Chapter 1

Introduction

This introductory chapter overviews the research undertaken, summarizes the academic contributions of the dissertation and describes the structure of the discussion.

1.1 Complex networks

Complex networks are networks which are more complex than classical random graphs [6, 7]. Here the term 'complex' means more complex organization (more complex distribution of connection). In particular, a degree distribution may be more complex than Poisson, and/or various correlations may be essential. Real-world nets (Social networks, biological networks, etc.) are complex networks, usually with fat-tailed degree distributions, usually with strong correlations of degrees of connected vertices, usually with an essential role of loops.

With the advancement of computational capability, the study of complex networks [8] has gained substantial attention from several groups of researchers in exploring and understanding the characteristics of real-world networks. Complex networks mostly refers to real networks which are often characterized by several attributes such as large

number of interacting entities, heterogeneity, evolution, self-organization, etc. These networks contain large number of entities with complex structural relationship among them. Some examples of complex real networks are Internet, social networks, biological networks, etc. Literature suggests that complex networks show some basic properties like power law degree distribution [9], community structure [10], high clustering coefficient [11, 12], low average path lengths [11], existence of motifs [13] etc. These properties have motivated to apply complex network to solve several real problems like, understanding the spreading phenomenon of infectious diseases (i.e. epidemic) [14], identifying functional groups in metabolic networks [15], fault detection grid networks [16], link prediction in different networks [17], etc.

1.2 Link prediction

A social network (a more general form is a complex network) is a standard approach to model communication in a group or community of persons. Such networks can be represented as a graphical model in which a node maps to a person or social entity, and a link corresponds to an association or collaboration between corresponding persons or social entities. The relationships among individuals are continuously changing, so the addition and/or deletion of several links and vertices take place. It results in social networks to be highly dynamic and complex. Lots of issues arise when we study a social network, some of which are changing association patterns over time, factors that drive those associations, and the effects of those associations to other nodes. Here, we address a specific problem termed as link prediction. Informally, link prediction is characterized as follows. Consider a simple undirected network $G(V,E)$ (Refer to the Figure 1.1), where V characterizes a vertex-set and E, the link-set. A simple graph is considered throughout the dissertation, i.e., parallel links and self-loops are not permitted. 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 $n = |V|$

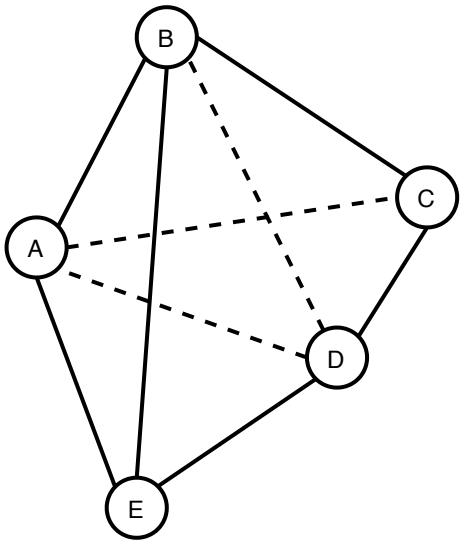

FIGURE 1.1: The Link Prediction (LP) finds missing links (i.e., AD, AC, and BD) in this observed network.

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., AC, BD, and AD) is the aim of link prediction [18].

Formally, Liben-Nowell et al. [17] 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 $[t_0,t_1] \leq [t_0',t_1']$. The link prediction idea is useful in several domains of application. Examples include automatic hyperlink creation [19], website hyper-link prediction [20] in the Internet and web science domain, and friend recommendation on Facebook. Building a recommendation system [21, 22] in e-commerce is an essential task that uses link prediction as a basic building block. In Bio-informatics, protein-protein interactions (PPI) also have been implemented using link prediction [23]. In security concern areas, link prediction is used to distinguish hidden links among terrorists and their organizations.

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

1.3 Evaluation criteria

The objective of this research is not only limited to present and discuss several approaches of link prediction available in the literature but also it explores some new techniques to maximize the accuracy. Earlier studies present that, in general, no approach is better for every types of networks i.e., there is a lack of systematic research that would enable to reveal the reason why the methods are good predictors when it comes to some of the networks but very bad when other networks are considered. Lichtenwalter and Chawla [24] presented some guidelines for fair and effective evaluation of link prediction. They suggest to use precision-recall curve, ROC curve with area under these curve as optional and avoid fixed thresholds. This problem is also tackled to some extent by Gao et al. [25] where they tried to a solve this issue by presenting a correlation between several network metrics and different approaches. In this thesis, we have used several evaluation measures mostly focused on threshold curve that are effective for imbalanced datasets.

