Chapter 1

Introduction

Many real-world systems are described as networks, with different system components acting as nodes and interactions between them as connections. Examples include social networks like acquaintance networks [4], professional networks [5], and collaboration [6], biological networks like protein-interaction networks [7], gene networks [8], and metabolic networks [9], and ecological networks like supply-chain networks [10]. Network analysis is an effective tool that provides us with a useful framework to explain phenomena connected to social, technological, and many other complex systems in the real world. In order to better explore and comprehend the properties of real-world networks, the study of complex networks [11] has attracted significant attention from a number of research groups as computer capability has increased. Most commonly, the term "complex networks" refers to actual networks, which are frequently defined by a number of characteristics, including a high number of interacting entities, heterogeneity, evolution, self-organization, etc. With intricate structural relationships between them, these networks contain several elements. According to published research, complex networks exhibit some fundamental characteristics, such as a power law degree distribution, community structure, a high clustering coefficient, short average path lengths, and the presence of motifs.

In order to better explore and comprehend the properties of real-world networks , the study of complex networks [11] has attracted significant attention from a number of research groups as computer capability has increased. Most commonly, the term "complex networks" refers to actual networks, which are frequently defined by a number of characteristics, including a high number of interacting entities, heterogeneity, evolution, self-organization, etc. With intricate structural relationships between them, these networks contain several elements. The Internet, social networks, biological networks, and other real-world instances of complex networks are some examples. Complex networks, according to the literature, exhibit some fundamental characteristics like a power-law degree distribution [12], community structure [13], a high clustering coefficient [14, 15], short average path lengths [14], and the presence of motifs [16]. These characteristics have inspired the use of complex networks to address a number of real-world issues, including understanding the spread of infectious diseases (also known as epidemics) [17], identifying functional groups in metabolic networks [18], fault detection grid networks [19], link prediction in various networks [20], etc.

A dynamic network permits a more precise depiction of complex interactions with temporal attributes than a static network. Dynamic network analysis is used in several disciplines. To predict linkages in dynamic networks, a number of techniques have been presented in the literature [1, 21, 22]. The link prediction problem identifies linkages that will change over the following time slot t_{n+1} for a given series of network snapshots captured at distinct time intervals t_1, t_2,... t_n. Let $G = (V, E)$ be a dynamic network, with V representing the set of vertices and each edge $(u, v) \in E$ representing the relation or link between nodes u and v. Let G_1,G_2,G_3.....G_n be the networks at various snapshots t_1, t_2, t_3,....., t_n, and we must predict G_{t+1} at $t+1$ time. To predict links in dynamic networks, numerous approaches have recently been proposed [22–25].

1.1 Link prediction

A common method of simulating communication in a group or community of people is the use of social networks, of which complex networks are a more general variant. Such networks can be seen as a graphical model where each node represents to a person or other social entity and each link corresponds to an association or collaboration between corresponding persons or social entities. The addition and/or deletion of numerous links and vertices occur as a result of the ongoing changes in the relationships between people. Social networks become more complex and dynamic as a result. When we analyze a social network, various problems come up, some of which are the dynamics of associations, the causes of associations, and the impacts of associations on other nodes. Here, we deal with a specific problem known as link prediction.

Link prediction is defined as follows informally. Consider a simple undirected network $G(V,E)$, where V indicates a vertex-set and E a link-set (Refer to the Figure 1.1). Parallel links and self-loops are not allowed because the dissertation only considers simple graphs. We use (vertex $\equiv$ node), (link $\equiv$ edge) and (graph $\equiv$ network) interchangeably. In the graph, a universal set U contains a total of $\frac{n(n-1)}{2}$ links (total node-pairs), where $n = |V|$ represents the number of total vertices of the graph. $(|U| - |E|)$ [1] number of links are termed as the non-existing links, and some of these links may appear in the near future. Finding such missing links (i.e., $1-2$, $2-7$, and $5-6$) is the aim of link prediction [26].

Formally, Liben-Nowell et al. [20] defined the link prediction problem as: suppose a graph $G_{t_0-t_1}(V,E)$ represents a snapshot of a network during time interval $[t_0,t_1]$ and $E_{t_0-t_1}$, a set of links present in that snapshot. The task of link prediction is to find set of links $E_{t_0^{'}-t_1^{'}}$ during the time interval $[t_0^{'},t_1^{'}]$ where $[t_0,t_1] \leq [t_0^{'},t_1^{'}]$.

