with most of the cluster of nodes along the equal rank line. The second panel for the Andhra data set shows more variations and lower quality. The third panel for the Bodhgaya data set shows no correlation between the two EC ranks and underlines the presence of a heavy section of colluding users. This was verified by semantic analysis of tweets of the users.

5.6 Summary

In this chapter we have analysed deliberate spread of information in OSNs using the proposed integrated model. The modelling of cognitive decision process of users was done by combining principles of Cognitive Psychology to cater to internal factors of decision making and evolutionary game theory to cater to social influences. We modeled the spread of new information as a mutant in the population playing cooperator defector game with the resident population. We concluded using theoretical concepts based on evolutionary graph theory that structure of OSNs support large scale spread of mutants when they act in clusters or collude with each other. This would result in formation of isothermal subgraphs which could be detected using k-core decomposition algorithm. The formation of bi-level evolutionary graphs in the form of trusted communities in the repropagation graph would enable study of spread of multiple mutants. Psychometric analysis of these communities based on repropagation and trust values proved the social computing properties of users of OSNs. It established the importance of analysing the communities in OSNs.

Prevention is better than cure and measures to limit spread of false information would require development of trust relationships between users in the network in the form of trusted contents and trusted communities. We have analysed the credibility of information propagation in OSNs using behavioural trust characteristics. We have proposed metrics for measurement of social capital of OSNs. Trust communities were identified using metrics of entropy to evaluate the quantum of information flow and gini coefficient to estimate the quality of information. The methodology makes use of information external to the data and not the semantic contents of messages. This would ensure scalability of the algorithm to measure credibility of information in large networks common in online social media in an efficient manner. We have also established the use of eigen centrality scores to detect possible sources of disinformation and estimate the quality of data sets. The methodology could be effectively used to rate sources of information. The use of Psychometric analysis to establish social computing properties of users has been published in [Pub1]. The use of Behavioural trust model has been published in [Pub5].

Chapter 6

Process of Modelling of Semantic Attacks

"Mis-information is rampant in this great age of mass-information. The amount of (mis)information at everyone's fingertips has lured us into a false sense of knowing. Whether it be information about science, politics, or theology, our society is suffering from an inability to research, process, filter, and apply."
David D. Flowers

6.1 Introduction

Spread of information has been investigated in social networks. Independent Cascade Model (ICM) [57] [58] and Linear Threshold Model (LTM) [60] [61] are two of the commonly used models to study information diffusion. We had given a review of information diffusion models in Section 2.2.7 of Chapter 2. However, the models fail to take into account semantic nature of contents as well as users' behaviour. Modelling spread of misinformation by users in a coordinated manner would not be possible with the basic models.

In general, we propose user centric models to perform better when semantic contents of messages as well coordinated efforts to spread information are considered. The traditional models like ICM and LTM could be suitably modified to cater for this. In this chapter, based on our earlier results, we propose a modification of basic ICM to take into consideration the collusion and coordinated efforts of users in spreading non credible information.

6.2 3-Dimensional Model for Diffusion of Disinformation

6.2.1 Basic ICM

The basic ICM consists initially of a set of *active* nodes A_0 at time t_0. The rest of the nodes are considered as *inactive*. At each of the discrete steps of time, an active node v at time t is given a single chance to activate an inactive node w. The probability of

success of the interaction is determined by the probability $p_{v,w}$. The probability $p_{v,w}$ is a parameter independent of any other parameters or history of interactions. When an inactive node has multiple active nodes trying to activate it, possibly with different probabilities, the interactions are randomly sequenced. Once the node *v* succeeds in activating the node *w* at time *t*, the newly activated node will try to become active at time *t+1*. Once a node becomes active, it is assumed that the node never becomes inactive again. The process is continued till no more activations happen in a single step, either due to lack of new susceptibles or failure to activate any particular node in any step. We call this 1-dimensional ICM.

Steps in Basic ICM

The step wise propagation of information using basic ICM is outlined below and depicted in Figure 6.1. There are four users who are initially infected in Step1. They independently infect others in the subsequent steps. Once infected, they remain active throughout. The final infected population by a single user who was infected in Step 1 is limited.

Initial condition:
At time, *t* = t_0, $A_0 \leftarrow$ Set of *active* nodes. All other nodes are *inactive*.

Information propagation:
At time, *t* = t_1, node $v \in A_0$ tries to activate neighbour node *w* with probability $p_{v,w}$.
$p_{v,w}$ is independent of any other parameters and history of interactions.
Once a node becomes active, it is never inactive again.
Process is repeated at subsequent intervals of time.

Termination condition:
Information propagation stop when no new nodes are activated in a single step.
i.e., At time t = t_k, $A_k \leftarrow = \phi$

6.2.2 3-Dimensional ICM

The type of deliberate spread of false information or disinformation efforts seen in the real world data sets could be more accurately modeled if collusion of users is also considered. Credibility of information plays an important role in their spread in social networks. Social networks are social computing systems and the decision making process of an individual is a cognitive process which estimates the credibility of information to spread. We explored the feasibility of spread of less credible information. The spread of less credible information was expressed with very low activation probability of a node while modelling using ICM. The spread of less credible information becomes possible under two circumstances - (a) Collusion of users in terms of space, *Spatial collusion* - i.e., number of users in social

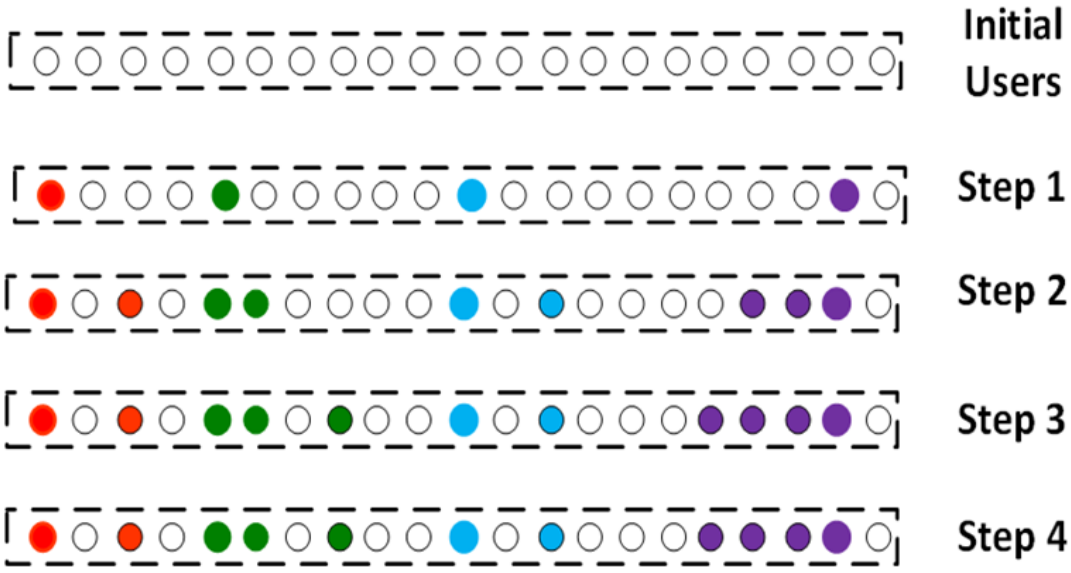


Figure 6.1: Step wise propagation of information in OSNs using 1-dimensional ICM.

network acting together, (b) Collusion of the users in terms of time, *Temporal collusion* - colluding users repeatedly sending similar messages where the subsequent messages support and further reinforce the initial messages. The addition of space and time as two more dimensions in deliberate spread of disinformation would result in 3-dimensional diffusion of information.

We describe 3-dimensional ICM as follows. Let C be a set of colluding nodes involved in the spread of information. An active node $v \in$ C at time t_0 tries to activate its neighbours with a message $m_0 \in$ M, with probability, p. Here M, is the set of all messages which are similar and convey similar disinforming information. All nodes $\in$ C are activated. Other nodes are activated with probability, p. We would like to keep p very low to indicate less credibility of information. At time t_1, all nodes which were activated at t_0, including the other colluding nodes $\in$ C, would try to activate their neighbours as per ICM. The presence of colluding nodes have increased the chances of activation of greater number of nodes. The process continues in similar manner till there are no new nodes activated in a particular step.

We define a set, I_0 consisting of all nodes which have been activated in the process of propagation of message m_0. At a subsequent time, t_1, one of the nodes $\in$ C, propagates another message $m_1 \in$ M and the whole process is repeated. The set of infected nodes, I_1 consists of all nodes infected by message m_1. We define I = $I_0 \cup I_1$. We call each of the propagation of messages as one *wave*, when a single message $m_i \in$ M is propagated, where i =1,2,3,.....N. So at the end of the process, the set of infected nodes, I = $I_1 \cup I_2 \cup I_1$ .....$\cup$ I_N, where N is the total number of waves. We consider infections by multiple messages at different time intervals to provide the third dimension in the proposed 3-dimensional ICM.

Steps in 3-Dimensional ICM

The step wise propagation of information using 3-dimensional ICM is outlined below and depicted in Figure 6.2. In the figure, all the users shown in red circles are in collusion. When a single such user gets infected in one step, all others are also infected in the next step. The final infected population is a union of all infected nodes by all the colluding nodes. This is shown in a single 'wave' of propagation. In the subsequent wave shown, a similar or the same message is propagated again. This brings in the aspect of time domain and the final infected population would be the union of all infected nodes in all such waves of propagation. The steps are outlined below.

Initial condition:
At time, *t* = t_0, $A_0 \leftarrow$ Set of *active* nodes. All other nodes are *inactive.*
Let C $\leftarrow$ Set of colluding nodes.
Let M $\leftarrow$ Set of similar misinforming messages.
Let node $c_i \in$ C be activated at time, *t* = t_0. i.e., $c_i \in A_0$
c_0 propagates message $m_0 \in$ M.

Information propagation:
At time, *t* = t_1, node *v* $\in A_0$ tries to activate neighbour node *w* with probability $p_{v,w}$.
$p_{v,w}$ is independent of any other parameters and history of interactions.
But all nodes, $c_i \in$ C get activated at time, *t* = t_1 by message, *m* = m_0 with $p_{v,w}$ = 1.
Once a node becomes active, it is never inactive again.
Let $I_0 \leftarrow$ Set of all activated nodes by node c_0.
At time, t_1, one of the nodes $c_1 \in$ C, propagates another message $m_1 \in$ M.
Let *wave* $\leftarrow$ propagation of a single message $m_i \in$ M.
Let $I_1 \leftarrow$ Set of all activated nodes by node c_1.
Let I = $I_0 \cup I_1$, at the end of *wave* = 1.
Process is repeated at subsequent intervals of time.

Termination condition:
Information propagation stop when no new nodes are activated in a single step.
i.e., At time t = t_k, $A_k \leftarrow = \phi$.
Let I $\leftarrow$ Set of all infected nodes by any message $m_i \in$ M.
I = $I_0 \cup I_1 \cup I_2 \cup I_1$ .....$\cup I_{N-1}$, where N is the total number of waves.

The above hypothesis has support from Cognitive Psychology where the decision making process of an individual is analysed [25]. The four factors or 4Cs involved in the decision making of a person were described earlier. They are *Coherency of message*, *Consistency of message*, *Credibility of source* and *Social Consensus.* The aspect of social consensus, where perceived social acceptability, pluralistic ignorance or false consensus could play an

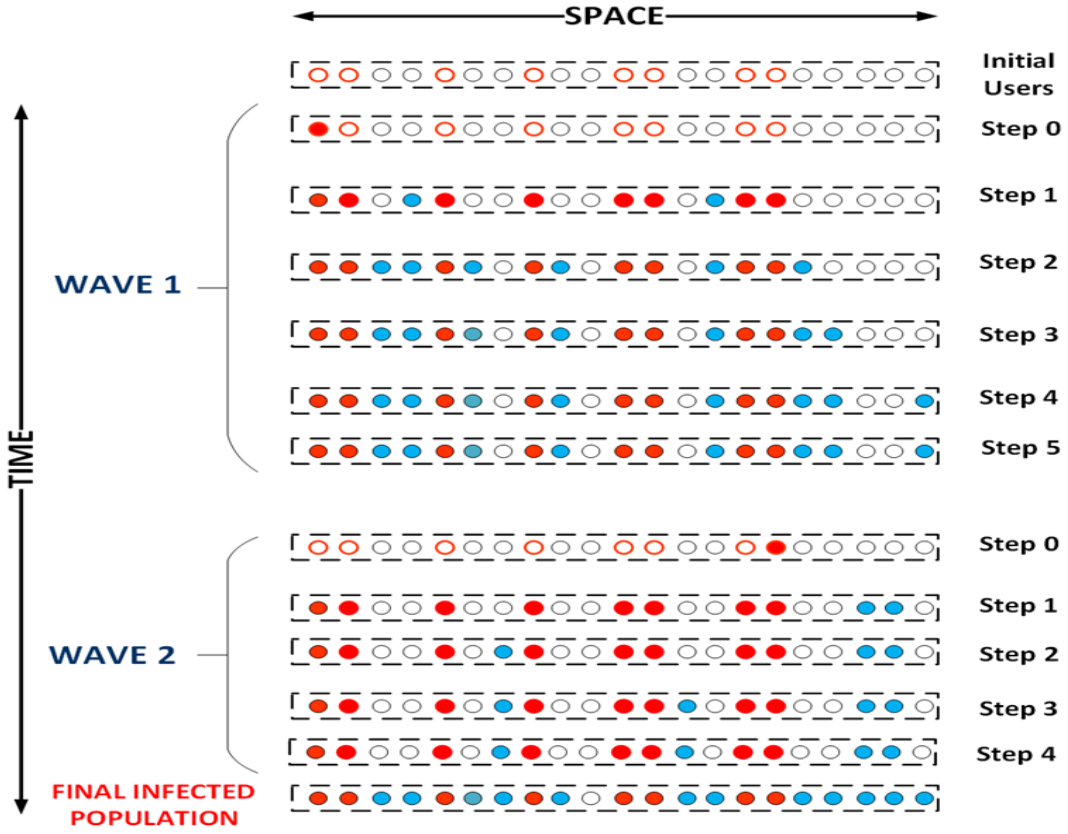


Figure 6.2: Step wise propagation of information in OSNs using 3-dimensional ICM.

important part in the decision making of individuals is especially true in social networks. The effect of multiple reinforcing messages and perceived acceptability of them by others would cause more users to accept the messages due to the factor of *Social Consensus*.

6.2.3 Simulation Results

We carried out our simulation of 3-dimensional ICM in different data sets to study the effect of collusion of sources in the spread of information as per details given in Table 6.1. For each data set, we varied the size of colluding users called cluster size from 1 to 50, in multiples of 5. The probability of activation, p, was randomly generated with a mean varying from 0.05 to 0.3. The dimension of time was considered with simulations done with varying values of waves as 1, 5 and 10. We considered the following types of real world and synthetic data sets. The results of information propagation using 3-dimensional ICM to cause information cascades are shown.

- **Synthetic Data Set.** We used preferential attachment model proposed by Barbarasi and Albert [118] to generate a scale free network of 1000 nodes.

- **Social Network Data Set.** Facebook like network of an online community with 1899 nodes representing students of University of California and edges representing the messages sent between them during the period from April to October 2004 [119].
- **Facebook Data Set.** This data set consists of 'friends lists' from Facebook [120]. The data set consists of 4039 node and 88234 edges.

