Statistica Sinica **35** (2025), 1-22
doi:https://doi.org/10.5705/ss.202022.0341

PERFECT SPECTRAL CLUSTERING WITH DISCRETE COVARIATES

Jonathan Hehir, Xiaoyue Niu* and Aleksandra Slavković

Penn State University

Abstract: Among community detection methods, spectral clustering enjoys two desirable properties: computational efficiency and theoretical guarantees of consistency. While most studies of spectral clustering consider only the edges of a network as input to the algorithm, we consider the problem of performing community detection in the presence of discrete node covariates, with network structure determined by a combination of a latent block model structure and homophily on the observed covariates. We propose a spectral algorithm that we prove achieves perfect clustering with high probability on a class of large, sparse networks with discrete covariates, effectively separating latent network structure from homophily on observed covariates. We apply this method to a network of online friendships among university students to uncover community structure not explained by covariates. To our knowledge, our method is the first to offer a guarantee of consistent latent structure recovery using spectral clustering in the setting where edge formation is dependent on both latent and observed factors.

Key words and phrases: Community detection, homophily, spectral clustering, stochastic block model.

1. Introduction

A structural pattern commonly observed in social networks is *homophily*, the tendency for two nodes sharing a certain trait to be more (or sometimes less) likely to form a connection (McPherson, Smith-Lovin and Cook, 2001). Homophily may occur on any number of traits, observed or latent, and is known to confound problems of causal inference in the social sciences (Smith and Christakis, 2008; Shalizi and Thomas, 2011; Goldsmith-Pinkham and Imbens, 2013; Lee and Ogburn, 2021). Homophily, meanwhile, lies at the heart of such issues as segregation (Shrum, Cheek Jr. and Hunter, 1988; Henry, Prałat and Zhang, 2011), job access (Ibarra, 1992), and political partisanship (Huber and Malhotra, 2017), where homophily on observed traits may be the subject of estimation in its own right. In order to fully understand the effects of network patterns like observed homophily, we first need to separate them from further latent network structure.

In the literature on community detection, latent structure is frequently

*Corresponding author. E-mail: xiaoyue@psu.edu