Statistica Sinica **35** (2025), 1-21
doi:https://doi.org/10.5705/ss.202022.0372

COMMUNITY EXTRACTION OF NETWORK DATA UNDER STOCHASTIC BLOCK MODELS

Quan Yuan[1], Binghui Liu*[1], Danning Li*[1] and Yanyuan Ma[2]

[1]*Northeast Normal University and* [2]*Pennsylvania State University*

Abstract: Most existing community discovery methods focus on partitioning all nodes of the network into communities. However, many real networks contain background nodes that do not belong to any community. In such a situation, typical methods tend to artificially split the background nodes and group them together with communities with relatively stronger connection, hence lead to distorted results. To avoid this, some community extraction methods have been developed to achieve community discovery with background nodes, which are based on searching algorithms, hence have difficulties in handling large-scale networks due to high computational complexity. To this end, in this paper we propose some algorithms with polynomial complexity to achieve community extraction of large-scale networks. We rigorously show that the proposed algorithms have attractive theoretical properties. In particular, the estimators of the community labels using the proposed algorithms reaches the asymptotic minimax risk under the community extraction model, a specific stochastic block model. Then, we illustrate the advantages and feasibility of the proposed algorithms via extensive simulated networks and a political blog network.

Key words and phrases: Background nodes, community extraction, refinement algorithm.

1. Introduction

Networks are widely used to represent and analyze the relationship between interacting units in complex systems (Goldenberg et al., 2010; Wasserman and Faust, 1994). In network data analysis, community discovery is a fundamental problem, which aims to divide the nodes of the network into communities, so that the nodes in the same community are closely connected, while the nodes from different communities are loosely connected. Identifying communities can provide important insights into network organizations. There is a large number of literature on community discovery from different research fields, such as computer science (Flake et al., 2002), social science (Moody and White, 2003), and genetics (Spirin and Mirny, 2003). We refer to Fortunato (2010), Fortunato and Hric (2016), and Zhao (2017) for comprehensive reviews on this topic.

Most literatures on community discovery study the problem without "background nodes", where the background nodes are defined as the weakly connected

*Corresponding author. E-mail: (Liu) liubh100@nenu.edu.cn, (Li) lidn040@nenu.edu.cn.