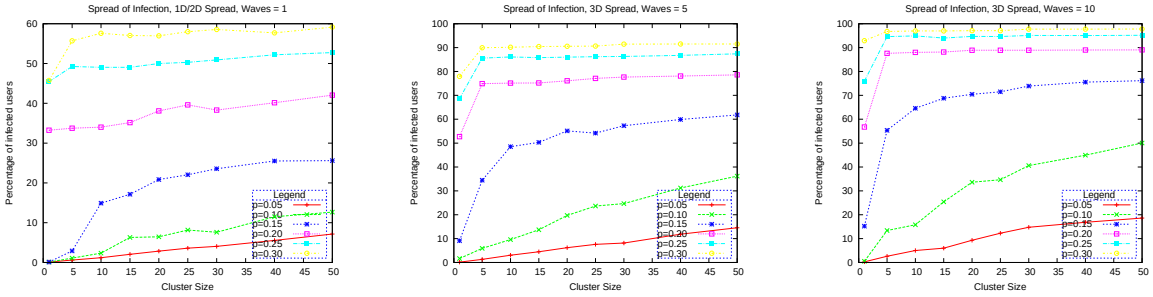
Table 6.1: Details of OSN data sets.

Data set	Nodes	Edges	Model	Type	Reference
Synthetic Network	1000	2500 (approx)	Preferential attachment	Undirected	[118]
Social Network	1899	20296	Facebook like network	Directed	[119]
Facebook Network	4039	88234	Facebook friends network	Undirected	[120]

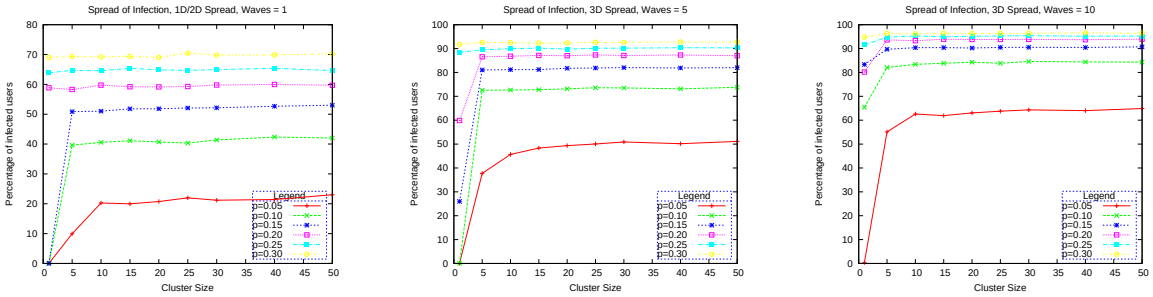
The results of simulation on the three data sets are shown in Figure 6.3. The figures show the number of infected users under normal ICM and under 3-dimensional ICM. Each row represents the simulation results for each of the data sets. When the cluster size is one and a single wave is considered, it is the traditional ICM. The left most figures in each row show the spread of infection with a single *wave* for different sizes of clusters from 1 to 50, showing the second dimension of diffusion. When the mean probability of infection, p is less, the number of infected users is very limited without clustering of users. As the cluster size is increased, the number of infected users have also increased. When credibility of news items is high, indicated by higher values of probabilities of propagation, clustering of users is not required for large scale propagation of information. The middle and right most figures in each row represent effect of multiple waves of 5 and 10 respectively. The effect of clustering at lower probabilities is enhanced when the number of waves is increased, giving a third dimension of time for diffusion. The results confirm that the quantum of infection is substantially increased when collusion of users takes place and are considered in *space* and *time* dimensions.

The extent of spread of messages would depend on the level of their acceptance by users. In the absence of collusion between users in terms of time and space, the messages have to be really credible to spread very far. However, with collusion, the spread could be more even in case of not so credible information. This is mainly due to the *social consensus* effect. In other words, for misinformation which is not credible to spread, collusion becomes a necessary condition. All collusions may not lead to extensive spread, but for extensive spread of less credible information to occur, collusion is necessary.

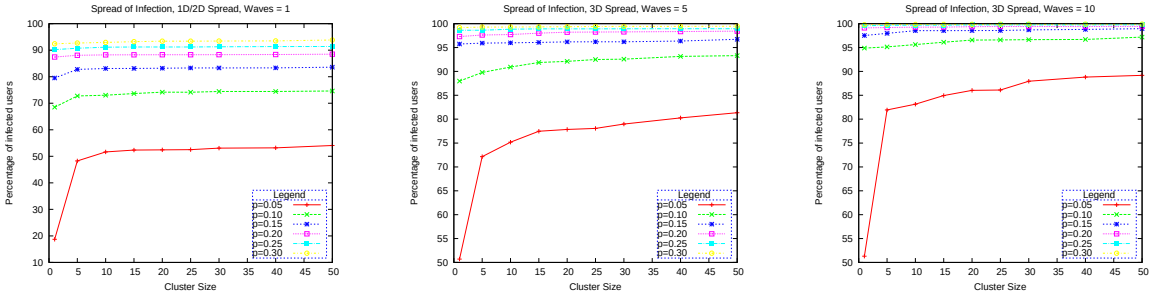
Consider a user who wants to spread some desired information which has less credibility. Given the options available with the user to infect maximum number of nodes, his best strategy would be to enlist the support of other users (shill attacks) or create multiple user profiles (sybil attacks). The presence of such collusion between users could be simulated in networks using 3-dimensional ICM.



(a) Synthetic Network Data set



(b) Social Network Data set



(c) Facebook Network Data set

Figure 6.3: Spread of infection in different data sets using 3-dimensional ICM.

6.3 Summary

We have used the integrated model to analyse the flow of non credible information in OSNs and proposed a variation of the basic diffusion model to cater for collusion and coordination of users in spreading information. The spread of deliberate false information would rarely be in isolation and hence considering the effects of a group of users sending a set of similar messages would be more accurate. The proposed diffusion model would more accurately depict real world information diffusion. Similar variation can be proposed for LTM also. The findings of the proposed model have been published in [Pub3].

Chapter 7

Results, Analysis and Recommendations

"An overwhelming majority of voters said that they encountered misleading or false information in the last election, with a majority saying that this occurred frequently and occurred more frequently than usual." Misinformation and the 2010 Election, A study of US Electorate, WorldPublicOpinion.org,2010.

7.1 Introduction

We have analysed a number of OSN data sets and real world data sets using the integrated model based on principles of Behavioural Sciences and Computer Science. The results obtained clearly brings out the potential of a system which can aggregate the responses of individual users to determine credibility of information. Any social media monitoring system would need to be scalable. In this chapter, we have analysed semantic attacks in the form of propaganda in OSNs during elections in India. We discuss the results obtained from different real world data sets of Twitter captured during the period of Parliamentary elections held in India during April to May 2014. We propose a framework for a scalable social media monitoring system which utilises social computing properties of users to filter potential sources of propaganda and possible non credible information.

7.2 Spread of Information in OSNs during Electioneering in India

Elections in the largest democracy in the world are studied with respect to use of OSNs for spreading information and propaganda. The elections for the constitution of 16th Lok Sabha of the Parliament in India provided a unique opportunity for the study of use of OSNs by various political parties and their leaders to spread propaganda. We studied the usage of micro blogging site Twitter by three major political parties and their leaders in the run up to the elections in April 2014 and also during the conduct of elections spread over 45 days.

It is a known fact that political parties use different methods to influence voters. For this

they use all media including OSNs to spread information supporting them and discrediting other political parties. Studying online behaviour of these parties and their leaders would help us to understand patterns of spread of propaganda - one of the counterfeits of information. Deliberate efforts made by political parties in Twitter were studied using models explained in the previous chapters. We analysed the importance of perceived social consensus in the acceptance of information by users of social networks. Parties which were perceived to run coordinated campaigns could be identified by these patterns. The integrated modelling using principles of Cognitive Psychology and Computer Science throw new light into understanding patterns of information propagation in OSNs in the form of disinformation and propaganda.

We studied dissemination of news items in Twitter in respect of three political parties and their leaders during elections. In a vast country like India, the general elections for Parliament were conducted in 9 phases between 7 Apr 2014 and 12 May 2014 across 28 states and 9 union territories. With a total electorate of 81.45 crores, the elections were the largest ever in the world. The number of voters between 18-19 years of age was 2.7% of the total eligible voters. The average election turnout was around 66.4%. The results of the elections were declared on 16 May 2014 [134].

7.3 Experiment Results and Analysis

We selected three political parties which contested elections from most parts of the country. We name these parties A, B and C and their leaders L_A, L_B, and L_C. We studied the use of Twitter by these parties. We tracked the presence of their names in the tweets during the period of data collection. This resulted in collection of tweets pertaining to six different contexts. We collected tweets over the entire election schedule. The details of tweets collected during the period is given in Table 7.1. The data sets show the quantum of tweets, the sources involved, the number of retweets of the messages and the number of user nodes retweeting them whom we name as 'retweeters'.

Table 7.1: Details of data collected during the elections.

Data set	Tweets	Retweets	Sources	Retweeters
Party A	967347	568550	23203	68469
Party B	944281	567232	19116	64009
Party C	275979	133603	6319	32502
Leader L_A	556994	345946	11703	56473
Leader L_B	855953	524187	12508	76037
Leader L_C	397180	257139	10284	55206

7.3.1 Segregation and Prediction of Most Repropagated Sources

We analysed the data sets using two different methods. We monitored the retweet behaviour continuously over the 45 day period in the first case. We also divided the whole period into phases or periods of five days each. Each phase was analysed independently to discern patterns of propagation. We carried out analysis of patterns of propagation in the retweet graph. Further analysis of sources of information propagation was done by core analysis algorithms.

We wanted to segregate sources whose messages are most heavily retweeted by looking at the inner cores of the repropagation graph. We initially determined the sources and their tweets which were most frequently tweeted. We then used iterative k-core decomposition algorithm given in Algorithm 2 of Chapter 5 on the data sets for all phases of elections to segregate sources in the inner cores. We compared the two list of sources to determine the effectiveness of the proposed methodology. The results are shown in Figure 7.1. We validated the results using measures of 'precision' and 'recall'. While the figure of 'recall' indicated the ability to segregate the most frequently repropagated sources, measure of 'precision' was used to measure as to how many of the segregated sources were actually heavily repropagated. The figures of recall and precision of results are given in Table 7.2.

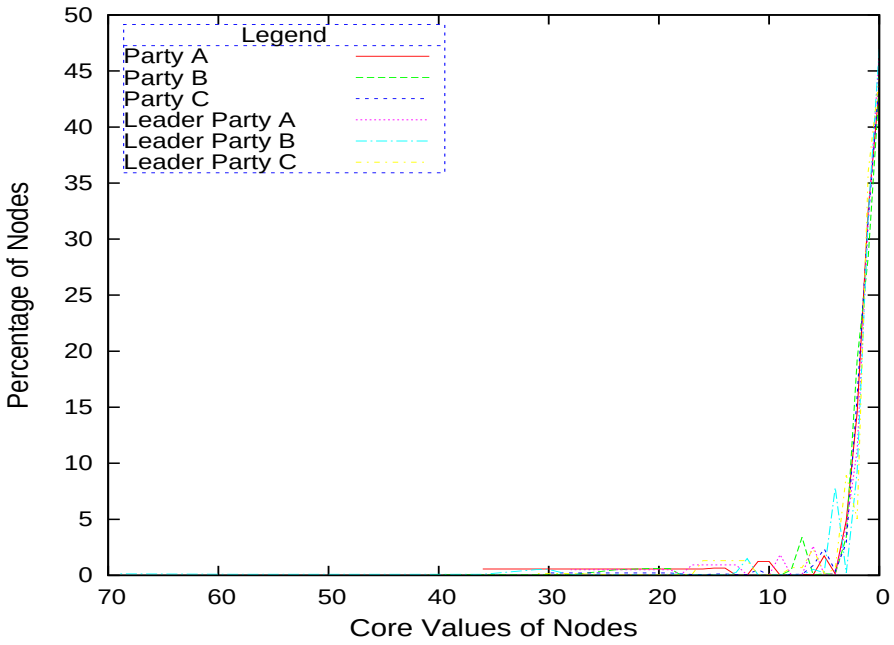


Figure 7.1: Segregation of sources of messages using iterative k-core decomposition algorithm on all the data sets of elections.

The results show that in order to achieve 100% recall of sources who are most frequently repropagated, it is sufficient if we look at the source nodes in the inner cores. The segregated source nodes in the inner cores shown in repropagation graph constituted less than 5% of the total number of nodes. The presence of sources in the inner cores would be a necessary

Table 7.2: Analysis of results of segregation of sources.

	Recall	Precision	Percentage of Nodes
Party A	100%	70.4%	1.11%
Party B	100%	88.5%	1.50%
Party C	100%	51.9%	1.81%
Leader Party A	100%	91.5%	2.52%
Leader Party B	100%	96.5%	2.1%
Leader Party C	100%	58.9%	3.78%

condition for being heavily repropagated, though not a sufficient condition. The results confirm our proposal of 'inside-to-outside' methodology where the sources in the inner cores are the ones whose messages are most frequently repropagated. Hence any sources of disinformation whose messages have the potential to spread in the population would also be present in the inner cores and could also be segregated in this manner.

The comparison of these cores across all phases showed that the users forming the inner most cores remained generally the same. More importantly, we wanted to predict the likely sources of messages who would be most frequently repropagated at the end of 45 days of elections, by looking at the initial data sets for say 5 or 10 days. If the sources in the cores are formed early, we would have an ideal mechanism for monitoring and detecting sources of misinformation who are likely to be heavily repropagated by looking at only a small percentage of the sources in the data sets. If their messages include efforts to spread rumours, propaganda or disinformation in OSNs, they could be monitored. Once these active users are separated out, other participating users in the spread of such information and messages involved could easily be separated by using standard community detection algorithms. For this, we analysed the data sets for the formation of inner cores in different phases. The aim was to find out exactly when the sources in the inner cores were active. We used the measures of 'recall' and 'precision' again. Here we wanted to achieve 100% recall of sources who are frequently repropagated at the end of 45 days by looking at the inner most cores alone of the data sets for the initial 5 or 10 days. The precision figures indicate how many of the segregated sources were actually most heavily repropagated. The application of Algorithm 2 of Chapter 5 for data sets in the different initial phases of data collection is shown in Figure 7.2. The results are given in Table 7.3.

We notice that the sources which were subsequently heavily repropagated were present quite early in the period. All such sources could be identified by their presence in the inner most cores of the data sets of the initial period. The number of sources which were segregated to achieve 100% recall again was a small fraction of the sources in the data sets. The data is for sources identified in the data sets for 5 days only. Similar results were obtained for data sets of longer periods of 10 days also.

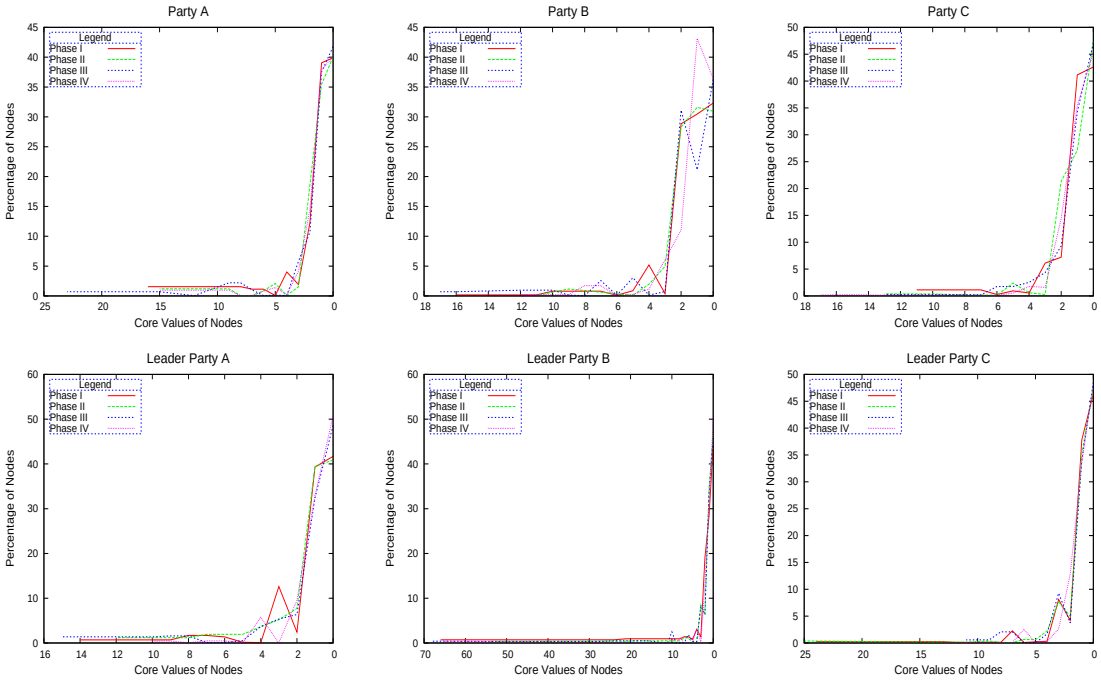


Figure 7.2: Segregation of sources of messages during initial phases of elections using iterative k-core decomposition algorithm.

