1. INTRODUCTION

Data mining has been leading research in handling huge amounts of data for solving real-world problems. Data mining seek to extract meaningful information from a data set that is not readily apparent and not always easily obtainable. With the ubiquitous effect of social media via the internet, an unprecedented quantity of data is available and of interest to many fields of study including sociology, business, psychology, entertainment, politics, news, and other cultural aspects of societies.

Application of data mining to social media can yield interesting perspectives on human behavior and human interaction. Data mining based techniques are confirmed to be useful for analysis of social network data, especially for large datasets that cannot be handled by traditional methods. Data mining can be used in conjunction with social media to better understand the opinions people have about a subject, identify groups of people amongst the masses of a population, study group changes over time, and find influential people, or even recommend a product or activity to an individual. It is an important fact that these networks have become a substantial pool of unstructured data that belong to a host of domains, including business, governments and health. Data mining techniques facilitate reforming the unstructured data and place them within a systematic pattern.

1.1 DATA MINING AND SOCIAL NETWORK ANALYSIS

Data mining is the computer-assisted process of mining through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools expect behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that were traditionally too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that specialists may miss because it lies outside their expectations. Data mining takes its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable one. Both processes require either sifting through an enormous amount of material, or intelligently probing it to discover where the value resides.

Encyclopedia Britannica defines Data Mining as the method of discovering interesting and useful patterns and relationships in large volumes of data with the subfields of predictive modeling, descriptive modeling, pattern mining and anomalies mining. It is popularly known

as Knowledge Discovery in Databases (KDD)**,** refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases are regularly treated as synonyms, data mining is part of the knowledge discovery process [1].

Data mining is one of the fastest emergent fields in the computer industry. Once, a small interest region within computer science and statistics, it has rapidly expanded into a field of its own. One of the greatest strengths of data mining is reflected in its wide range of methodologies and techniques that can be applied to host of problem sets.

Data Mining Techniques

Data mining gratify its main goal by identifying valid, potentially useful, and easily understandable correlations and patterns present in existing data. This objective of data mining can be satisfied by modeling it as either predictive or descriptive nature. The predictive model works by building a prediction about values of data, which uses known results found from different datasets. The predictive data mining model contains classification, prediction, regression and analysis of time series. The descriptive models mostly identify patterns or relationships in datasets. It serves as an easy way to explore the properties of the data examined earlier and not to predict new properties. Data mining systems are devised and well equipped with all possible techniques that can satisfy the user's requirements to large extent. Some of the most popular and commonly used techniques are explained below, which are broadly organized into four categories: classification, clustering, prediction and association rules.

Classification

Classification is a common supervised approach and is appropriate when the data set has labels or a small portion of the data has labels. Classification task can be seen as a supervised technique where each request belongs to a class. Classification algorithms begin with a set of training data which includes class labels for each instance. The algorithm study from the training data and builds a model that will automatically categorize new object into one of the distinct classes provided with the training data. Classification rules and decision trees are examples of supervised classification techniques. The main goals of a classification algorithm are to maximize the predictive accuracy attained by the classification model. There are several model techniques used for classification, some of them are decision tree, k-nearest neighbor, support vector machines, naive bayesian classifiers and neural networks [2].

Clustering

Clustering is a division of data into groups of related objects. Clustering is a common unsupervised data mining technique that is useful when confronting data sets without labels. Unlike classification algorithms, clustering algorithms do not depend on labeled training data to develop a model. Instead, clustering algorithms decide which elements in the data set are related to each other based on the similarity of the data elements. Similarity can be defined as euclidian distance for some numerical data sets but often in data associated with social media, cluster techniques has the ability to deal with text [3].

The clustering algorithm can be divided into five categories, viz, Hierarchical, Partition, Spectral, Grid based and Density based clustering algorithms.

Hierarchical Clustering Algorithm: The hierarchical clustering algorithm is a set of data objects forming a tree shaped structure. It can be mostly classified into agglomerative hierarchical clustering and divisive hierarchical clustering. In the agglomerative approach, which is also called as the bottom up approach, each data point is considered to be a separate cluster, and on each iteration the clusters are merged, based on a criterion. The merging can be done by using the single link, complete link, centroid method. In the divisive approach all data points are considered as a single cluster, and it is split into a number of clusters, based on certain criteria, and this is called as the top down approach [4]. Examples of this methods are LEGCLUST [5], BRICH [6] (Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using REpresentatives) [7], and Chemeleon [8].

Spectral Clustering Algorithm: Spectral clustering refers to a class of techniques, which relies on the eigen structure of a similarity matrix. Clusters are formed by partitioning data points using the similarity matrix. Any spectral clustering algorithm will have three main stages [9]. They are preprocessing, spectral mapping and post mapping. Preprocessing contracts with the construction of the similarity matrix. Spectral mapping deals with the building of eigen vectors for the similarity matrix. Post processing deals with the grouping of data points. Examples of this algorithm are, SM (Shi and Malik) algorithm, KVV (Kannan, VempalaandVetta) algorithm, and NJW (Ng, Jordan and Weiss) algorithm [10].

Grid based Clustering Algorithm: The grid based algorithm devise sizes the object space into a finite number of cells that forms a grid structure [11]. Operations are done on these grids. The advantage of this method is its lower processing time. Clustering complexity is supported on the number of populated grid cells, and does not depend on the number of objects in the dataset. The major features of this algorithm are, no distance a computation,

clustering is performed on summarized data points, shapes are limited to the union of grid-cells, and the difficulty of the algorithm is frequently O (Number of occupied grid-cells). STING (STatiscal INformation Grid) is an example of this algorithm [8].

Density based Clustering Algorithm: The density based algorithm allows the given cluster to continue to grow as long as the density in the neighborhood exceeds a certain threshold. This algorithm is suitable for managing noise in the dataset. It handles clusters of arbitrary shape, handles noise, needs only one scan of the input dataset, and the density parameters to be initialized. DBSCAN (Density-Based Spatial Clustering of Applications with Noise), DENCLUE (DENsity-based CLUstEring) and OPTICS (Ordering Points to Identify the Clustering Structure) are examples of this algorithm [12].

Partition Clustering Algorithm: Partitioning methods generally result in a set of M clusters, where each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative which is considered as a summary description of all the objects contained in a cluster.

