Accepted Manuscript

A group evolving-based framework with perturbations for link prediction

Cuiqi Si, Licheng Jiao, Jianshe Wu, Jin Zhao

PII: S0378-4371(17)30114-0
DOI: http://dx.doi.org/10.1016/j.physa.2017.01.087
Reference: PHYSA 17980

To appear in: *Physica A*

Received date: 30 September 2016
Revised date: 29 December 2016

Please cite this article as: C. Si, L. Jiao, J. Wu, J. Zhao, A group evolving-based framework with perturbations for link prediction, *Physica A* (2017), http://dx.doi.org/10.1016/j.physa.2017.01.087

A group evolving-based framework with perturbations for link prediction☆

Cuiqi Si[a,1,*], Licheng Jiao[a], Jianshe Wu[a], Jin Zhao[a]

[a]*Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an, Shaanxi Province 710071, China*

Abstract

Link prediction is a ubiquitous application in many fields which uses partially observed information to predict absence or presence of links between node pairs. The group evolving study provides reasonable explanations on the behaviors of nodes, relations between nodes and community formation in a network. Possible events in group evolution include continuing, growing, splitting, forming and so on. The changes discovered in networks are to some extent the result of these events. In this work, we present a group evolving-based characterization of node's behaviorial patterns, and via which we can estimate the probability they tend to interact. In general, the primary aim of this paper is to offer a minimal toy model to detect missing links based on evolution of groups and give a simpler explanation on the rationality of the model. We first introduce perturbations into networks to obtain stable cluster structures, and the stable clusters determine the stability of each node. Then fluctuations, another node behavior, are assumed by the participation of each node to its own belonging group. Finally, we demonstrate that such characteristics allow us to predict link existence and propose a model for link prediction which outperforms many classical methods with a decreasing computational time in large scales. Encouraging experimental results obtained on real networks show that our approach can effectively predict missing links in network, and even when nearly 40% of the edges are missing, it also retains stationary performance.

Keywords: Link prediction, Group evolution, Perturbation, Sociability

1. Introduction

As a fundamental problem in complex network analysis, link prediction is universally used in many applications including predicting interactions between pairs of proteins and recommending friends in online social networks. It is worth noting that there exists two types of link prediction [1]: temporal and structural. In the first one, we have a sequence of fully observed graphs at several time stamps as input, and we wish to predict the graph state at the future time steps. In the other one, the input is a partially observed graph, and our goal is to discover the edge status for unobserved node pairs, especially the missing ones. This work focuses on the structural version of the problem, in which the likelihood of edge existence can be estimated. To be specific, consider an undirected simple network or graph $G(V,E)$, where V represents nodes and E the edges. We split the edge sets into two parts: the observed edges E^P, and the probe set E^T for testing, $E^P \cup E^T = E, E^P \cap E^T = \emptyset$. The universal set U contains all the possible links for any node pairs, $|U| = |V||V-1|/2$. With the known information, every node pair (x,y) in the non-observed set U/E^P is assigned a score s_{xy} to denote the possibility of edge existence. Take Fig 1 as an example, the true network has 5 nodes and 6 links, where the links (1, 3), (3, 4) are missing. Our task is to find out these missing links in set U/E^P.

This problem has been widely studied from diverse perspectives. One of the popular philosophy is neighborhood-based algorithm, which performs excellent in networks with large clustering coefficients. Yet in some real networks, the known information are limited, and this limitation affects the sparseness of the networks. The performance of neighborhood-based models is not effective enough for the clustering coefficients reduction caused by spasification

☆Fully documented templates are available in the elsarticle package on CTAN.

*Corresponding author

Email address: sqhrek@sina.com (Cuiqi Si)