Link prediction has many applications in various domains. It includes automatic hyperlink creation [27], website hyper-link prediction [28] in the Internet and web science domain, as well as friend recommendation on online social networks such as

[1] Existing links=$|E| = m$

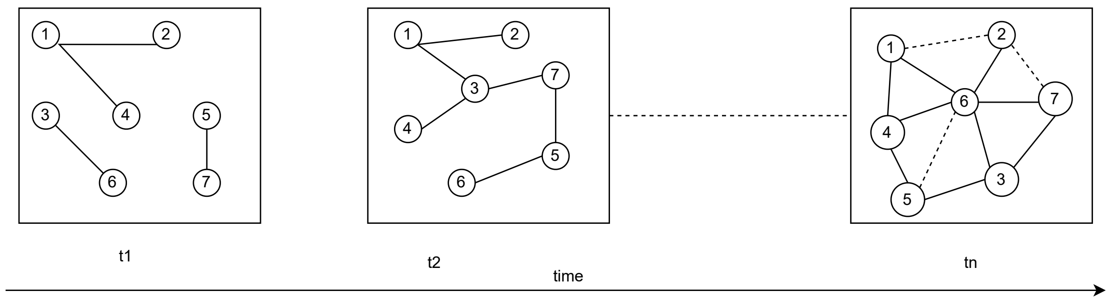

FIGURE 1.1: The Link Prediction (LP) finds missing links (i.e., $1-2$, $2-7$, and $5-6$) in this observed network.

Facebook and Instagram. Building a recommendation system [29, 30] in e-commerce websites that make new item suggestions to users is an essential task that uses link prediction. In the bio-informatics domain, protein-protein interactions (PPI) also have been analyzed using link prediction [31]. In security related areas, link prediction can be applied to distinguish hidden links among terrorists and their organizations [32].

1.2 Challenges in Link Prediction

The link prediction problem faces three major challenges:

- The nodes contain some additional covariates in addition to topological information that is useful to predict links for sparsely connected nodes. However, covariate and topology of graphs encode different variants of information. Therefore, combining both the information improves the overall model performance [1, 21].
- Due to the imbalance characteristics of link prediction datasets, the known number of present edges is smaller than the known number of absent edges. As a result, for link prediction tasks, the area under the ROC curve (AUC) has become the de facto performance metric [33, 34].

- For the dense graph, models are computationally efficient. Such graphs can be seen in a variety of real-world applications. Facebook, friend-recommendation in social networks are some examples of such graphs [21, 35].

1.3 Motivation of the thesis

This thesis is focused on four objectives which are discussed below -

- **Studying the effects of combination of different local, global and quasi-local similarity-based features on link prediction in dynamic networks.**

  Similarity-based link prediction can be classified into three broad categories based on the information used for index calculation [1, 21]. These are local information which is derived from the immediate neighborhood of nodes, global information, which is estimated using the whole graph structure, and quasi-local information, which attempts to combine attributes of both local and global information. Different similarity methods have their own advantages and disadvantages. Compared to the similarity-based approaches, machine learning techniques have been used to improve prediction accuracy as they are more useful for finding changes in patterns over time which is a basic requirement for dynamic graphs. It is our belief that all three types of methods, local similarity-based indices, global similarity-based indices, and quasi-local similarity-based indices, can be combined to create a feature set that would be more useful for link prediction than its components individually. This feature set would be used for link probability estimation using different machine learning algorithms.

- **Studying the effect of clustering-based similarity features in providing more feature rich information for link prediction in dynamic graphs.**

  Common neighborhood-based feature estimation is used by many local and quasi-local similarity-based link prediction algorithms. Effectively these measures count the number of available paths between nodes for a fixed hop neighborhood. But the relative contribution of individual nodes in graphs either remains the same or is otherwise modeled as a resource allocation problem based on node degree. Clusters in graphs provide a more effective manner of quantifying the relative stability that a node provides for possible information passage. The smallest clusters in graphs are triangles, and they have been widely used for the formulation of link prediction indexes in static networks [36–39]. It is our belief that on combining such features with other similarity-based link prediction features, machine learning-based link probability estimation can be enhanced.

- **Contrasting the effect of different individual link prediction features with each other as well as on the combined feature set using feature selection**

  Different link prediction methods use different information for the calculation of node pair similarity scores. In our research, we study the performance of these individual features on different datasets to better understand which features work best on which dataset. When these individual features are combined to create a feature set with rich information to improve the link prediction accuracy, we also use feature selection to optimize the groups of individual features which should be used. The information gleaned from this feature selection step results in a reduced feature set, decreasing overall computational complexity. It also helps prevent redundancy such that similar kinds of individual features are not used in a grouping along with performance improvement.