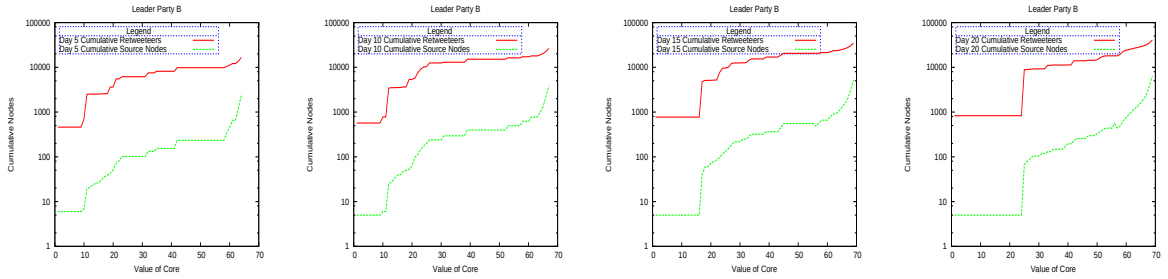
Table 7.3: Analysis of results of prediction of most repropagated sources.

	Recall	Precision	Percentage of Nodes
Party A	100%	66.19%	1.21%
Party B	100%	81.4%	1.63%
Party C	100%	57.06%	2.47%
Leader Party A	100%	60%	3.84%
Leader Party B	100%	78.29%	2.61%
Leader Party C	100%	76.89%	2.44%

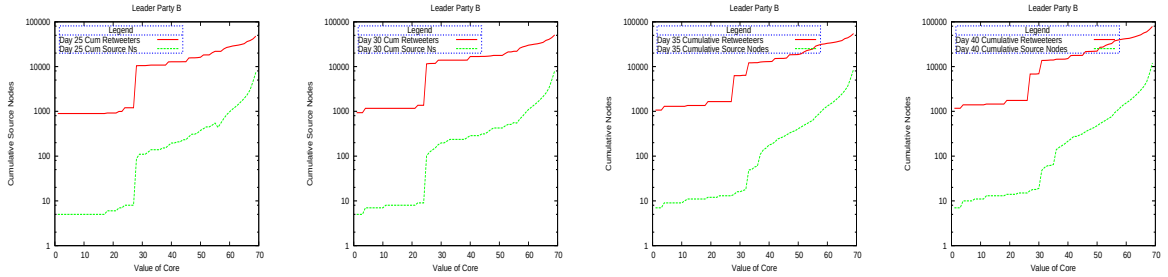
7.3.2 Evolution of User Behaviour

The progressive evolution of behaviour of users was studied over the complete period. The graphs in Figure 7.3 show the cumulative behaviour of users in one of the data sets, that of Leader of Party B, at intervals of five days each. We are depicting the results at intervals of five days to make the analysis easier and meaningful. Log scale is used for the Y-axis.

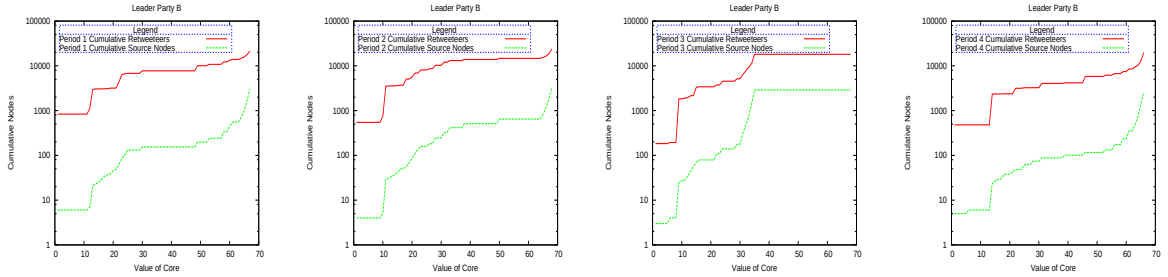
We analysed the retweet behaviour based on each of the source nodes. We calculated the coreness of the source nodes and gave them values from inner most to the outermost ones in the retweet graph, with the inner most core being given a value of 1. The retweeters of the source nodes were also separated and arranged with their corresponding source nodes.



(a) Cumulative Core wise Plot (Day wise) of Leader Party B (Day 5 to Day 20)



(b) Cumulative Core wise Plot (Day wise) of Leader Party B (Day 25 to Day40)



(c) Cumulative Core wise Plot (Period wise) of Leader Party B (Phase 1 to Phase 4)

Figure 7.3: Core wise correlation between source nodes and retweeters for L_B data set.

The cumulative number of source nodes and their corresponding cumulative number of retweeters were plotted against the core values. The graphs show the excellent correlation between the two. The high correlation shows that the quality of information spreading in the network could be studied by starting with the source nodes in the innermost cores. The analysis reveals that the core wise segregation of sources and their cumulative behaviour would be an accurate indicator of the retweet behaviour of users in the data set.

Segregation of source nodes in the inner cores would also reveal those users who coordinate with each other to spread propaganda or disinformation. The first two rows in Figure 7.3 show the distribution for cumulative day wise distribution of source nodes and retweeters at intervals of five days. The third row shows the evolution of users when the phases are considered independently. We show the results only for four initial phases. We notice that the two curves show excellent correlation in all the graphs. We obtained similar results for other data sets also.

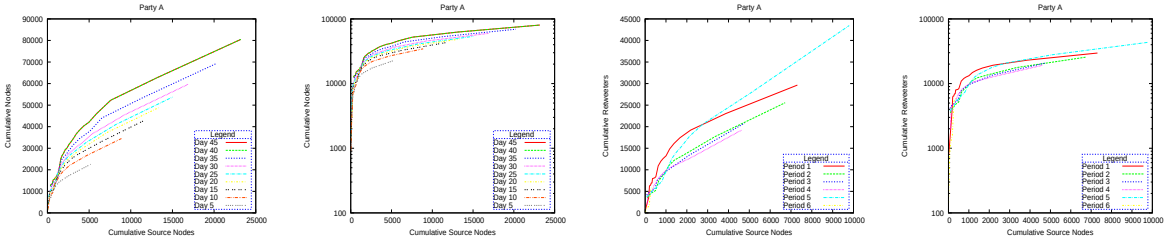
In order to further study the correlation between the sources and their retweeters, we

plotted cumulative number of source nodes arranged from inner most to the outer most cores and their corresponding retweeters in Figure 7.4. The results are for all the data sets with one row for each of them. Further in each row, the first two figures show day wise cumulative distribution of source nodes and retweet nodes, one using normal scale and the other using log scale. The third and fourth graphs in each row show similar results for different phases as discussed earlier. Each of the phase is considered independently. For each of the curve in the graphs, there is an initial steep linear portion, followed by another linear portion with distinct reduction in slope. The initial steep portion reveals that fewer source nodes are involved in the steeper increase in the number of retweeters. These source nodes, which are smaller in number, could be easily separated. Further increase in retweeters is linear showing similar increases in source nodes and retweeters. Similar results were obtained for the distribution of source nodes and retweeters for all the data sets.

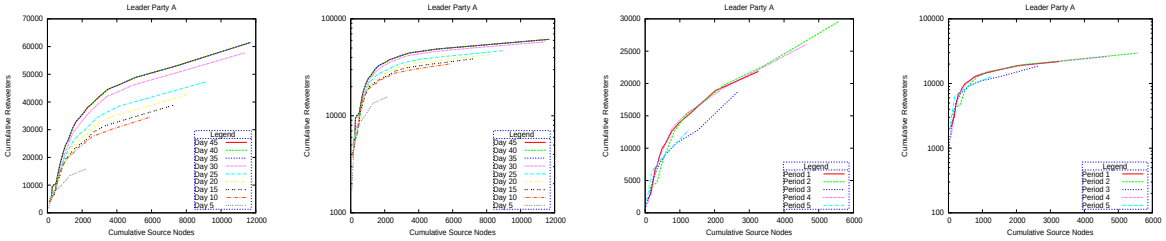
We analysed the tweets of the source nodes in the inner most cores which were separated. All the sources of information which were identified to spread biased information were part of these inner cores. All the source nodes in inner cores were not involved in the spread of biased information. But all sources involved in deliberate spread of possibly less credible information were found in these inner cores. The slopes of the curves in Figure 7.4 between cumulative source nodes and cumulative retweeters would give the increase in the number of retweeters with a corresponding increase in the number of source nodes. The graphs show the correlation between the two for all data sets.

The analysis reveals the usefulness of isolating the source nodes in this manner. In order to quickly segregate possible spread of disinformation and biased information including propaganda, an *inside-out* approach would be most suitable. The coordination between users to spread information would result in higher value of coreness of the sources. All sources with higher coreness values were not involved in coordinated spread of information. However, all source nodes who made substantial efforts to coordinate amongst each other to spread information were found in these inner cores. The number of source nodes in the inner cores was much smaller as compared to the total number of source nodes and the retweeters, reducing the amount of data which would need to be further analysed. The percentage of source nodes thus separated is less than 2% of the total number of source nodes. Content analysis of their tweets could be done much faster than resorting to content analysis of the entire set of tweets to segregate possible source nodes spreading propaganda or disinformation. The separation of inner most cores for each of the period helped us to isolate source nodes making coordinated efforts to spread information. Such isolation which happens in the earlier period of data collection would enable us to monitor these source nodes and take effective counter action if necessary. This, we believe would form the basis of any OSN monitoring system.

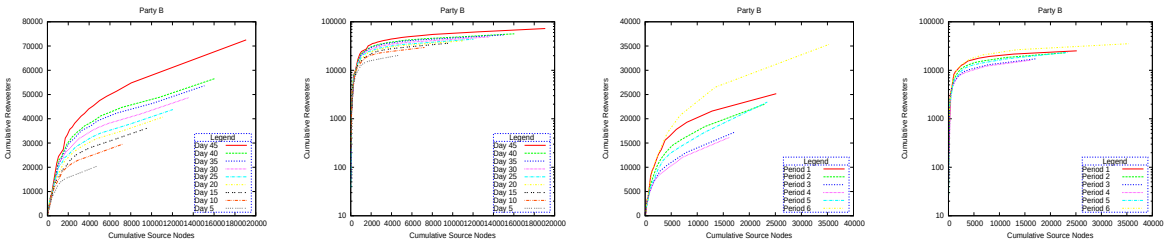
The linear variation between log of cumulative number of source nodes and the log of the cumulative number of retweeters is obvious from Figure 7.5. The source nodes were arranged from the innermost core to the outermost in the graph. The first two rows in the



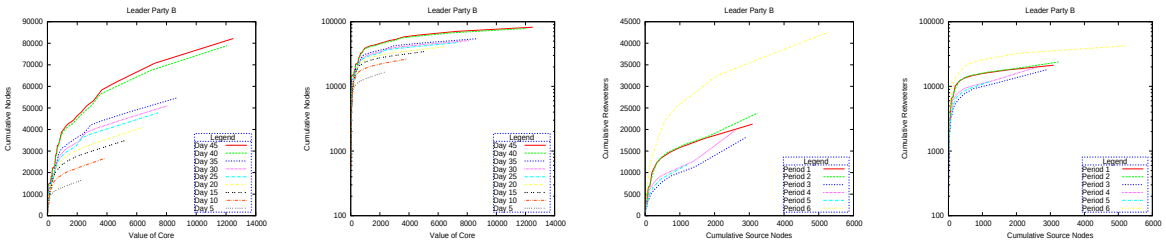
(a) Party A: Day wise **(b) Party A: Period wise**



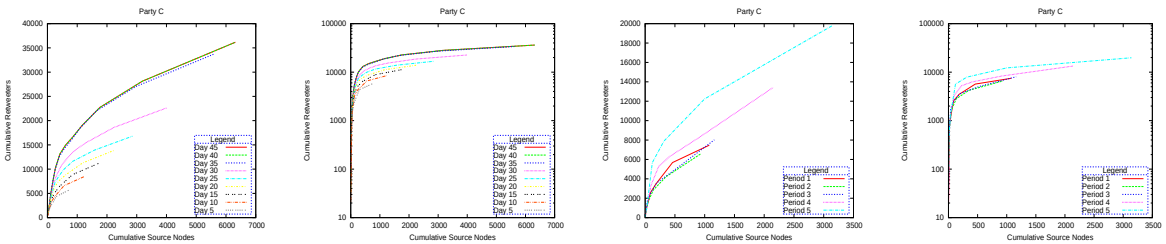
(c) Leader Party A: Day wise **(d) Leader Party A: Period wise**



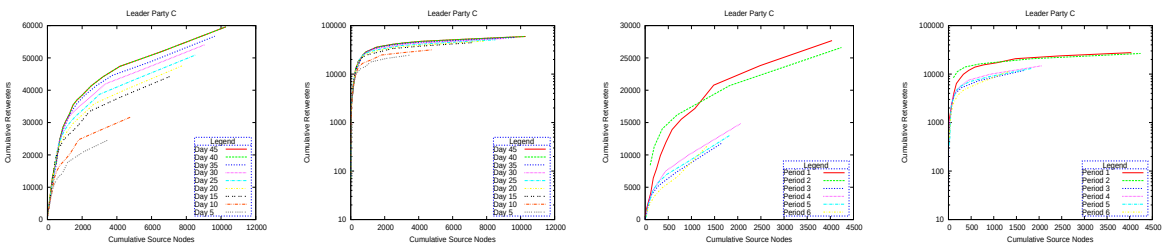
(e) Party B: Day wise **(f) Party B: Period wise**



(g) Leader Party B: Day wise **(h) Leader Party B: Period wise**



(i) Party C: Day wise **(j) Party C: Period wise**



(k) Leader Party C: Day wise **(l) Leader Party C: Period wise**

Figure 7.4: Corewise correlation between source Nodes and retweeters in all data sets (day wise and period wise).