Association Rule mining

Association rule mining technique is one of the most efficient methods of data mining to search unseen or desired pattern among the vast amount of data. Association rule learning is a rule-based machine learning methods for establishing interesting relations between variables in large databases. It is intended to recognize strong rules hidden in databases using some measures of interestingness. In this method, the focus is on discovery relationships between the unlike items in a transactional database. Association rules are used to discover out elements that co-occur repeatedly within a dataset consisting of many independent selections of elements, and to discover rules. A typical example is market basket, where a pattern has been established showing that buyers who bought bread and butter also bought milk. The most common apriori algorithm provides two input parameters: rule support and confidence. Association rule confidence is the proportion of the data set to which this rule applies. For example, an 80 % rule confidence would mean that 80 % of customers who bought bread and butter also bought milk. Association rule support is the proportion of data set which provides the condition for the rule. For example, 20 % rule support would mean that a total of 20 % of customers bought bread and butter [13].

Prediction

Prediction is nothing but finding out the knowledge or some pattern from the large amounts of data. Prediction in data mining is to recognize data points purely on the

description of another related data value. The prediction in data mining is identified as numeric prediction. Generally regression analysis is used for prediction. For example prediction models in data mining are employed by a marketing manager who predict that how much quantity a particular customer will use during a sale, so that upcoming sale amount can be intended accordingly [14].

Regression is widely used for prediction. Regression algorithm estimates the value of the target as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different set in which the target values are unknown. Regression models are tested by computing various statistics that compute the difference between the predicted values and the expected values. The historical data for a regression project is typically divided into two datasets for building the model and testing the model. The various regression algorithms are discussed. Regression analysis assists to understand how the value of the dependent variable changes when any one of the independent variables changes while the other independent variables are held fixed. In regression the dependent variable is estimated as function of independent variables which is called regression function.

Regression is a prediction method that is supported on an assumed or known numerical output value. This output value is the result of a sequence of recursive partitioning; with every step include one numerical value and another group of dependent variables which branch out to another pair such as this. The regression tree establishes with one or more precursor variables, and terminates with one final output variable. The dependent variables are either continuous or discrete numerical variables. Regression is concerned with modeling the relationship between variables that is iteratively developed using a measure of error in the predictions made by the model [15].

Summarization

Summarization is a key data mining concept which involves techniques for discovery a compact description of a dataset. Simple summarization techniques such as tabulating the mean and standard deviations are often related for data analysis, data visualization and automated report generation. For example, centroids of document clusters derived for a collection of text documents can supply a good indication of the topics being covered in the collection. The clustering supported approach is effective in domains where the features are continuous or asymmetric binary, and hence cluster centroids are a meaningful description of the clusters. If the data has categorical attributes, then the typical techniques for computing a

cluster centroid are not appropriate and hence clustering cannot directly be concerned for summarization [16].

Type of Data mining

Data mining is divided into many areas like text mining, spatial mining, web mining, graph mining, biological data mining, sequential data mining etc.,.

Text Mining: Text mining or text analytics is a process in which information is extracted from the written sources. It also transforms the unstructured text into the structured data for better analysis. The main purpose of text mining is to identify facts and relationships from the large textual data. It helps businesses and organizations to get valuable insights useful for their business. The main processes in text mining include information retrieval, lexical analysis, pattern recognition, and predictive analytics. The foremost step in text mining is to organize the data into a more structured form by involving the use of natural language processing technology. Text mining finds its application in sentiment analysis. Other important applications include social media monitoring, bioinformatics, scientific discovery, competitive intelligence. It is a popular area for research in data mining.

Web Mining: Web mining is the application of data mining techniques to data originating from the web. The sources are web pages, web data repositories, web logs, click-streams, web traffic and web links. Web mining provides the capability of mining the web documents and contents to extract information in an efficient manner to generate extraordinary results. The World Wide Web (WWW) is huge collection of globally distributed set of news, advertisements, consumer records, financial, education, government, e-commerce and many other services. The WWW also contains huge and dynamic collection of hyper linked information, providing a huge source for data mining. Based on the above facts, the web also poses great challenges for efficient resource and knowledge discovery. Web mining describes the application of traditional data mining techniques onto the web resources and has facilitated the further development of these techniques to consider the specific structures of web data [17].

The different dimensions of web mining are web structure mining, web content mining and web usage mining. Various research experts have utilized these techniques enabling the web document publishers and web users to get closer with increased efficiency and better experience. It aims to discover models of objects, processes, interactions and relationships on the web.

Web content mining is the procedure of extracting helpful information from the contents of web documents. Content data communicates to the collection of facts a web page was designed to convey to the users. It can consist of text, images, audio, video, or structured records such as lists and tables. Research activities in this field also use techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). Web content mining processes the content of the web page and extracts knowledge which helps to retrieve data relevant to be presented for a user query. Web data are mainly semi-structured or unstructured. This is widely used to improve web search engine performance. The different types of users accessing a web page more number of times on different context are considerably increasing [18].

The construction of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting between two related pages. In addition, the content within a web page can also be planned in a tree-structured format, based on the various HTML and XML tags within the page. Web structure mining can be regarded as the procedure of discovering structure information from the web. This type of mining can be performed either at the intra-page in document level or at the inter-page in hyperlink level. Web structure mining is the application of data mining techniques on web pages to determine the relationship between them and the manner in which they are linked. It deals with link structure analysis. Web structure mining is the procedure of using graph theory to evaluate the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two different types. They are extracting patterns from hyperlinks in the web and mining the document structure [19].

Web usage mining is the application of data mining methods to discover interesting usage patterns from web data, in order to understand and better serve the needs of web-based applications. Usage data confines the identity or origin of web users along with their browsing behaviour at a web site. Some of the typical usage data collected at a web site includes IP addresses, page references, and access time of the users. Web usage mining constructs patterns out of usage data and browsing behavior of the web. This analyses the web server log data for knowledge discovery. It enables to improve its marketing by attracting and retaining customers. The different kinds of usage data considered are web server data where the user logs are collected by the web server. The others are application server data and application level data. The user sessions are cached using log files. Researchers derive the secondary data from the usage pattern of surfers from which

predictions are made for the interactions. The application area which is steadily gaining interest is e-commerce [20].

Biological Data Mining: Data mining find its application in bioinformatics. It is a field that deals in the collection, processing, and collection of the biological data. There are various applications of data mining in bioinformatics such as gene finding, protein function domain detection, protein function interference. Data mining also offers a solution for analyzing large-scale biological data. It helps in the prediction of functions of anonymous genes. Clustering and classification methods of data mining help in microarray data and protein array data analysis.

Sequential *Data Mining:* Sequential data mining is a discipline of data mining, which aim is to extract frequent subsequences, patterns and sequence rules from given sequences. Sequential mining technique will find the complete set of patterns while supporting the minimum support threshold within given constrains like length constraint, type constraint, gap constraint, etc. In sequential data mining, it is important to keep good identification of sequence states, distinguishing between sequences and order of states in each sequence. Sequential data mining not only provides information about patterns that do occur together, but also distinguishes the order and time difference between each event in given sequences. Sequential data mining is used to discover relationships between occurrences of states to find specific order of the occurrences.

