Chapter 2

Social Networks Analysis Background

This chapter gives a brief introductory theory of social network analysis with focus on link prediction. It emphasis on general characteristics, some common studied problems and applications related to social networks.

2.1 Social networks and graph Theory

The Social network is a Sociology term studied first in Sociology, then in Physics, and more recently in Economics, Biology, and other diverse area of research. Social networks are important and affect our lives in several ways. For example, the decision to whether buy a product or not, advertise a product or service from which personality, finding the status of a person whether he/she is criminal or not, and several such decisions are influenced by their neighbors/acquaintances. Such scenario can be represented and studied by social networks. These networks consists of group of social actors/entities and some types of connection among them. The graph theory concepts, an existing field of mathematics, have been utilized to study structure and dynamics of social networks.

A social network is represented by a graph with nodes corresponding to entities/actors and links corresponding to the relationships among them. A graph is formally represented by $G = (V, E)$ where $V = \{V_1, V_2, ..., V_n\}$ is the set of nodes with cardinality n. A link $(e \in E)$ is considered to be a set of two nodes from node set V with a connection or arc between them. Total number of possible links U is $\frac{n \times (n-1)}{2}$ (for undirected graph) and $n \times (n-1)$ (for directed graph). At a given instant of time, a graph can have several node pairs with at least one link are known as existing links and vice-versa with non-existing links i.e., $\{U - E\}$. Finding these links is the link prediction problem in networks.

2.2 General characteristics of social networks

2.2.1 Triadic closure and clustering coefficient

Triad is group of three entities (or nodes in network) that can be open or closed as shown in the Figure 2.1. There are two links out of threes available in an open triad (left figure),

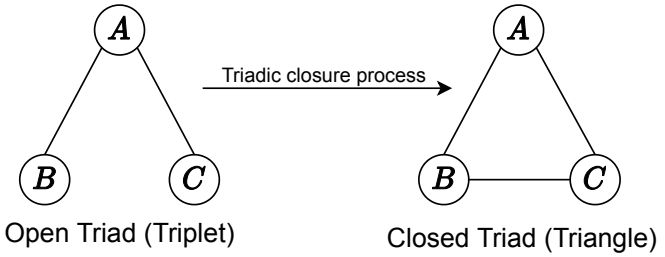

FIGURE 2.1: Triads in networks

while closed triad is completely connected (right figure). Formation of closed triads developed from open triads in networks is known as triadic closure process.

Triadic closure is a key concept to understand the fundamental mechanism of formation and evolution of networks [36]. In particular, what are the mechanism by which nodes and links arrive and depart in/from the networks.

The clustering coefficient [8, 11] and the transitivity [12] are used to quantify triadic closure in the networks i.e., These are two statistical measures to count the number of triangles in a network. It is defined as the probability that randomly picking triplet of three nodes to become a triangle. These two measures differ in how they sample the triplets in the network. In a graph $G(V,E)$ with n nodes and m links, the transitivity can be expressed as

$$Transitivity(G) = \frac{3 \times \text{number of closed triplets or triangles in G}}{\text{Total number of connected triplets in G}} \tag{2.1}$$

In the equation, the number 3 represents triple counting of a triangle, one with each node. The local clustering coefficient $C(i)$ is defined using the following equation.

$$C(i) = \frac{|(j,k) \in E \ni j,k \in \Gamma(i)|}{\binom{k_i}{2}} \tag{2.2}$$

and the average clustering coefficient of the network is

$$C(G) = \langle C \rangle = \frac{1}{n}\sum_i C(i) \tag{2.3}$$

2.2.2 Small world phenomenon

Most social networks reveal that the average path length (distance) between each pair of nodes is short i.e. one can reach from one node to other nodes in few steps in the network and this number of steps is, usually, 6. In 1960's, Milgram [29] experimentally showed this phenomenon where he found that randomly selected an individual could send a letter/package to a chosen recipient (separated geographically at a large distance) through a short distance (on average 6 people) short chain of acquaintances.

