Chapter 1

Introduction

This chapter establishes the key concepts and vocabulary used in the rest of the thesis. We begin with a general introduction to the network science in Section 1.1, followed by a brief description of social networks, link prediction, and influence maximization in sections 1.2, 1.4, and 1.5, respectively. In Section 1.6, the general limitations of existing methods of link prediction and influence maximization in online social networks are illustrated. In Section 1.7, we provide motivation for our work. Section 1.8 summarise the main contributions of the thesis. Finally, Section 1.9 presents the layout of the rest of the thesis.

1.1 Network Science

Network Science has been getting a lot of attention recently. Broadly, it deals with complex networks such as telecommunication networks, computer networks, biological networks, cognitive and semantic networks, and social networks, considering distinct elements or actors represented by nodes and the connections between the elements or actors as links. This field of research is interdisciplinary based on the concepts and algorithms of graph theory [1] from mathematics, statistical mechanics [2] from physics, data mining [3] and information visualization [4] from computer science, inferential modelling [5] from statistics, and social structure [6] from sociology.

The United States National Research Council defines network science as "*the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena*" [7].

Various kinds of data can be naturally mapped to graph/network structures [8, 9]. The scientific study of networks has greatly benefited from a broad range of ideas brought by specialists from different disciplines. Information networks such as web pages linked together by hyperlinks [10], virtual networks of computers that allow sharing of files between computer users [11], and citation networks between academic papers [12] have become essential for information dispensation. Biological networks, such as the network of metabolic pathways [13], the network of interactions between proteins [14], or genetic regulatory networks [15], can be used to understand the phenomena in nature better. The technological networks include the artificial networks designed for resource distribution such as Electric Power Grids [16], the networks of Airline Routes [17], Roads [18], and of course, the Internet [19].

Probably the most studied network these days is the social network, representing a set of people with some contacts, such as friendships, business relationships, etc. In this thesis, we mainly deal with these social networks and issues associated with them.

1.2 Social Networks

A social network is a social structure made up of social actors (such as individuals or organizations), dyadic ties, and other social interactions between actors. The social network outlook offers a set of methods for analyzing the structure of complete social entities and a variation of theories explaining the patterns observed in these structures. The first representation of social networks was the sociograms developed by Jacob Moreno in 1933 to learn interpersonal relationships and was mathematically introduced in 1954 by the sociologist J. A. Barnes.

A new form of social network has emerged recently. These are Online Social Networks (OSN) such as Twitter, Facebook, YouTube, Google+, Instagram, Snapchat, Telegram, and LinkedIn, which have appeared after the development of the internet and computing technology. They provide an online platform that enables electronic communication between the users. Millions of users have joined these networks forming relationships and groups based on their interests and requirements. These networks can be accessed through computers, mobile phones, tablets, etc., and provide many services like chatting, content sharing, profile management, online forums, and instant messaging. Advancement in internet technology boosted the development of these online social networks.

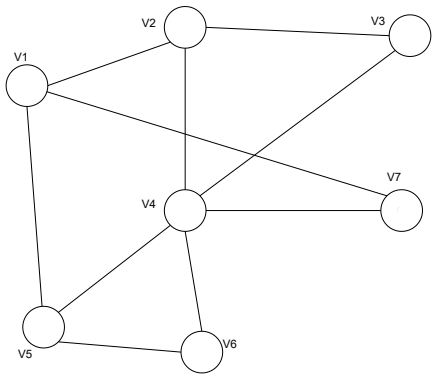
1.2.1 Social Network Representation

Social network representation [20] is based on the concept of classical graph theory. A network of n users and m connections is mathematically represented by $G(V, E)$, where V is the set of users $\{v_1, v_2, \ldots, v_n\}$ and E is the set of links $\{e_1, e_2, \ldots, e_m\}$. The network can be directed or undirected, weighted or unweighted. The weighted network has edge weights which may represent the strength, probability, or frequency of the edge. $N(u)$ denotes the set of neighbour nodes of node u.