Spatial Data Mining: The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. Spatial data mining is the process of discovering interesting and previously unknowns, but potentially useful patterns from spatial databases. The complexity of spatial data and intrinsic spatial relationships limits the useful of conventional data mining techniques for extracting spatial patterns. Spatial data are the data associated to objects that occupy space. Spatial database collects spatial objects represented by spatial data types and spatial relationship among such objects. Spatial data brings topological and/or distance information and it is often organized by spatial indexing structures and accessed by spatial access methods. These methods distinct features of a spatial database pose challenges and bring opportunities for mining information from spatial data. Spatial data mining or knowledge discovery in spatial data base refers to the extraction of implicit knowledge, spatial relations or other patterns not explicitly stored in spatial databases.

Graph Mining: Graph mining is another good area in data mining for research and thesis. It is a process in which patterns are extracted from the graphs that represent the underlying data. Graph mining discovers its applications in various problem domains, including bioinformatics, chemical reactions, program flow structures, computer networks, social networks etc. Different data mining approaches are used for mining the graph based data and performing useful analysis on these mined data [21].

Social Network Analysis

Social network analysis (SNA) is the mapping and determining of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. The nodes in the network are the community and groups while the links show relationships or flows between the nodes. SNA provides both a visual and mathematical analysis of human relationships.

SNA is also one of the popular topics in data mining for thesis and research. It is a quantitative and qualitative process that measures the flow of relationship in a social network. The relationship is represented in the form of nodes and links where nodes represent the people and links represent the relationships between the nodes. Mathematical and visual analysis of the human relationship is represented by social network analysis [22].

A social network is a term used to describe web-based services that allow individuals to create a public/semi-public profile within a domain such that they can communicatively connect with other users within the network. The social network has improved on the concept and technology of Web 2.0, by enabling the formation and exchange of user-generated content. In simple, the social network is a graph consisting of nodes and links used to represent social relations on social network sites [23]. The nodes include entities and the relationships between them form the links as given in Fig. 1.1.

Fig. 1.1 Social Network Showing Nodes and Links

Social networks are important sources of online interactions and contents sharing, subjectivity, assessments, approaches, valuation, influences, observations, feelings, opinions and sentiments expressions borne out in the text, reviews, blogs, discussions, news, remarks, reactions, or some other documents. Before the advent of the social network, the homepages was popularly used in the late 1990s which made it possible for average internet users to share information. However, the activities on the social network in recent times seem to have transformed the WWW into its intended original creation. Social network platforms facilitate rapid information exchange between users regardless of the location. Many organizations, individuals and even government of countries now follow the activities on the social network. The network enables big organizations, celebrities, government official and government bodies to obtain knowledge on how their audience reacts to postings that concerns them out of the enormous data generated on the social network. The network permits the effective collection of large-scale data which gives rise to major computational challenges [24].

The growing use of the internet has led to the development of networked interaction environments such as social networks. Social networks have acquired much attention recently, largely due to the success of online social networking sites and media sharing sites. In such networks, rigorous and complex interactions occur among several different entities, leading to huge information networks with outstanding business potential. Researchers are increasingly interested in addressing a wide range of challenges exist in these social network systems.

Social networks are graph structures whose nodes represent people, organizations or other entities, and whose edges represent a relationship, interaction, collaboration, or

influence between entities. The edges in the network connecting the entities may have a direction indicating the flow from one entity to the other and the strength of the edge. Social networks need not be always social in context. Real-world networks like WWW, electrical power grids, the spread of computer viruses, telephone call graphs, and co-authorship and citation networks of scientists, customer networks are instances of technological, business, economic, and biologic social networks. Epidemiological networks, cellular and metabolic networks, food webs, are some of the examples of biological networks.

Social networks are highly dynamic in nature. The network grows and changes quickly over time through the addition of new nodes and edges, signifying the social structure. The number of degrees grows linearly in the number of nodes. It has been experimentally shown that when the network grows, the closeness of the nodes increases, resulting in shrinking diameter of the network. The dynamic, dense, reduced diameter properties of the graph show that the social network exhibit heavy-tailed out-degree and in-degree distributions [25].

The dynamic property of such large, heterogeneous, multi-relational social networks has led to an interesting field of study known as social network analysis. Social network analysis examines the structure and composition of links in a given network and provides insights into its structural characteristics. SNA is the study of the evolution of structures i.e., how the networks change over time, and how information propagates within the networks. SNA assumes that relationships are important and focuses on the structure of relationships. It also includes understanding of the general properties of networks by analyzing large datasets collected with the aid of technology.

Social network analysis has emerged as a key technique in modern sociology and has become a popular topic of study in areas like business and economics, geography, information science, organizational studies, social psychology, sociolinguistics. For example, SNA has been used in epidemiology to understand the pattern of human contacts that cause the spread of diseases in a population. SNA can be used as a tool for market analysis based on opinions about products or brand to market products and services. SNA can also be an effective tool for mass surveillance - for example to determine whether or not a particular individual has criminal tendencies.

Structural characteristics of social networks can be explored using socio metrics. Socio metrics are measures used to understand the structure of the network, the properties of links, the roles of entities, information flows, and evolution of networks,

clusters/communities in a network, nodes in a cluster, a center node of the cluster/network, and nodes on the periphery. Some of the commonly used measures for analysis of social networks are,

- Centrality – Node's relative importance within a community
- Prestige- Central nodes in the network
- Prominence – Nodes with the most incoming connections
- Influence – Nodes with most outgoing connections
- Outliers - Nodes with the least connections
- Clique - How connected are neighbors of an entity
- Community - Nodes that are communicating more often with each other
- Path length - Nodes that are involved in passing information through the network
- Density - Proportion of possible links that actually exist in the network

1.2 DATA MINING IN SOCIAL NETWORKS

Social media refers to a variety of information services used collaboratively by many people placed into the subcategories as shown in Table I. The data offered via social media can give us insights into social networks and societies that were not previously likely in both scale and extent. It is extremely difficult to gain useful information from social media data without applying data mining technologies due to unique challenges.

Table I Blogs of Social Media

Category	**Examples**
Blogs	Blogger, LiveJournal, WordPress
Microblogs	Twitter, GoogleBuzz
Opinion mining	Epinions, Yelp
Photo and video Sharing	Flickr, YouTube
Social bookmarking	Delicious, StumbleUpon
Social networking sites	Facebook, LinkedIn, MySpace, Orkut
Social news	Digg, Slashdot
Wikis	Scholarpedia, Wikihow, Wikipedia, Event maps

Data mining techniques can help effectively deal with the three main challenges with social media data. First, social media data sets are large. Without computerized information

processing for evaluating social media, social network data analytics becomes an unattainable in any reasonable amount of time. Second, social media data sets are noisy. For example, spam blogs or splogs are abundant in the blogosphere, as well as excessive trivial tweets on Twitter. Third, data from online social media is dynamic such that frequent changes and updates over short periods of time are not only common but an important dimension to consider in dealing with social media data [26].