- **Enhancing link prediction in dynamic networks using snapshot-impartial features which process information on multiple snapshots into one identifying feature.**

  The link prediction issue in dynamic networks aims to identify future network linkages based on the relative behavior of previous network updates. To create feature sets for machine learning-based link estimation, different categories of similarity-based link prediction methods can be used as a feature. The basic limitation of such a method is that the task of estimating temporal changes to edges is left entirely to the machine learning algorithm. It is our belief that feature-based solutions that consider both individual snapshots and the overall network throughout the full-time span (called snapshot-impartial features) would lead to better link prediction accuracy.

- **Understanding the effect of combination of different community detection and their correlated link estimations with other similarity-based features for link prediction on dynamic graphs.**

  One of the well-known classes of methods in link prediction is quasi-local similarity-based methods, which use both local and global topological information of the network to predict missing links. Such methods attempt to achieve a trade-off between computational complexity and accuracy to consider both local neighborhoods as well as overall graph structure for link prediction. The task of community detection can also be generalized as a method of considering both local and global information to create such communities in which intra-community differences between nodes are minimized, and the inter-community difference is maximized. Using these communities, it is our belief that link estimation features can be defined which can enhance link prediction in dynamic networks. Some research that uses communities for link prediction in static and multiplex networks has already been done by Singh et al. [40], and Karimi et al. [41].

- **Studying the effect of quantum kernel-based feature transformation for link prediction in dynamic networks.**

  Quantum systems are distinguished by a generalization of probability theory that allows for unique phenomena such as superposition and entanglement that are impossible to simulate with a standard computer [42]. Quantum computers can perform rapid linear algebra on a state space that expands exponentially with the number of “qubits”. This is one of the main breakthroughs that has led to their use in machine learning. Hsin-Yuan Huang et al. [3] have shown that quantum machine learning outperforms classical machine learning in some cases. Quantum models have a mathematical framework that is quite similar to kernel methods, i.e., they evaluate data in high-dimensional Hilbert spaces to which we can only acquire access via inner products disclosed by measurements [43]. Kernel-based methods have been employed with good results in link prediction recently [44–46]. By projecting back from the quantum space to a classical one in the projected quantum kernel model [3], it is our belief that the underlying patterns in data can be enhanced.

1.4 Contribution of the thesis

Major contributions of the thesis[2] are point-wise discussed below.

1.4.1 Feature fusion based link prediction in dynamic network

This work comprehensively explores and analyses various link prediction methods grouped into several categories. One of the well-known methods for link prediction is the similarity-based method, which uses a similarity-based score. The three widely used similarity-based indices are Local (L), Global (G), and Quasi-local (Q) for calculating similarity scores. In this work a new LGQ model for link prediction is proposed and these wide categories of indices are used in different combinations (L, G, Q, LG, LQ, GQ, LGQ) for feature set generation that can be used with various machine learning techniques for link prediction. In the snapshot-based paradigm for link prediction on dynamic graphs, feature sets are used to track the behavioral changes in edges. These features are then used to make predictions for non-existing edges. Machine learning algorithms are used on these feature sets to build prediction models [47, 48]. Compared to the similarity-based approaches, machine learning techniques have been used to

[2] Thesis and Dissertation are interchangeably used

improve prediction accuracy as they are more useful for finding changes in patterns over time. By literature survey on link prediction [1, 21, 35, 49], different similarity methods have their own advantages and disadvantages. Therefore, all three types of methods, Local similarity indices, Global similarity indices, and Quasi-local similarity indices, can be combined to create a feature set that we believe would be more useful for link prediction than its components individually. This work proposes a novel model for building feature vectors for machine learning model-based link prediction that combines the properties of three types of similarity indices. Apart from the individual groups, in order to maximize the information represented by feature sets, we have created and tested different combinations of these groups with each other. This is done in order to create a feature set that represents all the relevant properties of the graph, which may be useful for link prediction and to test the performance of these variations with different machine learning models. The combination of groups is inspired by the methodology of unique quasi-local indices, which aim to highlight both local and global features of the graph. As far as we know, this combination approach has not been attempted for feature sets generation using dynamic graphs. After creating these feature sets, for the task of link prediction, we have also compared two different learning models for predicting actual edge probabilities.

