Neurocomputing 172 (2016) 71–83

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Influence of edge weight on node proximity based link prediction methods: An empirical analysis

Niladri Sett*, Sanasam Ranbir Singh, Sukumar Nandi

Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India

ARTICLE INFO

Article history:
Received 1 November 2013
Received in revised form
22 July 2014
Accepted 5 November 2014
Available online 23 May 2015

Keywords:
Link prediction
Link weight
Node proximity
Strength of tie

ABSTRACT

Tie weight plays an important role in maintaining cohesiveness of social networks. However, influence of the tie weight on link prediction has not been clearly understood. In few of the previous studies, conflicting observations have been reported. In this paper, we revisit the study of the effect of tie weight on link prediction. Previous studies have focused on additive weighting model. However, the additive model is not suitable for all node proximity based prediction methods. For understanding the effect of weighting models on different prediction methods, we propose two new weighting models namely, *min-flow* and *multiplicative*. The effect of all three weighting models on node proximity based prediction methods over ten datasets of different characteristics is thoroughly investigated. From several experiments, we observe that the response of different weighting models varies, subject to prediction methods and datasets. Empirically, we further show that with the right choice of a weighting model, weighted versions may perform better than their unweighted counterparts.

We further extend the study to show that proper tuning of the weighting function also influences the prediction performance. We also present an analysis based on the properties of the underlying graph to justify our result. Finally, we perform an analysis of the *weak tie theory*, and observe that unweighted models are suitable for inter-community link prediction, and weighted models are suitable for intra-community link prediction.

1. Introduction

Given a social network graph, the task of link prediction problem can be broadly categorized into two (i) prediction of existing links, but yet unknown [1] and (ii) prediction of non-existing links, but likely to appear in the future [2]. Though, originally it started with social networks [2] and biological networks [3], recently the problem has attracted many other domains such as information retrieval [4], where the problem is to predict the missing relationship between words and documents from a word-document graph; and recommendation system [5], where the problem is to predict relationship between products and users. Initial studies [2,6,7] in link prediction methods such as *Common Neighbor* (*CN*), *Jaccard's coefficient* (*JC*), *Adamic/Adar* (*AA*), *Resource Allocation* (*RA*) have explored the topological characteristics of graphs by performing local analysis on node proximity. Majority of studies on the local analysis based link prediction methods consider unweighted graphs. Links in typical social networks, such as message passing, co-authorship and friendship are weighted in nature i.e., links are characterized by *strong* and *weak* ties [8]. It is natural to take tie weights into consideration for the link prediction problem. Traditionally, tie weights are represented by the frequency of interaction between two actors. A different way of representing the characteristics of the ties is also available in the literature [9], where the authors have used the content of a tie to improve the community detection performance. However, influence of tie weights on the local analysis based link prediction methods is not clearly understood. Few studies have been reported in the literature. First of such study has been reported in [10], where the authors have observed positive influence of weight on the above prediction methods. However later in studies [11,12], conflicting results are observed, where the former observes a negative influence and the later observes a positive influence. In all these studies, an additive weighting model has been used. It has not been thoroughly analyzed whether the reported weighting model in its current form is the best estimate of incorporating weights. This has motivated us to propose different forms of weighting models and investigate their performance on prediction methods and datasets. Like [10–12], this paper focuses only on the local analysis based link prediction methods.

From several experiments using ten datasets constructed from eight social networks, it is observed that different weighting models respond differently on prediction methods and datasets.

* Corresponding author. Tel.: +91 9678554481.
E-mail addresses: niladri@iitg.ernet.in (N. Sett), ranbir@iitg.ernet.in (S. Ranbir Singh), sukumar@iitg.ernet.in (S. Nandi).

http://dx.doi.org/10.1016/j.neucom.2014.11.089
0925-2312/

To study the influence deeper, we systematically explore it in three levels. First, plain weighting models are applied over datasets. On the first set of datasets, the weighted model consistently outperforms its unweighted counterparts, and on the second set of datasets, unweighted model consistently outperforms its weighted counterparts. For other datasets, their performances are only marginally different. Second, we modify the weighting models by applying tuning parameters. With a proper selection of tuning parameters, a significant boost in the performance of the weighted models is observed. After tuning, the weighted models perform better than its unweighted counterparts even for some of the second set of datasets, which has been observed otherwise before tuning. Third, from the experimental observations it is evident that an effect of the tie weight depends on the characteristics of the dataset. In the light of this, we present a neighborhood based analysis of datasets to find out the reason behind the diversified effect of the tie weight on the node proximity based link prediction methods. We further extend the analysis based on density of the neighborhood of participating nodes,[1] by introducing *odd ratio* over the node degree.[2] It is interesting to observe that for the participating nodes with low average odd ratio, weighted models are suitable, and for the nodes with high average odd ratio, unweighted methods are suitable. In short, we can summarize our contributions as follows:

- Propose min-flow and multiplicative weighting models.
- Investigate the effect of three weighting models; additive, min-flow and multiplicative on the prediction methods and datasets.
- Systematically study the effect of weight on prediction methods by introducing weighted links iteratively.
- Analyze the effect of two weight tuning methods and apply over RA.
- Present a node proximity based analysis of underlying graph to justify our results.
- Define degree odd-ratio and use it to propose a directive model for effective prediction.

