Chapter 1

Introduction

The 21^{st} century might well be considered the time when the internet revolution swept the globe. The internet has transformed every aspect of human contact, from introductory emails to complex financial transactions. Social networks are among the most significant new internet applications, having an impact on billions of individuals across the globe. These technological behemoths have given many chances to academia to investigate human interaction dynamics across many media (text, image, audio, video, etc.). Graphs with participants as nodes and their interactions as edges may depict these intricate networks. How new linkages emerge and what causes them to form are two key issues that academics are particularly interested in. The link prediction issue may be interpreted specifically regarding Facebook as the issue with new friend suggestions based on shared friends and interests. The extensive research in the field of link prediction has been thoroughly reviewed by several in-depth investigations. One of the most recent surveys was carried out by Ajay et al. [2]. Several studies have also been focused on the applicability of link prediction in various fields [3–8].

Link prediction is a basic problem in network research since it seeks to estimate missing or unseen linkages in a network based on its present state (observed links). This gives a theoretical basis for evaluating network modelling hypotheses that aim to explain the genesis and evolution of a particular network. Practically, link prediction can be used in various fields (biological, citation, etc.) to predict new interactions which are most likely to occur. These interactions can be of any type, such as drug-target [9] and protein-protein [6] interactions, as well as friend and product recommendations [10]. Recently, link prediction in dynamic networks has also seen increased relevance in the field of research where multiple features and machine learning algorithms have been used [11]. Liben-Nowell and Kleinberg [12] characterized the first challenge of link prediction in the present setting as the inference of new or unknown interactions between entities. This may be accomplished utilizing either the entities themselves or their historical interactions. Various approaches for link prediction have been developed, which may be roughly categorized as similarity-based [12], learning-based [13], probabilistic and maximum likelihood-based [14, 15], etc. In techniques using similarity-based algorithms, the likelihood of edges between node pairs is calculated based on the network's structural features. These attributes may be taken straight from the network, simplifying the calculation of similarity scores. The attributes may also be classified depending on the neighborhood that is considered while calculating them, i.e., they can be local, which requires information extracted from the immediate neighborhood of the nodes themselves, or they can be global such that the entire structure of the network is processed to extract relevant information. Some local information based similarity scores (indices) are Common Neighbors [16], Adamic/Adar [17], Resource Allocation [18], Preferential Attachment [16], Jaccard [19], CAR-based Common Neighbor Index (CAR) [20], and Local Naive Bayes-based Common Neighbors (LNBCN) [21]. Katz index [22], Rooted PageRank [23], SimRank [24], and others are global similarity indices that are calculated based on the complete physical structure of the network. Quasi-local similarity indices are based on the balance between local and global information, and they attempt to combine the most significant aspects of both methodologies. Some examples include , Node and Link Clustering coefficient (NLC) [25], Local Path [26] and L3 [27].

1.1 Link prediction

A standard method of simulating communication in a group or community is social networks, of which complex networks are a more generic variant. These networks may be seen as a graphical model where each node relates to a person or other social entity and connects to an affiliation or cooperative effort between the associated nodes or social entities. The addition and deletion of many connections and vertices occur as a result of the ongoing changes in the relationships between people. Social networks, as a consequence, become very dynamic and complicated. When studying a social network, various problems arise, including shifting connection patterns over time, the causes of those links, and the impact of those linkages on other nodes. Here in this thesis, link prediction is the issue under focus.

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. In this thesis, (vertex $\equiv$ node), (link $\equiv$ edge) and (graph $\equiv$ network) are used 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., AC, BD, and AD) is the aim of link prediction.

Formally, Liben-Nowell et al. [12] 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 $[t_0,t_1] \leq [t_0^{'},t_1^{'}]$.

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

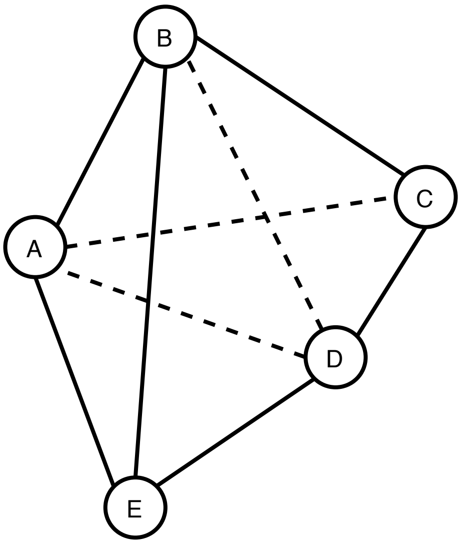

FIGURE 1.1: The Link Prediction (LP) finds missing links (i.e., AD, AC, and BD) in this observed network.

