CHAPTER 2

RELATED WORK AND LITERATURE REVIEW

2.1 Introduction

Social Network Analysis (SNA) is a research area that tries to analyse and model, actor behaviour based on his or her connections or relations to other members of a group [1]. SNA is an interdisciplinary area including such fields as social psychology, sociology, statistics, and graph theory. The methodology followed, originated from graph theory and extended its application to Social Science domains.

2.2 Introduction to Graph Theory and Complex Networks

A Social structure comprising of a set of participants (individuals, organizations) and the mutual ties between these participants is a social network. Understanding and analysing the structure of the social network is necessary for identifying local and global characteristics, influential participants and the dynamics of networks [1]. An introduction to graphs and networks from a more theoretical viewpoint is found in [2]. This gives

This gives description and reference for common graph metrics and topologies. Definitions are given for directed and undirected graphs, unconnected graphs and connected components, complete and star graphs and lattices. The metrics considered are, clustering coefficient, average path length, centrality, and degree distribution function. A graph taxonomy is categorized as planar/non planar, cyclic, acyclic, transitive, Eulerian, Hamiltonian, bipartite, polyhedral, n-connected, cubic, complete, complete bipartite, isospectral, endspectral, cages, hypercubes, saturated and maximally saturated. Newman [3] describes the structure and function of complex networks. The definitions of metrics and topologies are explained by Mislov et al. [4]. Some distinct definitions are made by Newman [5]. A hyperedge is defined as an edge which joins more than two vertices together, and a hypergraph is defined as a graph which contains one or more hyperedges. A bipartite graph is defined as a graph which contains vertices of two distinct types, with edges running only between unlike types. A component is defined as a set of vertices that can be reached from a given node by paths running along edges of the graph.

2.3 Properties of Complex Networks

The networks with connections neither regular nor random belong to the category of Complex Networks [3]. All networks based on interpersonal communication, form clusters. The first known study by a sociologist on social networks is by Milgram et al. [6] [7]. Investigating the "*small-world Problem*". Milgram [6] studied the relational information based on an

experiment of forwarding of letters to the acquaintances. They considered two nodes A and B of the social network as connected if either A sent or received a letter from B. The concept of a small-world network structure allowed a connection between two random individuals in the network via a few individuals, established a theory that we live in a small-world with "six degrees of separation" [6]. Study on complex networks contributed additional knowledge on the structural properties and dynamics of the networks. This theory was reinforced by another notable study by Watts & Strogratz [8] investigating the structure of the social world of Hollywood. They defined two actors as connected if they acted in the same film. Analysing the information present in the Internet Movie Database, they concluded that the 225.000 actors were separated from each other by only four steps.

In online networks, equal attention is given for every connection whereas this need not be true in the case of offline interactions. Granovetter [9] describes the union of factors like time of interaction, intimacy, emotional attachment and reciprocated interaction as the strength of a tie and the distinction between strong and weak ties.

Network theory made significant progress during the last decade, as a result of the publications of Watts & Strogatz [8] and Barabasi & Reka [10]. The pure mathematical Network theory became a part of everyday life of a commoner fitting theoretical expectation to empirical findings. The analysis of real-world networks with the scale-free paradigm is very powerful in

explanation and prediction. This captures the important part of natural mechanisms that lead to higher efficiency and robustness. Mendes [11] gives many perspectives of real-world applications.

The "small-world phenomenon" explained above, is the characteristic that differentiates Online Social Network (OSN) graphs from the general graphs. The phenomenon represents the observation that only a small number of connections are necessary to link two nodes. In human relation terms, it means that two people, highly differentiated socio-economically and geographically, are often only a small number of links away from each other in an OSN graph (on an average, six). This is studied by Kleinberg [12] in which an algorithmic perspective is presented, in order to analyse and explain why it occurs. Some of the interesting observations are

a) "Why doesn't this overload/saturate the network?" The answer is that although a node is potentially just six steps away from any other node, the probability that a given node sends a message tries to contact a node at distance 6, and is very low.

b) "Why the number 6?" Is another phenomenon which the author proposes that it has to do with the inverse square law?

