Physica A 390 (2011) 1150–1170

Contents lists available at ScienceDirect

Physica A

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

Minireview

Link prediction in complex networks: A survey

Linyuan Lü [a,b,c], Tao Zhou [a,d,*]

[a] Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 610054, People's Republic of China
[b] Research Center for Complex System Science, University of Shanghai for Science and Technology, Shanghai 200093, People's Republic of China
[c] Department of Physics, University of Fribourg, Chemin du Musée 3, CH-1700 Fribourg, Switzerland
[d] Department of Modern Physics, University of Science and Technology of China, Hefei 230026, People's Republic of China

ARTICLE INFO

Article history:
Received 5 October 2010
Received in revised form 10 November 2010
Available online 2 December 2010

Keywords:
Link prediction
Complex networks
Node similarity
Maximum likelihood methods
Probabilistic models

ABSTRACT

Link prediction in complex networks has attracted increasing attention from both physical and computer science communities. The algorithms can be used to extract missing information, identify spurious interactions, evaluate network evolving mechanisms, and so on. This article summaries recent progress about link prediction algorithms, emphasizing on the contributions from physical perspectives and approaches, such as the random-walk-based methods and the maximum likelihood methods. We also introduce three typical applications: reconstruction of networks, evaluation of network evolving mechanism and classification of partially labeled networks. Finally, we introduce some applications and outline future challenges of link prediction algorithms.

Contents

* Corresponding author at: Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 610054, People's Republic of China.
E-mail addresses: linyuan.lue@unifr.ch (L. Lü), zhutou@ustc.edu (T. Zhou).

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.11.027

1. Introduction

Many social, biological, and information systems can be well described by networks, where nodes represent individuals, biological elements (proteins, genes, etc.), computers, web users, and so on, and links denote the relations or interactions between nodes. The study of complex networks has therefore become a common focus of many branches of science. Great efforts have been made to understand the evolution of networks [1,2], the relations between topologies and functions [3,4], and the network characteristics [5]. An important scientific issue relevant to network analysis is the so-called *information retrieval* [6,7], which aims at finding material of an unstructured nature that satisfies an information needed from large collections [8]. It can also be viewed as prediction of relations between words and documents and is now further extended to stand for a number of problems on link mining, wherein *link prediction* is the most fundamental problem that attempts to estimate the likelihood of the existence of a link between two nodes, based on observed links and the attributes of nodes [9].

In many biological networks, such as food webs, protein–protein interaction networks and metabolic networks, whether a link between two nodes exists must be demonstrated by field and/or laboratorial experiments, which are usually very costly. Our knowledge of these networks is very limited, for example, 80% of the molecular interactions in cells of Yeast [10] and 99.7% of human [11,12] are still unknown. Instead of blindly checking all possible interactions, to predict based on known interactions and focus on those links most likely to exist can sharply reduce the experimental costs if the predictions are accurate enough. Social network analysis also comes up against the missing data problem [13–15], where link prediction algorithms may play a role. In addition, the data in constructing biological and social networks may contain inaccurate information, resulting in spurious links [16,17]. Link prediction algorithms can be applied in identifying these spurious links [18]. Readers should be warned that some "unexpected" links may be incorrectly identified as spurious links and thus the removal of these links may lead to biased understanding of the system's structure and function. Actually, as we will show in Section 6.1, the method by Guimerà and Sales-Pardo [18] can find out most of the spurious links yet incorrectly remove some real links. As a whole we believe that these kinds of methods are helpful because the reconstructed network is shown to have closer functionality to the real network.

Besides helping in analyzing networks with missing data, the link prediction algorithms can be used to predict the links that may appear in the future of evolving networks. For example, in online social networks, very likely but not-yet-existent links can be recommended as promising friendships, which can help users in finding new friends and thus enhance their loyalties to the web sites. Similar techniques can be applied to evaluate the evolving mechanism for given networks. For example, many evolving models for the Internet topology have been proposed: some more accurately reproduce the degree distribution and the disassortative mixing pattern [19], some better characterize the k-core structure [20], and so on. Since there are too many topological features and it is very hard to put weights on them, we are not easy to judge which model (i.e., which evolving mechanism) is better than the others. Note that, each model in principle corresponds to a link prediction algorithm, and thus we can use the metrics on prediction accuracy to evaluate the performance of different models.

Link prediction problem is a long-standing challenge in modern information science, and a lot of algorithms based on Markov chains and statistical models have been proposed by computer science community. However, their works have not caught up the current progress of the study of complex networks, especially, they lack serious consideration of the structural characteristics of networks, like the hierarchical organization [21] and community structure [22], which may indeed provide useful information and insights for link prediction. Recently, some physical approaches, such as random walk processes and maximum likelihood methods, have found applications in link prediction. This article will give detailed discussion on these new developments.

This article is organized as follows. In the next section, we will present the link prediction problem and the standard metrics for performance evaluation. Our tour of link prediction algorithms starts with the mainstreaming class of algorithms, the so-called *similarity-based algorithms*,[1] which are further classified into three categories according to the information used by the similarity indices: local indices, global indices and quasi-local indices. In Sections 4 and 5, we introduce the maximum likelihood algorithms and probabilistic models for link prediction. The applications of link prediction algorithms are presented in Section 6, including the reconstruction of networks, the evaluation of network evolving mechanism and the classification of partially labeled networks. Finally, we outline some future challenges of link prediction algorithms.

2. Problem description and evaluation metrics

Consider an undirected network $G(V, E)$, where V is the set of nodes and E is the set of links. Multiple links and self-connections are not allowed.[2] Denote by U, the universal set containing all $\frac{|V|\cdot(|V|-1)}{2}$ possible links, where $|V|$ denotes the number of elements in set V. Then, the set of nonexistent links is $U - E$. We assume that there are some missing links (or the links that will appear in the future) in the set $U - E$, and the task of link prediction is to find out these links.

Generally, we do not know which links are the missing or future links, otherwise we do not need to do prediction. Therefore, to test the algorithm's accuracy, the observed links, E, is randomly divided into two parts: the training set, E^T,

[1] The similarity indices between nodes are also called kernels on graphs in some literature of computer science community [23].

[2] A network with multiple links can be represented by a weighted network where the weight of a link connecting two nodes equals the number of links between these two nodes [24]. We will discuss the problem of link prediction on weighted networks in Section 7.