A review on Link Prediction in Social Networks

Pham Thi Viet Huong

I. Introduction

In recent years, the emergence and popularity of social networks has been growing strongly, leading to a more collaborative environment. A social network is a social structure comprising of persons or organizations, which usually represented as node, while their social relations are represented as edges among nodes [1]. The social relation could be explicit, such as colleagues and classmates, or it could be implicit, such as friendship and common interest.

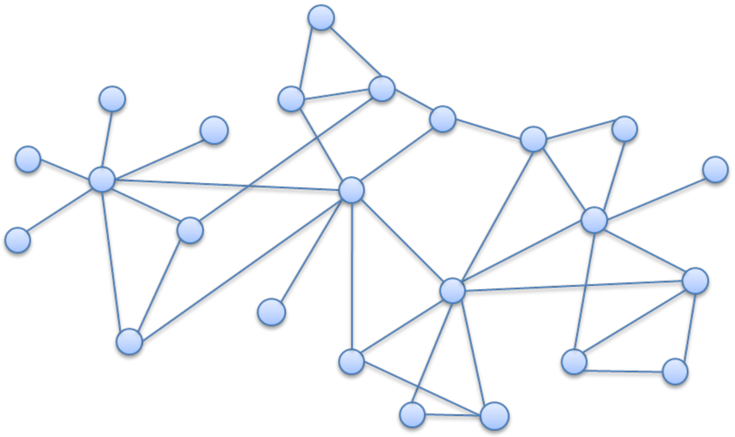

Fig. 1 An example of social network

In Figure 1, for example, each node represents an employee. The edge between two nodes means these two employees have some connection in work, and the weight of the edge is the communication frequency. Social networks are highly dynamic; they expand and vary quickly over time. This change is caused by the addition of new edges, signifying the appearance of new collaboration in the social structure. Given these dynamic networks, can we infer which new interactions among its members are likely to occur in the near future? This question is formalized as the link prediction problem in social networks.

Link prediction is an important task for analyzing social networks. It is applicable to a wide variety of applications. In addition to its role of understanding social network evolution, there are many other applications on link prediction. In e-commerce, one of the most prominent usages of link prediction is to build the recommendation systems ([2], [3], [13]). In the domain of healthcare, a prediction model was built to predict future association between doctors which can be very helpful to reduce the referral traffic; and in stock market, it has successfully predicted the financial profitability to consider which company's stock would be more profitable to hold and for how long to hold it [14]. Effective methods for link prediction could suggest promising interactions or collaborations that have not yet been utilized within the company/organization. In security related application, we can predict connections between members of terrorist organizations who have not been directly observed to work together. Link prediction has potential use for recommending new friends in online social networks, or to determine whether two people are family members or not. Or we can predict the disease type of a patient based on the characteristics of the patient and his connections.

Traditionally, a social network is represented as a graph $G = (V, E, X^V, X^E)$, where V is a set of nodes which are in the network. Nodes can be people, places, events, etc. E is the set of links, which represent relationship among those entities in V. X^V is a set of features about the entities in V. Similarly, X^E is a set of features providing information about links among entities in V. For example, Facebook is one of the most popular social network ([15]). Its set V includes users and E represents the relationship among users. X^V is a set of features about users in Facebook, it can be name, age, gender, school, favorite movie, relationship status, etc. X^E represents the information about the links in E such as the time of formation, the number of times users interact with others, or possibly the contents of messages that was exchanged among the users. The conventional prediction tasks in data mining deal with the graph of $G = (V, X^V)$, which is based on the characteristics of the entities (users) in the graph to analyze the network. However, social network is not only about people and entities but it is also about relationship and connection. Moreover, social networks are very dynamic, since new edges and vertices are added to the graph over time. Understanding the dynamics that drive the evolution of social networks is a complex problem due to a large number of variable parameter.

The problem can be formulated as: Given a snapshot of a social network at time *t*, we seek to accurately predict the edges that will be added to the network during the interval from time *t* to a given future time t_0. In other words, the purpose is to predict new interactions among users at time $(t_0 - t)$.