Adjacency Matrix: An Adjacency Matrix $A[|V|][|V|]$ is a $2D$ array of size $|V| \times |V|$ where $|V|$ is the number of vertices in a undirected graph. If there is an edge between V_x to V_y then the value of $A[v_x][v_y] = 1$ and $A[v_y][v_x] = 1$, otherwise the value will be zero. For a directed graph, if there is an edge between v_x to v_y, then the value of $A[v_x][v_y] = 1$, otherwise the value will be zero. For a weighted graph, if there is an edge between v_x to v_y then the value of $A[v_x][v_y] = w_{xy}$ and $A[v_y][v_x] = w_{yx}$, otherwise the value will be zero. Here, w_{xy} and w_{yx} are the weights associated with the edges between v_x to v_y, and v_y to v_x, respectively.

Adjacency List: An adjacency list is a collection of unordered lists used to represent a graph. Each unordered list within an adjacency list describes the set of neighbours of a

particular vertex in the graph.



(a) Undirected Graph

Node	Adjacent to
v_1	v_2, v_5, v_7
v_2	v_1, v_3, v_4
v_3	v_2, v_4
v_4	v_2, v_3, v_5, v_6, v_7
v_5	v_1, v_4, v_6
v_6	v_4, v_5
v_7	v_1, v_4

(b) Adjacency List of above undirected graph

	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1	0	1	0	0	1	0	1
v_2	1	0	1	1	0	0	0
v_3	0	1	0	1	0	0	0
v_4	0	1	1	0	1	1	1
v_5	1	0	0	1	0	1	0
v_6	0	0	0	1	1	0	0
v_7	1	0	0	1	0	0	0

(c) Adjacency Matrix of above undirected graph

FIGURE 1.1: Social Network Representation as (a) Undirected Graph, (b) Adjacency List, and (c) Adjacency Matrix of the Undirected Graph

Graph data is stored either using matrices or lists. Matrix representation is suitable for dense networks. Generally, for sparse social networks, list representation is favoured.

Figure 1.1 (a) represents the social network with seven users and ten connections between them. Figure 1.1 (b) and (c) show the list and matrix representation of the given graph, respectively.

1.2.2 Social Networks Properties

Social Networks have a unique structure that often differentiates them from random mathematical networks. Basically, there are three important properties of social networks that make them different from random networks. These properties are:

1. **Small-Word Networks** [16] phenomenon claims that real networks often have very short paths (in terms of the number of hops) between any connected network members (the six handshakes theory [21]). This applies to both the real networks (networks of airports or electric grids) and the virtual social networks (communication networks).

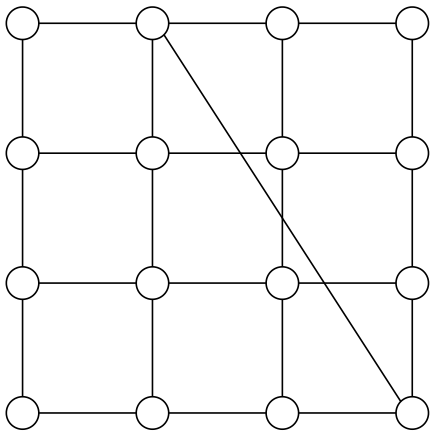
2. **Scale-Free Networks** [22] property tend to have a *power-law degree distribution* [23]. These have a skewed population with a few highly connected nodes (such as social influences) and many loosely connected nodes.

   Power-Law Distribution: A power law is a functional relationship between two quantities, where a relative change in one quantity results in a relatively proportional change in the other quantity, independent of the initial size of those quantities: one quantity varies as a power of another. Mathematically, a quantity x obeys a power law if it is drawn from a probability distribution

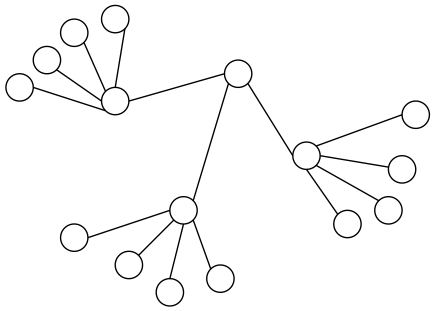
   $$p(x) \propto x^{-\alpha}, \tag{1.1}$$

   where α is a constant parameter of the distribution known as the *exponent* or *scaling parameter*. The scaling parameter typically lies in the range $2 < \alpha < 3$, although there are occasional exceptions [23].

