Physica A 456 (2016) 294–301

Contents lists available at ScienceDirect

Physica A

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

Link prediction based on path entropy

Zhongqi Xu, Cunlai Pu *, Jian Yang

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

HIGHLIGHTS

- We study the information entropies of paths in networks.
- We propose a new similarity index based on the path entropy.
- Prediction accuracy of our index is higher than the mainstream of similarity indices.

ARTICLE INFO

Article history:
Received 21 December 2015
Received in revised form 8 March 2016
Available online 4 April 2016

Keywords:
Link prediction
Complex networks
Information entropy

ABSTRACT

Information theory has been taken as a prospective tool for quantifying the complexity of complex networks. In this paper, first we study the information entropy or uncertainty of a path using the information theory. After that, we apply the path entropy to the link prediction problem in real-world networks. Specifically, we propose a new similarity index, namely Path Entropy (PE) index, which considers the information entropies of shortest paths between node pairs with penalization to long paths. Empirical experiments demonstrate that PE index outperforms the mainstream of link predictors.

1. Introduction

Fundamental principles underlying various complex systems, such as social, biological, technological systems, have attracted lots of attention from the network science community over the past two decades [1–4]. It has been demonstrated that plenty of real-world networks have scale-free degree distributions [5–8], small-world effects [9–12], and high clustering properties [13,14]. Generally, social and collaboration networks are assortative mixing [15,16], while biological and technological networks are disassortative mixing [17,18]. Scale-free networks are very robust to random attacks, but are fragile to target attacks [19–22]. Also, scale-free networks facilitate epidemic spreading since the epidemic threshold for scale-free networks approximates to zero [23–27]. The deep understanding of network structures and dynamics helps us to make critical predictions of complex networks [28]. Link prediction [29–32] is to estimate the existence possibility of links between disconnected nodes based on the network structures, nodes attributes, and many others. Generally, there are two kinds of special links: one is the missing links in the current network, and the other one is the future links emerging in the evolution of the network. Link prediction has both scientific meaning and broad applications. On the one hand, prediction methods usually echo the fundamental organization rules of complex networks, and the prediction performance indicates the predictability of complex networks [33]. For example, common neighborhood (CN) based indices [34–36] are based on the high clustering property of complex networks. High prediction accuracy of CN-based indices indicates that the network has a strong clustering property and a large predictability. Preferential attachment (PA) index [5] reflects the rich-get-richer mechanism of social networks. In addition, link prediction provides us a natural standard for the comparison

* Correspondence to: 200 Xiaolingwei, Nanjing 210094, China. Tel.: +86 13915966537.
E-mail address: pucunlai@njust.edu.cn (C. Pu).

http://dx.doi.org/10.1016/j.physa.2016.03.091
0378-4371/

of various network models [29]. On the other hand, link prediction is widely used in various applications, for example discovering potential interactions in protein–protein interaction networks [37], recommending goods and friends in social networks [31,38], exploring coauthor relationships in collaboration networks [39], and even revealing hidden relations in terrorist networks [40].

Link prediction has long been discussed in computer science, but is booming recently in network science [29]. The reason is that structural similarity indices are generally simpler with lower computational cost than machine learning based prediction methods. Specifically, structural similarity based methods can be classified into three groups: local indices [34–36,5,41,42], global indices [43,44] and quasi-local indices [45,46]. Local indices are usually defined by using the knowledge of common neighbors and node degree, which include CN, PA, Adamic–Adar (AA) [41], resource allocation (RA) [42], etc. Global indices are defined based on the whole network topological information, such as Katz Index [43], Leicht–Holme–Newman (LHN) Index [44], and so on. Quasi-local indices are between local indices and global indices since the network topological information used in quasi-local indices is more than local indices, but less than global indices. Quasi-local indices contain local path (LP) index [45], local random walk (LRW) index [46], Superposed Random Walk (SRW) [46], etc. Generally, the prediction accuracy of local indices is the lowest among the three groups of indices. However, the computational cost of local indices is the smallest among three. Global indices are the opposite of local indices, while quasi-local indices fall in between them. In addition, information of hierarchical and community structures [47,48] has been referred to link prediction which further improves the prediction accuracy with additional computational cost.

Recently, information theory has been employed to quantify the complexity of complex networks structures with various scales [49,50]. The Von Neumann entropy [51] and Shannon entropy [49] of a network are defined respectively. Bauer et al. [52] used the maximum entropy principle in their construction of random graphs with arbitrary degree distribution. Bianconi [53] studied the entropy of randomized network ensembles and found that network ensembles with fixed scale-free degree distribution have smaller entropy than homogeneous degree distribution. She [54] further provided the expression of the entropy of multiplex networks ensembles. Halu et al. [55] further studied the maximal entropy ensembles of spatial multiplex and spatial interacting networks. Entropies of network dynamics such as diffusion process [56] and random walks [57] are also discussed. Network entropy measures have been applied to community detection [58], aging and cancer progression characterization [59], and very recently link prediction [60].

So far, the information entropy or uncertainty embodied in a path has not been explored specifically yet. In complex networks, heterogeneity of paths can be further quantified by the path entropy or uncertainty. With path entropy, we can study how the path heterogeneity affects network properties and dynamics. In this paper, first we study the path entropy and obtain an approximate expression of path entropy which is based on the entropies of links in the path. After that, we apply path entropy to the link prediction problems and propose a new similarity index based on path entropy. The outline of the article is as follows. Section 2 provides a detailed derivation of the entropy of a path. Section 3 gives the new similarity index. Section 4 introduces the basic link prediction framework and some traditional similarity indices which are used in our experiments for comparison purpose. Section 5 presents the experiment results, and finally Section 6 provides the conclusion.

2. Information entropy of a path

In information theory, the uncertainty of an event depends on the probability of its occurrence. Given an event Q with occurrence possibility $P(Q)$, its information entropy or uncertainty $I(Q)$ is defined as [61]:

$$I(Q) = -\log P(Q), \tag{1}$$

where the base of the logarithm is 2, the same in the following. Apparently, the larger occurrence possibility, the smaller entropy of event Q. For a node pair (a, b) in a network, let's denote L^1_{ab} (L^0_{ab}), which means that there is (not) a link between a and b. Assuming that there is no degree correlation among nodes in the network, the probability of L^1_{ab} is calculated as follows:

$$P(L^1_{ab}) = 1 - P(L^0_{ab}) = 1 - \prod_{i=1}^{k_b} \frac{(M-k_a)-i+1}{M-i+1} = 1 - \frac{C^{k_b}_{M-k_a}}{C^{k_b}_{M}}, \tag{2}$$

where k_a and k_b are the degrees of a and b. M is the number of edges in the network. $C^{k_b}_{M}$ denotes the number of candidate link sets for b, in which all links are incident with b. $C^{k_b}_{M-k_a}$ denotes the number of candidate link sets for b, in which all links are incident with b, but none of them is incident with a. Thus, $C^{k_b}_{M-k_a}/C^{k_b}_{M}$ estimates the possibility that node a and b are not connected by a link. Combing Eqs. (1) and (2), we get the entropy of L^1_{ab} as:

$$I(L^1_{ab}) = -\log(P(L^1_{ab})) = -\log\left(1 - \frac{C^{k_b}_{M-k_a}}{C^{k_b}_{M}}\right). \tag{3}$$

Through the above derivation, we infer that $I(L^1_{ab}) = I(L^1_{ba})$. Assuming the network is sparse, we have $M \gg k_{\max}$, where $k_{\max}$ is the maximum node degree. Then, let's consider a simple path $D = v_0 v_1 \cdots v_\delta$ of length δ. The occurrence probability of