In order to show this, a formula is derived in terms of powers of two which includes the number six (the maximum number of steps from one node to another), as an upper bound for the inverse-square distribution, thus:

[4 log (6n) d (u, v) 2] −1 (2) where log is the natural (e) log, n is the

number of individuals, u is a given node, v is another (target) node in the graph, and d (u, v) is the distance between the two.

In contrast to the particular topology of OSNs, the topology of the WWW is also studied. The topology of WWW has a distinctive structure, defined by Broder et al. in [13] as looking like a 'bowtie', made up of a central strongly connected component (SCC), one side of the bow being an 'IN' component, the other side the 'OUT' component, and with 'Tendril' components attached to the IN/OUT components. The Web has a very large SCC, hence very resilient to node deletions.

2.4 Social Network Sites and Social Media

A social network provides a variety of mechanisms for users to share data with other users. Boyd & Ellison [14] defines Social networking sites as a web-based service that allows individuals to construct a public or semi-public profile within the system, manage a list of other users with whom they share a connection, and is able to view and traverse their list of connections. The rise and growth of social networking sites like Orkut, Facebook has been a milestone in World Wide Web history.

The emergence of many social networking sites has increased the participation of people's activities producing large data set of information like interactions, reviews, comments, posts etc. This huge volume of information available has attracted commercial establishments to advertise and promote their products leading the researchers to have access to this

information and analyse.

Many studies are conducted on how information gained from different websites can be enhanced to become knowledge using methods and models from the research field of Social Network Analysis. This is context sensitive. Six Degrees [15] was one among the first few to incorporate contemporary OSN functionality to manage user profiles and friend lists. Other notable applications were Friendster [16] in 2002, followed by MySpace [17] and LinkedIn [18] in 2003. Facebook [19] launched in 2004 was the last to join this revolution, and by 2009 it became the largest social networking site. We could say that the great period of growth in the use of OSN applications manifested itself throughout the years 2005–2010. Open to the public in 2006, Facebook reported serving one billion monthly active users at the end of 2012.

Social media is defined by Kempe [21] as a social structure for a group of individuals that connects both directly or indirectly, based on each interest or importance and allow the creation and exchange of user-generated content [22, 23]. Kaplan & Haenlein [24] defines social media as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 that allow the creation and exchange of User Generated Content. The popularity of social media has grown extensively over the past few years.

2.5 Diffusion of Information

Networks represent a fundamental medium for the emergence and diffusion of information [26]. For example, we often think of information, a rumour, or a piece of content as being passed over the edges of the underlying social network [15, 27]. The way information spreads over the edges of the network like an epidemic [15]. Work on the diffusion of innovations [26] provided a conceptual framework to study the emergence of information in networks. Conceptually, in a network (often implicit) each node is considered either active (infected, influenced) or inactive. Active nodes spread the contagion (information, disease) along the edges of the underlying network. A rich set of models were developed to describe different mechanisms by which the contagion spreads from the infected to an uninfected node [27, 28, 29, and 30]. In the earlier studies, the models focus on the diffusive part of the contagion adoption process, neglecting the external influence. But the later studies found out that the activity and the influence of the external source are important and not to be neglected. External influence in networks has been considered in the case of the popularity of YouTube videos [12]. A simple model of information diffusion on an implicit completely connected network is considered in this case. Some videos became popular, quicker than their model predicted; the argument is that popularity is the result of external influence. This model is built on the notion of exposure curves, which was proposed and studied by Romero et al. [30]. There is also an argument [31] that it is the shape of

exposure curves, that stop the information from spreading. It is proven that external influence models are more accurate than the exposure curve method proposed [30, 32].