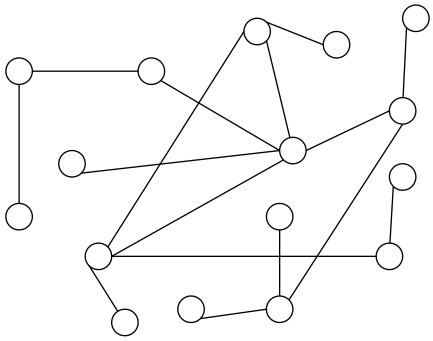
3. **Homophily** [24] is the tendency of individuals to associate and bond with similar people. This results in similar properties among neighbours in the network.



(a) Small-Word Network



(b) Scale Free Network



(c) Random Network

FIGURE 1.2: Real-World Networks (a) Small-World Network, (b) Scale-Free Network, and (c) Random Network

Figure 1.2 (a), (b), and (c) represents the real-world networks showing: small-world network, scale-free network, and random network, respectively. We can use these properties

of social networks for analysis and in developing algorithms.

1.2.3 Social Network Analysis

Social network analysis (SNA) is the process of investigating social structures through networks and graph theory. Social network analysis has emerged as a key technique in modern sociology. It has also gained significant popularity in the field of anthropology, biology, demography, communication studies, economics, geography, history, information science, organizational studies, political science, public health, social psychology, development studies, socio-linguistics, marketing, finance, and computer science and is now commonly available as a consumer tool.

Social network analysis has its theoretical roots in the work of early sociologists such as Georg Simmel and Emile Durkheim, who wrote about the importance of studying patterns of relationships that connect social actors. Social scientists have used the concept of social networks since the early 20^{th} century to study complex sets of relationships between members of social systems at all scales, from interpersonal to international.

1.2.3.1 Measures used in Social Network Analysis

Networks can be characterized by various statistical measures such as nodes' average degree, shortest path length between a pair of nodes, network density, diameter, number of triangles, clustering coefficient, and number of isomorphisms. Popular social network analysis measures include degree centrality, closeness centrality, and betweenness centrality [25]. *Centrality measures* evaluate the impact of a node (i.e., an individual actor) in the network, and they index the representativeness and proximity of a node towards others. Few of the popular centrality measures are defined as follows:

Degree Centrality: It identifies the structural centrality of an actor. It represents the number of links that a node is a part of. Degree centrality is easy to compute since only the local structure around the node is considered. Though simple, it is effective

when evaluating the importance of a node in the network. For example, in several social networks, people with many connections are likely to be more visible and important. The degree centrality of a vertex v, for a given graph $G = (V, E)$ with $|V|$ vertices and $|E|$ edges, is represented by d_v

Closeness Centrality: It considers the global structure of the network. It measures how close a node is to all the others. It computes the average shortest path between two actors in the network. For instance, the more a node is central, the lower its total distance to other nodes is. Mathematically, it can be represented as:

$$C_C(u) = \frac{1}{\sum_v d(v, u)} \tag{1.2}$$

where $d(v, u)$ is the distance between vertices u and v.

Betweenness Centrality: measures the proportion of geodesics, i.e., the shortest paths between two nodes, passing through a particular node in the network. The higher the number of shortest paths, the more is the centrality of the node in this network. Nodes with high betweenness are often called information gatekeepers or brokers. Mathematically, the betweenness centrality of a vertex v, for a given graph $G = (V, E)$ with $|V|$ vertices and $|E|$ edges are defined as:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{1.3}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v.