The rest of the paper is organized as follows. Section 2 presents the existing node proximity based prediction methods and the proposed weighting models. The experimental datasets are discussed in Section 3. Sections 4–6 discuss different observations and analysis in details. Finally, Section 7 concludes the paper.

2. Prediction methods: weighted and unweighted

The prediction methods CN, JC, AA and RA explore the local proximity of two nodes to estimate a predicted score. A classical comparative study of various prediction methods (including the first three) has been reported by Liben-Nowell and Kleinberg in [2]. Later in 2009, the resource allocation measure has been introduced by Zhou et al. [7]. All these measures assign a positive score to a node pair, if and only if there is at least one 2-length path between the participating nodes, i.e., the participating nodes have at least one common neighbor. Among these four methods, RA is reported to perform better in several studies [7,11,13], and all these studies except [11] have considered only unweighted networks.

Study on the effect of tie weights over the local analysis based node proximity measures is still not explored much. The first such study has been presented by Murata et al. [10]. Authors have investigated the effect on three measures; CN, JC and AA using Yahoo! Chiebukuro social graph. Their results indicate a positive influence of tie weights on the link prediction. However in [11], the authors revisited the problem and observed conflicting results i.e., the performance of weighted measures of almost all proximity measures (CN, AA and RA) are worsen in all of the three datasets; USAir (US air transportation network), C.elegans (neural network of the nematode worms) and CGScience (co-authorship network of computational geometry). Lü et al. have further extended the study to investigate the role of weak ties and concluded that their results have been influenced by Granovetter's weak tie theory [8], i.e., weak ties play an important role in the information dissemination in social networks. In [12], the authors have not found significant improvement in performance, while experimenting with weighted co-authorship networks. However, the performance improved when they have applied supervised approach to the weighted measures. In another recent study [14], the authors have explored face to face interaction network among researchers and have observed that the weighted methods outperform their unweighted counterparts.

In all these studies, only an additive (linear summation) model has been used to incorporate weights. However, the additive model may not have equal effect on different prediction methods. Like existing studies, this paper also focuses on CN, JC, AA and RA, but investigates the responses of three different weighting models: (i) additive, (ii) min-flow and (iii) multiplicative.

2.1. Three weighting models

If x and y are the two participating nodes, the *additive* strength between x and y is bound by a common neighbor z, which is defined by a linear model $w(x,z)+w(z,y)$, where $w(-,-)$ is the symmetric edge weight connecting two nodes. If we assume that two nodes x and y have infinite supply of information through channels connecting them, $w(x,z)+w(z,y)$ represents the aggregate information received by node z from nodes x and y. The higher the volume of information z receives, the tighter is the bond that z holds between x and y. Fig. 1(a) shows the graphical representation of the additive model, where z acts as an information aggregator. Thickness of the edges represents the strength of the tie.

Unlike additive model, *min-flow* defines the bonding between x and y by the channel capacity, $\min(w(x,z), w(z,y))$, through z. Considering channels of different capacities, the information received by one node from another node is defined by the channel of lower capacity. Fig. 1(b) shows the graphical representation of min-flow measure, where z acts as a flow control node between x and y.

In *multiplicative* model, node z acts as a flow booster. The incoming flow is exaggerated by many folds defined by the outflow channel capacity i.e., $w(x,z) \times w(z,y)$. In the following subsections, we incorporate the above three weighting models with each of the prediction methods and define the weighted versions.

2.2. Prediction methods

In this section, we incorporate above three weighting models with each of the prediction methods and define their weighted estimates.

2.2.1. Common Neighbor (CN)

The idea behind the common neighbor index in a social network graph is that – if two actors (nodes) x and y have many friends in common, they are more likely to form a link in the future. If $\Gamma(x)$ and $\Gamma(y)$ denote the set of neighbors of x and y

[1] Nodes connected by strong ties are considered to belong to the same region or community and nodes connected by weak ties are considered to belong to different regions or communities.

[2] Ratio between unweighted and weighted.