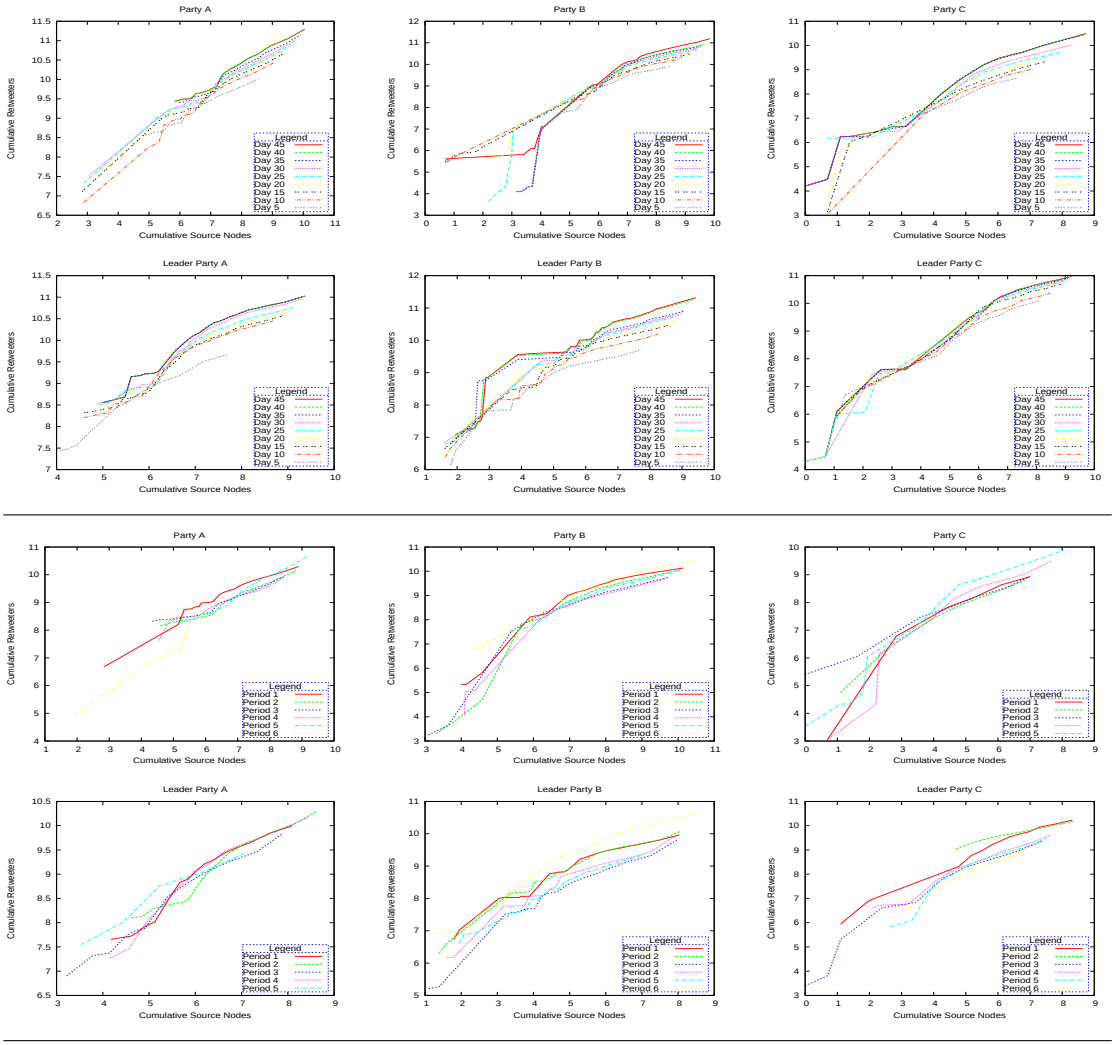


Figure 7.5: Correlation between logarithm of cumulative number of source nodes and logarithm of cumulative number of corresponding retweeters.

figure show the plot for all data sets when cumulative day wise distribution of source nodes and retweet nodes are considered. The third and fourth rows in the figure show the same when all the phases were considered independently. The lines for the cumulative distribution day wise and phase wise show excellent correlation. The Pearson's correlation coefficients obtained for the graphs were in excess of 0.92 in all cases and going up to 0.99 in many cases. The figure indicates that the extent of spread of information could be estimated by studying the quantum of source nodes in the inner cores. The excellent correlation between the initial graphs obtained during the process - both in the first dew days as well as during the first period, and their subsequent distribution during the remaining days of electioneering gives us the ability to predict the likely final spread of retweeters, in the event that the stimulus for the spread exists throughout the period. This is an important observation as it enables us to predict the spread of information diffusion of a set of similar messages by studying the initial repropagation behaviour of users. In order to prevent formation of

disinformation cascades, ability to accurately estimate the likely final spread of information in the population would be an important step to control the spread by adopting suitable counter measures.

7.3.3 Detection and Analysis of User Communities

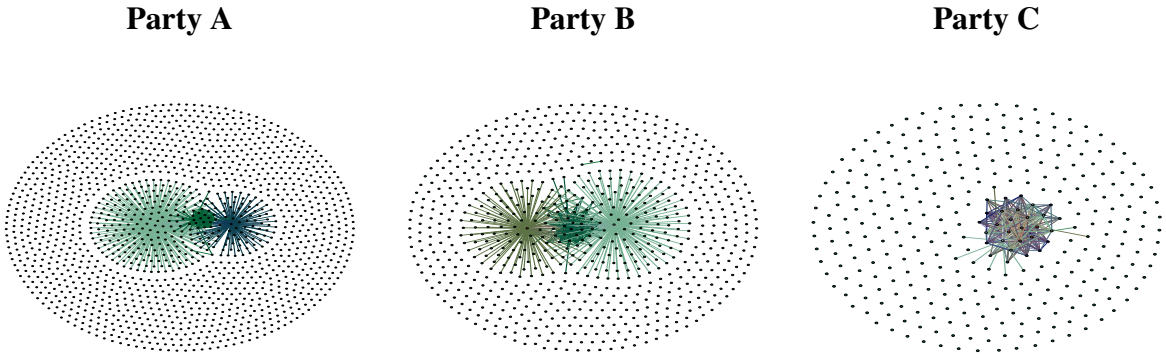
We analysed the connectivity of the retweet graph and formation of homogenous communities based on types of messages being retweeted. We used standard community detection algorithms based on modularity [103] to detect the presence of homogenous communities of user and retweet nodes. The analysis supports our earlier results. There were large number of isolated communities with very few of them having edges between them as shown in Figure 7.6. The size of the communities were also not uniform and exhibited a tailed distribution confirming the fact that the political ideologies have created strong communities and users were involved in the repropagation of selected tweets only. The classification of such tweets as true or false information could be done after detection of deliberateness in their efforts. The probability of such tweets reaching a large section of the population is much larger when supported by deliberate actions by users.

We semantically analysed the communities. Most of the communities had a strong presence of users who were either propagating 'For' a particular political party and leader or propagating messages 'Against' them exclusively. There were 'Neutral' communities also giving balanced views. There were a few communities which were interconnected for all the data sets. The size and number of such communities were a small percentage of all the communities. The existence of such closely linked communities enables us to group together users of similar profile and obtain an aggregate view of semantic contents in their messages to be analysed easily. By initial segregation of sources and their messages based on core analysis and further analysis of users in the core based on their communities would help us to systematically study information diffusion - especially orchestrated spread of information to include disinformation, propaganda etc.

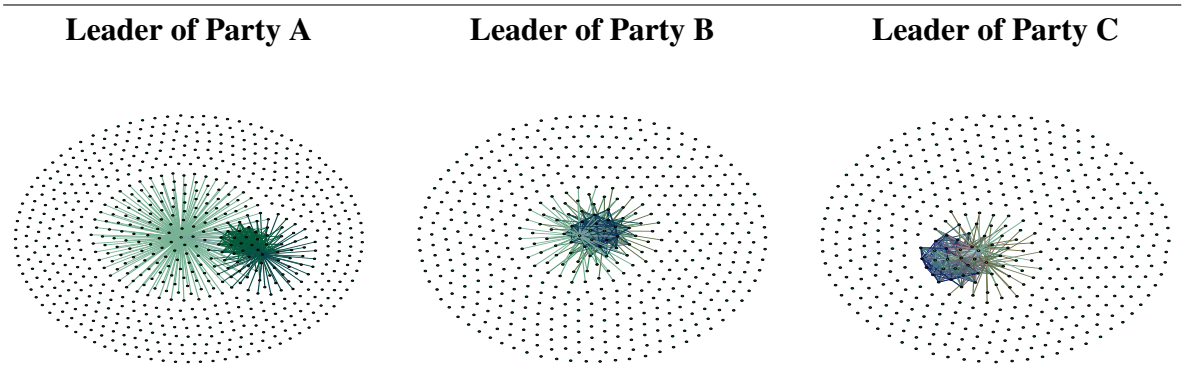
7.3.4 Psychometric Analysis

Psychometric analysis of users based on dichotomous responses

The process of separating the sources spreading misinformation or propaganda and tweets could be done using Psychometric analysis. We had seen earlier it clearly establishes the social computing properties of users of OSNs. We carried out Psychometric analysis based on dichotomous responses of the communities in the repropagation graph. Sample results obtained from each of the data sets are given at Figure 7.7. The results obtained were matched against our observed results and we found the method to be effective with an accuracy of over 92%.



(a) Communities in data sets of the three political parties



(b) Communities in data sets of the three leaders of political parties

Figure 7.6: Presence of large number of disconnected communities in the election data sets.

Psychometric analysis of behavioural trust relationships between users based on polytomous responses

We evaluated the behavioural trust relationships in communities with high gini coefficients and entropy values. The psychometric analysis of trust relationships would help us separate sources to be monitored along with other users involved. A set of item response characteristic curves, item information curves and test information curves from the data sets is given in Figure 7.8. The identification of sources and users spreading misinformation or propaganda could be done with an accuracy of over 90%.

7.3.5 Social Capital in the Communities

Having verified the social computing properties of users to detect misinformation or propaganda, we carried out the analysis as per Behaviour Trust model. We constructed retweet graph as the repropagation graph or the trust graph. The social capital of the communities

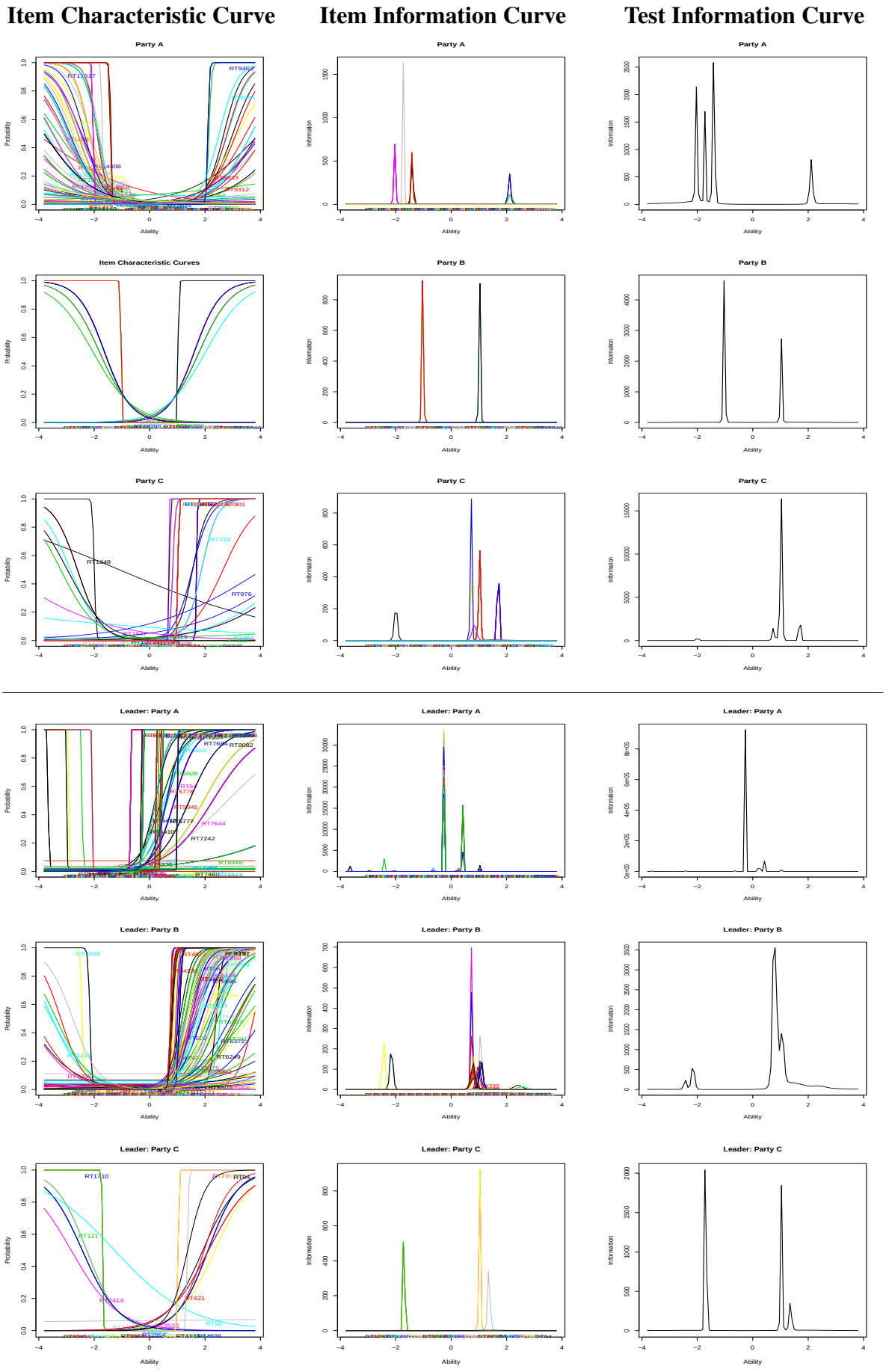


Figure 7.7: Representative item characteristics curves, item information curves and test information curves based on dichotomous responses in all the data sets pertaining to elections.

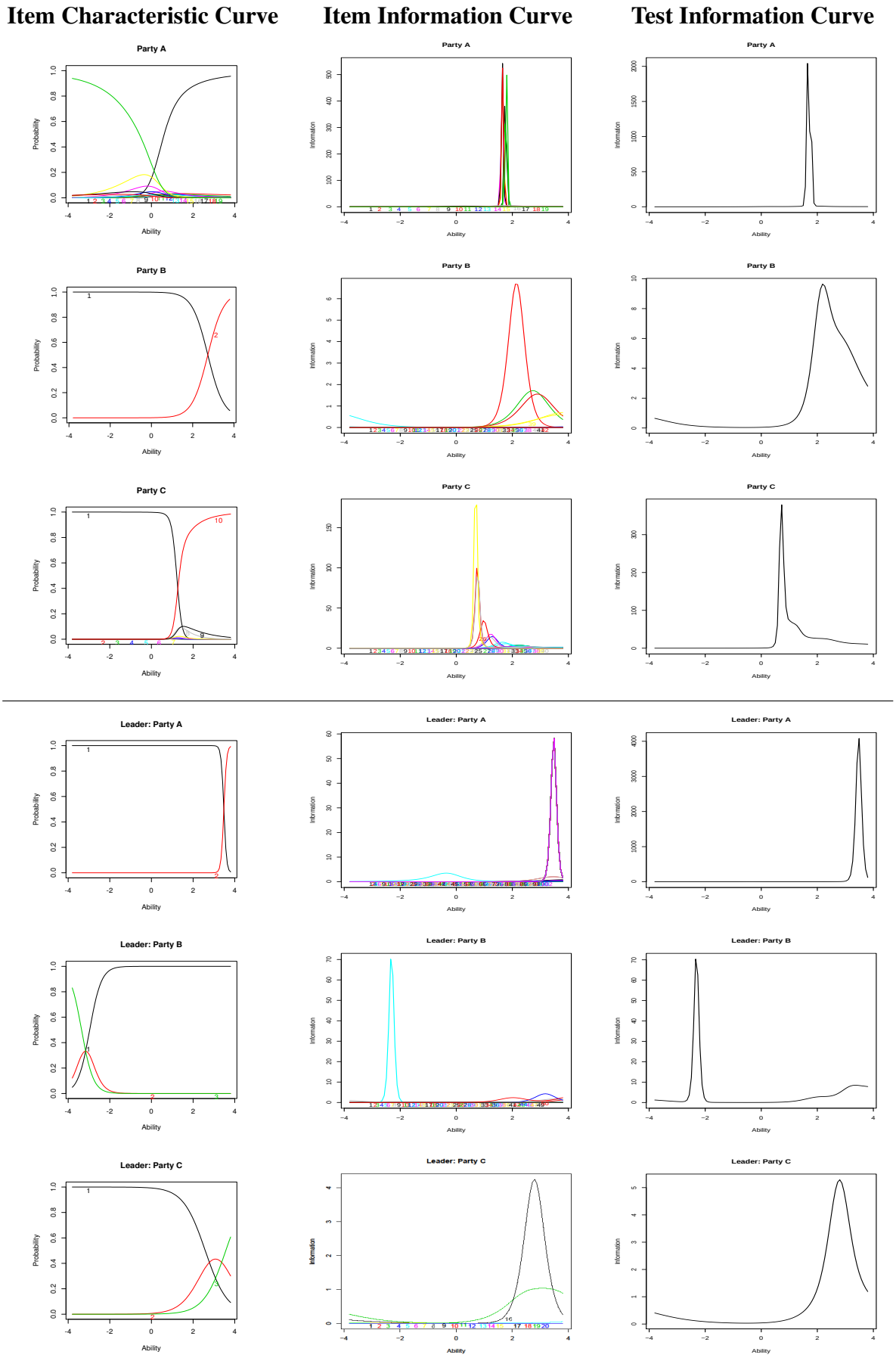


Figure 7.8: Representative item response characteristics curves, item information curves and test information curves based on polytomous responses in all the data sets pertaining to elections.

in the graphs was estimated using Algorithm 5 in Chapter 5. The distribution of entropy of communities and their gini coefficients showed a tailed distribution with few communities having higher values for both metrics and most of them having very low values. Plots of distribution of entropy and gini coefficient for the six data sets are given in Figure 7.9.