This can be understood using the following Figure 2.2 [36]. Suppose each individual knows 100 friends (acquaintance) excluding itself. In two steps, one can reach up to $100 * 100 = 10,000$ acquaintances and in 3 steps, it goes to $10000 * 100 = 1,000,000$

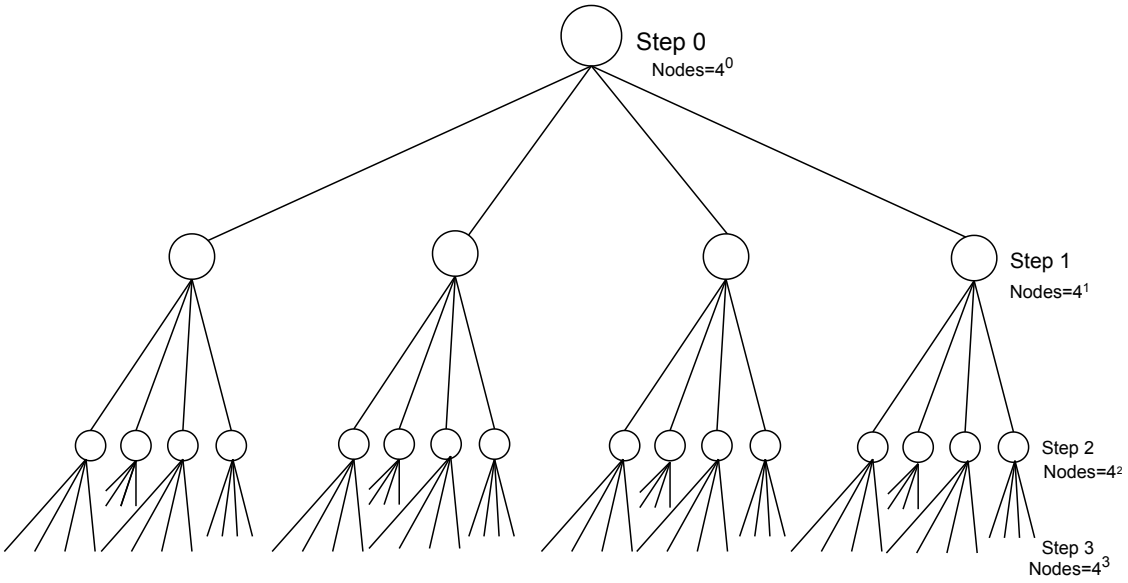

FIGURE 2.2: Small world phenomenon with pure essential growth

people. In other words, this number increases in power of 100 with each step. It reaches to 10 million in 4^{th} step and 10 billion after five steps, accumulating the entire world population. However, in real world, this is not the case as most friends in second step mutually know each other rather than different other friends in third step. In other words, many nodes in higher steps may form triangles that reduces this growth significantly. However, Duncan Watts and Steve Strogatz [11] argued that this growth can be achieved by following two ideas; homophily principle (that creates many triangles) and weak ties (still produce the kind of widely branching structure that reaches many nodes in a few steps).

2.2.3 Scale-free networks and preferential attachment

A scale-free network is the one where degree distribution follows a power law, at least asymptotically (means for large k). That is, fraction of nodes with degree k can be can be estimated using the equation 2.4

$$P(k) \sim k^{-\gamma}, \tag{2.4}$$

where $P(k)$ is fraction of nodes with degree k in the network, and γ is a constant with $2 < \gamma < 3$. Barabasi et al. [9] coined the term scale-free for networks though, Derek de Solla Price, in 1965, stated that the number of citations received (links) by papers follow a heavy tailed distribution in turn followed power law. Barabasi and his colleagues mapped the topology of a portion of World Wide Web and observed that few nodes are having many more connections (known as hubs) compared to others and network as a whole followed a power law distribution.

Growth and preferential attachment [9] are two major components that explains the emergence of the scale-free behavior of complex networks. The first component shows the growth process where a new nodes joins to an existing system over an extended period of time. The preferential attachment component gives a way of new nodes to join other existing nodes in the system based on certain number of links to others. That is, probability of joining new links to an existing higher order nodes (hubs) are higher.

2.2.4 Homophily or assortative mixing

Homophily is a characteristic of networks where high degree nodes interact or connect to other higher degree nodes. In other words, people of a network tend to make friends with similar to themselves, and this similarity can be in terms of interest, age, location, job, etc. Opposite behavior of homophily is disassortative mixing where higher degree nodes tries to connect with lower degree nodes. Newman [37] analyze homophily nature of several networks and found that social networks (actor networks and collaboration network of academician) exhibit assortative mixing where as technological networks (the Internet and the World Wide Web network), and biological networks (protein interactions, food webs and neural networks) exhibit disassortative mixing. Moreover, he also observed the percolation theory or effects of removal of nodes in networks. He stated that there is little effect of removing higher degree nodes to the resilience of assortative network because of the redundancy of such nodes where as it affects largely to the networks with disassortative mixing.