Given today‘s rapid dissemination of social media platforms such as Twitter and Facebook, scholars from different fields of studies have investigated those sites, and the challenges they pose for society in general, for interpersonal relations and psychological well-being, for political participation and civic engagement, and for media organizations and online journalism [14, 33, 34]. The flow of information is affected by social structure and knowledge of communities in the network. This information is useful for searching for individuals, resources as well as navigating the networks. Using the HP email network, Adamic & Adar [35] has simulated Milgram‘s small-world experiment to study this concept.

Social network analysis assumes that individuals are interdependent and that the communication between individuals, defines this interdependence [1]. The data of an organization were collected through surveys, where individuals indicate the others with whom they communicate. These data define the structure of communication or the relationships between the actors (individuals) within the organization.

Over the past forty years, traditional methods of studying social processes such as information diffusion, expert identification or community detection were focused on longitudinal studies of relatively small groups. With the widespread proliferation of social media like Facebook, Twitter, Digg,

Flickr, and YouTube provided many avenues for researchers to study such processes at very large scales. This is because data could be acquired and stored over extended time intervals and for very large populations. The result is that study of social processes on a scale of million nodes, which was inconceivable a decade ago, has become a routine.

The pervasive use of social media, made the cost involved in propagating a piece of information to a large audience, extremely negligible, providing extensive evidence of large-scale social contagion. This helped researchers to conduct many empirical studies on diffusion in a different context. Cascading behaviour of information propagation on blog spaces is studied by Gruhl et al. [36]. To understand the spreading patterns in the Internet, chain letter networks were studied by Liben-Nowell & Kleiberg [37]. They found that a chain letter to someone six steps away could even pass through 100 intermediaries before reaching the destination. Social tagging experiments by Anagnostopoulos et al. [38] revealed that social co-relation and social influence are two different concepts. The empirical investigation of information diffusion through Facebook news feed by Sun et al. [39] identified factors affecting large chain diffusion. The social influence of user content in a virtual world with the data from Linden Lab is investigated by Bakshy et al. [40]. They report the influential role of weak ties in the process

2.6 Metrics

There are many metrics for gauging, transforming and representing graphs, the standard format being the averaged statistics derived from the degree, clustering coefficient and average path length of the nodes and edges.

Watts & Strogatz [8] defined the clustering coefficient for node v as the ratio of the number I of existing nodes to the possible number of edges between the direct neighbours of node v [1]:

$$C(0) = 3(K-2)/4(K-1)$$

C(1)= K/N, Given the desired number of nodes N and the mean degree K

As for the degree distribution, the clustering coefficient plays an important role when analysing networks in terms of important properties like diffusion, which is discussed in the subsequent paragraphs [41]. The path length between two nodes of a network is defined as the number of edges between them. The minimal path length is the shortest path between two nodes. The average path length is the average of all the minimal path lengths between all pairs of nodes in a network.

Newman [5] has identified that social networks differ from most other types of networks, including technological and biological networks, in two important ways. First, they have non-trivial clustering or network transitivity, and second, they show positive correlations, also called assortative mixing, between the degrees of adjacent vertices. Social

networks are often divided into groups or communities, and it has recently been suggested, that this division could account for the observed clustering. The group structure in networks accounts for degree correlations. A simple model is used to confirm, that we should expect assortative mixing in such networks, whenever there is variation in the sizes of the groups, and that the predicted level of assortative mixing, compares well with the observations in real-world networks. Social networks exhibit a positive correlation between the degrees of adjacent vertices (assortativity), whereas most non-social networks are disassortative.

In Chakrabarti & Faloutsos [42], the review of graph metrics is conducted, defining typical graph metrics such as: number of nodes and edges in the graph; degree of each node; average degree for all nodes in graph; cc, clustering coefficient for the whole graph; cc (k), clustering coefficient for all nodes of degree k; power law exponent; and time/iterations since start of processing. The graph is assumed to be being generated and processed by some algorithm. Some of the main data mining problems considered in [41] are as follows:

- Detection of abnormal sub graphs, edges, and nodes.
- Simulation studies on synthetic graphs generated to be as close as possible to the real equivalent;
- Sampling on large graphs— the smaller graph has to match the patterns of the large graph or it will not be realistic;

- Graph compression—data can be compressed using graph patterns which represent regularities in the data.

