Modeling Homophily in ERGMs for Bipartite Networks

Rashmi P. Bomiriya[1], Shweta Bansal[2] and **David R. Hunter**[3*]

[1] *Remote Sensing Metrics Asia (Pvt) Ltd.; rashmi.bomiriya@gmail.com.*
[2] *Georgetown University; shweta@sbansal.com.*
[3] *Pennsylvania State University; dhunter@stat.psu.edu*
[*] *Presenting author*

Keywords. *Exponential-family random graph model; Bipartite network; Homophily*

Bipartite networks, in which edges only exist between two disjoint sets of nodes, represent an important tool for modeling processes such as affiliations, collaborations, and co-location. Frequently, we would like to model the propensity of similar nodes to form links among themselves, a property referred to as homophily. Modeling homophily in a bipartite network is complicated by the prohibition of direct ties between nodes in the same subset. This paper introduces a method for modeling homophily in the commonly used exponential-family random graph model (ERGM) framework.

If we allow the $n \times n$ matrix Y to encode the status of all the binary edges of the network, where Y_{ij} equals 0 or 1 according to whether the (i, j)th edge is absent or present, then the basic ERGM may be written

$$P_\theta(Y = y) = \frac{\exp\{\sum_{i=1}^{p} \theta_i s_i(y)\}}{\kappa(\theta)}, \quad y \in \mathcal{Y}, \tag{1}$$

where $s_1(y), \ldots, s_p(y)$ are user-defined statistics measured on the network y and we denote the vector of all network statistics by $s(y)$. When covariates X should be included in the model, we may add X to the notation and write $s(y, X)$, where we allow these statistics to depend on any available known covariates. The parameters $\theta_1, \ldots, \theta_p$ are the corresponding unknown coefficients to be estimated, $\mathcal{Y}$ is the set of all allowable networks, and $\kappa(\theta)$ is merely a normalizer necessary to ensure that Equation (**??**) defines a legitimate probability distribution.

We argue that the "natural" approach to modeling homophily in a bipartite network, in which we incorporate the total count of matching two-stars into an ERGM as one of the $s(y, X)$ statistics, might be problematic due to potential degeneracy issues. We introduce a new set of model terms for the ergm package in R designed to model homophily while mitigating such issues. We demonstrate that these model

terms can be expressed in a curved exponential family form, then discuss real-world applications.