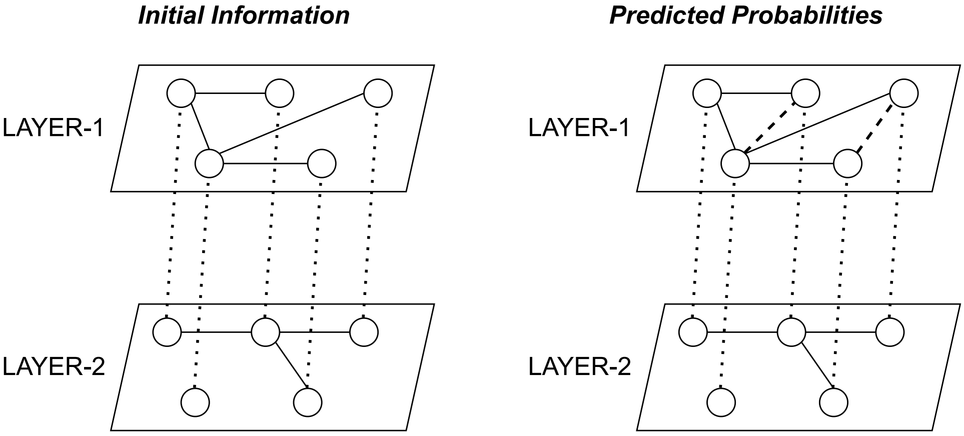

FIGURE 1.2: Link Prediction in Multiplex network (dashed line in Predictions for Layer-1)

1.1.1 Problem Definition in Multiplex Networks

In multiplex networks, nodes can have multiple types of relationships (links) encoded into different layers such that each layer represents a single type of link. Even though the nature of links in different layers may differ, the nodes remain the same, and so do their underlying relations among themselves. From the general definition of link prediction in static networks, the problem can be inferred as the task of calculating the link likelihood score for target links. In a single-layered network, this can be represented as the task of calculating $LS(x,y)$ for determining the chance of link creation between nodes $x\&y$. For the multiplex network setting, this task is extended to the layer-specific paradigm such that link prediction in multiplex networks becomes the task of calculating $LS_{\beta}(x,y)$. Here β represents a specific layer, and $x\&y$ are the nodes in the network between which link existence has to be evaluated for the specific layer. For the task of link prediction in the target layer, the higher the likelihood score for an unconnected pair, the higher the chance of a link between them. A visual representation for this task can be found in Fig. 1.2 in which the initial network information is used to make predictions about probabilities of possible edges (dashed edges in Layer-1 of network).

Definition 1.1.1. (Link prediction in Multiplex network). For a given graph $G(V,E)$ with each layer graph being $G_{\beta}(V,E_{\beta})$, link prediction problem in multiplex network estimates the likelihood score $LS_{\beta}(x,y)$ for each target link $(x,y) \in U \setminus E_{\beta}$ using information from the whole graph $G(V,E)$.

1.2 Motivation and Objectives of the thesis

Different link prediction methods have been proposed in literature whose performance varies depending upon the properties of the graphs themselves. The graphs vary with size, edge densities, clustering coefficients and their overall connectivities. Quasi-local link prediction methods have been shown to produce better results than either individual local or global similarity-based ones on a wide variety of graphs. Social network graphs are sparse by nature but even then the importance of one edge is assumed to be less than one node. But this edge may prove to be of greater importance depending on its relevance in the surrounding community. Link prediction methods usually do not focus on the edge perspective and local similarity-based methods especially consider all edges to be of equal relevance in an unweighted graph. For example, an edge that connects a centralized community with its leaf node is considerably less relevant than one that acts as a bridge between two separate communities. Contemporary edge relevance quantification approaches are based on computing complex edge-based centralities. These centralities consider the number of shortest paths, simple paths, and edge's contribution to overall graph connectivity. Such approaches fail to consider the relative sphere of influence of the edge itself, which is much more constrained than the whole graph according to the 3-Degree of Influence theory [28, 29]. Depending on different types of edge relevance quantifications, the underlying calculation can usually be categorized as a global information-based task.