The typical graph characteristics which occur in naturally occurring graphs are

a) *power laws **(**of degree distributions, and other values**)**,*

b) *small diameters (OSNs ≈ 6)*

c) *Community effects (high clustering coefficients**).***

Power laws can be traditional or can be skewed, for example, as a consequence of the presence of a significant sub-community within the global community. In order to show this, the authors plotted the power law distributions of the 'in-degree' and 'out-degree' for the 'Epinions' and 'click stream' datasets. In the case of click stream, the plot showed a skewed effect, given that this data was known a priori to contain a significant sub-community. The authors commented that two of the most common deviations are exponential cut-offs and log normal.

As different interpretations of the 'centrality' of a node, are considered:

a) A 'centrality metric', in which a high degree for a node implies it is more central;
b) Degree of indirect neighbours;
c) 'Closeness centrality', defined as the inverse of the average path length of a node to other nodes;

d) 'Betweenness centrality', defined as the number of shortest paths which pass through a node;
e) 'Flow centrality', defined as the number of all paths which pass through a node.

Another reference in this section on metrics, Mislove et al., in [43], defines a series of properties for OSNs, such as the power-law distribution, the small-world phenomenon, and the scale-free characteristics. Mislove et al. [43] statistically analysed four OSNs (Flickr, YouTube, LiveJournal and Orkut) in terms of these properties. They observed that in OSNs, the 'in degree' of nodes tends to match the 'out degree'; that OSN networks contain densely connected cores of high degree nodes, and that this core links small groups of strongly clustered, low degree nodes at the fringes of the network.

The Power Law degree distribution is given P (k) directly proportional to c * k^(-y), The Power law defines that $k^{-\gamma}$, for large k and $\gamma > 1$.

Scale-free networks are defined as a class of power-law networks in which the high-degree nodes tend to be connected to other high-degree nodes. The properties of small-world networks are small diameter and high clustering.

The most interesting part is Misloves[43] analysis on WCC (Weakly Connected Component). The correlation of 'indegree' and 'outdegree' are considered along with Joint degree distribution (JDD), the frequency with which nodes of different degrees are connected are taken as measures. The

scale-free behaviour where the graph has hub-like core to which high degree nodes connected to other high degree nodes are studied. The degree correlation compares a node‘s degree against those of its neighbours and tells whether a hub is likely to connect other hubs rather than low-degree nodes in an undirected network. The positive trend in degree correlation is called assortativity and is known as one of the characteristic features of human social networks [21]. This is feasible only in undirected graphs and does not apply to Twitter. When the mean of average numbers of followers of friends, is plotted against the number of followers, a positive correlation is seen slightly below x = 1, 000, and dispersion beyond that point.

H.Kwak et al. [44] in his paper has explained crawling by studying the entire Twitter sphere, obtaining 41.7 million user profiles, 1.47 billion social relations, 4,262 trending topics, and 106 million tweets. In its follower-following topology analysis, they have found a non-power-law follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from known characteristics of human social networks [21].

2.7 Growth of Online Social Networks

Study of evolution (growth) of social networks is an interesting research problem and can be empirically measured. A diversity of approaches to the analysis and modelling of evolution in OSN graphs can be found in the literature. There are authors who statistically analysed evolution in real