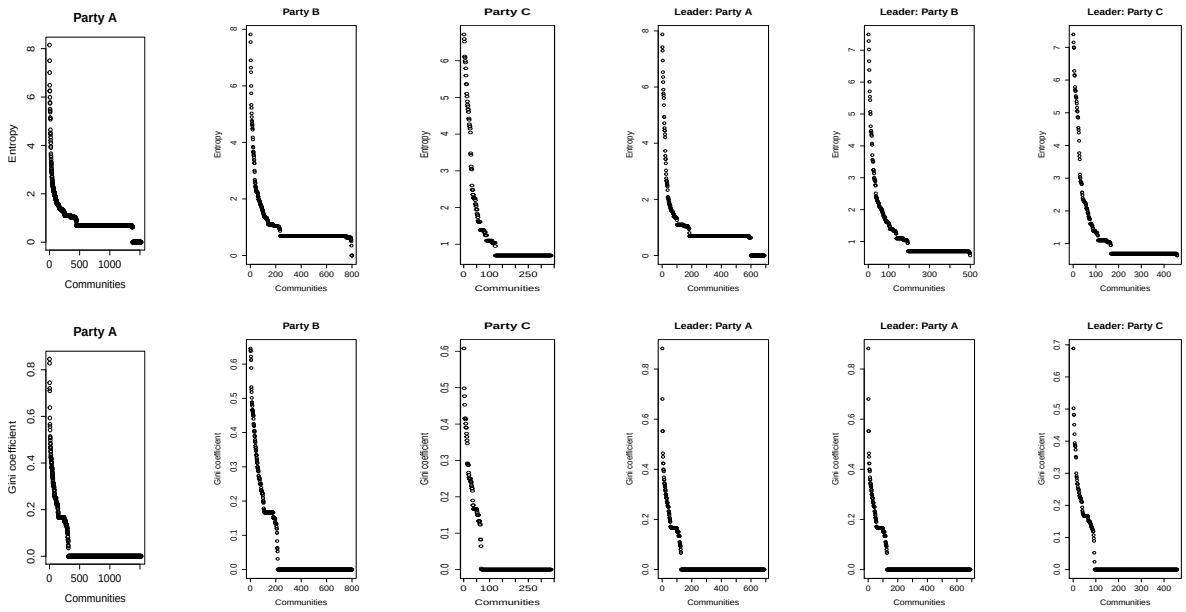


Figure 7.9: Distribution of entropy and gini coefficients of communities in the election data sets.

We studied the variation of entropy and gini coefficient across all communities. A plot of entropy values and gini coefficients of the communities was made. In order to segregate the communities based on entropy and gini values, we used spatial clustering algorithm proposed in [133], which cluster points based on medoids. This enabled us to spatially cluster communities with similar entropy and gini coefficients in the plot. The plots for the various data sets are given at Figure 7.10. The possible interpretation of the values of gini coefficient and entropy was given in Table 5.7. Higher entropy value indicates greater interactions in the community and lower gini coefficient indicates more even distribution of repropagated news items among the nodes and hence greater credibility of information.

The classification into four quadrants in each of the data sets was done based on *k*-medoid spatial clustering algorithm [133] and approximate threshold value of gini coefficient of 0.45 observed in all data sets for credible information. The different clusters of communities are shown in the graph. The communities in clusters shown in (H,H) and (L,H) quadrants have a high probability of less credible information. The credibility of information was highest in the communities in (H,L) quadrant. The information in communities in (L,L) were not misinformation, but possessed less value. The measure of 'recall', for identifying potential misinforming communities in the (H,H) and (L,H) communities was over 95%. The 'precision' was between 65 to 70%. We propose the use of this method as

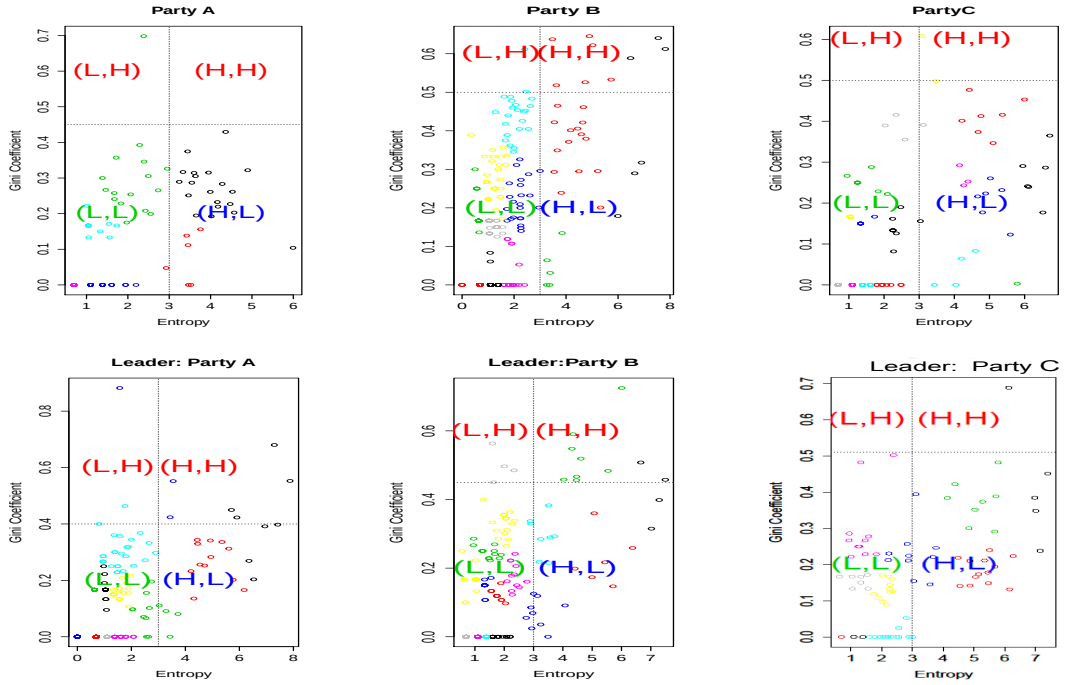


Figure 7.10: Variation of Entropy and Gini coefficients of communities in the election data sets.

the first stage of the social media monitoring system before more computationally intensive semantic analysis is carried out. Towards this end, the methodology adopted has proved to be very efficient in terms of computation, especially for large networks and when a near real time monitoring system is required. The number of communities in the (H,H) and the (L,H) quadrants, which requires further semantic analysis is less than 5% of the total communities in the graph.

7.3.6 Evaluation of Credibility of Sources

We evaluated the credibility of sources using Algorithm 6 in Chapter 5. The use of gini coefficients of the degree distribution of the sub graph rooted at the sources of information would indicate the credibility of sources as shown in the previous section. We improve the accuracy of calculating the *Credibility scores* of sources by using their weighted and unweighted Eigen Centrality (EC) ranks. The results obtained for the graphs of weighted and unweighted EC ranks for all the data sets are shown in Figure 7.11. The colors of the nodes indicate the clusters identified by the clustering algorithm [133]. If both types of EC ranks were comparable, the distribution of nodes would lie on the line where the two EC ranks are equal (shown in dotted lines in the figures). The panels for the Party A and Leader of Party A data sets show most of the cluster of nodes along the equal rank line. The panels for the Party B and Leader of Party B data sets also show concentration of sources

of information along the equal rank line. The panels for Party C and Leader of Party C show little correlation between the two EC ranks and underline the presence of a heavy section of colluding users. This was verified by semantic analysis of tweets of the users. The homogenous nature of users in data sets of Party A, Leader of Party A, Party B and Leader of Party B would require further analysis for the type of information propagation in them.

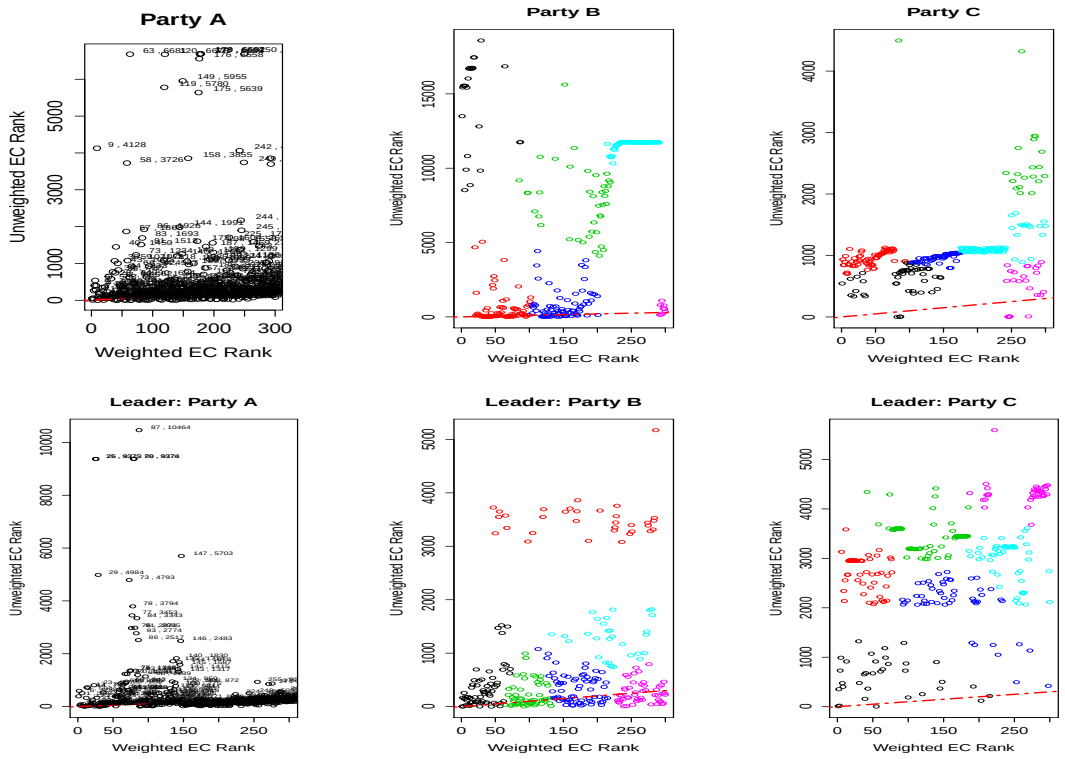


Figure 7.11: Comparison of weighted and unweighted eigen centrality rankings of nodes in the election data sets.

7.3.7 User Behaviour and Information Propagation during Elections

The summary of our analysis of retweet behaviour is given below. The results clearly bring out the fact that for carrying our cyber security analysis of networks involving people, Psychological studies can aid in providing faster and efficient solutions. Methodologies combining Computer Science and Psychology are better at solving problems created by Big Data in an acceptable time frame so as to take suitable counter measures. We summarise below our observations of user behaviour and analysis of information propagation during elections in India.

- Deliberate efforts to spread information or propaganda in the form of coordinated campaign were observed.

- Efforts to coordinate the spread of news in OSNs varied across political parties and their leaders.
- The patterns of propagation reveal artificial and coordinated efforts by the political parties rather than spontaneous involvement from the public.
- The parties have made effective use of principles of Cognitive Psychology to build up social consensus by repeated exposure to same or similar news items.
- Analysis of core-periphery model of retweet behaviour resulted in segregation of three different types of user behaviour - active 'For' and 'Against' types forming the inner cores, while bulk of the retweeting users were in the outer layers or the periphery.
- The formation of retweet communities displayed the political affiliations with lots of 'islands of communities' with little interconnections between them.
- The evolution of patterns of repropagation during different phases of elections remained the same. The regularity of patterns pointed towards coordinated campaigns to reach maximum possible users.
- More deliberate efforts by the parties resulted in the formation of deeper cores. Segregation and analysis of users and tweets of the inner cores brought out such efforts.
- The core wise plot of source nodes and their corresponding retweeters showed excellent correlation which could be used to predict final spread of information based on initial propagation patterns.
- The 'inside-out' strategy of concentrating on the fewer number of users in the inner most cores for studying deliberate spread of information and propaganda has been proved to be effective. The analysis of the larger numbers in the outer cores could be done subsequently. This has important lessons for the analysis of large scale spread of misinformation and disinformation in OSNs for the purpose of their early detection and launching preventive measures.

7.4 Framework to Prevent Spread of False Information

Prevention is better than cure. Based on the analysis of cognitive process, it becomes clear that the receiver obtains cues to deception or misinformation from the OSN to decide on the credibility of information. The same cues could be used by a social media monitoring system to detect spread of misinformation, disinformation or propaganda. The proposed framework for such a system to help a user to make informed decisions is given in Figure 7.12.

The user would employ his cognitive capabilities to determine the *cognitive consistency* and *cognitive coherency* of the messages received. The algorithms proposed in this work

would help to quantify the *credibility of the source* and the general acceptability of message in terms of *social consensus* to the user. With non biased input of these two factors, a user should be able to make informed decisions regarding the truthfulness of messages and applying his own coherency and consistency values.

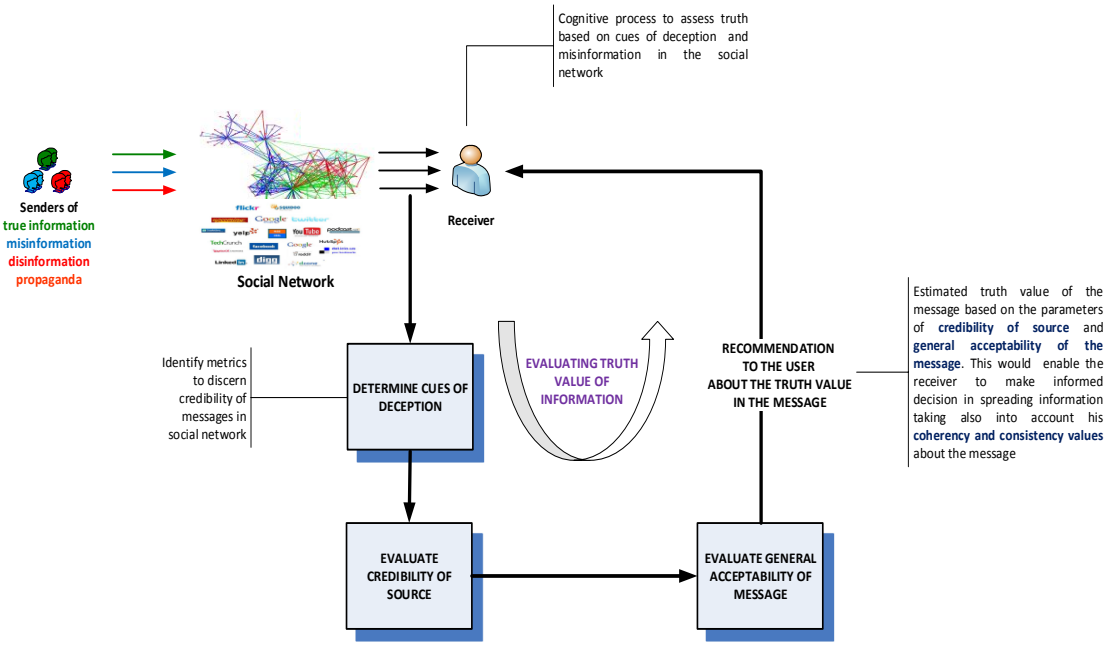


Figure 7.12: A generic framework for a user of OSN to evaluate credibility of information.