Centralization: The centralization of any network is a measure of how central its most central node is in relation to how central all the other nodes are. Centralization measures then (a) calculate the sum in differences in centrality between the most central node in a network and all other nodes, and (b) divide this quantity by the theoretically largest such sum of differences in any network of the same size [26]. Thus, every centrality measure

can have its own centralization measure. Mathematically, it can be represented as:

$$C_v = \frac{\sum_{i=1}^{N} C_v(v_*) - C_v(v_i)}{\max \sum_{i=1}^{N} C_v(v_*) - C_v(v_i)}, \tag{1.4}$$

where $C_v(v_i)$ is any centrality measure of vertex v_i and $C_v(v_*)$ is the largest such measure in the network.

1.2.4 Applications of SNA

Some of the applications of social network analysis are in the behaviour Analysis of nodes, Information Diffusion, Community Detection, Social Sharing and Filtering, Recommender Systems, Data Aggregation and Mining, Link Prediction, and Influence Maximization. In this thesis, we consider the task of link prediction and influence maximization in online social networks.

1.3 Dynamic Social Networks

In this thesis, we are addressing the issues associated with *Dynamic Social Networks*. We are also considering the attributes and location information associated with dynamic social networks. We give formal definitions of dynamic social networks and location-aware dynamic attributed networks in the following subsections.

1.3.1 Dynamic Social Networks

Formally, at given timestamp $T = \{1, 2, 3, \ldots, t, \ldots\}$, we define the dynamic social networks as follows:

Dynamic Social Networks: At a particular timestamp t, the corresponding dynamic social network is represented as $G^t = (V^t, E^t)$, where vertices V^t denotes the set of users at time t and $E^t \subseteq (V^t \times V^t)$ denotes the pairs of users having a friendship relationship at t.

These networks evolve with time in terms of the number of users and the relationships among them. In this thesis, we consider the dynamic behaviour of social networks in terms of the addition of relationships among the user with respect to the change of time. We are also using location-aware dynamic social networks, which can be defined as:

1.3.2 Location-aware Dynamic Attributed Networks

Formally, at given timestamp $T = \{1, 2, 3, \ldots, t, \ldots\}$, we define the location-aware dynamic attributed networks as follows:

Location-aware Dynamic Attributed Networks: At a particular timestamp t, the corresponding location-aware dynamic attributed network is represented as $G^t = (V^t, E^t, A^t, L^t)$, where vertices V^t denotes the set of users, $E^t \subseteq (V^t \times V^t)$ denotes the pairs of users having a friendship relationship at t, $A^t = [a^1, a^2, \ldots, a^{n_t}]^t$ denotes the node attributes and $L^t = [l_1, l_2, \ldots, l_{n_t}]^t$ denotes the nodes location check-in information.

In these networks, along with the users and their relationship information, users' attributes and their location check-in information is also given.

1.4 Link Prediction

In the Link Prediction (LP) problem, we need to predict the edges that are expected to be added to the network at a future time t' given the snapshot of the network at time t [27]. Link prediction is a fundamental problem in social network analysis [28] and knowledge graph completion [29]. This has many applications [30], such as friend recommendation in social networks [31], click-through prediction for target marketing [32], academic recommender systems [33, 34], electric grid network [35], finding new connections in protein-protein interaction networks [36], metabolic network reconstruction [37]

and missing link completion in the knowledge graph [38].

Many link prediction methods are based on some heuristics such as Common Neighbours, Jaccard coefficient, Adamic-Adar [31], Preferential Attachment [39], Katz coefficient [40], PageRank [41], SimAttri [42] and their numerous variants. Various data mining and machine learning algorithms have been used to predict the future or missing links in the network with the knowledge of existing links and nodes [43, 32, 44].

Models like probabilistic matrix factorization [45, 46], network embedding based models [47, 48], graph neural network (GNN) models [49], and stochastic block models [50] have also been developed.