II. Existing approaches to link prediction

1. Based on similarity between nodes

1.1 Feature based Link Prediction

The link prediction problem can be modeled as a supervised classification task, by considering each data point is a pair of vertices in social networks. The link prediction problem is formulated by choosing a training interval $[t_0, t'_0]$ and a test interval $[t_1, t'_1]$ where $t'_0 < t_1$, then giving an algorithm access to the network $G[t_0, t'_0]$. It must then output a list of edges which are not presented in $G[t_0, t'_0]$ but are predicted to appear in the network $G[t_1, t'_1]$ [7]. In more details, assume $u, v \in V$ are two vertices in the graph $G = (V, E)$ and the data point is labelled as x. Assume the interaction between u and v are symmetric (the pair (u, v) and (v, u) are the same data point). Then data points are formulated as $x = \{\begin{matrix} +1, & (u,v) \in E \\ -1, & (u,v) \notin E \end{matrix}$

This is a typical binary classification task that can be performed with any of the popular supervised classification tools, such as naïve Bayes, support vector machine (SVM), or *k* nearest neighbors. But the critical challenge in this approach is how to choose an appropriate feature set. Below are the most common methods that have been used in similarity-based link prediction.

Common Neighbors

It is intuitive to note that in social networks, if node x is connected to z, node y is connected to z, then there is a high probability that x will be connected to y in the future. As the number of common neighbors grows higher, the chance that x and y will have a link between them increases. Based on this idea, Newman [12] has computed this quantity in the context of collaboration networks to show that a positive correlation exists between the number of common neighbors of x and y at time t, and the probability that they will collaborate together. The common neighbors (CN) measure for unweighted networks is defined as the number of nodes with direct relationship with both x and y. $CN(x, y) = |\Gamma(x) \cap \Gamma(y)|$. The higher the $CN(x, y)$, the more likely x and y will connect in the near future.

Jaccard Coefficient

Jaccard Coefficient is the normalization of the Common Neighbors. It defines the probability that a common neighbor of a pair of x and y would be selected if the selection is made randomly from the union of the neighbor sets of x and y. Jaccard Coefficient $(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$. However, the experimental result in [7] showed that the performance of Jaccard coefficient is worse in comparison to the number of common neighbors.

Adamic/Adar

In [16], Adamic and Adar proposed this score as a metric of similarity between two webpages. They summed up the number of items the two items is sharing. Items that are unique to a few users are weighted more than commonly occurring items. The weighting scheme is the inverse log frequency of the occurrence. For a set of feature z, it is defined as below

$$\sum_{z:feature\ shared\ by\ x,y} \frac{1}{\log\left(frequency(z)\right)}$$

For example, if only two people mentioned an item, then the weight of that item is $\frac{1}{\log(2)} = 1.4$, if 5 people mentioned this item, its weight decreased to $\frac{1}{\log(5)} = 0.62$. For link prediction, in [7], they customized this metric as *adamic/adar*$(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log|\Gamma(z)|}$, where the common neighbors are considered as features. In this way, Adamic/Adar defines a higher importance to the common neighbors which have fewer neighbors. Hence, it measures how strong is the relationship between a common neighbor and the evaluated pair of nodes. From the reported results of the existing works on link prediction, Adamic/Adar works better than the previous two metrics.

1.2 Path based Features

a. Shortest Path Distance

This is based on the fact that the friends of a friend can become a friend; hence the path distance between two nodes in a social network can influence the formation of a link between them. The shorter the distance, the higher chance it will be linked. But poor performance by this feature is reported in [8]

b. Katz

Leo Katz proposed this metric in [9]. Katz value is the summation of all the paths that exist between a pair of vertices x and y. But, to penalize the contribution of longer paths in the computation, it exponentially damps the contribution of a path by a factor of β^l, where l is the path length.

$$katz(x, y) = \sum_{l=1}^{\infty} \beta^l \left|path_{x,y}^{<l>}\right|$$

Where $\left|path_{x,y}^{<l>}\right|$ is the set of all paths of length l from x to y. The parameter β (≤ 1) is chosen to be a very small value (for dampening). For weighted networks, $\left|path_{x,y}^{<l>}\right|$ is the number of collaboration between x and y. For unweighted networks, $\left|path_{x,y}^{<l>}\right| = 1$ if x and y collaborate.