7.4.1 Reputation System for OSN

We propose the requirement of a reputation system for OSNs similar to the existing ones in other social computing systems like e-commerce portals. OSNs are important sources of information. Every user is also a creator of information and detecting false information using a centralised system would prove to be near impossible in an effective time frame. The social network user would have to be diligent enough to use cues to credibility to detect non credible information. In this context, information external to the messages, in terms of the reputation of users and manipulation of spread of messages could be used to determine the credibility of messages. Computationally also, it is less expensive to monitor sources and study propagation based features than carry out semantic analysis of contents of messages.

We propose a reputation system for social network users to determine 'Quality scores' or 'Q-scores' of sources of messages based on ratings obtained using social computing properties of the network. Such ratings of sources of information would help users in expressing caution while repropagating messages. The methodology for determination of Q-scores in OSNs for all sources of information is depicted in Figure 7.13. It combines the ratings obtained by users based on the manner in which their messages are repropagated. The

proposed metrics could also be extended to include other Natural Language Programming (NLP) techniques to further refine the credibility of sources. The availability of credibility rating of a source along with the messages themselves would help users to make better decisions while participating in the repropagation process in OSNs. The distribution of Q-scores of different data sets calculated based on our rankings is given in Figure 7.14.

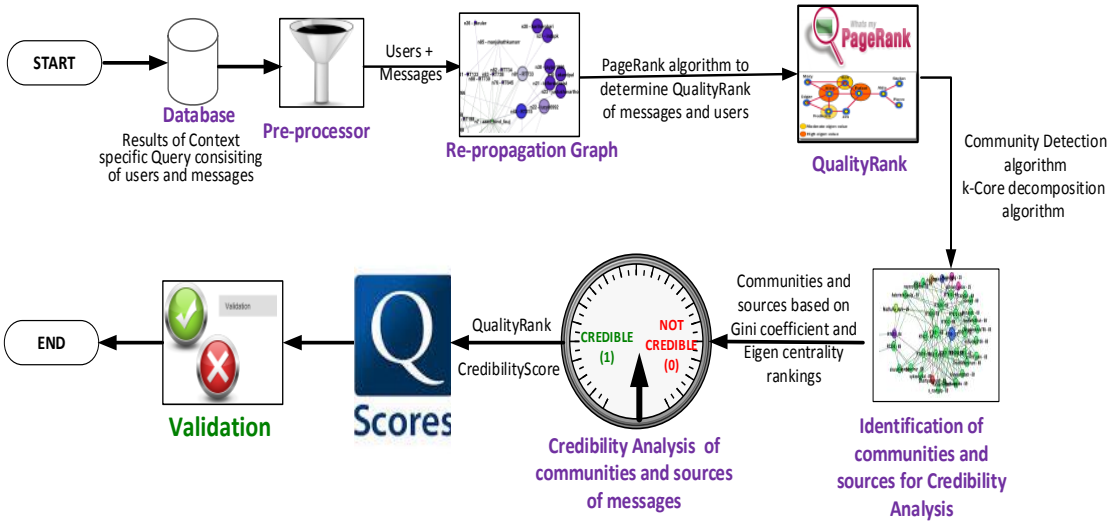


Figure 7.13: Methodology for evaluation of Quality scores of sources of information.

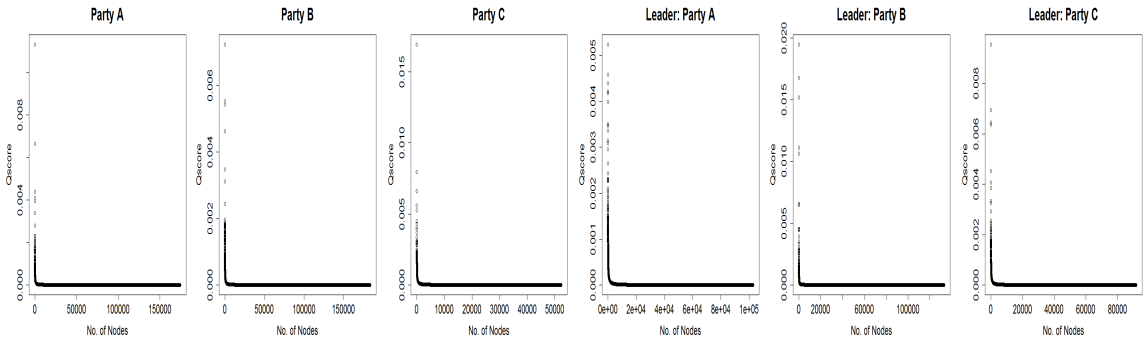


Figure 7.14: Distribution of Quality scores in data sets showing a tailed distribution.

The determination of Q-score is a function of *Page Rank* of the sources as well as their *Credibility Scores*. While the use of page rank is nothing new, we use credibility scores to isolate sources intending to game the system to improve their rankings. The Quality ranks derived from PageRank algorithm [56] need to be modified to filter coordinated efforts made by users to spread their messages. We have proposed two measures of establishing

credibility of sources in terms of gini coefficient and variation in weighted and unweighted EC rankings of the sources. Higher values of gini coefficient is an indication of lower credibility of the sources. Greater difference in weighted and unweighted EC ranks would also indicate lower credibility of sources. While all deliberate efforts to spread certain types of messages may not be misinformation, simulation results prove that spread of less credible messages are more likely to be successful when supported by coordinated actions of sections of users. Such efforts if made known to the users, who would also use their cognitive powers to understand the contents of the messages, would enable them to make informed decisions while participating in the repropagation process.

7.4.2 Proposed Framework for Cyber Surveillance

The adoption of information by a user and its subsequent diffusion in OSNs have been studied from Cognitive Psychology and Behaviour trust model. Based on the model, we summarise our proposal of a framework for the development of a social media monitoring system as shown in Figure 7.15. The proposed model integrates propagation based and content based features for an OSN monitoring system.

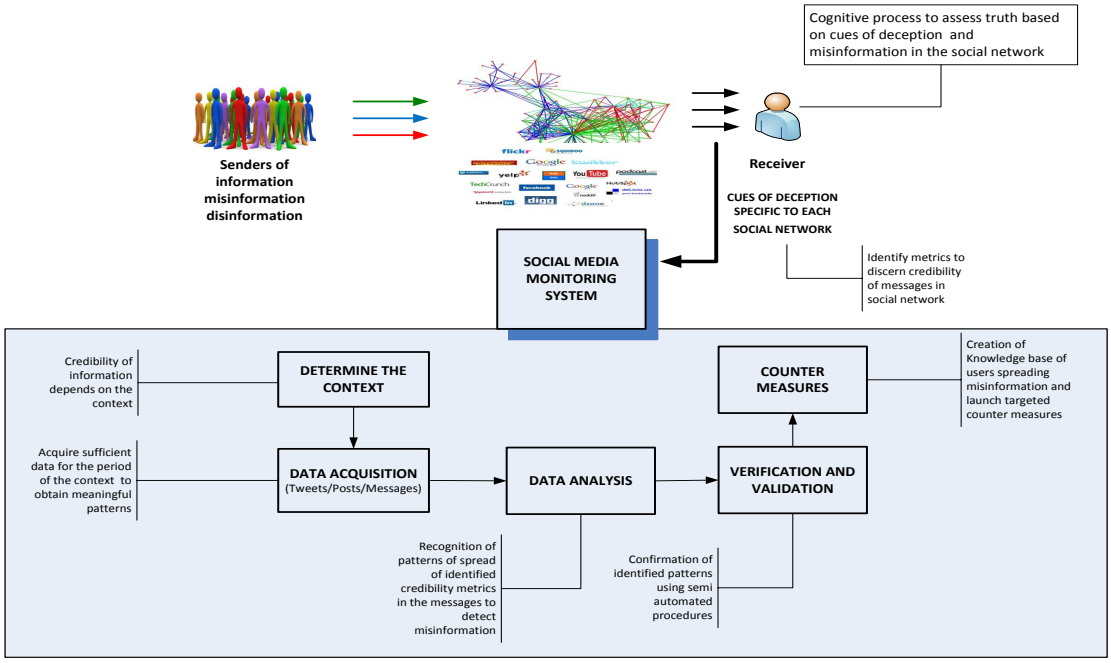


Figure 7.15: A generic framework for detection of the spread of misinformation in OSNs.

The detection process begins with identification of suitable credibility metrics of the social network being studied to evaluate the cues by which a user would have made the decision of evaluating the accuracy of information. The subsequent stages of the social media monitoring system involve **acquisition of data** from the social network and to separate out the identified metrics based on the **context** being studied. The segregation of messages

based on context could be done using keyword search of the messages based on keywords provided by subject matter experts. The **data analysis** stage would use suitable algorithms to identify patterns in the spread of identified metrics to detect misinformation. The segregated contents need **verification and validation**. The whole process would be a continuous loop to identify the misinforming messages as well as the users spreading misinformation. Suitable **counter measures** could be launched in the next stage to prevent further spread of misinformation. The whole loop indicates a continuous process of knowledge building whereby the history of users and patterns of spread of misinformation are stored and analysed to help improve the accuracy of the process.

Data analysis would use the core-periphery and the community structure of the repropagation graph to detect the potential to spread throughout the population. The intrinsic attributes of messages may not be sufficient to estimate the effect of social influence in their spread. The spread of less credible information is similar to diffusion of innovations, which requires a certain number of early adopters to accept the innovation and cause its subsequent spread. Any social media monitoring system would be scalable if it could use the social computing properties of the network. Collaborative filter mechanisms prior to more computationally intensive semantic analysis of messages would provide solutions for effective cyber surveillance. The proposed methodology is summarised in Figure 7.16. The OSN reputation system would form part of the counter measures to limit the effects of deliberate spread of false information.

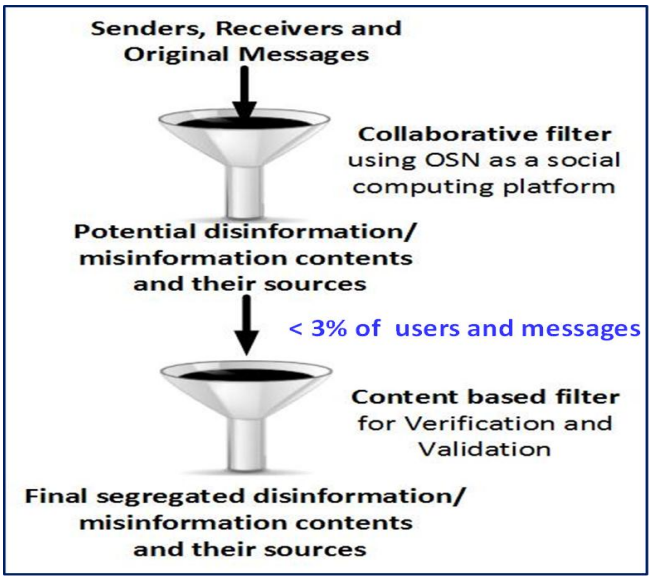


Figure 7.16: Proposed methodology to use collaborative filter algorithms prior to any content based filtering.

7.5 Summary

We have analysed the use of social media to spread propaganda during elections in the largest democracy in the world. The largest democracy also has the maximum population of youth in the world. The use of OSNs to reach out to all users during elections, especially the youth was studied in this research work. All parties and their leaders had their presence in OSNs for propagating their ideas and views. All of them made deliberate efforts to spread them as much as possible to influence voters. Psychological analysis of users spreading information along with social network analysis algorithms have proved effective in understanding their behaviour. Such analysis also highlighted the efforts taken by sources of information to ensure greater spread of their messages.

The deliberate attempts seen during elections may not always be to spread false information, but the information was in support of one party or the other. Deliberate attempts of similar kind could be made to spread biased information and proper planning and effective implementation would result in the same being disseminated to a wide audience in OSNs and being accepted as true. It is because of this reason, that monitoring of OSNs has become an important task of law enforcement agencies world wide. Monitoring the quality of information in such networks is very important to prevent semantic attacks of any kind in them.

Analysis of retweet behaviour has revealed highly regular patterns in the repropagation graph due to action of dedicated set of users to spread information. The users form strong communities spreading similar messages in OSNs and supporting the theory of 'cyber-ghettos'. Analysis of communities of users have revealed the social computing properties of OSNs where users act as filters for deciding the quality of information. Development of early warning system in OSNs to prevent spread of non-credible information would have to use the social computing capabilities of such networks and integrating research work in the fields of Psychology and Computer Science.

Chapter 8

Conclusions and Future Scope

In this research work we explored the development of suitable cyber surveillance systems to monitor large scale spread of false information in the form of disinformation, propaganda and misinformation in OSNs. We have proposed an integrated model consolidating work done in the fields of and Computer Science and Behaviour Sciences to include Cognitive Psychology and Sociology.

Semantic attacks in OSNs have serious consequences. The major contribution of our work is the capability of the proposed system to carry out near real time monitoring of activities in OSNs using computationally efficient algorithms. The system also prevents spread of misinformation by enabling end users to make informed decisions while propagating information. We summarise the important deductions from our work in the next section.

8.1 Summary of Deductions

Spread of non credible information in the form of misinformation, disinformation and propaganda have the potential to cause semantic attacks and influence decision making of users of OSNs. Deliberate efforts to spread false information would be higher when the intention is to reach a wide audience. Semantic analysis of contents of messages alone would not be able to detect the spread of misinformation in an acceptable time frame. Propagation based features and other methods are required to be integrated.

Integrating researches in the fields of Computer Science and Behavioural Sciences to include Cognitive Psychology and Sociology have the potential to yield better results in analysing OSNs. Analysis of literature using principles of Cognitive Psychology has brought out ‘repropagation’ as a suitable measure to study adoption and diffusion of information. The use of ‘repropagation’ as a common metric of analysis has been found to be accurate and the availability of this feature in all OSNs makes it an ideal choice for analysis.

Gini coefficient of degree distribution of user nodes in a repropagation graph could be used as a metric to determine credibility of sources of messages and their acceptability. Diffusion of information by profiling of users based on semantic contents of messages and

their acceptance could be done using evolutionary game theory as a cooperator-defector game. Evolutionary games played on network graphs are studied using evolutionary graph theory. Analysing OSNs using isothermal graphs and bi-level evolutionary graphs yielded detection of core set of users involved in deliberate spread of information. Bi-level evolutionary graphs were constructed using community detection algorithms. The detection of communities of spread of information and their analysis resulted in segregation of misinforming sources and other users and messages involved. The core-periphery structure of repropagation graph provided the best means of studying these communities.

The essence of using information diffusion modelling to detect credibility of information is the social computing properties of users to determine the credibility of information. This inherent ability of social network users has been proved using Psychometric analysis of users. Using dichotomous responses to study quality of information and polytomous responses to qualify trust and credibility of sources of information have proved to be effective. The communities were further analysed using behaviour trust models. Metrics of entropy and gini coefficient were found useful in measuring social capital of these communities and segregating them.

The research work has proved that the collective intelligence of users of OSNs could be used to decide the credibility of information and reliability of sources of news items. We have proposed a framework for an online reputation system in OSNs which uses the integrated model approach proposed in our research work. The reputation system would prevent the spread of less credible information by aiding users to make informed decisions while repropagating information. For any cyber surveillance system, it would segregate the communities to be monitored for spread of misinformation, disinformation and propaganda.