OSN datasets (typically over time), like Viswanath et al. in [45], Kossinets & Watts in [46] and also authors who studied specific aspects and modelled the evolution process. With reference to the latter, Tang, et al., in [47] tried modelling a 'multi-mode' network, containing different types of user and actions by those users. Leskovec et al. in [48] defined a graph generator called the "forest fire" model which reflected the way link creation propagates through the network. Finally, [49] focuses on the disconnected components of the graph and the incorporation of weights, and the authors propose an improved version of the forest fire model. The theme of community evolution, in Dynamic Multi-Mode networks, is studied by Tang et al. [47]. The authors found that an understanding of the structural properties of a network will help in balancing problems and identifying key influential factors. A crucial aspect of modelling evolution is, of course, the temporal dimension. They have also given a theoretical presentation and pseudo code for modelling a multi-mode network, a network which has different types of user and actions by those users. Models are progressively refined using data with ordered time stamps, and weighted attribute values and were found working well with the medium level of noise. To tune the model's parameters, online clustering and evolutionary multi-mode clustering were applied. Different methods were used to evaluate the results, although it is stated that the true community clusters could not be exactly known apriori. One drawback of the method is the requirement of an apriori definition of the number of communities and the weights for temporal and interaction information. The structure and evolution of online social

networks can also be analysed, from data logs of applications like Yahoo360 and Flickr. In the study carried out by Kumar et al. in [49], the authors discovered three regions:

(a) Singletons which do not participate in the network

(b) Isolated communities which display a dominant star structure

(c) A giant component anchored by a well-connected core region.

The authors present a simple model which captures these three structural aspects. Their model parameters are

i. User type distribution (passive, inviter, and linker)
ii. Preference for giant component over the middle region
iii. Edges per time step

A specific data log of the Facebook application, corresponding to the New Orleans geographical region, was collected and analysed by Viswanath et al. in [45]. Their study on the evolution of user interaction found that the structural network (links between accepted friends) is not a very true picture of the real friends of an individual. This is because many of the users (in Facebook) are not very discriminative when they aggregate persons as "friends". Thus, Viswanath et al. [45] proposed that the measure of "activity" will give a much better picture of who communicates with whom, where the intensity of "activity" is proportional to the strength of the relation. However, the activity measure used is that of "writes to the wall",

and many users of Facebook also use others communications channels, such as the chat box, sending an email, and so on. Also, the dataset used is skewed with respect to the general Facebook community because the users were selected by geographical region of New Orleans, USA. But this study derived some useful conclusions and implications. The authors collected their data, from a university faculty environment is described by Kossinets et al. in [46] for an empirical analysis of the evolution of a social network. They constructed their dataset from emails and other data about students, faculty and staff of a large university of one year period. The authors have used three types of data:

i. Registry of e-mail interactions—the timestamp, sender and list of recipients of each mail, but not the content;

ii. Personal attributes information like status, gender, age, department affiliation, and number of years in the community;

iii. Complete lists of classes attended or taught, for each semester.

It was found that the network was influenced by the organizational structure of the environment and by the network topology. The authors found that the general network characteristics tend to reach equilibrium whereas the individuals are much more volatile.

A multivariate survival analysis was conducted using the following attributes: ‘strong indirect‘, ‘classes‘, ‘acquaintances‘, ‘same age‘, and ‘same year‘. The effect of gender was studied by comparing pairings of

male–male with female–male, and female–female with female–male. It was found that the ‘average vertex degree‘, the ‘fractional size of the largest component’ and the ‘mean shortest path length’ exhibited seasonal changes. On the other hand, the distribution of ‘tie strength’ was found to be stable in the network as a whole over time. The users who were part of bridges also had a tendency to be transient. Although the bridges may act to diffuse information across whole communities, Kossinets et al. [46] found them to be unstable and not permanently represented by particular individuals. It was found that users did not ‘strategically manipulate’ their networks, even though it was technically possible because there was no motivation. These results were interesting but within the specific context of the study data and the environment, which cannot be generalised to other OSN domains.

Different kinds of networks like social, informational, technological and biological are considered. Network resilience, the ability to retain its basic functionality when failures and or any changes occur is an important property in the complex network. Barabasi & Reka [10] has proved that a scale-free network is resilient to random failures but the targeted attack can fragment to different disconnected sub-networks. Community structure is also considered and the dendrogram (hierarchical clustering) is described as a way of identifying communities.