Studies indicate that the network structure evolution depends both on the dynamics of the structure as well as the attributes of the nodes [51, 52, 53, 24]. Incorporating the node attributes for link prediction proves to be helpful in achieving better performance in link prediction, especially for the sparse graphs. In evolving networks, as the structure of the network changes with time, the respective attributes of the nodes also change with time [54, 55, 56]; examples of these include modification of posts/comments/reviews, updating educational qualification, job organization, political party, relationship status, and age [57, 58].

In this thesis, we propose models to improve the efficiency and effectiveness of link prediction in dynamic social networks.

1.5 Influence Maximization

Influence Maximization (IM) is a problem of detecting a small set of highly influential users in a social network so that under an explicit propagation model, the spread of influence

is maximum [59]. The set of highly influential users who initiates the influence spread is termed as *seed set* or *seeds*. The *spread of influence/influence spread* is the total number of users reached with a news/message/product/information in the social networks. The process of influence maximization in social networks has applications in various domains such as social media [60], epidemiology [61], viral marketing [62, 63, 64], political campaigning [65], fake news containment [66], and many more.

Several models have been proposed during the last two decades to formulate the information diffusion process. Independent Cascade (IC) model and Linear Threshold (LT) model are the two elementary diffusion models proposed by David Kempe et al. [59]. In the *IC model*, a node has a probability of convincing each of its neighbours. And in the *LT model,* a node accepts a new idea if the influence from all its neighbours has crossed a threshold. Here, the probability of convincing a neighbour node v by a node u is known as **propagation probability** and it is denoted as P_{uv}. Basically, all the diffusion models use this propagation probability concept to evaluate the diffusion in the networks. Most of the traditional techniques use a uniform distribution model; however, few of them uses edge weights evaluated by considering various properties/information of the networks.

Initially, David Kempe et al. in [59] formally defined the *IM problem* as the problem of finding the seed set S of size k to maximize the influence spread in the network using the IC/LT diffusion model. They also proved that the IM problem is *NP-hard*, and the corresponding objective function is monotone and submodular. They also proposed a hill-climbing greedy algorithm to solve the IM problem, which is quite close to the optimal solution.

The Influence Maximization algorithms can be categorised as centrality based [67, 68, 69, 70], sub-modularity based [59, 71, 72, 73], or influence path based [74, 75, 76, 77]. These approaches are significantly faster than the traditional greedy approach. Recently, various algorithms introduced under a new category of context/topic-aware Influence Maximization techniques [78, 79, 80, 81, 82, 83, 84] have gained popularity. These methods improve

the quality of seed set as compared with traditional structure-based approaches.

In this thesis, we consider the problem of influence maximization in online social networks and propose efficient novel techniques to solve them.

1.6 Major Challenges

There exist many challenges and issues in Link Prediction and Influence Maximization in Online Social Networks. Some of them are listed as:

1.6.1 Challenges in Link Prediction and Influence Maximization

1. **Incompletion:** Almost all obtained social network data is incomplete since only part of social information can be collected from social network platforms.
2. **Non-linearity:** The social network data is highly non-linear in terms of users and the relationship between the users.
3. **Dynamic:** Social networks are highly dynamic, which might lead the nodes and edges to appear or disappear in the future.
4. **Ill behaved nodes:** In social networks, few nodes might be very unpredictable, i.e., their behaviour differs from the usual nodes. Predicting the behaviour of such nodes is a challenge in the social network.
5. **Large size of the network:** In social networks analysis, the size of the networks is always an issue. It affects the scalability and complexities of the algorithms involved.
6. **Feature extraction:** It is difficult to understand the factors responsible for link prediction in dynamic social networks. This is due to the non-linear behaviour of the networks. Thus, extracting the appropriate features responsible for link prediction is still a challenge.

7. **Ignorance of Context:** Most of the link prediction and influence maximization techniques uses the structure of the network and ignore the contexts and users' interest. Considering all these factors collectively is still a challenge.