8.2 Future Scope of Work

We have made initial efforts towards providing a computationally efficient solution in carrying out cyber surveillance of OSNs. The results obtained through our work have confirmed the potential of integrating research works in the fields of Behavioural Sciences, game theory and Computer Science to develop practical solutions to problems which involve human beings. We would like to experiment further to develop a full fledged reputation system for OSNs which could work in near real time basis. Development of a system at the client end in the form of an app or browser plugin which could work as a personalised reputation system to rank sources of information could be undertaken based on the results obtained.

We would like to further develop on the Psychometric analysis to detect misinforming sources. Quantification of latent abilities of individuals and their utilisation to provide workable solutions for problems involving human beings seem to be a more effective way of solving problems in the social computing domain. To that extent, the work done could be extended to other social computing platforms like recommender systems and reputation systems to improve the results in those fields.

The social capital of OSNs is an important parameter which defines the usefulness of the media for the purposes it is meant for. Freedom of expression, quality of information and privacy are a few concerns which affect proper utilisation of the networks. We would like to address the privacy issues of individuals in our future work.

8.3 Concluding Remarks

The role of psychological studies to improve cyber security has found strong evidence in this work. Semantic security is a very important dimension of cyber security. Misuse of media like OSNs needs to be prevented to ensure that social capital of such networks is maintained. People would trust such networks only when they receive information which is credible. An integrated solution to detect deliberate spread of information in OSNs prior to their classification as true or false information would help in near real time processing of credibility of messages.

The sustained growth of OSNs as a media for personal communication, e-commerce, brand building, and source of reliable and real time information would depend on the credibility of information presented by them and methodologies adopted to prevent manipulation of information. A reliable, scalable, effective and near real time reputation system for sources of information for each context has become a necessity. Any such system would have to make use of the social computing properties of users themselves as brought out in our work.

The results clearly brings out the requirement for greater vigilance and monitoring of OSNs. We do not advocate restriction of access or freedom of expression of users. However, a form of ‘citizen-policing’ where each user contributes towards the safety and security of the network would be necessary to maintain autonomy of such systems. A reputation system which can quantify the online behaviour of sources of information and consolidate responses of individuals to different messages in conjunction with other propagation based features would help in keeping OSNs as a media for reliable and real time information.

List of Publications Published/Accepted

International Journals

[Pub1] KP Krishna Kumar, Agrima Srivastava and G. Geethakumari (2014). "Psychometric Analysis of Information Propagation in Online Social Networks using Latent Trait Theory" *Journal of Computing by Springer.* Accepted for publication.

[Pub2] KP Krishna Kumar, and G. Geethakumari (2014). "Detecting Misinformation in Online Social Networks using Cognitive Psychology." *Human-centric Computing and Information Sciences - HCIS-D-14-00002 by Springer.* ISSN:2192-1962, Vol 4, No.14, Sep 2014.

[Pub3] KP Krishna Kumar, and G. Geethakumari (2014). "Analysis and Modelling of Semantic Attacks in Online Social Networks" *International Journal of Trust Management in Computing and Communications (IJTMCC), Inder Science Publishers. ISSN:2048-8378 (Print), 2048-8386 (Online), Vol. 2, No. 3, 2014, pp 207 – 228.*

[Pub4] KP Krishna Kumar, and G. Geethakumari (2014). "A Taxonomy for Modelling and Analysis of Diffusion of (mis)information in Social Networks." *International Journal of Communication Networks and Distributed Systems, Inderscience Publishers*, Vol. 13, No.2, 2014, pp 119 – 143.

International Conferences

[Pub5] KP Krishna Kumar, Agrima Srivastava and G. Geethakumari (2014). "Preventing Disinformation Cascades using Behavioural Trust in Online Social Networks." *International Conference on Advances in Computing, Communications and Information Science, ACCIS - 14, June 26 - 28, 2014, Kollam, India.* Proceedings in Elsevier Publications, ISBN: 9789351072478, pp 113–123.

[Pub6] KP Krishna Kumar, and G. Geethakumari (2014). "Identifying Sources of Misinformation in Online Social Networks." *International Symposium on Signal Processing and Intelligent Recognition Systems, (SIRS-2014).,* March 13-15, 2014 Technopark, Trivandrum, India. Proceedings in Springer Series: Advances in Intelligent and Soft Computing Journal, Special Volume, pp 417–428. *Best paper award.*

[Pub7] KP Krishna Kumar, and G. Geethakumari (2014). "Analysis of Semantic Attacks in Online Social Networks" *International Conference on Security in Computer Networks and Distributed Systems (SNDS-2014), March 13-14, 2014 Technopark, Trivandrum, India*. Proceedings in Springer Series: Recent Trends in Computer Networks and Distributed Systems Security, Communications in Computer and Information Science Series (CCIS), ISSN: 1865:0929, pp 45–56.

[Pub8] KP Krishna Kumar, and G. Geethakumari (2013). "Information Diffusion Model for Spread of Misinformation in Online Social Networks." *Proceedings of the IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI-2013), August 22-25, 2013, Mysore, India.* pp 1172–1177.

[Pub9] KP Krishna Kumar, and G. Geethakumari 2013). "Analysing Spread of Misinformation in Online Social Networks using Cognitive Psychology." *International Conference on Behavioral, Cognitive and Psychological Sciences (BCPS 2013), November 18-19, 2013, London, UK.* Accepted paper.

[Pub10] KP Krishna Kumar, and G. Geethakumari 2013). "Modeling Semantic Attacks in Social Networks." *IEEE International Conference on Informatics, Electronics and Vision (ICIEV), May 17 - 18, 2013, Dhaka, Bangladesh.* Accepted paper.

Bibliography

[1] Martin C Libiki. *Cyberdeterrence and cyberwar*. Blackstone Audiobooks, 2010.

[2] Martin C Libiicki. *The Mesh and the Net: Speculations on Armed Conflict in a Time of Free Silicon*. University Press of the Pacific, 2004.

[3] Bruce Schneir. *Secrets and lies: digital security in a networked world*. Wiley. com, 2011.

[4] Bruce Schnier. The psychology of security. In *Progress in Cryptology–AFRICACRYPT 2008*, pages 50–79. Springer, 2008.

[5] Bruce Schneier. Semantic attacks: The third wave of network attacks. *Crypto-gram Newsletter*, 2000.

[6] Bruce Schneir. The psychology of security. In *Progress in Cryptology–AFRICACRYPT 2008*, pages 50–79. Springer, 2008.

[7] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:4, 2009.

[8] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.

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

[10] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.

[11] E. Morozov. Swine flu: Twitter's power to misinform. In *Foreign Policy Magazine Website Post*, 2009.

[12] Reuters. Netizens feel blaming social media for rumours incorrect. *Hindustan Times. http://www.hindustantimes.com/India-news/NewDelhi/Government-cracks-down-on-internet-after-northeast-exodus/Article1-917424.aspx. Last accessed 10 Jan 2015*.

[13] George Cybenko, Annarita Giani, and Paul Thompson. Cognitive hacking: A battle for the mind. *Computer*, 35(8):50–56, 2002.

[14] Elizabeth E Kirk. Evaluating information found on the internet. *http://guides.library.jhu.edu/evaluatinginformation*, 10:2006, 1996.

[15] Ivan Enrici, Mario Ancilli, and Antonio Lioy. A psychological approach to information technology security. In *Proceedings of the 3rd Conference on Human System Interactions (HSI)*, pages 459–466. IEEE, 2010.

[16] Brenda K Wiederhold. The role of psychology in enhancing cybersecurity. *Cyberpsychology, Behavior, and Social Networking*, 17(3):131–132, 2014.

[17] Hsien-Ming Chou and Lina Zhou. A game theory approach to deception strategy in computer mediated communication. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 7–11. IEEE, 2012.

[18] Lina Zhou, Yongmei Shi, and Dongsong Zhang. A statistical language modeling approach to online deception detection. *Transactions on Knowledge and Data Engineering*, 20(8):1077–1081, 2008.

[19] Frank J Stech, Kristin E Heckman, Phil Hilliard, and Janice R Ballo. Scientometrics of deception, counter-deception, and deception detection in cyber-space. *PsychNology Journal*, 9(2):79–122, 2011.

[20] Dhoha Almazro, Ghadeer Shahatah, Lamia Albdulkarim, Mona Kherees, Romy Martinez, and William Nzoukou. A survey paper on recommender systems. *arXiv preprint arXiv:1006.5278*, 2010.

[21] Natascha A Karlova and Karen E Fisher. Plz RT: A social diffusion model of misinformation and disinformation for understanding human information behaviour. *Information Research*, 18(1):1–17, 2013.

[22] Bernd Carsten Stahl. On the difference or equality of information, misinformation, and disinformation: A critical research perspective. *Informing Science: International Journal of an Emerging Transdiscipline*, 9:83–96, 2006.

[23] Don Fallis. A conceptual analysis of disinformation. *iConference, Chapel Hill, NC, California*, 2009.

[24] Martin C Libiki. Conquest in cyberspace. *National Security and Information Warfare*, 2007.

[25] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, 2012.

[26] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flamini, and Filippo Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. *arXiv preprint arXiv:1011.3768*, 2010.

[27] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web*, pages 249–252. ACM, 2011.

[28] Jayson Harsin et al. The rumour bomb: Theorising the convergence of new and old trends in mediated us politics. *Southern Review: Communication, Politics & Culture*, 39(1):84, 2006.

[29] Daniel Leonard Bernardi, Pauline Hope Cheong, Chris Lundry, and Scott W Ruston. *Narrative landmines: rumors, Islamist extremism, and the struggle for strategic influence*. Rutgers University Press, 2012.

[30] W De Neys, S Cromheeke, and M Osman. Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, 6(1):e15954, 2011.

[31] Markus Prior. *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press, 2007.

[32] Thomas J Johnson, Shannon L Bichard, and Weiwu Zhang. Communication communities or cyberghettos?: A path analysis model examining factors that explain selective exposure to blogs. *Journal of Computer-Mediated Communication*, 15(1):60–82, 2009.

[33] George Cybenko, Annarita Giani, and Paul Thompson. *Cognitive Hacking*, chapter 2. Elsevier Academic Press, New York, NY, 2003.

[34] Mary Ann Fitzgerald. Misinformation on the internet: Applying evaluation skills to online information. *Emergency Librarian*, 24(3):9–14, 1997.

[35] Anna Kata. A postmodern pandora's box: Anti-vaccination misinformation on the internet. *Vaccine*, 28(7):1709–1716, 2010.

[36] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2451–2460. ACM, 2012.

[37] David Westerman, Patric R Spence, and Brandon Van Der Heide. Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication*, 19(2):171–183, 2014.

[38] Washington Post. Market quavers after fake ap tweet says obama was hurt in white house explosion. *http://www.washingtonpost.com/business/economy/market-quavers-after-fake-ap-tweet-says-obama-was-hurt-in-white-house-explosions/2013/04/23/*$d96d2dc6 - ac4d - 11e2 - a8b9 - 2a63d75b5459_{s}tory.html$. *Last accessed 10 Jan 2015*.

[39] BBC. Boston bombing: How internet detectives got it very wrong. *BBC News. http://http://www.bbc.com/news/technology-22214511. Last accessed 10 Jan 2015*.

[40] Channel4. Unmasked: The man behind top Islamic State Twitter account. *Channel4. http://www.channel4.com/news/unmasked-the-man-behind-top-islamic-state-twitter-account-shami-witness-mehdi. Last accessed 10 Jan 2015*.

[41] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860. ACM, 2010.

[42] Kate Starbird and Leysia Palen. (How) will the revolution be retweeted?: Information diffusion and the 2011 Egyptian uprising. In *Proceedings of the International Conference on Computer Supported Cooperative Work*, pages 7–16. ACM, 2012.

[43] Brian Solis. The information divide: The socialisation of news. *http://www.briansolis.com/2010/02/the-information-divide-the-socialization-of-news-and-dissemination/*, 2010.

[44] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the 4th International Conference on Web Search and Data Mining*, pages 65–74. ACM, 2011.

[45] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600. ACM, 2010.

[46] Nam P Nguyen, Guanhua Yan, and My T Thai. Analysis of misinformation spread containment in online social networks. *Computer Networks*, 57(10):2133–2146, 2013.

[47] Nam P Nguyen, Guanhua Yan, My T Thai, and Stephan Eidenbenz. Containment of misinformation spread in online social networks. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 213–222. ACM, 2012.

[48] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International Conference on World Wide Web*, pages 665–674. ACM, 2011.

[49] Dung T Nguyen, Nam P Nguyen, and My T Thai. Sources of misinformation in online social networks: Who to suspect? In *Military Communications Conference, MILCOM 2012*, pages 1–6. IEEE, 2012.

[50] Reuters, IANS. Ethnic riots sweep Assam, at least 30 killed. *http://in.reuters.com/article/2012/07/24/india-assam-riots-floods-idINDEE86N04520120724. Last accessed 10 Jan 2015*.

[51] Bamshad Mobasher, Robin Burke, Runa Bhaumik, and JJ Sandvig. Attacks and remedies in collaborative recommendation. *Intelligent Systems*, 22(3):56–63, 2007.

[52] Bamshad Mobasher, Robin Burke, Runa Bhaumik, and Chad Williams. Effective attack models for shilling item-based collaborative filtering systems. In *Proceedings of the 2005 WebKDD Workshop*, 2005.

[53] Audun Jøsang and Jennifer Golbeck. Challenges for robust trust and reputation systems. In *Proceedings of the 5th International Workshop on Security and Trust Management (SMT 2009)*, 2009.

[54] Kevin Hoffman, David Zage, and Cristina Nita-Rotaru. A survey of attack and defense techniques for reputation systems. *ACM Computing Surveys (CSUR)*, 42(1):1, 2009.

[55] Richard Colbaugh and Kristin Glass. Early warning analysis for social diffusion events. *Security Informatics*, 1(1):1–26, 2012.

[56] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[57] Richard Durrett. *Lecture Notes on Particle Systems and Percolation*. Wadsworth & Brooks/Cole Advanced Books & Software, California, 1988.

[58] Thomas M Liggett. *Interacting Particle Systems*. Springer-Verlag, Berlin, 2005.

[59] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.

[60] Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, pages 1420–1443, 1978.

[61] Thomas C Schelling. *Micromotives and macrobehavior*. WW Norton, 2006.

[62] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

[63] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining*, pages 137–146. ACM SIGKDD, 2003.

[64] Rick Durrett. *Random graph dynamics*, volume 20. Cambridge University Press, New York, 2007.

[65] Elihu Katz and Paul Felix Lazarsfeld. *Personal influence: The part played by people in the flow of mass communications*. Transaction Publishers, 2006.

[66] Daniel Gruhl, Ramanathan Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, pages 491–501. ACM, 2004.

[67] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge University Press, 1994.

[68] Ramasuri Narayanam and Yadati Narahari. A shapley value-based approach to discover influential nodes in social networks. *IEEE Transactions on Automation Science and Engineering*, 8(1):130–147, 2011.

[69] Eyal Even-Dar and Asaf Shapira. A note on maximizing the spread of influence in social networks. *Internet and Network Economics*, pages 281–286, 2007.

[70] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining*, pages 420–429. ACM SIGKDD, 2007.

[71] Eytan Adar and Lada A Adamic. Tracking information epidemics in blogspace. In *Proceedings of the International Conference on Web Intelligence*, pages 207–214. IEEE, 2005.

[72] Akshay Java, Pranam Kolari, Tim Finin, and Tim Oates. Modeling the spread of influence on the blogosphere. In *Proceedings of the 15th International Conference on World Wide Web*. ACM, 2006.

[73] Masahiro Kimura, Kazumi Saito, Ryohei Nakano, and Hiroshi Motoda. Extracting influential nodes on a social network for information diffusion. *Data Mining and Knowledge Discovery*, 20(1):70–97, 2010.

[74] Masahiro Kimura and Kazumi Saito. Tractable models for information diffusion in social networks. *Knowledge Discovery in Databases*, LNAI 4213:259–271, 2006.

[75] Masahiro Kimura, Kazumi Saito, and Ryohei Nakano. Extracting influential nodes for information diffusion on a social network. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, pages 1371–1376. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press, 2007.