2.3 Common studied problems of social network analysis

2.3.1 Link-based object ranking (LBR)

As the name suggests, LBR is a problem of ranking the objects (or entities or vertices) in the networks based on link structure [38]. LBR algorithms assign relevance score to each entity of the network based on the patterns of its relationship (link). Most popular LBR algorithms are HITS [39] and PageRank [40]. The real time application of PageRank algorithm is in the largest search engine named, Google.

2.3.2 Community detection

This problem aims at identifying highly connected groups (communities) of entities or nodes in the network. In other words, community detection is clustering of nodes based on their structural properties. The organization of the node in the clusters is such that the intra-community links (i.e. links withing the same group) are dense compared to the inter-community links (i.e. links between groups). Community detection is a non-trivial problem of networks where lots of algorithms [41, 42] to find these communities are available in the literature.

2.3.3 Finding central nodes

Identification of central nodes is a non-trivial and important task in networks. These nodes can be used to disseminate information to large portion of the network as well as to prevent or control spreading of disease using efficient utilization of vaccines. Most popular algorithms [43] to calculate centrality measures are degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality, etc.

2.3.4 Influence maximization

Influence maximization is the problem of finding a small set of most influential nodes in the network so that the aggregate influence of the set can be maximized. Study of the problem is very crucial as most influential nodes can be used in advertising any product or services to a large portion of the population, this is known as viral marketing. Online networking websites like Facebook, Amazon, Flipkart, LinkedIn, etc., have applied influence maximization to attract millions of people, so these networks are viewed as effective platforms for implementing viral marketing strategy. Kempe et al. [44] introduced a first systematic study of maximizing the spread of influence in networks.

2.3.5 Link prediction

Link Prediction finds missing links in static networks or future links in dynamic networks. It is an important task in network science where lots of seminal surveys papers [17, 18, 27, 45] exploring this area extensively are available in the literature. We have also experimentally explored it further in our recent publication [46]. Chapter 3 is dedicated to analyze the existing and current state-of-the-art of link prediction in detail.

Most link prediction algorithms are classified in following different groups namely similarity-based approaches, maximum likelihood, and probabilistic models. Based on links in the network, a probabilistic or statistical relational model builds joint probability distribution. Given this distribution and inference algorithms, likelihood of missing links are estimated [47]. Maximum likelihood algorithms assumes the networks to be organized in hierarchy or community, etc. For example, hierarchical random graph [1] represented graph as hierarchical structure in dendrogram model and a set of connection probability is inferred from the consensus dendrogram that most accurately represent the graph hierarchically. Similarity-based approaches compute relevance score of each node-pair of the graph. This score represents some types of similarity between the pair. Edges are arranged in decreasing order based on the relevance score and then top-l links

are extracted as predicted links. Similarity-based approaches are scalable to large networks due to less computational time. Several other groups of algorithms with recent approaches are depicted in detail in Chapter 3.

2.4 Applications of social network analysis

Real world is very complex and hence most of the applications are represented by complex networks and analyze their behavior by performing simulation. For example, biological networks of human and animal can be represented and simulated by complex networks, and interaction among proteins can be utilized to identify several functions in the body. E-commerce can be simulated by social networks and contents (object, product, services, etc) can be recommended based on some characteristics. For example, products and services recommendation in Flipkart, amazon, etc, friends recommendation in Facebook and so on. Working strategy of these recommendation systems [21, 22, 48, 49] are based on collaborative filtering where a system automatically predicts (filters) about interests of a user by collecting preferences of many users (collaboration) in the same area. The recommendation systems suffer from a problem called cold-start or bootstrapping which is the case when no preference available for the first time when the product or service is launched. Huang et al. [21] solve this problem using link prediction to make the graph or network denser and hence increasing user preferences. Social networks play a pivotal role in advertising a product or service called viral marketing. For example, advertisement of several products and services appear on Facebook, Youtube, and so on. These companies earn lots of revenue from advertising the products. Thus, social networks have been the sources of income to many companies and users (as these company pay some amount to users for their video in Youtube). Specialized application of link prediction is expressed in a special section in the Chapter 3.

