Weights and Multi-Edges in Link Prediction

Victor F. Cavalcante
IBM Research Brazil

Ana Paula Appel
IBM Research Brazil

INTRODUCTION

Over the past years the amount of data collected has increased substantially and it has been powered mainly by the World Wide Web expansion. The outbreak of novel and increasingly more powerful analytical approaches makes the extraction and production of knowledge from big amounts of data possible. Also, the spread of online social networks use is one of the factors responsible for the current high interest in complex network analyses.

Graphs may express and model mathematically complex network structures whenever it is useful to represent how things are either physically or logically linked to one another in a network structure. A graph G is mathematically represented as G = <V,E>, on which V represents a set of |V| nodes (or vertices), and E represents a set of |E| edges (or links), and a relation that associates with each edge two vertices (West, 2001). Two nodes are neighbors if they are connected by an edge.

For example, in social networks, nodes represent people or groups of people, and edges correspond to some kind of social interaction (Backstrom & Leskovec, 2011). In information networks, nodes correspond to information resources such as Web pages or documents, and edges represent logical connections such as hyperlinks, citations (Shi, Leskovec, & McFarland, 2010) or cross-references, and so on.

Complex networks are graph-based representations used for data have substantial non-trivial topological features. The study of complex networks brought important properties such as the power-law degree distributions (L. A. Adamic et al., 2000) and Small World phenomenon (Milgram, 1967), which help us to understand the interaction among human being, the dissemination of information and intrusion detection (Newman, 2010).

One of the most interesting task in network mining is *link prediction*, defined as: "Given a snapshot of a graph G, predict accurately which edges will appear in the near future of network." (Liben-Nowell & Kleinberg, 2003). While in social network this definition is represented by "People you may know" feature, in online commerce this represents recommendations "Costumers who buy X tend to buy Y." The question "*How people get connected"* is relevant not only in the context of social networks, but also in work social network inside companies (Paula, Appel, Pinhanez, Cavalcante, & Andrade, 2012).

Predicting interaction and collaboration among people in organizations can help manage companies in a productive way. The task of recommending unknowns but "similar" people is quite different from possible friend recommendation tasks, which focus on recommending individuals who have friends in common (Guy, Ur, Ronen, Perer, & Jacovi, 2011). In the context of organizations, introducing people with similar skills, profiles or common interests can be valuable for employees in many ways. For instance, searching for people with similar skills can facilitate issues related to problem solving. Through networking, people are able to offer advice to one another and help each other out with new projects or career development (Burke, Marlow, & Lento, 2010).

Weights and directions in networks have, with some exceptions (Akoglu, McGlohon, & Faloutsos, 2010), received relatively little atten-

DOI: 10.4018/978-1-4666-5202-6.ch242

tion, meaning that all connections in the network have the same importance, which is not true. An excellent reason for this is that the simple cases are usually investigated first (unweighted networks), before moving on to more complex ones (weighted networks). On the other hand, there are many cases where edge weights are known for networks, and to ignore them is to throw out a lot of data that might help us to understand these systems better. For instance, in a social network not all connections (friends) have the same importance; weights might indicate the strength of a relationship. The same happens with direction. In some social networks, such as Facebook, reciprocity is common, but in others, such as Twitter, this is not true and this should not be ignored.

In this article, we briefly summarized the progress of studies on link prediction, emphasizing the recent contributions of weighted and multi-edges network. Although link prediction is not a new problem in information science, traditional methods have not caught up with the new development of network science; especially the new perspectives and tools that resulted from the studies of complex networks with respect of weight and/or multiple edges.

BACKGROUND

There are many reasons, exogenous to the network itself, as to why two individuals will be connected in the near future: they may end up geographically close to one another after one of them moves to the same city or neighborhood, or they may attend the same conference or go to the same university. Despite these types of interaction are hard to predict, one also senses that a large number of new interactions are hinted at by the topology of the network: two individuals, who are "close" in the network, will have people in common suggesting that they are more likely to become a connected in the near future.

Traditional linking prediction methods are based on graph structural properties by assigning a connection value, called score(u, w), to pairs of unconnected nodes <u, w> based on a desired graph G. The scores are ranked in a list in decreasing order of score(u,w) and then predictions are made according to this list.

$\Gamma(u):=(u \in V:\exists\ (v,u) \in E)$ of node *u* is defined to be the set of nodes in V that are adjacent to *u*. For a node *u*, let $\Gamma(u)$ denote the set of neighbors of *u* in G. Usually link prediction approaches are based on the idea that two nodes *u* and *w* are more likely to form a link, in the future, if their sets of neighbors $\Gamma(u)$ and $\Gamma(w)$ have large overlap. The most direct implementation of this idea for the link-prediction problem is the common-neighbors predictor, under which we define:

$$\text{score}(u, w) = |\Gamma(u) \cap \Gamma(w)|$$

The *common-neighbors* predictor captures the notion that the probability of two people becoming connected increases with the number of common friends. The *Jaccard coefficient* is used to measure the overlap that both *u* and *w* share an attribute *f*.. Formally it is defined as:

$$\text{score}(u, w) = |\Gamma(u) \cap \Gamma(w)| \,/\, |\Gamma(u) \cup \Gamma(w)|.$$

The meaning of *Jaccard coefficient* is that the number of common friends is important but it is dependent on the number of friends a person has. Thus, if two people have a few friends in common, but, in general they also have a few friends connected to them, they are more likely to become friends than two people who have a few friends in common but have a lot of friends connected to them.

The *Adamic/Adar predictor* (L. Adamic & Adar, 2003) evaluates the neighborhood of the common neighbors and emphasizes the nodes that share neighbors within a small neighborhood. This is because a highly-connected person has a higher chance to be in the common neighborhood of others. This method computes features of the nodes, and defines the similarity between two nodes to be the following: