

Directed Network Community Detection: A Popularity and Productivity Link Model

Tianbao Yang¹ Yun Chi² Shenghuo Zhu² Yihong Gong² Rong Jin¹

¹Department of Computer Science and Engineering, Michigan State University, MI 48824, USA

²NEC Laboratories America, 10080 N. Wolfe Rd, SW3-350, Cupertino, CA 95014, USA

¹{yangtia1,rongjin}@msu.edu, ²{ychi,zsh,ygong}@sv.nec-labs.com

Abstract

In this paper, we consider the problem of community detection in directed networks by using probabilistic models. Most existing probabilistic models for community detection are either *symmetric* in which incoming links and outgoing links are treated equally or *conditional* in which only one type (i.e., either incoming or outgoing) of links is modeled. We present a probabilistic model for directed network community detection that aims to model both incoming links and outgoing links *simultaneously* and *differentially*. In particular, we introduce latent variables *node productivity* and *node popularity* to explicitly capture outgoing links and incoming links, respectively. We demonstrate the generality of the proposed framework by showing that both symmetric models and conditional models for community detection can be derived from the proposed framework as special cases, leading to better understanding of the existing models. We derive efficient EM algorithms for computing the maximum likelihood solutions to the proposed models. Extensive empirical studies verify the effectiveness of the new models as well as the insights obtained from the unified framework.

keywords: community detection, popularity, productivity, stochastic block model, directed network

1 Introduction

Community detection is an important topic in analyzing networked data because it reveals underlying structures in a complex network, a key to the network analysis. A community can be intuitively considered as a set of nodes that are densely connected with each other while sparsely connected with other nodes in the network. Based on this intuition, many previous studies (e.g., [7, 13, 15]) focused on defining appropriate metrics to quantify the connection and efficient algorithms to optimize the defined metrics. These approaches usually rely on some heuristics and lack a rigorous mathematical model.

More recently, various probabilistic models have been proposed for community detection. Among them, stochastic block models [10, 14, 17, 3, 16, 2, 12] are probably the most successful ones in terms of capturing meaningful communities, producing good performance, and offering probabilistic interpretations. The basic idea is first to define a generative process where links are generated based on latent community memberships of nodes, and then to infer the community memberships from the links by either maximizing the data likelihood or computing the posterior distribution for community memberships.

Most stochastic block models can be classified into two categories: the *symmetric* approaches [14, 17] that model links by symmetric joint probabilities, and the *conditional* approaches [3, 16] that focus on the conditional probability of receiving links. Neither of these models is satisfying: a symmetric model misses the semantics of link directions, a key factor that distinguishes directed networks from undirected networks; a conditional model only captures one type of links, either incoming links or outgoing links, and therefore is unable to characterize nodes in a full spectrum. As an example, in a blog readership network, there are two types of bloggers: “writers” who generate influential blogs read by many, and “readers” who read a lot but seldom write anything for others to read. Evidently, to characterize these two types of bloggers, it is important to examine both incoming links and outgoing links of the network.

In this work, we propose a novel probabilistic framework for directed network community detection, termed **Popularity and Productivity Link** model or **PPL** for short, that explicitly addresses the shortcomings of the existing stochastic block models. In particular, we model both outgoing links and incoming links by the introduction of the latent variables *productivity* and *popularity*. We demonstrate the generality of the proposed framework by showing that both the symmetric models and the conditional models can be derived from the pro-