Global information-based edge relevance quantification has high computational complexity. The motivation of the work in this thesis is to use quasi-local edge relevance-based link prediction methods to improve the overall accuracy of link prediction. In this thesis, a localized area of node is taken into consideration where edge relevance is quantified to improve upon the complexity of global approaches. In this research, local and global information have been exploited to create quasi-local similarity-based link prediction approaches that apply to single-layered and multiplex networks. The proposed approaches attempt to achieve a trade-off between local and global approaches such that the region of relevance for link prediction is more significant than local similarity-based approaches. However, the processing complexity would be less than in global similarity-based approaches.

The simple calculation strategies and above-par performance of local similarity-based algorithms are sometimes used as a preprocessing step for more complex algorithms to eliminate highly improbable edges from candidate sets. However, they treat the relevance of all edges as equal, which is a simplistic methodology. When these algorithms are employed for simple weighted networks, their formulations use edge weights. However, those weights are the inherent properties of the datasets themselves and hence cannot be referenced as edge relevance obtained from the graph structure itself. Usually, this edge relevance is calculated using intensive computational strategies such as testing the overall connectivity of the graph with and without the edge or checking the edge effect on all shortest paths. This leads to first objective of this thesis.

Objective 1 : To employ efficient edge relevance quantification to suggest improved quasi-local extensions to local similarity-based link prediction algorithms.

Further, the aggregation methodology can also be used on multiplex networks, and this leads to the backdrop of second objective of the thesis. The layer aggregation is used to create a complete overview of the network. Link prediction algorithms for multiplex networks are proposed in a layer-specific methodology. Such methods use current layer information along with layer differences in the pre-calculation phase itself, making each prediction cycle relevant for an individual layer only [30, 31]. The core concept of a multiplex network is that it represents nodes that can have different types of relationships between themselves. If a node pair has many different types of relationships, it is trivial to assume that more types of relationships have a high probability of forming. Layer aggregation can create weighted graphs, which can be construed as a summary of the entire multiplex network. Suppose link prediction is performed on this summary network. In that case, the result will hold different levels of similarities to the link prediction task on specific layers depending on the proportionate matching structure of the layer and summary graph. In such a manner, aggregation of multiplex network along with de-aggregation with respect to layer relevance can provide us with a more efficient framework than the individual layer-based one. This leads to second objective of the thesis.

Objective 2: To employ layer aggregation as a method for simplifying link prediction in multiplex networks.

1.3 Contributions of this thesis

Based on these objectives, novel similarity-based methods of link prediction for both single and multiplex networks have been proposed. The main contributions of the thesis are divided into four works addressing the aforementioned two major objectives. The contributions of the thesis are as follows -

- Addressing Objective 1, an Ego-based Link Prediction algorithm (*ELP*) is proposed in Chapter 3, which uses an ego-based [32] link strength estimation perspective to predict target links. Classical algorithms do not consider the cumulative effect of node-based strength propagation on edges to predict target links, but that is the specialty of proposed *ELP* algorithm. The closest comparison to this approach can be found in path-counting algorithms dependent on adjacency matrix-based operations, which are computationally expensive. The full algorithm has few essential steps - at first the algorithm computes each existing edge's ego strength using ego networks, which can be construed as regions of influence of specific nodes. These ego strengths can be abstracted as the total effect of all local nodes on a particular edge. Then a topological feature set is utilized to estimate the prediction scores for target links which can be construed as an edge relevance-based extension of local similarity-based link prediction methods. *ELP* performs exceptionally well in the Accuracy metric, a combined representation of the prediction performance of both existent and non-existent edges. For other metrics, i.e., AUPR and AUC, it can be observed that *ELP*'s performance is better on datasets with an average degree more significant than 10. This makes *ELP* algorithm more suitable for link prediction of networks with the magnitude of edges much larger than nodes.