Data mining techniques developed for other problem domains can be applied to social media data without having to start from scratch. Applying data mining techniques to large social media data sets has the potential to continue to improve search results for everyday search engines, realize specialized target marketing for businesses, help psychologist study behavior, provide new insights into social structure for sociologists, personalize web services for consumers, and even help detect and prevent spam for all of us. The advancement of the data mining field itself relies on large data sets and social media is an ideal data source in the frontier of data mining for developing and testing new data mining techniques, algorithms and methodologies to perform social network analysis [27].

Data mining techniques can be concerned to social media to understand data better and to make use of data for research and business purposes. Representative areas include community or group detection, information diffusion, influence propagation, topic detection and monitoring, individual behavior analysis, group behavior analysis, and marketing research for businesses. The graph representation enables the application of classic mathematical graph theory, traditional social network analysis methods, and work on mining graph data.

Social media mining is a process of visualizing, evaluating and extracting applicable patterns over the social network. Social media mining defines the basic principle and concepts for investigating a huge amount of social media data. Social media sites generate user data which is different from traditional attribute-values of data for hellenic data mining. The data which is generated from social sites is noisy, distributed, not in proper structure and frequent. All the characteristics of social media data pose challenges for data mining task and for that new techniques and algorithm has been developed [28].

Community detection is one of the most popular applications of data mining to social networking sites which finds and identifies a community or group. An example social network community structure is shown in Fig 1.2. Community detection applied to social networking sites is based on analyzing the structure of the network and finding individuals that associate more with each other than with other users. Understanding what groups an

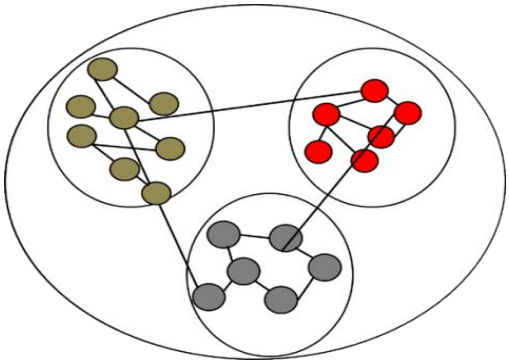

Fig. 1.2 Social Network Community Structure

individual belongs to can help lead to insights about the individual such as what activities, goods, and services, an individual might be interested in. Community detection can also yield interesting perspectives about the social networking site, such as how many different groups are using the social networking site. Community detection has been approached using various strategies including node-centric, group-centric, network-centric, and hierarchy-centric. Large number of persons participating in social network sites severely limits the ability to detect groups without computational processes [29].

Applications of Social Network Mining

Social network and its analysis is an important field and it is widely spread among many young researchers. Social networks research emerged from psychology, sociology, statistics and graph theory. Based on graph theoretical concepts a social network interprets the social relationships of individuals as points and their relationships as the lines connecting them.

A number of research issues and challenges facing the realization of utilizing data mining techniques in social network analysis could be identified as follow in link mining, viral marketing, recommended the system, semantic web, community, news groups, opinion mining. Having presented some of the research issues and challenges in social network

analysis, the following sections and sub-sections present the overview of different data mining approaches used in analyzing social network data.

Link mining: Traditionally data mining and machine learning tasks have been carried out using a single relation of homogenous objects. The data comprising social networks is heterogeneous, multi relational, and semi-structured, thus a new field of research called link mining has emerged. Various mining tasks can be performed by considering only links - the relationships between objects. Both object attributes and link information are made use of for link mining process and being applied in various domains like www, business, bibliography, and epidemiology. Link based object classification, object type prediction, link type prediction, link existence prediction, link cardinality prediction, object reconciliation, group and subgroup detection, metadata mining are common link mining tasks. For example, a typical data mining task is to predict the edges that will be added to the network from a particular time to a given future time [30].

Mining Customer Networks for Viral Marketing: Viral marketing is a new marketing strategy using social network mining that explores how individuals can influence the buying behavior of others. The basic principles of viral marketing are social profile gathering, proximity market analysis, real-time key word density analysis. The development of social networks like e-mail mailing lists, usenet groups, on-line forums, instant relay chat (IRC), instant messaging, collaborative filtering systems, and knowledge-sharing sites facilitates mining the buying pattern of customers for viral marketing. Such sites allow users to offer an opinion about products to help customers to rate the products. Viral marketing aims to optimize the positive word-of-mouth effect among customers. The customer's network value is considered as most important for viral marketing. Based on the interactions between customers, viral marketing can produce higher profits than traditional marketing such as direct marketing, mass marketing which ignores such interactions. Finally, for example, a data mining task would be finding the optimal set of customers that maximizes the net profits. Viral marketing techniques can also be applied to other areas like reducing the spread of HIV, combating teenage smoking, and grass-roots political initiative [31].

Mining Newsgroups using Networks: Newsgroups are rich sources of openly available discussions on any conceivable topic wherein the arguments are mostly open, frank, and unmodified. Newsgroup postings can provide a quick pulse on any topic. Extracting hidden information from newsgroups is another area of social network mining. A newsgroup discussion on a topic consists of seed postings and a large number of additional postings that

are responses to a seed posting or responses to responses. Responses typically quote clear passages from earlier postings. Such quoted responses from quotation link and create a network in which the vertices represent individuals and the links represent responded-to relationships. It is also true that people more frequently respond to a statement when they disagree than when they agree. This behavior exists in many newsgroups, based on which, one can effectively classify and partition authors in the newsgroup into opposite groups by analyzing the graph structure of the responses. The graph structure is constructed by creating a quotation link between person i and person j if i has quoted from an earlier posting written by j.

Opinion Mining: The web is a wealthy source of information. Web 2.0 provides ample opportunities to express personal experiences and opinions on almost anything at review sites, forums, discussion groups, blogs etc. People express their emotions and opinions about various topics like arts, literature, financial markets, about individuals, organizations, ideologies, and consumer goods. When people make decisions to buy products or use services, they search for these opinions instead of searching for facts. 84 percent of millennial say that user-generated content has at least some influence on what they buy. Organizations also look up to opinions regarding their products to be aware of the market trends and changes. Hence a system to identify and classify opinions expressed in electronic text and to find valuable and interesting information is essential. Sentiment analysis or opinion mining is the computational learning of people's opinions, appraisals, and emotions toward entities, events and their attributes. It involves the application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in source materials [32].