8. **Running Time:** Designing efficient algorithms for link prediction and influence maximization is still a challenge.

In this thesis, we consider the challenges such as dynamics of the network, the large size of the networks, ignorance of the contexts and time efficiency of link prediction and influence maximization models. We address these challenges and propose models to improve the efficiency and effectiveness of link prediction and influence maximization in online social networks.

1.7 Motivation and Scope of the Thesis

In recent decades, Internet technology has grown significantly, and the digital world has affected human activities and lifestyle in various aspects. Communication, the spread of news, trade, propagation of information have taken new forms [85]. Different forms of social network platforms have become popular in society. Users communicate, share data, or exchange views in these networks. A large number of users and rapid exchange of information among them has made social networks a powerful means for information spreading. Various companies use these networks for the recommending and marketing [86] of their products. However, it is impractical for them to connect with each user individually. We can also observe that the users often decide to purchase a product on the advice of their friends [87]. So, social network analysis is a potential field of research due to its power of connection between the users. We have chosen this area of research as there exists lots of possibility of developing new solutions which improve the state-of-the-art and have many applications like online advertisements where these solutions can be deployed.

Motivated by various social behaviour, we have added mobility, popularity, and similar interests of the nodes as additional factors along with the structural attributes to predict

the network evolution pattern. We have then used it for link prediction and influence maximization tasks in the evolving social networks.

1.8 Contributions of the Thesis

This thesis contributes algorithms, tools, models, and new insights to problems that arise in the area of graph mining for Link Prediction and Influence Maximization in online social networks. The main contributions of the thesis can be divided into three parts which are as follows:

1.8.1 Objective 1: Link Prediction based Influence Maximization in Online Social Networks

Here we define a novel Influential Node Tracking problem to maximize the influence spread in an online social network. We propose a framework for efficient and effective influence maximization in dynamic social networks. The proposed model uses the link prediction technique to predict the upcoming snapshot of the graph and then computes the seed set for influence maximization. We experimentally show that the proposed framework performs better in terms of influence spread in comparison to the considered baseline techniques on many real-world datasets.

1.8.2 Objective 2: Multifeature Analysis based Link Prediction in Dynamic Social Networks

Here, we formally define the problem of link prediction in location-aware dynamic attributed social networks. To the best of our knowledge, this is the first attempt to solve the link prediction problem by considering multiple key factors responsible for predicting new links in advance for the upcoming snapshot of the networks. We propose a novel framework for solving the link prediction problem using the Hidden Naive Bayesian (HNB)-based link prediction model employing users' relationships and behaviour patterns derived

from attributes, geographical location, popularity, and interests of the users. We experimentally evaluate the effectiveness and efficiency of our proposed framework on various real-world location-aware dynamic attributed networks for link prediction.

1.8.3 Objective 3: Context-aware Influential Nodes Tracking in Online Social Networks

In this, we formally formulate the problem of topic-aware influence maximization in dynamic social networks where the influence spread depends on multiple features, and seed nodes are discovered according to the topic of interest of users and message/product. Here, we propose a novel multi-feature-based diffusion model named CIC, which is a modified version of the traditional IC model. The proposed model considers the similarity of topic-of-interest between users and also between users and messages. It also considers the popularity and location information of the users to perform the diffusion process. Further, we propose a novel topic-aware influence maximization algorithm based on the CIC model named MINT algorithm for topic-aware seed set selection. We have experimentally evaluated the effectiveness and efficiency of our proposed framework on six real-world network datasets. Experiments show that the proposed MINT algorithm performs better in comparison to the considered baseline algorithms in terms of both influence spread and time.

1.9 Organization of the Thesis

The thesis is organised into seven chapters. Figure 1.3 provides an overview of the thesis and the connection among the chapters.

Chapter 2 presents a detailed literature survey on the state-of-the-art on link prediction and influence maximization in online social networks and their usage in different applications.

Chapter 3 centres around our first objective of the thesis, i.e., link prediction-based influence maximization in online social networks.