[76] Masahiro Kimura, Kazumi Saito, Hiroshi Motoda, and Ryohei Nakano. Finding influential nodes in a social network from information diffusion data. In *Proceedings of the International Workshop on Social Computing and Behavioral Modeling*, pages 138–145. Springer, 2009.

[77] Shishir Bharathi, David Kempe, and Mahyar Salek. Competitive influence maximization in social networks. *WINE 2007, LNCS*, 4858:306–311, 2007.

[78] Everett M Rogers. *Diffusion of innovations*. Free Press, New York, USA, 2010.

[79] Ceren Budak, Divyakant Agrawal, and Amr El Abadi. Diffusion of information in social networks: Is it all local? In *Proceedings of the 12th International Conference on Data Mining (ICDM)*, pages 121–130. IEEE, 2012.

[80] Frank M Bass. Comments on a new product growth for model consumer durables the bass model. *Management Science*, 50(12_supplement):1833–1840, 2004.

[81] Christopher Griffin and Kathleen Moore. A framework for modeling decision making and deception with semantic information. In *Proceedings of the International Symposium on Security and Privacy Workshops (SPW)*, pages 68–74. IEEE, 2012.

[82] Paulo Shakarian, Patrick Roos, and Anthony Johnson. A review of evolutionary graph theory with applications to game theory. *Biosystems*, 107(2):66–80, 2012.

[83] Chunxiao Jiang, Yan Chen, and KJ Liu. Evolutionary dynamics of information diffusion over social networks. *IEEE Transactions on Signal Processing*, 62:4573–4586, 2014.

[84] Adrien Guille, Hakim Hacid, Cécile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(1):17–28, 2013.

[85] Seth A Myers and Jure Leskovec. Stanford univ., stanford, ca, usa. In *Proceedings of IEEE 12th International Conference on Data Mining (ICDM)*, pages 539–548. IEEE, 2012.

[86] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.

[87] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 675–684. ACM, 2011.

[88] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, pages 71–79. ACM, 2010.

[89] Aditi Gupta and Ponnurangam Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, pages 2–. ACM, 2012.

[90] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[91] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Readings in information retrieval*, 24(5):355–363, 1997.

[92] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics, 2011.

[93] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM)*, volume 14, pages 10–17. AAAI, 2010.

[94] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 441–450. ACM, 2012.

[95] Eunsoo Seo, Prasant Mohapatra, and Tarek Abdelzaher. Identifying rumors and their sources in social networks. In *SPIE Defense, Security, and Sensing*, pages 83891I–83891I. International Society for Optics and Photonics, 2012.

[96] Rudra M Tripathy, Amitabha Bagchi, and Sameep Mehta. A study of rumor control strategies on social networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1817–1820. ACM, 2010.

[97] Rudra M Tripathy, Amitabha Bagchi, and Sameep Mehta. Towards combating rumors in social networks: Models and metrics. *Intelligent Data Analysis*, 17(1):149–175, 2013.

[98] Eni Mustafaraj and P Takis Metaxas. From obscurity to prominence in minutes: Political speech and real-time search. In *WebSci10: Extending the Frontiers of Society On-Line*, page 317. The Web Science Trust, 2010.

[99] P Hedstrom. Explaining the growth patterns of social movements. *Understanding Choice, Explaining Behaviour. Oslo Academic Press, Oslo*, 2006.

[100] Martin Hawksey. Twitter Archiving Google Spreadsheet TAGS v5. JISC CETIS MASHe: The Musing of Martin Hawksey (EdTech Explorer). http://mashe.hawksey.info/2013/02/twitter-archive-tagsv5/, 2013. [Online; last accessed 10 Jan 2015].

[101] M De Domenico, A Lima, P Mougel, and M Musolesi. The anatomy of a scientific rumor. *Scientific reports*, 3, 2013.

[102] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: an open source software for exploring and manipulating networks. 2009. In *International AAAI Conference on Weblogs and Social Media*, pages 361–362, 2011.

[103] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[104] Shyong K Lam and John Riedl. Shilling recommender systems for fun and profit. In *Proceedings of the 13th international conference on World Wide Web*, pages 393–402. ACM, 2004.

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

[106] Robin Burke, Bamshad Mobasher, Chad Williams, and Runa Bhaumik. Classification features for attack detection in collaborative recommender systems. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*, pages 542–547. ACM SIGKDD, 2006.

[107] Michael P OMahony, Neil J Hurley, and Guenole CM Silvestre. Attacking recommender systems: The cost of promotion. In *Proceedings of the Workshop on Recommender Systems, in conjunction with the 17th European Conference on Artificial Intelligence*, pages 24–28. Citeseer, 2006.

[108] Fuguo Zhang. Reverse bandwagon profile inject attack against recommender systems. In *Proceedings of the 2nd International Symposium on Computational Intelligence and Design (ISCID)*, volume 1, pages 15–18. IEEE, 2009.

[109] Fugao Zhang. Analysis of bandwagon and average hybrid attack model against trust-based recommender systems. In *Proceedings of the 5th International Conference on Management of e-Commerce and e-Government*, pages 269–273. IEEE, 2011.

[110] Zunping Cheng and Neil Hurley. Robust collaborative recommendation by least trimmed squares matrix factorization. In *Proceedings of the 22nd International Conference on Tools with Artificial Intelligence (ICTAI)*, volume 2, pages 105–112. IEEE, 2010.

[111] Robin Burke, Bamshad Mobasher, Runa Bhaumik, and Chad Williams. Segment-based injection attacks against collaborative filtering recommender systems. In *Proceedings of the 5th International Conference on Data Mining*, pages 4–pp. IEEE, 2005.

[112] J. McKenzie Alexander. Evolutionary game theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 2009.

[113] Josef Hofbauer and Karl Sigmund. Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, 40(4):479–519, 2003.

[114] Erez Lieberman, Christoph Hauert, and Martin A Nowak. Evolutionary dynamics on graphs. *Nature*, 433(7023):312–316, 2005.

[115] Martin A Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563, 2006.

[116] Hisashi Ohtsuki, Jorge M Pacheco, and Martin A Nowak. Evolutionary graph theory: breaking the symmetry between interaction and replacement. *Journal of Theoretical Biology*, 246(4):681, 2007.

[117] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, page 8. ACM, 2013.

[118] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[119] Tore Opsahl and Pietro Panzarasa. Clustering in weighted networks. *Social networks*, 31(2):155–163, 2009.

[120] Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.

[121] Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. *Wilensky, U. NetLogo. http://ccl.northwestern.edu/netlogo/*, 1999. Release 5.0.3.

[122] Pu-yan Nie. Evolutionary graphs on two levels. *Ars Combinatoria*, 86:115–120, 2008.

[123] P-A Zhang, P-Y Nie, and D-Q Hu. Bi-level evolutionary graphs with multi-fitness. *IET systems biology*, 4(1):33–38, 2010.

[124] Frank B Baker and Seock-Ho Kim. *Item Response Theory: Parameter estimation techniques*. CRC Press, 2004.

[125] Shirley J Behrens. A conceptual analysis and historical overview of information literacy. *College and research libraries*, 55(4):309–22, 1994.

[126] Sibel Adali, Robert Escriva, Mark K Goldberg, Mykola Hayvanovych, Malik Magdon-Ismail, Boleslaw K Szymanski, William A Wallace, and Gregory Williams. Measuring behavioral trust in social networks. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 150–152. IEEE, 2010.

[127] Ivailo Partchev. A visual guide to item response theory. *Friedrich Schiller Universität Jena*, 2004.

[128] Jennifer Golbeck, Bijan Parsia, and James Hendler. *Trust networks on the semantic web*. Springer, 2003.

[129] Fumiko Samejima. Graded response model. In *Handbook of modern item response theory*, pages 85–100. Springer, 1997.

[130] Eiji Muraki. A generalized partial credit model: Application of an em algorithm. *Applied psychological measurement*, 16(2):159–176, 1992.

[131] Aurélie Brunie. Meaningful distinctions within a concept: Relational, collective, and generalized social capital. *Social Science Research*, 38(2):251–265, 2009.

[132] Sepandar D Kamvar, Mario T Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*, pages 640–651. ACM, 2003.

[133] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. Wiley. com, 2009.

[134] Election commision of India. General election 2014. *http://eci.nic.in/eci_main1/GE2014/ge.html. Last accessed 11 Mar 2015.*

Glossary of terms used in Online Social Networks

Avatar. An avatar is an image or username that represents a person online within forums and social networks.

Collaboration: Social media tools offer enormous scope for collaboration. Collaboration is being able to discuss and work with people across boundaries of organisation, time and space. Low-risk activities like commenting, social bookmarking, chatting and blogging help develop the trust necessary for collaboration.

Collective intelligence has been defined by George Pór as the capacity of a human community to evolve toward higher order complexity thought, problem-solving and integration through collaboration and innovation. For a network to develop this "mind of its own" there needs to be a willingness among members to share and collaborate. It is a shared or group intelligence that emerges from the collaboration and competition of many individuals and appears in consensus decision-making in social networks.

Content. Any kind of meaningful information to include text, photos, videos, audios, etc on the Internet.

Crowdsource. Use of collective brain power, skills and opinion of large number of people to solve problems or build solutions.

Folksonomy. Taxonomies are centralised ways of classifying information - as in libraries. Folksonomies are the way folk create less structured ways of classifying by adding tags.

Like. The Like option allows to acknowledge content on social media in a positive way without needing to add actual comments. Facebook uses this feature.

Metadata. Refers to information including titles, descriptions, tags and captions that describes a media item such as a video, photo or blog post.

Microblog. Short message postings from a social network account. Twitter posts, Facebook statuses are examples of microblogs.

MySpace. An online social network. MySpace caters to artists and bands, who enjoy the flexibility of creating an individual "look" for their page. MySpace allows users to "friend"

each other and create groups.

News Feed. A list of updates on a user's Facebook home page. The updates include status updates by friends as well as by official pages.

Online Social Networks. An online community of people who are socializing with each other via a particular web site. It helps to connect socially or professionally with other people. They provide users an online community to share and explore common interests and activities.

Profiles. Profiles are the information that one provide when signing up for a social networking site. This may include a picture and basic information like name, age, gender etc. apart from personal and business interests.

Share. The act of sharing a piece of content with specific friends so that those friends particularly interested will read it.

Social Capital. Social capital is a concept used in business, non profits and other arenas that refers to the good will and positive reputation that flows to a person through his or her relationships with others in social networks.

Social Media. Social media refers to works of user-created video, audio, text or multimedia that are published and shared in a social environment, such as a blog, podcast, forum, wiki or video hosting site. More broadly, social media refers to any online technology that lets people publish, converse and share content online.

Tags. Keywords that describe the content of a web site, bookmark, photo or blog post. Multiple tags can be assigned to the same online resource.

Taxonomy. Taxonomy is an organised way of classifying content, as in a library.

Threads. Threads are strands of conversation. On an email list or web forum they will be defined by messages that use the same subject.

User Generated Content (UGC). It refers to all forms of user-created materials such as blog posts, reviews, podcasts, videos, comments, etc.

Viral. Contents shared in OSNs that get passed along rapidly. Tweets of celebrities, You Tube videos are examples.

Web 2.0. Second generation of the Web. Web 2.0 is a term coined by O'Reilly Media in 2004 to describe blogs, wikis, social networking sites and other Internet-based services that emphasize collaboration and sharing, People can blog, create web sites without requiring specialized technical knowledge and training. Users become content creators.

Whiteboards. Whiteboards are online tools that enable anyone to write or sketch on a web page. They are useful in online collaboration.

Twitter

Direct message (DM)]. An instant, direct and private message from one Twitter user to another that appears in a users "messages" box. A tweet, on the other hand appears on users timelines and is usually public.

Follow. Subscribing to the updates of other users.

Follower. A subscriber of another users Twitter feed.

Following. Users whose twitter feeds you subscribe to.

Handle. Unique name applied to each Twitter user. Handles are typically prepended with the @ symbol.

Hashtag. A word or a string of characters that start with '#'. A mechanism in Twitter used to group posts under the same topic, by including a specific word preceded by the # symbol (a word, or tag, denoted with a hash: hashtag). Messages containing identical hashtags are grouped together into a search thread.

Lists. A grouping mechanism where users can group other users into manageable sets.

Live-tweeting. The same as live blogging, but using tweets to tell the story in real-time instead of blog posts.

Modified tweet (MT). Same as a retweet, but with text thats been slightly changed, hence the word modified.

Protected tweets. When a Twitter user restricts viewing of their tweets to approved followers only.

Retweet (RT). When Twitter users re-post a post made by another user.

Tweet. A post on micro-blogging site Twitter.

Twitter API. The protocol that allows software and third party clients to interact with and collect data from Twitter. The open nature of Twitters original API gave rise to a large number of third party Twitter clients, which allowed users to bypass the Twitter website. Twitter has since restricted its API, making it more difficult for these clients to operate. Version 1.1 of Twitters API was released in 2012.

Timeline. A news feed of updates posted or retweeted by those a user follows.

Biography: K P Krishna Kumar

K P Krishna Kumar completed his ME(Computer Science and Engg) from Indian Institute of Science in 2005. His research interests include cyber security, information diffusion, semantic security, collaborative filtering systems, online social networks and game theory.

K P Krishna Kumar has been a faculty member and Head of Computer Systems Department in Military College of Electronics and Mechanical Engineering (MCEME), Secunderabad. He was project officer in Army Software Development Centre. He has been actively involved in the design and implementation of numerous IT projects in the Indian Army.

Biography: Dr. G. Geethakumari

Dr G Geethakumari is Asst.Professor, Dept. of Computer Science and Information Systems at BITS Pilani, Hyderabad Campus. Before joining BITS, she worked as a faculty in the CSE Dept. at the National Institute of Technology, Warangal. Dr Geetha received her Ph.D. from University of Hyderabad. Her Ph.D. thesis was titled 'Grid Computing Security through Access Control Modelling'.

Dr. Geetha has many international publications to her credit. Her areas of research interests include: Information security, cloud computing and security, cloud forensics, enterprise security challenges and data analysis, cloud authentication techniques, cyber security, semantic attacks and privacy in online social networks. She has been in the forefront of technical activities at BITS-Pilani, Hyderabad Campus. She has been the Faculty Advisor for Computer Science Association during 2008-2011. Presently she is the IEEE Student Branch Counselor, BITS-Pilani, Hyderabad Campus. She is also the Coordinator for the Linux User Group, BITS Pilani, Hyderabad Campus.

Dr. Geetha is a Member, IEEE as well as Member, IEEE Computer Society. She is also a Professional Member, ACM. She was the Organizing Committee Member for the IEEE INDICON Conference conducted in BITS Pilani, Hyderabad Campus during December 16-18, 2011. Dr Geetha was the Publicity Co-Chair for the IEEE Prime Asia Conference hosted by BITS Pilani, Hyderabad Campus during December 5-7, 2012.

Dr Geetha was the Publicity Co-Chair for the IEEE Prime Asia Conference hosted by BITS Pilani, Hyderabad Campus during December 5-7, 2012. She was the Organizing Committee Member for the Workshop on Advances in Image Processing and Applications held in BITS Pilani, Hyderabad Campus during October 26 - 27, 2013. She was part of the Organizing Committee for the National Seminar on Indian Space Technology - Present and Future (NSIST-2014) held at BITS Pilani Hyderabad Campus on 1st May, 2014.

Dr Geetha has given many guest lectures on topics in emerging areas such as cyber security, cloud computing and cloud security. She has been a member of the Technical Program Committees of various IEEE International Conferences. An extract from the paper 'A taxonomy for modelling and analysis of diffusion of (mis)information in social networks', co-authored by Dr Geetha and published in the International Journal of Communication Networks and Distributed Systems, Vol. 13, No. 2, 2014, pp.119-143, by Inderscience Publishers, was selected for a press release on 'Semantic attacks in online social media'.