Social Recommendation: Traditional recommendation systems effort to recommend items based on aggregated ratings of objects from users or past acquires histories of users. A social recommendation system builds use of user's social network and related information in count to the traditional recommendation means. Social recommendation is based on the hypothesis that people who are socially connected are more likely to share the same or similar interests and users can be easily influenced by the friends they trust and favor their friends' recommendations to random recommendations. Objectives of social recommendation systems are to improve the quality of recommendation and alleviate the problem of information overload. Examples of social recommendation systems are volume suggestions

based on friends' reading lists on Amazon or friend recommendations on Twitter and Facebook [33].

Community Mining: Community mining is one of the major directions in social network analysis. A community can be defined as a group of objects sharing some common properties. Community mining can be thought of as subgraph identification. In real social networks, there always exist various kinds of relationships between the objects. Each relation can be viewed as a relation network or relation graph and multiple relations form a multi relational social network called as heterogeneous social network. Each kind of relation may play a distinct role in a particular task. The different relation graphs can provide us with different communities. The relation that plays an important role in a community is to be identified in order to determine a community with certain characteristics. This leads to the problem of multi relational community mining, which involves the mining of hidden communities on heterogeneous social networks. For example, in www, two Web pages or objects are related if there is a hyperlink between them. A graph of web page linkages can be mined to identify a community or set of web pages on a particular topic [34].

1.3 COMMUNITY DETECTION APPROACHES

Community is most often defined as a group of individuals living in the same geographical location. It can also be used to portray a group of people with a shared characteristic or common interest. In a social network context, community is a set of actors interacting with each other frequently among which there are relatively strong, direct, intense, frequent or positive ties.

Community detection is the process of discovering groups in a network where individuals group memberships are not explicitly given. The problem of community detection in real-world graphs that involves large social networks, web graphs etc. Communities are the output of a community detection procedure, hoping that communities bear some relationship to a set of nodes. Once extracted, such clusters of nodes are often interpreted as organizational units of social networks.

Identifying groups of vertices in a network based on structural properties is known as community detection. Methods to recognize such groups take a wide variety of approaches, mirroring the diversity in fields where an accurate view of structural communities is useful. Each definitions of community detection have resulted in a number of methods which aim to produce the best set of communities relative to the definition chosen. The criteria of

communities can be categorized into four categories: node-centric, group-centric, network-centric, and hierarchy-centric. Some available methods for community detection are stated below.

Node-Centric Community Detection: Node-centric community detection is based on nodes in a network where each node in a group requires satisfying certain properties such as mutuality, reachability and degree. In case of community detection with complete mutuality, the communities are formed by considering more than two nodes and all are adjacent to each other which are termed as a clique. Reachablilty property is satisfied when there exists a path between two actors or nodes in a community. Otherwise, two nodes can be considered as part of one community if there is a path between these two nodes. Nodal degree is used to check whether the actors within a group are adjacent to a relatively large number of group members or not. Two commonly studied sub-structures are k-plex and k-core.

Group-Centric Community Detection: Group-centric community detection considers connection inside a group as a whole. In this case, it is acceptable to have some nodes in the group with some low connectivity as long as the group overall satisfies certain requirements. It only focuses on connection of nodes only inside a particular group. It has no guarantee whether the reachability for each node in a group [35].

Network-Centric Community Detection: The Network-centric community detection considers the global topology of a network. It aims to partition nodes of a network into a number of disjoint sets. Network-centric-community detection aims to optimize a criterion defined over a network partition rather than over one group. It detects the complete connection of the whole network by partitioning the actors into a number of small disjoint sets [35].

Hierarchy-Centric Community Detection: Based on the network topology, hierarchy-centric community detection builds a hierarchical structure of communities. There are three types of hierarchical clustering: divisive, agglomerative, and structure search. Divisive clustering first partitions the actors into several disjoint sets. Then each set is further divided into smaller ones which contain only a small number of actors. Agglomerative hierarchical clustering starts with each node as a separate community and merges them successively to form a larger community. The hierarchical clustering is based on modularity criteria [35].

Challenges in Community Detection

The major challenges usually encountered in the problem of community detection in social media data are scalability, heterogeneity, evolution, evaluation, collective intelligence and privacy.

Scalability: The amount of online social media content over the internet is raising everyday at a tremendous rate. The sizes of social networks are in scale of billions of nodes and connections. As the network is expanding, both the space requirement to store the network and time complexity to process the network would increase exponentially. This imposes a great challenge to the conventional community detection algorithms. A traditional community detection method often deals with thousands of nodes or more.

Heterogeneity: Raw social media networks comprise multiple types of edges and vertices. Usually, it is represented as hypergraphs or k-partite graphs. Majority of community detection algorithms are not applicable to hypergraphs or k-partite graphs. It is common practice to extract simplified network forms that depict partial aspects of the complex interactions of the original network.

Evolution: Due to the highly dynamic nature of social media data, the evolving nature of the network should be taken into account for network analysis applications. Community detection has progressed under the silent assumption that the network under consideration is static. The time factor is to be incorporated in the community detection approaches.

Evaluation: The lack of reliable ground-truth makes the evaluation extremely difficult. Currently, the performance of community detection methods is evaluated by manual inspection. Such anecdotal evaluation procedures involve extensive manual effort, are non-comprehensive and imperfect to small networks.

Collective Intelligence: People share their thoughts online in the form of comments, reviews, ratings, etc. Such meta-information is useful for many applications. Collecting the intelligence from such data efficiently is not a straightforward job but it is very necessary because this intelligence is very expensive.

Privacy: Privacy is a big concern in social media. Facebook, Google often appear in debates about privacy. Simple anonymization does not necessarily protect privacy. As private information is involved, a secure and trustable system is dangerous. Hence, a lot of valuable information is not made available due to security concerns [35].

1.4 REVIEW OF LITERATURE

One of the key research fronts in social network domain is community structures. Community structure is the most widely considered structural features of complex networks. Communities in a network are the dense groups of the vertices, which are tightly coupled to each other contained by the group and loosely coupled to the rest of the vertices in the network. Community detection plays a key task in understanding the functionality of complex networks. To afford insightful information about community detection, much research has been conducted in the form of surveys, systematic literature reviews and visual studies. Analysis of social networks and in particular discovering communities within social networks has been a focus of recent work in several fields and has diverse applications. In order to analyze complex networks to find significant communities, subgroup communities, and overlapping communities, several methods have been proposed in the literature and few prominent works are presented below.