1.4 Motivation of the thesis

Motivation of the thesis lies in the homophily principle (i.e., Similarity breeds connection) in networks, a seminal work entitled "Birds of a Feather: Homophily in Social Networks" published by Miller McPherson [26]. Homophily is commonly used term by sociologists to refer similarity between two social entities. Based on similarity, relationships build among these entities and building such connection refers to link prediction. Probability of forming new relationships (links) is positively correlated with similarity between pair of nodes and hence follow homophily. Similarity between a pair of nodes can be computed based on several topological or structural features, node features, etc. Exploiting these features from networks are prevalent in the literature and building blocks of several similarity based link prediction methods [17, 27].

Studies reveal many real networks (especially, social networks) show some topological properties (scale-free [9], clustering [11, 12], and small-world [28, 29]) that are consistent across different domains. This dissertation attempts to show how measures of these features are useful in different network evolution and how they contribute to link prediction. It shows how higher level clustering and path information enhance the prediction accuracy. Link prediction is a mathematical and statistical technique that can be applied to many business, social and software problems. Link prediction has many applications. Some examples include:

- Predicting links among members of a terrorist network or identifying the structure of a criminal network [30].
- Overcoming the data-sparsity problem in recommender systems using collaborative filtering [31].
- Predicting which web pages users will next visit in order to improve the efficiency and effectiveness of a site's navigation [32].
- Monitoring and controlling computer viruses that use email as a source [33].
- Improving hypertext analysis for information retrieval and search engines [34].

1.5 Contribution of the thesis

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

1.5.1 A comprehensive experimental survey of link prediction

This work comprehensively explores link prediction methods grouped in to several categories. Similarity based methods are presented in three categories namely, local,

[2]Thesis and Dissertation are interchangeably used

global and quasi local methods. Similarity scores of these methods are calculated using structural information of underlying network. Local and quasi local methods are feasible for large networks as they extract somehow only local structural information. Most global methods are not feasible for large networks. Next, probabilistic and maximum likelihood methods are presented where hierarchical random graph (HRG), stochastic block models (SBM), local probabilistic model, probabilistic relational model (PRM), and exponential random graph (ERG) are explored comprehensively. Further, dimensional reduction based methods are presented in one group with some more methods in another group namely, "Other approaches". Similarity based methods and other representative methods are experimentally analyzed in terms of accuracy and efficiency. Some recent works are explored under deep learning and fuzzy groups. Finally, different variations of link prediction problems with their applications are explored to complete the survey.

1.5.2 Influence of higher level clustering features towards link prediction

This work proposes a new similarity index that explores clustering information of nodes in the networks. It extends the existing work [35] based on node clustering coefficient to higher level. It defines the notion of level-2 common node and its corresponding clustering coefficient that extracts clustering information of level-2 common neighbors of the seed node pair and computes the similarity score based on this information. The proposed index experimentally evaluated on several network datasets against some existing classical and some recent methods. Computational complexity of the proposed index is also given. Statistical analysis has also been performed to show its difference with the other methods.

1.5.3 Influence of higher order paths indices towards link prediction

This work proposes a new similarity method that explores path of different lengths to compute similarity score of a node pair. It searches for paths up to length 6 (due to six degree of separation). Similarity score calculation is based on the idea of resource allocation process where the amount of information received by the receiver node derives the similarity score between the sender node and the receiver node. The proposed work maximizes this information by minimizing the information leaks through their neighbors in the path. An iterative procedure to calculate this score is also proposed in the work with its computational complexity. Prediction accuracy is computed in terms of different metrics against several datasets of diverse areas. This method is empirically evaluated and compared with several existing methods in the literature. Further, effects of different path lengths to prediction accuracy are also presented. Also, effects of penalization factor towards longer paths for different path based methods are presented.

1.6 Organization of the thesis

This thesis is organized as follows. **Chapter** 2 presents brief overview social network and link prediction. It includes preliminary concepts related to these two terms. **Chapter** 3 is focused on extensive and experimental survey of link prediction. **Chapter** 4 exploit clustering information of nodes as discriminating features and extended higher level clustering coefficient employed to perform link prediction in networks. **Chapter** 5 discuss path and degree as features in link prediction framework. Finally, **Chapter** 6 draws conclusion of the work done with some future directions.