c. Hitting Time

The concept of hitting time comes from random walks on a graph. The hitting time, $H_{x,y}$, is defined as the expected number of steps required for a random walk from x to y. Shorter hitting time denotes that the nodes are similar to each other, so they have higher chance of linking together in the future. The commute time from x to y is defined as $C_{x,y} = H_{x,y} + H_{y,x}$. The advantage of this metric is that it is easy to compute. On the downside, its value can have high variance, so prediction by this feature can be poor [8]. For example, the hitting time between two nodes x and y can be affected by a vertex z, which is far away from x and y. If z has high stationary probability, then it could be hard for a random walk to escape from the neighborhood of z. To prevent this problem, we can use random walks with restart, where we reset the random walk by returning x with a fixed probability α in each step. Due to the scale free nature of a social network some of the vertices may have very high stationary probability (π) in a random walk. To protect against this problem, the hitting time can be normalized as

$$normalized - hitting - time(x, y) = H_{x,y}\pi_y + H_{y,x}\pi_x$$

d. Rooted Pagerank

The original definition of pagerank denotes the importance of a vertex under two assumptions: for some fixed probability α, a surfer at a web-page jumps to a random webpage with probability α and follows a linked hyperlink with probability $1 - \alpha$. Under this random walk, the importance of an webpage u is the expected sum of the importance of all the webpages v that link to v. For link prediction, the random walk assumption of the original pagerank can be altered as below: similarity score between two vertices x and y can be measured as the probability of y in a random walk that returns to x with probability $1-\beta$ in each step, moving to a random neighbor with probability β. In [10], it is name as rooted pagerank (RPR) and can be derived as follows. Let D be a diagonal degree matrix with $D(i,i)=\sum_j A[i,j]$. Let $N=D^{-1}A$ be the adjacency matrix with row sums normalized to 1. Then $RPR=(1-\beta)(I-\beta N)^{-1}$

2. Based on Bayesian probabilistic models, and probabilistic relational models

2.1 Bayesian Probabilistic Models

a. Link prediction by Local Probabilistic Model

In [11], the author proposed a local probabilistic model for link prediction that uses Markov Random Field (MRF), an undirected graphical model. To predict the link between a pair of nodes x and y, it introduces the concept of central neighborhood set, which consists of other nodes that appear in the local neighborhood of x or y. The central neighborhood sets of x and y can be found in many ways. The most natural way is to find a shortest path between x and y and then all the nodes along this path can belong to one central neighborhood set. If there exists many shortest paths of the same length, all of them can be included in the collection. Let $\{\mathrm{w},\mathrm{x},\mathrm{y},\mathrm{z}\}$ be the central neighborhood set of x and y, then the main objective of this model is to compute the join probability $\mathrm{P}(\{\mathrm{w},\mathrm{x},\mathrm{y},\mathrm{z}\})$, which represents the probability of co-occurrence of the objects in this set. This probability can be marginalized (in this example, over all possible w and z) to find the co-occurrence probability between x and y. There can be many such central neighborhood sets (of varying size) for the pair x and y, which makes learning the marginal probability (p(x, y)) tricky.

b. Link prediction by Hierarchical Probabilistic Model

In [6], they proposed a probabilistic model which considers the hierarchical organization in the network, where vertices divide into groups that further subdivide into groups of groups and so forth over multiple scales. This hierarchical model can be used to predict missing links [6]. The learning task is to use the observed network data to fit the most likely hierarchical structure through statistical inference – a combination of the maximum likelihood approach and a Monte Carlo sampling algorithm.

Let G be a graph with n vertices. A dendogram D is a binary tree with n leaves corresponding to the vertices of G. Each of the $n - 1$ internal nodes of D corresponds to the group of vertices that are descended from it. A probability p_r is associated with each internal node r. Then, given two vertices i,j of G, the probability p_{ij} that they are connected by an edge is $p_{ij} = p_r$ where r is the lowest common ancestor in D. The combination, $(D, \{p_r\})$ of the dendogram and the set of probabilities then defines a hierarchical random graph. The learning task is to find the hierarchical random graph or graphs that best fits the observed real world network data.

For the task of link prediction, a set of sample dendograms are obtained at regular intervals once the Monte Carlo Markov Chain random walk reaches an equilibrium. Then, for the pair of vertices x and y for which no connection exists, the model computes a mean probability p_{xy} that they are connected by averaging over the corresponding probability p_{xy} in each of the sampled dendograms.