Application specific to link prediction

- Building recommendation system and solution to the data sparsity problem in it [21].
- Identifying missing links among terrorists and their organization [30].
- Protein-protein interaction in biological network.
- Improving the efficiency of search engines by predicting which web pages will be visiting next [40].
- Improving hypertext analysis for information retrieval and search engines [20].
- It helps predicting the spread of entities like disease spreading in epidemic [14], rumour spreading among people, etc.

2.5 Computational complexity of social network analysis

Time and storage are biggest issues when dealing with complex networks, especially social networks where size (number of nodes) and volume (number of links) are increasing continuously. In general, volume of the networks increases exponentially. The computational issue regarding this can be illustrated using some metrics like betweenness centrality (a monadic metric, applicable to single entity), link prediction (a dyadic metric that requires two entities). Betweenness centrality [50] of a node is the percentage of shortest path between all nodes containing the specific node. This process requires computing the shortest path between every pair of nodes in the network. Time complexity to do this is $O(n^3)$ using Flyod-Warshall algorithm but more efficient Brandes algorithm [51] with time complexity $O(n^2 \log n + nm)$ and space complexity $O(n+m)$ can be found for sparse networks. In case of link prediction, the total number of existing links are $O(n)$ in sparse graph, resulting in $O(n^2)$ non-existing links for which similarity scores are computed. A simple intuitive method (e.g., Common Neighbor index [17]) require $O(n)$ to calculate similarity score of a link, which results a total of

$O(n^3)$ time for a total. Also, very complex methods (that require more than $O(n)$ time) are available, that result in very time consuming operations. Hence, many methods are not feasible to large real world networks.

Computational Complexity of Similarity-based Link Prediction Scalability is a key issue in most of the problems in social network analysis especially link prediction where $O(n^2)$ links need to be considered. Size and volume of online social networks are increasing rapidly and hence networks with size millions of nodes need to be dealt that require algorithms to be scalable. Time and space complexities of link prediction depends on the way of representation of graph or network. In general, adjacency list and adjacency matrix are used to represent a graph. When we concentrate on efficient use of space, adjacency list representation is appropriate as it requires $O(m)$ for a graph with m existing edges. In adjacency matrix representation, the cost becomes $O(n^2) \gg O(m)$ which is very larger in case of sparse graph. When time efficiency is considered, adjacency matrix is better representation because elements of the matrix can be directly accessed (e.g., get[x][y]). Adjacency list representation requires search operation to access an element (e.g., list[x].search(y)). Clearly, the choice of graph representation depends on the problem at hand.

In general, link prediction approaches are evaluated using area under the ROC curve [52], average precision, etc., that require to access all edges of the graph. The time complexity of all possible edges in a graph with n vertices is $O(n^2)$ as $\frac{n(n-1)}{2}$ edges are possible in undirected case and $n(n-1)$ in directed case. This is the scenario for a complete graph, but real-world networks are sparse (mostly) where only K edges (i.e., average degree) need to be accessed per node and $2(K-1)$ vertices per edge with a total of $\frac{nK}{2}$ edges (undirected case). Computing local score (where direct neighbors are accessed) for a node pair is $O(K)$ as an edge can access $2(K-1)$ nodes in the graph. Therefore, total time complexity is $O(nK^2)$.

Quasi-local similarity approaches access deeper section of a graph but less compared to global. These methods [53, 54] enhance link prediction accuracy by increasing time complexity compared to local approaches. The computational cost can be reduced if the number of steps to explore, is known beforehand. Time needed to compute quasi-local similarity score of a node pair is $O(n^s), s \geq 2$ in the worst case, where s is the number of steps to explore the graph. It complexity degenerates to local similarity if $s = 1$. The total computational time is, therefore, $O(n^2) * O(n^s) = O(n^{s+2})$ in worst case. In realistic, it is $O(nK^2 * K^s) = O(n * K^{s+2})$ for $s \geq 2$.

Global similarity approaches explore the network completely. Computational costs of these algorithms are of cubic nature in general, hence infeasible for large real-world network with millions of nodes. For example, Katz index [55] requires $O(n^3)$.