2.8 Information Diffusion Models

Information diffusion models were studied and discussed through many research publications [27], with applications like spreading of epidemics and spreading rumours in the conventional human networks. It was noticed that the study of the same through online social media was rarely attempted in the earlier period. The specific domain of epidemiological processes is discussed in the context of the spread of viruses. The SIR and SIS models are mentioned. With reference to network search, Newman [3] proposes using Web keywords or making use of the skewed degree distribution to find results more quickly. The Phase transition is also considered on networks modelled by statistical mechanical models. It is concluded that the network structure can change the physical properties of the system, thus affecting the visible, the critical phenomena and the phase transition. It was commented that in the limit $n- > \infty$, a model has a finite-term Newman & Park in [5], consider non-trivial clustering (network transitivity) and positive correlations (assortative mixing) between degrees of adjacent vertices. They comment that social networks are often divided into groups or communities. The 'small-world' effect is mentioned, together with the skewed degree distributions, and positive degree correlations between adjacent vertices (in most other networks they have negative correlations). They also mention 'network transitivity', that is, the propensity for vertex pairs to be connected if they share a mutual neighbour. Clustering within networks is another important factor when analysing networks. It is interesting to note the interconnectivity of nodes within a specified area of

the network. Using the ratio between existing and possible relations, a clustering coefficient may be computed. If a node has z nearest neighbours, a maximum of zðz # 1Þ=2 edges is possible between them.

With such a vast amount of information generated through interactions, this can be exploited to generate recommendations or influence for other users. Which persons are the most influential in an OSN? Which Internet Web pages are most influential in a given topic? These subjects are of commercial interest, as well as being key aspects which help us to understand interactions and information flow in an OSN graph. In this section, we refer to three key papers which consider, respectively, authoritative sources in a Hyperlinked web environment [21], finding 'influential' individuals [21, 24] and how to maximize the spread of influence through a social network [21]. In the context of Web pages [83], the theme of authoritative sources in a Hyperlinked environment is considered. This paper is related to Kleinberg's work with co-author Gibson on the HITS algorithm [45], which defined and considered Hubs and Authorities. An iterative algorithm for computing hubs and authorities is defined. In the social network context, the concepts of 'standing', 'impact' and 'influence' are also theoretically defined. In the scientific citations domain, the 'impact factor' was considered, and an improvement was proposed, called the 'influence weight'. The 'influence weight' considers a journal to be 'influential' if it is heavily cited by other influential journals defined in the same manner, and so on, recursively. Other concepts considered were 'diffusion' and 'generalization'.

In the context of finding 'influential' individuals (that is, individuals who are able to influence other individuals in a social network), Kempe et al. [21] studied how to maximize the spread of influence through a social network. The initial problem defined is to choose a subset of N individuals from the whole graph, who are decision makers. Finding these individuals stated as being NP-hard problem. In order to reduce the computational search cost, the 'degree centrality' and 'distance centrality' metrics were proposed as search heuristics. Linear Threshold and Independent Cascade diffusion models which represent two different approaches to solve the influence maximization problem are considered. All empirical tests were carried out using the Arrive high energy physics citation dataset. This filtered dataset has 10,748 nodes and 26,500 edges. The high degree heuristic (based on node centrality) chooses nodes v in the order of decreasing degree size d. The centrality measure assumes that a node with short paths to the other nodes in a network will have a higher chance of influencing them. Four variants were tried for finding the most influential nodes: 'greedy search', 'high degree', 'degree centrality' and 'random'. Greedy combined with degree centrality was found to give the best results, whereas degree centrality on its own worked well, to choose the first node but after that showed little or no improvement. This is because the first node tends to be connected to other candidate nodes. However, combining degree centrality with a greedy search, in which the already chosen nodes were excluded as candidates, gave significantly better results. All the search methods were applied using a weighted cascade model.