David R Wood introduced a branch-and-bound algorithm for the maximum clique problem which was applied existing clique finding and vertex coloring heuristics to determine lower and upper bounds for the size of a maximum clique. Computational results on a variety of graphs indicated the proposed procedure in most instances outperformed leading algorithms [36].

Vladimir Batagelj et al. studied the k-core decomposition applied to visualization problems, introducing some graphical tools to analyze the cores, mainly based on the visualization of the adjacency matrix of certain k-cores. It was found that k-core decomposition can be used for quick identification of important parts of a network. Since other types of subgraphs like k-clique, k-connected component are contained in the k-core and it can be used to speed-up the corresponding search algorithms [37].

Michelle Girvan et al. highlighted the property of community structure, in which network nodes are joined together in tightly-knit groups between which there are only looser connections. A new method was proposed for detecting such communities, built around the idea of using centrality indices to find community boundaries. This method was tested on computer generated and real-world graphs whose community structure was already known, and found that it detected this known structure with high sensitivity and reliability. The method was applied to two networks, a collaboration network and a food web whose community structure was not well-known and found that it detected significant and informative community divisions in both cases [38].

M. Girvan, M. E. J. Newman, and A. Clauset investigated complex networks that possess many distinctive properties, of which community structure is one of the most studied. The Edge betweenness community detection method proposed in classified 90% or more of the vertices correctly [39]. In [40] the authors analyzed a network of items for sale on the website of a large online retailer items in the network being linked if the items are frequently purchased by the same buyer. The network has more than 400000 vertices and 2 million edges. The algorithm extracted meaningful communities from this network, revealing large-scale patterns present in the purchasing habits of customers.

M. E. J. Newman et al. proposed algorithm for discovering community structure in networks. The algorithm involve iterative removal of edges from the network to split it into communities, the edges removed being identified using any one of a number of possible betweenness measures and these measures are, crucially recalculated after each removal. A measure to determine the strength of the community structure found by algorithms is proposed which gives us an objective metric for choosing the number of communities into which a network is divided. It was demonstrated that the proposed algorithms are highly effective at discovering community structure in both computer-generated and real-world network data [41].

Palla et al. in 2005 presented the first overlapping community detection algorithm. In this approach, communities were identified based on the k-cliques. According to this algorithm node, may belongs to many communities. Clique is a subset of nodes where every node is adjacent to every other node. K-clique represents size of the clique. For exampl, 4-clique indicates a sub graph having 4 nodes [42].

Baumes et al. described models and efficient algorithms for detecting groups functioning in communication networks which attempt to hide their functionality i.e. hidden groups. In the algorithm, communication networks were viewed as random graphs where the nodes are actors of the network and an edge represents a communication between the corresponding actors. The authors assumed that an approach to detect hidden groups should not rely on the semantic information contained in the communications since it is usually encrypted or unavailable. The authors pointed out that hidden group communication arises out of necessity and certainly nonrandom in behaviour. In order to identify the presence of a hidden group and its members, the authors used three notions namely internally connected, externally connected and disconnected. A group is said to be internally or externally

connected if a message can be passed between any two group members without or with the use of outside third parties. A group is disconnected if it is not externally connected [43].

E. A. Leicht et al. considered the problem of finding communities or modules in directed networks. It was shown that the widely used community finding technique of modularity maximization can be generalized in a principled fashion to incorporate information contained in edge directions. An explicit algorithm based on spectral optimization of the modularity was described and shown that it gives demonstrably better results than previous methods on a variety of test networks, both real and computer generated [44].

Kumpula et al. presented a fast community detection algorithm called Sequential Clique Percolation method (SCP) for weighted and unweighted networks with cliques of a chosen size. It sequentially inserts links to the network and keeps track of the emerging community structure. When links are inserted in order of decreasing weight, the algorithm detects k-clique communities at chosen threshold levels in a single run and simultaneously produces a dendrogram representation of hierarchical community structure. This algorithm has been specifically designed for dense weighted networks containing hierarchical communities where weight-based thresholding of either the links or the cliques formed by them is necessary for obtaining meaningful information on the structure. The computational time of the SCP algorithm scales linearly with the number of k-cliques in the network. SCP is faster than CPM and allows multiple weight thresholds in a single run [45].

Hua-Wei Shen proposed a metric to address the problem of overlapping communities. Metric supposes that a maximal clique only belongs to one community. A maximal clique network from the original network was constructed and proved that the optimization of the metric on the original network was equivalent to the optimization of Newman's modularity on the maximal clique network. Thus, the overlapping community structure can be recognized through partitioning the maximal clique network using any modularity optimization method. The effectiveness of the metric was demonstrated by extensive tests on both artificial networks and real-world networks with a known community structure. A measure Qc for the quality of a cover of network was proposed to quantify the overlapping community structure and it was 0.385 and 0.490 for both artificial networks and real-world networks [46].

Balasundaram et al. formulated the maximum k-plex problem as an integer linear program and used valid inequalities based on independent sets of size at least k. These

inequalities were generated by a simple greedy algorithm both for the whole problem and at local branches in the search tree. It successfully ran a branch and cut algorithm on large-scale instances of real-life social networks known as the Erdos graphs and successfully solved the maximum 2-plex problem on graphs with 80% density and 350 vertices in less than 8 hours [47].

I.Psorakis has identified a probabilistic approach to community detection that utilizes a Bayesian non-negative matrix factorization (NMF) model to extract overlapping modules from a network. This algorithm uses a generative model in a probabilistic framework in which priors exist over the model parameters. The number of latent communities or classes of nodes is used as model order selection in this framework. The authors showed that the degree of participation of two individuals in various communities is a latent generator of the expected number of interactions between them. The algorithm demonstrated how NMF not only captures the membership of a node in multiple communities but also quantifies how strongly that individual participates in each of the groups. By using the entropy of the node membership distribution, core nodes in each community or inversely, broker nodes that act as mediators between different groups were identified. The authors emphasized that the mean entropy of the membership distributions can help to quantify the degree of fuzziness in the network or the clarity of community structure. The limitation of NMF is, it assumes a fully observed adjacency matrix which is not applicable to many real-world networks [48].

JieruiXie et al. proposed a framework for evaluating the ability of algorithms to detect overlapping nodes, which helps to assess over-detection and under-detection. After considering community level detection performance measured by Normalized Mutual Information, the Omega index, and node level detection performance measured by F-score it was found that Speaker-listener Label Propagation Algorithm (SLPA), Order Statistics Local Optimization Method (OSLOM), Game and Community Overlap PRopagation Algorithm (COPRA) offered better performance for low overlapping density networks than the other tested algorithms [49].