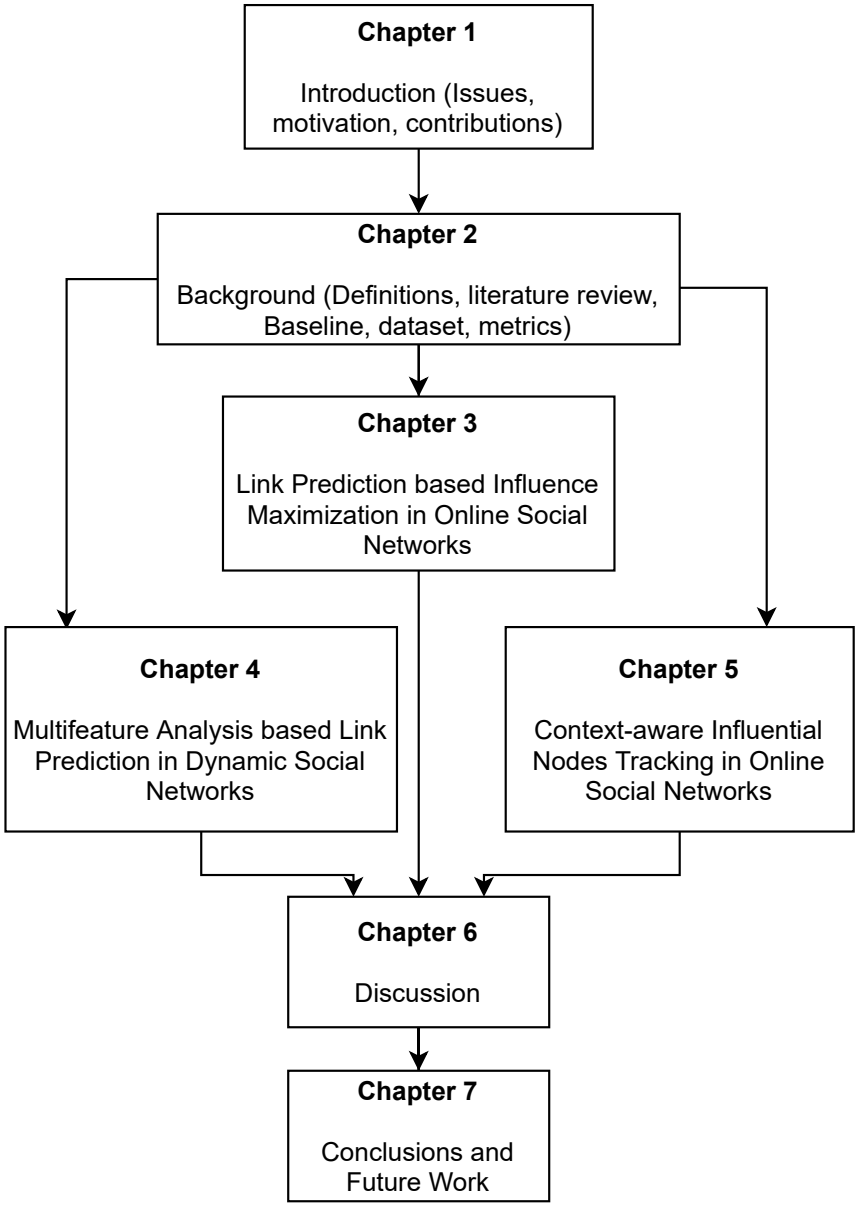


FIGURE 1.3: Structure of the Thesis

Chapter 4 investigates in detail our second objective of the thesis, i.e., popularity, interests, location used hidden naive bayesian-based model for link prediction in dynamic social networks.

Chapter 5 explains our third objective of the thesis i.e., context-aware influential nodes tracking in online social networks.

Chapter 6 summarizes the entire thesis work highlighting the key insights obtained along

with discussions and limitations.

Chapter 7 concludes the thesis by summarizing the contributions and pointing to a few topics for future research that have arisen up from this work.