It is difficult to obtain large-scale diffusion data and to identify and track on a large scale the elements, such as recommendations [32]. Many researchers have used mathematical models to predict information diffusion over a time period in online social networks. But very less has been attempted to understand temporal and spatial dimensions. A Linear Influence Model to predict the number of newly infected nodes is introduced based on the time when the previous set of nodes are infected by Yang & Leskovec in [60]. The literature says the nodes with the highest follower count is to be most influential, but the findings [4] confirm that always the users with the highest follower count were not the most influential in terms of information diffusion

The main interest in the study of Information diffusion is its application in a commercial environment. This could be either legitimate (solicited) or illegitimate (unwanted or spam) diffusion. Here the focus of research is to understand and analyse how users propagate information from one to another (analogous to 'word of mouth') with semantics [40]

It is studied how viral marketing uses bonafide users to spread a message by 'word of mouth'. In a recent MIT study [31], advertisers are recommended to be social rather than commercial in their marketing messages. For example, advertising programs should use phrases such as 'be like your friend', 'your friend knows this is a good cause', 'learn from your friend' and 'don't be left out', in order to provoke 'viral' propagation through the network. The information diffusion is researched as a model for

representing information in OSNs [50] and the analysis of spammers social networks in order to identify criminal individuals and groups by [51]. The role of social networks in information diffusion is studied in [52] and [53] has detected Internet buzzes and amplification phenomena. Tracking of the contextual entity using 'memes' in [54], its application to Twitter [55] and categorization of Tweets in a large Twitter dataset are done by [56]. Agrawal et al. in [50] present a model for representing information in OSNs, which assigns two parameters to each information item, called endogeneity and exogeneity. The receptivity of a node is introduced as an additional parameter in the model. In a social network data related to the ordering of the adoption of information items by nodes, a maximum-likelihood based method is defined. This is for estimating the endogeneity, exogeneity and receptivity parameters explained by Bakshy et al. in [52]. The information diffusion in a social network like Facebook will be a large scale field experiment using 250 million users. This experiment randomizes exposure to signals, about friend's information sharing. The finding is that those who are exposed are significantly more likely to spread the information, than those who are not exposed. The second finding is, with respect to the role of strong and weak ties in information propagation. The authors confirm that stronger ties are individually more influential. However, they find that are the weak ties, which are much more frequent, which account for the propagation of novel information. The users were demographically identified by gender, age, and country. The specific data analysis technique used was "temporal clustering", which was used to

identify the degree of proximity of the actions of the users.

Tie strength was measured in terms of four types of interactions:

(i) The frequency of private online communication between the two users in the form of Facebook messages;

(ii) The frequency of public online interaction in the form of comments left by one user on another user's posts;

(iii) The number of real-world coincidences captured on Facebook in terms of both users being labelled as appearing in the same photograph;

(iv) The number of online coincidences in terms of both users responding to the same Facebook post with a comment [59].

These four types of interactions are summarized as: comments received, messages received, photo coincidences and thread coincidences. There is also an interest in how OSNs communities are related to information diffusion [51], viral type diffusion [52] and memes as transferable information units in [53, 54] influence and recommendation.

2.9 Visualisation Tools

Visualisation facilitates the analysis and interpretation of different model output and ease in comparing the models. Many visual analytic tools are available, that provide insights into the comparative analytics of different models. Applications and software for social network analysis with respect to software for OSN analysis, on the one hand, there are the "off the shelf"

applications such as Gephi [23] for visualizing graphs and calculating different graph statistics, including community labelling. Gephi includes as standard the following metrics: node centrality, betweenness, closeness, density, path length, diameter, HITS, modularity, and clustering coefficient. Gephi also has a Java API interface for developers. Another popular application is Net Miner [25] which is a commercial software system with specific modules for Twitter data analysis. On the other hand, there are software development libraries and databases for programmers. 'Neo4' [110] is a graph database software for high performance processing with a Java API for 'big data' requirements. The 'Python Network X' graph library includes generators for classic graphs, random graphs, and synthetic networks, standard graph algorithms, and network structure/analysis measures. JUNG (the Java Universal Network/Graph Framework) [111] is an open source graph modelling and visualization framework written in Java. Finally, for those who prefer programming in the 'C' language, there is the 'igraph' library and API [114], and the Stanford Network Analysis Platform (SNAP) [113] is a general purpose, high performance system for analysis and manipulation of large networks, written in C++.