- Addressing Objective 2, different strategies of link prediction on the summary graph (a simple weighted graph) can be explored based on the quasi-local paradigm such that there is a trade-off between the costly calculation of global information and overly simplistic local information. Within this layer aggregation and likelihood de-aggregation framework, three methods of quasi-local link prediction have been proposed in this thesis -

1. Based on extended simple paths between nodes, an algorithm was proposed called Higher Order Path-based Link Prediction for Multiplex networks ($HOPLP - MUL$) in Chapter 4. The proposed method sought to anticipate linkages by including more information about nodes (considerably larger zones of influence) and applying appropriate damping and layer fusion procedures. Density-based proposed parameters and the modified initial significance play an essential role in the $HOPLP - MUL$ method. The findings reveal that localized neighborhood-based algorithms have a relatively limited picture of the routes connecting nodes, resulting in reduced accuracy. This fact has been capitalized on in this study. The proposed approach can be divided into three essential parts. To begin, an aggregation model is used that combines information from many layers into a single summary weighted static network while accounting for the relative density of the layers. Then, an algorithm is proposed which iteratively calculates link likelihoods taking longer paths between nodes into account. The concept of layer ranking based on densities is also incorporated as well as the dampening effect of longer paths on information flow. This solution beats existing link prediction algorithms for link prediction in multiplex networks.

Though the 3-Degree of Influence phenomenon leads to taking into account paths as long as six hops, the node's role in the entire graph structure is not taken into account. In order to improve upon this issue, we attempt to combine node and edge relevance to enhance link prediction in multiplex networks.

2. Based on both node and edge relevance, a novel method for link prediction in multiplex networks is proposed, called Merged Node and Edge Relevance based Link Prediction in Multiplex networks ($MNERLP - MUL$) in Chapter 5. The proposal aimed to predict links using more information between nodes (quasi-local approach) and to better predict links in specific layers from a summarized weighted graph. The results demonstrate that local neighborhood-based algorithms take a very restrained view of overall

network information to predict edges between nodes, resulting in lower accuracy. This fact has been improved upon. The variation of weightage of both edge and node relevance for link prediction has also been explored. Another characteristic is that only one round of link prediction should be performed (non-layer specific). Layer-specific link likelihoods can be calculated with just a simple multiplication with an unpacking constant. The full approach can be divided into three phases. First, an aggregation model is utilized that encodes the information from different layers into one summarized weighted static network, taking into account the relative density of the layers themselves. Then, an algorithm is presented which first calculates node and edge relevance based on the summarized graph, and then both these factors are combined to perform link prediction on unconnected pairs of nodes. The edge relevance is calculated using the information from the immediate vicinity of the edge (local information), while node relevance is calculated based on the node's importance to the overall structure of the graph (global information).

The node centrality calculation process involves possible discerning paths in a graph which is a computationally expensive operation for large graphs. In order to improve upon this issue, we attempt to use community detection to enhance link prediction in multiplex networks.

3. Based on community detection which in itself is a quasi-local information optimization task, called Community-based Link Prediction on Multiplex networks ($CLP-MUL$) in Chapter 6. The proposed algorithm predicts links that are not specific to a particular layer but are based on communities detected using the summarized information of all layers. The proposed approach for link prediction considers these communities to stretch across layers even if the edge structure of a particular layer may not totally agree with it. This approach can be divided into three essential phases. First, an aggregation model is presented that encapsulates the information from

several layers into a single weighted static network. Then, a modified clustering method is proposed and applied to this weighted graph. This method uses information diffusion for label propagation to determine the regions of influence (rigid communities/clusters) of different central nodes. Finally, these clusters are used for calculating intra-cluster and inter-cluster similarity between node pairs for link prediction. Experiments were performed on six real-world datasets, and the results indicate that the original argument was justified for datasets with low average shortest path length and relatively higher number of training edges.

1.4 Organization of the thesis

This thesis is organized as follows. **Chapter** 2 presents a brief overview of link prediction and the background of research done in this thesis. **Chapter** 3 is focused on *ELP*, an edge relevance-inspired quasi-local information-based extension of local similarity-based algorithms. In Chapters 4, 5, and 6, edge relevance quantification is employed to improve the link prediction task in multiplex networks using layer aggregation. **Chapter** 4 exploits maximum influence regions of three hops from the nodes themselves to evaluate paths of a maximum of six hops between nodes for link prediction. **Chapter** 5 exploits node centralities as global information variables to improve the accuracy of link prediction along with local information obtained for local region aggregation-based edge relevance. **Chapter** 6 community detection of aggregated multiplex graphs is exploited as a form of quasi-local information for link prediction. Finally, **Chapter** 7 concludes the work done with some future directions.