Kaikuo analyzed the association between cohesive subgraph visualization plot and k-clique community detection. An algorithm named LargeKCliqueCSV was proposed to detect k large communities that reduced search space. Experiments were conducted on stock market datasets Experimental results showed the good scalability of algorithm comparing with the state-of-art methods such as Clique Percolation Method and Sequential Clique Percolation

algorithm when *k* grows large. LargeKCliqueCSV is superior over both algorithms when *k* is larger than 7[50].

Yangyang Li et al. proposed to detect the community structure in complex networks for a spectral clustering-based adaptive hybrid multi-objective harmony search algorithm (SCAH-MOHSA) combined with a local search strategy. In this approach, an improved spectral method was employed to convert the community detection problem into a data clustering issue. An adaptive hybrid multi-objective harmony search algorithm is used to solve the multi-objective optimization problem so as to resolve the community structure. The algorithm was tested on both synthetic and real-world networks, and demonstration of that method achieves partition results which fit the real situation in an even better fashion [51].

Sundip Misra et al. proposed a community detection scheme in an integrated Internet of Things (IoT) and Social Network (SN) architecture. The integrated proposed method uses a graph mining approach in which the formation of community is considered only if the occurrences of two nodes are at most one hop apart and has at least two mutual friends. This approach has taken mutual friends as a metric for suggesting friends, and the suggestions for friends were generated based on the number of shared friends. In this approach, the smallest community is a sub-graph with a cycle of length four, and a node can be part of multiple communities, which works well for weighted graphs. The results for community detection approach in an integrated environment are more relevant for intra-community methods than inter-community methods [52].

Deepjyoti Choudhury et al. focused on the detection of communities as well as sub-communities occurring in a social network by applying Newman-Girvan algorithm. The implementation was completed on real-world networks like Zachary Karate Club, College Football Network and Bottlenose Dolphin Network, and it detects sub-communities in real world networks. The study proved that the Newman-Girvan algorithms are highly effective at discovering community structure in both computer-generated and real-world network data, and they can be used to shed light on the complex structure of networked systems [53].

Ahmed Ibrahem Hafez et al. introduced a statistical model of the interactions between actors in a social network and used Bayesian network to show the relation between model variables. A community detection algorithm was presented, and the model parameters were derived using Expectation Maximization. The approach worked well with directed and undirected networks, with weighted and un-weighted networks and yielded very promising results when applied to the community detection problem [54].

Bapuji Rao et al. proposed a new and simple algorithm for community detection as well as finding isolated communities in a social network using graph techniques. The scenario of a social graph that consists of various villages in a panchayat was considered. An algorithm was proposed to detect the community and graph underlying the original large community graph. The work analyzed the community network using the incidence matrix of an undirected graph. In this approach, a community or a group was detected based on complete mutuality, nodes reachability, and nodal degrees [55].

Jing Qiu et al. proposed a method to detect the number of communities based on spectral clustering. The bisection method based on the spectral clustering is one of the most common methods for community detection, based on graph spectral theory. The algorithm estimated the number of communities according to the eigen value distribution of the Laplacian matrix, and the k-means algorithm was applied for clustering. The algorithm based on spectral clustering was applicable to the network graph which was divided into two communities. Experimental results on Zachary Karate Club showed that the proposed method yielded high accuracy and effectiveness. [56].

Giulio Rossetti et al. proposed a novel approach to evaluate aimed at calculating the adherence of a community partition to the ground truth. The methodology provided more information than the state-of-the-art ones and is fast to compute on large-scale networks. It was evaluated by applying it to six popular community detection algorithms LOUVAIN, INFOHIERMAP, CFINDER, DEMON, ILCD and EGONETWORK on four large-scale network datasets Amazon, DBLP, YouTube and LiveJournal. The proposed network achieved F1 measure of 0.82 for DBLP network using CFinder [58].

Yue Wang et al. proposed a new k-plex based community model for community search. The authors formulated the maximum k-plex community query (MCKPQ) problem, that is, given a set of query nodes Q, searching for optimal k-plex containing Q. An efficient branch-and-bound (B&B) method was proposed and an effective upper bound function, pruning strategy was designed. Furthermore, the basic branch-and-bound method was optimized by fast candidate generation. The effectiveness of the model and the efficiency of the methods were verified by elaborate experiments [59].

Hossein Esfandiari et al. present the first distributed and the first streaming algorithms to compute and maintain approximate k-core decomposition. The algorithms achieved rigorous bounds on space complexity while bounding the number of passes or number of rounds of computation. A new powerful sketching technique for k-core decomposition was

computed efficiently in both streaming and MapReduce models. The effectiveness of the sketching technique was confirmed empirically on eight publicly available graphs and the median error was always below 50% [60].

The summary of the literature survey is presented in Table II.

Table II Summary of Literature Survey

Authors	Algorithms	Dataset	Advantages	Disadvantages
Hua-Wei Shen, Xue-Qi Cheng1 and Jia-Feng Guo (2009)	Maximal clique Algorithm	Word association network	A maximal clique network is constructed from the original network, and then the overlapping community structure can be identified using any modularity optimization method on the maximal clique network	
Sudip Misra, Romil Barthwal, Mohammad S. Obaidat (2012)	Community Detection in an Integrated Internet of Things and Social Network Architecture	complex networks into basic nodes and IoT nodes	An actor or a node in the network can be part of multiple communities, similar to a real-life situation. This approach can be used to suggest friends	This approach is not generalized to all the networks. It fails to give results for directed networks, such as twitter
Ganjaliyev. F (2012)	Community Detection in weighted network	data forwarding in Delay Tolerant Networks and worm containment in Online Social Networks	Total weight of all selected clusters in the network is calculated and also the similarity in between the clusters	The proposed approach is not cleared for the type of network, for it will work for all network
Deepjyoti Choudhury, Saprativa Bhattacharjee, Anirban Das (2013)	Community & sub – community detection using newman-girvan algorithm	captures the intuition of a network cluster	Author's have defined a new concept of detecting sub-communities	Computational cost is relatively high
Michael Girvan (2013)	Community detection for distributed environment in Web-Scale Networks	Complex smaller sub networks	The proposed approach makes possible to process a graph as large as ∼ 3.3 billion edges on small Hadoop cluster with 50 nodes in just a few hours	
Ahmed Ibrahem Hafez, Abaul ella Hassanien, Aly A. Fahm, and M.F. Talba (2013)	Community Detection using Bayesian network and Expectation Maximization technique	social network's actors	It works well with directed and undirected networks, and with weighted and un-weighted networks	It requires specifying the number of communities in advance

Bapuji Rao, Anirban M (2014)	Community detection in a Social Network, using graph mining technique	Consists of various villages in a panchayat	Useful for a very large number of nodes	
Yomna M. ElBarawy, Ramadan F. Mohamedt and Neveen I. Ghali (2014)	Community detection using DBSCAN algorithm	LFR1, LFR2, LFR3 and LFR4	DBSCAN algorithm has a notion of noise, and is robust to outliers	DBSCAN is not entirely deterministic, i.e. border points that are reachable from more than one cluster can be part of either cluster
Jing Qiu, Jing Peng Ying Zhai (2014) Yangyang Li, Ruochen Liu, and Jianshe Wu (2012)	Community detection based on spectral clustering	Social and economic networks, Information networks, The internet and the world wide web, Immunization and epidemiology networks, Molecular and gene regulatory networks, Sensor networks, Power grids amd transportation networks	It not only find the real or near the real partition but also is able to find out solutions with underling hierarchical structures and fuzzy nodes that can hardly be discovered by single objective optimization approaches	
Yue Wang, Xun Jian, Zhenhua Yang, Jia Li (2017)	Maximum k-plex Algorithm	DBLP, Amazon, Arxiv COND-MAT, Google-Web	propose an efficient branch-and-bound method and design an effective upper bound function and a pruning strategy.	A global optimal graph partition may not be an local optimal community for query sets
Zhenjun Li, Yunting Lu, Wei-Peng Zhang, Rong-Hua Li, Jun Guo, Xin Huang Rui Mao	K-Core-Truss algorithm	NotreDame, LiveJournal1, DBLP, Gowalla, wiki-Talk	k-core-truss decomposition algorithm to find all k-core-truss in a graph G by iteratively removing edges with the smallest degree-support.	To study k-core-truss decomposition and search in the environment of dynamic graphs, ie, the nodes/edges are frequently inserted /deleted.
Hossein Esfandiari, Silvio Lattanzi, Vahab Mirrokni (2018)	k-core decomposition algorithm	Stanford Large Network Dataset Library, Enron, Epinions	design efficient MapReduce and streaming algorithms to approximate the coreness number of all the nodes in a graph efficiently.	The most interesting open problem in the area is to design a fully dynamic algorithm to maintain the core-labeling of a graph by using only poly log n operations per update

Motivation

Ever since the internet became publicly available it has allowed users to interact with each other across virtual networks. An online social network facilitates the sharing of a lot of information, allows discussions on different topics and individuals can either post information or access any information. An online community provides openness, interoperability, scalability and extensibility of a traditional community. Community detection in social network analysis provides an approach to a real-world problem with immediate application. This is the need of the hour, especially with more and more of society being incorporated into a virtual world where high amounts of data are being recorded.

Most of the existing studies in community detection deal with benchmark datasets and few studies analyze the crawled dataset. A detailed analysis of the strength of a specific network, its community, sub-community, and overlapping community detection was not carried out. From the literature survey, it is observed that there is no agreement on a definition for a community in spite of excessive studies which have been performed on the community structure of real networks. Hence evaluating the quality of the communities identified by different community detection algorithms is difficult. Also, there exist challenges in using community detection for analysis of network data with regard to the scalability of the existing algorithms. The data which is generated from social sites is noisy, distributed, unstructured and dynamic. The detection of communities is a complex and challenging task due to scalability, heterogeneity, the security of the social network.

The motivation of the proposed work is to carry out certain investigations on community detection in social network and to develop efficient community detection algorithm by converting the social network analysis problem into graph partition problem since networks are modeled as graphs. Various community detection approaches like community-based, sub community based, overlapping community detection are experimented on a sample twitter network of a sports person and proposed two new approaches for overlapping community detection.

1.5 OBJECTIVES OF THE RESEARCH

Community detection in social networks plays a major role in solving real-world problems. Proper analysis of the network with its crucial properties provides a solution for community detection. Recent advancements in research offer powerful tools and techniques for effective community detection. The main aim of the research work is to analyze node-

centric community detection methods and to propose two hybrid approaches for efficient community detection in social network using graph partitioning techniques. The core objectives of the proposed work are as follows,

- To analyze twitter network data with its structure and graph properties such as closeness, degree, and betweenness
- To discover principal communities of the twitter network using edge betweenness and random walks of the directed network through graph-partitioning
- To detect and analyze communities using sub graph such as maximal k-clique, maximal k-core, maximum k-plex techniques
- To develop ground truth communities for evaluating the performance of proposed community detection methods
- To identify and analyze overlapping communities using clique percolation method and its variants such as optimization and parallel clique percolation
- To develop a hybrid clique percolation method with optimal z-score of k-core subgraph for detecting overlapping communities
- To develop an enhanced clique percolation method through the mining of frequent patterns of similar interest groups for detecting overlapping communities

1.6 ORGANIZATION OF THE THESIS

The rest of the thesis is organized as follows:

Chapter 2 includes the basic definitions and concepts of graph theory. A brief description of various community detection approaches in a social network such as Girvan-Newman edge betweenness and random walks community are also presented. The chapter elucidates various node-centric community detection algorithms.

A sample twitter network is drawn for the purpose of research and used for implementing various community detection approaches. The process of crawling sports person's twitter network is presented in chapter 3. Various representations of the network in such as adjacency matrix, nodelist, edgelists are illustrated. Network analysis of the input network using network properties are also described in the same chapter.

Chapter 4 presents the implementation of principal community detection. Experiments carried out using Girvan-Newman edge betweenness algorithm and random walk algorithms on real time twitter data are illustrated with tables and figures. The chapter also discusses various measures such as centrality measures and modularity scores used to find the strength

of group. The performance analysis of the results and findings of the experiments are reported in the same chapter.

In chapter 5, a brief introduction to sub-community detection of social networks is presented. It discusses the working principle of commonly used node-centric subgraph analysis algorithms like Maximal k-Clique, Maximal k-Core, and Maximal k-Plex algorithms. Also, the experiments on twitter data using these algorithms are illustrated and the results are reported in tables and charts in this chapter.

Chapter 6 provides an overview of the clique percolation method of finding overlapping community. Ground truth communities developed for the sample twitter dataset is illustrated. The experiments based on normal, optimized and parallel clique percolation methods on twitter dataset is presented in detail in the same chapter with an analysis of results using ground truth communities based on different quality measures.

Chapter 7 describes the methodology of two proposed hybrid approaches for community detection (i) using optimal k-core with clique percolation and (ii) enhanced clique percolation method using association rule mining. These overlapping community detection experiments are elaborated, and the effectiveness of the proposed methods is reported in this chapter with tables and figures.

Finally, in chapter 8 the research work is concluded by giving an outline of entire research work with various findings of the investigations on node-centric community detection. This chapter also summarizes the research achievements of the proposed community detection models and presents the scope for future research.