2.2 Probabilistic Relational Models

In earlier section, we discussed that the vertex attributes play a significant role in link prediction task. However, in most of the cases, these approaches are not generic and not applicable in all possible cases. Probabilistic Relational Models (PRM) is a concrete modeling tool that provides a systematic way to incorporate both vertex and edge attributes to model the joint probability distribution of a set of entities and the links that associate them. The PRM is better than a flat model in the sense that it considers the object-relational nature of structured data by capturing probabilistic interactions between entities and the link themselves. PRM was first designed for the attribute prediction problem in relational data. For link prediction task, the links are first-class citizens in the model, additional objects, named link objects are added in the relational schema. The link prediction task now reduces to the problem of predicting the existence attribute of these link objects.

In the training step of the model, a single probabilistic model is defined over the entire link graph, including both object labels and links between the objects. The model parameters are trained discriminatively, to maximize the probability of the (object) and the link labels given the known attributes. The learned model is then applied using probabilistic inference, to predict and classify links using observed attributes and links.

There are two pioneering approach of PRM, one based on Bayesian networks [5], which consider the relation links to be directed, the other based on Markov networks, which consider the relation links to be undirected [4]. Though both are suitable for link prediction task, the undirected model seems to be more appropriate due to its flexibility

References

[1] Sheng Y. and Subhash K., “A survey of Prediction Using Social Media”, *Computer Science - Social and Information Networks*, Physics - Physics and Society, 2012. http://arxiv.org/ftp/arxiv/papers/1203/1203.1647.pdf

[2] Huang, Zan, and Li, Xin, and Chen H., “Link Prediction approach to collaborative filtering,” *Proceedings of the fifth ACM/IEEE Joint Conference on Digital Libraries*, 2005.

[3] Li, Xin, Chen Hsinchun, “Recommendation as link prediction:a graph kernel-based machine learning approach,” *Proceedings of the ninth ACM/IEEE Joint Conference on Digital Libraries*, 2009.

[4] Tasker, Benjamin, and Wong, Ming F., and Abbeel, Pieter, and Koller, Daphne, “Link Prediction in Relational Data,” *NIPS’03: In Proceedings of Neural Information Processing Systems*, 2003.

[5] Getoor, Lise, and Friedman, Nir, and Koller, Dephne, and Taskar, Benjamin, “Learning Probabilistic Models of Link structure,” *Journal of Machine Learning Research,* 3:679-707, 2002.

[6] Clause, Aaron, and Moore, Christopher, and Newman, M. E. J, “Hierarchical structure and the prediction of missing links in network,” *Nature* 453:98-101, 2008.

[7] David L. and Jon K., “The link prediction problem for social networks,” *Proceedings of the twelfth international conference on Information and knowledge management*, pp. 556 – 559, ACM 2003.

[8] Liben-Nowell, David, and Kleinberg, Jon, “The Link Prediction Problem for Social Networks,” *Journal of the American Society for Information Science and Technology*, 58(7):1019-1031, 2007.

[9] Katz, Leo, “A new status index derived from sociometric analysis,” *Psychometrika*, 18(1):39-43, 1953.

[10] Liben-Nowell, David, and Kleinberg, Jon., “The Link Prediction Problem for Social Networks,” *Journal of the American Society for Information Science and Technology*, 58(7):1019-1031, 2007.

[11] Wang, Chao, and Satuluri, Venu, and Parthasarathy, Srinivasan, "Local Probabilistic Models for Link Prediction," *In Proceedings of International Conference on Data Mining*, 2007.

[12]. Newman, M. E. J., "Clustering and Preferential attachment in growing networks," *Physical Review Letters E*, 2001.

[13] Liu, Yan and Kou, Zhenzhen, "Predicting who rated what in large-scale datasets," *SIGKDD Exploration Newsletter*, Vol.9 No. 2, 2007.

[14] Wadhad A., Shang G. Tamer M. Reda A. and Jon R., "Link prediction and classification in social networks and its application in healthcare", *IEEE International Conference on Information Reuse and Integration (IRI),* 2011.

[15] Ryan A. Rossi, Luke K. McDowell, David W. Aha, Jenifer Neville, "Transforming Graph Data for Statistical Relational Learning," *Journal of Artificial Intelligence Research*, 45, pp. 363-441, 2012.

[16] Adamic, Lada A. and Adar Eytan, "Friends and neighbors on the web," *Social Networks*, 25 (3), pp. 211 – 230, 2003.