Supervised Random Walks: Predicting and Recommending Links in Social Networks

Lars Backstrom
Facebook
lars@facebook.com

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

Predicting the occurrence of links is a fundamental problem in networks. In the link prediction problem we are given a snapshot of a network and would like to infer which interactions among existing members are likely to occur in the near future or which existing interactions are we missing. Although this problem has been extensively studied, the challenge of how to effectively combine the information from the network structure with rich node and edge attribute data remains largely open.

We develop an algorithm based on *Supervised Random Walks* that naturally combines the information from the network structure with node and edge level attributes. We achieve this by using these attributes to guide a random walk on the graph. We formulate a supervised learning task where the goal is to learn a function that assigns strengths to edges in the network such that a random walker is more likely to visit the nodes to which new links will be created in the future. We develop an efficient training algorithm to directly learn the edge strength estimation function.

Our experiments on the Facebook social graph and large collaboration networks show that our approach outperforms state-of-the-art unsupervised approaches as well as approaches that are based on feature extraction.

Categories and Subject Descriptors: H.2.8 **[Database Management]**: Database applications—*Data mining*

General Terms: Algorithms; Experimentation.

Keywords: Link prediction, Social networks

1. INTRODUCTION

Large real-world networks exhibit a range of interesting properties and patterns [7, 20]. One of the recurring themes in this line of research is to design models that predict and reproduce the emergence of such network structures. Research then seeks to develop models that will accurately predict the global structure of the network [7, 20, 19, 6].

Many types of networks and especially social networks are highly dynamic; they grow and change quickly through the additions of new edges which signify the appearance of new interactions between the nodes of the network. Thus, studying the networks at a level of individual edge creations is also interesting and in some respects more difficult than global network modeling. Identifying the mechanisms by which such social networks evolve at the level of individual edges is a fundamental question that is still not well understood, and it forms the motivation for our work here.

We consider the classical problem of link prediction [21] where we are given a snapshot of a social network at time t, and 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'. More concretely, we are given a large network, say Facebook, at time t and for each user we would like to predict what new edges (friendships) that user will create between t and some future time t'. The problem can be also viewed as a *link recommendation* problem, where we aim to suggest to each user a list of people that the user is likely to create new connections to.

The processes guiding link creation are of interest from more than a purely scientific point of view. The current Facebook system for suggesting friends is responsible for a significant fraction of link creations, and adds value for Facebook users. By making better predictions, we will be able to increase the usage of this feature, and make it more useful to Facebook members.

Challenges. The link prediction and link recommendation problems are challenging from at least two points of view. First, real networks are extremely sparse, i.e., nodes have connections to only a very small fraction of all nodes in the network. For example, in the case of Facebook a typical user is connected to about 100 out of more than 500 million nodes of the network. Thus, a very good (but unfortunately useless) way to predict edges is to predict *no new edges* since this achieves near perfect predictive accuracy (i.e., out of 500 million possible predictions it makes only 100 mistakes).

The second challenge is more subtle; to what extent can the links of the social network be modeled using the features intrinsic to the network itself? Similarly, how do characteristics of users (*e.g.*, age, gender, home town) interact with the creation of new edges? Consider the Facebook social network, for example. There can be many reasons exogenous to the network for two users to become connected: it could be that they met at a party, and then connected on Facebook. However, since they met at a party they are likely to be about the same age, and they also probably live in the same town. Moreover, this link might also be hinted at by the structure of the network: two people are more likely to meet at the same party if they are "close" in the network. Such a pair of people likely has friends in common, and travel in similar social circles. Thus, despite the fact that they became friends due to the exogenous event (*i.e.*, a party) there are clues in their social networks which suggest a high probability of a future friendship.

Thus the question is how do network and node features interact

WSDM'11, February 9–12, 2011, Hong Kong, China.