2.10 Sentiment Analysis

Sentiment analysis has become a hot research topic in recent years because of the variety of applications. The popularity is because of the real time applications in classification and summarising reviews. Many tools are available for Sentiment analysis. The main focus areas are, lexicon

construction, feature extraction, and then determine the polarity which can be taken as feedback for improvements in many cases. Sever challenges are existing in understanding the vocabulary of the natural language and discovering their polarity in complex sentences. Machine learning challenges like the aspect and feature identification from different corpora, natural language understanding are some of them. The emergence of crowdsourcing created new opportunities in data collection and annotation methods.

Dave et al. [64] introduced the term opinion mining and further explains the opinion mining tool, capable of processing search results highlighting the attributes, combining all the opinions with qualitative featuring of product attribute. A detailed survey is presented by Pang and Lee [65] and Liu [67] [69]. They have detailed the techniques for solving problems in Sentiment Analysis focusing on applications. They have also discussed and challenges in Sentiment Analysis. User reviews are analysed using machine learning methods by Hu and Liu [69]. They also express the challenges in the analysis especially due to the presence of noise in the text and the complexities of natural language processing. The synonyms and antonyms in WorldNet are used by the set of opinion words tagged. Pang et al. [68] uses overall sentiment to classify documents. The well-known and much-cited paper of Hu and Liu [69] represent components like a product, a person, an event, etc. and associated set of attributes as aspects. The majority of the algorithms for aspect-level sentiment analysis use machine learning classifier. Hoogervorst et al. [70] employ a discourse parser

implementing Rhetorical Structure Theory (RST). In this case, the context of each aspect is determined from the parser and expressed sentiment is computed with respect to the weightage of the discourse relations between words. Determine the polarity of comments whether it is positive, negative or neutral by extracting features and components of the object is Opinion mining [67]. Pang et al. in [68] investigate the effectiveness of sentiment classification of the documents by machine learning techniques. It is demonstrated that human produced baseline for sentiment analysis on movie review data is found to be inferior to by machine learning techniques, but accuracy is better. The experiment was a review of movie corpus classified using SVM, maximum entropy classification and Naive Bayes and features based on unigrams and bigrams. Unsupervised classification and semantic orientation of the classification of reviews as positive or negative is discussed in by Turney [71]. The classification is done by finding out word‘s point wise mutual information (PMI) for their co-occurrence with positive or negative seed word. This method has 74% accuracy.

Sentiment classification can be lexicon based, machine learning or hybrid model. Medhat et al. [105] in their survey gives an overview of the important SA techniques and applications. This paper gives a fine categorization of the various Sentiment Analysis techniques. Gupta [108] discuss the importance of social media mining is reshaping business models especially in “viral” marketing, and promoting more interactions. The authors have presented a framework for Social Media Mining detailing the

underlying process.

Sentiment analysis is a measure of popularity. Sentiment Analysis studied the impact of 13 twitter accounts of celebrated persons on their followers by Bay & Lee[120]. The trend of followers are tracked through Sentiment analysis. About 3,000,000 tweets mentioning or replying to the 13 influential users were analysed. The users selected were those who ranked in the top 50 of online social influence services and having more than 1 million followers. They analysed over 3,000,000 tweets mentioning or replying to these users to determine audience sentiment. It was found that the audience responded by replying, mentioning, or retweeting them with a positive or negative sentiment.

…….&&.....