1.4.2 Path Weight Aggregation Feature for link prediction

This work explore a novel feature, Path Weight Aggregation Feature (PWAF) to address link prediction problem in dynamic networks. It is a new feature, based on ranking multi edge occurrences across the entire network. This work uses different topological aspects of the networks (Local, Global, and Quasi-local) as well as Clustering Coefficient based features are taken into consideration for feature generation, in addition to the suggested Path Weight-Based Aggregation Feature (PWAF). In this work clustering coefficient of Level-1 [36] and Level-2 [37] common neighbor of the seed node pair for more information for better accuracy is calculated. These clustering approaches provide more information about the nodes and edges, improving accuracy. Some researchers have recently sought to combine more information about network features and have achieved high accuracy in some networks [1, 35, 50]. The proposed approach employs feature vectors of node pairs generated from all snapshots to incorporate several types of structural information (topological based which includes Local, Global and Quasi-local similarity, clustering coefficient based algorithm, and PWAF-based). For classification, supervised machine learning (ML) can be used to address the link prediction problem.

1.4.3 A new cost based feature for link prediction

This work proposes a novel Cost-based Feature for Link Prediction (CFLP) to solve the link prediction challenge in dynamic networks, which uses all the information from previous snapshots instead of the individual snapshot. This work provides a link prediction framework that uses diverse topological and clustering data to improve link prediction. This feature-based solution to the link prediction problem uses a reward and penalty structure to summarize node activity across the entire network. In this work, similarity indices are classified into four major categories: local similarity, global similarity, quasi-local similarity, and clustering coefficient-based similarity to measure edge activity in individual snapshots. The relative influence of characteristics on one another and the overall link prediction problem should be accurately quantified using regression and mutual information-based scoring for feature selection. By using similarity indices with high scores, the feature selection approach optimizes the features.

1.4.4 Community Enhanced Link Prediction

The objective of this work to enhance link prediction accuracy by a community information-based feature estimation and link prediction method. This work proposes a community enhanced framework to predict missing links on dynamic social networks. First, a link prediction framework is presented to predict missing links using parameterized influence regions of nodes and their contribution in community partitions. Then, a unique feature set is generated using local, global, and quasi-local similarity-based as well as community information-based features. This feature set is further optimized using scoring-based feature selection methods to select only the most relevant features.

1.4.5 Projected Quantum Kernel based Link prediction

The objective of this work is to explore a quantum-enhanced feature-based framework for solving link prediction problems in dynamic networks. It combines the disciplines of social networks and quantum computing. It employed high-dimensional Hilbert spaces to enhance the prediction data in this model. This paper formalizes the use of QML techniques to solve the problem of link prediction in social networks. Using the Projected Quantum Kernel and machine learning models, this work proposes a novel approach, PQKLP, for solving the link prediction problem by employing both local and global information. Using PQK, it transformed the popularly used snapshot-based feature set form into quantum space such that the accuracy of machine learning-based link prediction can be enhanced.

1.5 Organization of the thesis

This thesis is organized into eight chapters.

Chapter 2 provides literature survey on dynamic networks and link prediction approaches in dynamic networks.

Chapter 3 deals with feature fusion based link prediction in dynamic networks. It provides a rich feature set for link prediction by combining Local, Global and Quasi-local similarity-based indices in dynamic networks.

Chapter 4 deals with a Path Weight Aggregation Feature (PWAF), which is a novel feature, based on ranking multi edge occurrences across the entire network. Different topological aspects of the networks (Local, Global, and Quasi - local) as well as Clustering Coefficient based features are taken into consideration for feature generation, in addition to the suggested Path Weight-Based Aggregation Feature (PWAF).

Chapter 5 discuss a novel feature called Cost-based feature for link prediction (CFLP) for estimating edge behavior throughout the entire network, which uses a reward and penalty structure to summarize node activity across the entire network. Incorporating local, global, quasi-local, and clustering coefficient-based similarity, to measure edge activity in individual snapshots. It has used regression and mutual information-based scoring for feature selection to correctly quantify the relative effect of features among themselves and the overall link prediction problem.

Chapter 6 studies Community Enhanced Link Prediction in Dynamic Networks. A community enhanced framework is proposed by incorporating a link prediction framework presented to predict missing links using parameterized influence regions of nodes and their contribution in community partitions.

Chapter 7 investigates quantum-enhanced feature-based framework for solving link prediction problems in dynamic networks. Projected Quantum Kernel (PQK) approach is used to enhance the feature set. It employs high-dimensional Hilbert spaces to enhance the prediction data in this model.

Chapter 8 concludes the contributions and details possible future directions in respect of each of the proposed works.