This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2017-7550C

# A Community and Node Attribute-Corrected Stochastic Blockmodel

Kristen M. Altenburger
Stanford University
kaltenb@stanford.edu

W. Philip Kegelmeyer, Ali Pinar, and Jeremy D. Wendt
Sandia National Laboratories
{wpk,apinar,jdwendt}@sandia.gov

## The Problem: A Statistical Network Model

Individual preferences and group-level structures are common in social networks as characterized by homophily and communities respectively. Given an observed social network with known attribute labels and an inferred community structure, our aim is to develop a statistical network model that preserves the joint structure of observed homophilous interactions and estimated community structures. We propose the community and node-attribute corrected stochastic blockmodel (canacSBM).

## The Set-up: Node- vs. Community-level Network Structure

Given G=(V,E) with |V|=N, we suppose there are known binary categorical attributes (●, ●) and an underlying community structure denoted by $\{\hat{c}_1,\hat{c}_2,\ldots,\hat{c}_N\}$ where we define the following notation and measures:

### Notation:

$\{a_1,a_2,\ldots,a_N\}$ observed attributes

$\{\hat{c}_1,\hat{c}_2,\ldots,\hat{c}_N\}$ estimated community

$d_i$ node i's total degree

$d_{i,\hat{c}_j}$ degree within community *j*

$d_{i,\hat{c}_j,a_i}$ degree with community *j* and attribute $a_i$

### Community-Homophily Index:

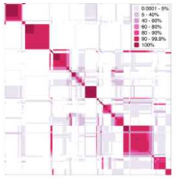
Based on an adaptation of the conventional homophily index [1] to account for community structure [2], we utilize the following community-homophily index, as illustrated below for a specific community and attribute value:

$$\hat{h}_{\hat{c}_i,a_i} := \frac{\Sigma_{k\in\hat{c}_i,a_i} d_{k,\hat{c}_i,a_i}}{\Sigma_{k\in\hat{c}_i,a_i} d_{k,\hat{c}_i}}$$

Supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

## The Proposed Model: canacSBM

**(1)**



**(2)**

Community Degree Sequence:

$$\{\{d_{1,\hat{c}_j}\},\{d_{2,\hat{c}_j}\},\ldots,\{d_{N,\hat{c}_j}\}\}$$

Community-Attribute Degree Sequence:

$$\{\{d_{1,\hat{c}_j,a_k}\},\{d_{2,\hat{c}_j,a_k}\},\ldots,\{d_{N,\hat{c}_j,a_k}\}\}$$

**(3)**

Compute edge probabilities via a Chung-Lu approach (based on [3,4]) but now accounting for the community structure.

## Analytical canacSBM Properties

The following analytical properties exist in expectation across the ensemble of graph instances created from canacSBM, assuming the community structure. We state the properties here and refer the interested reader to our paper draft for the details. Following notation in [5], let <x> represent the average of x in the ensemble of synthetic canacSBM graphs:

- Preserve expected within-community degrees.

$$\sum_{k\in\hat{c}_i}\langle A_{ik}\rangle = d_{i\hat{c}_i}$$

- Preserve expected between-community degrees

$$\sum_{k\in\hat{c}_j}\langle A_{ik}\rangle = d_{i\hat{c}_i}\cdot\frac{\sum_{i\in\hat{c}_i,i\in a_i} d_{i\hat{c}_j}}{\sum_{i\in\hat{c}_i,i\in a_i} d_{i\hat{c}_i}}$$

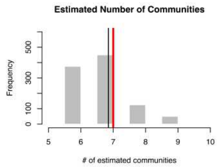
- Preserve attribute-specified edge counts

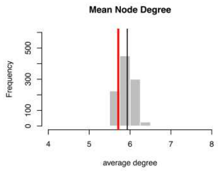
$$\sum_{i\in\hat{c}_i j\in\hat{c}_j}\langle A_{ij}\rangle\delta_{a_i r}\delta_{a_i s} = \sum_{k\in\hat{c}_i,k\in r} d_{k\hat{c}_j s}$$

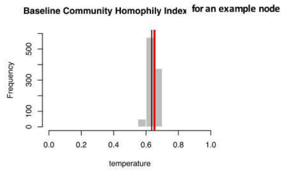
- Preserve the observed community-homophily index as a result of preserving the expected within-community degrees and the attribute-specified edge counts.

## Empirical Results

Our source dataset is a two hop ego network derived from crawling the web. The binary categorical attribute is based on a node metadata property. We evaluate the distribution of 3 network statistics across 1,000 simulated synthetic networks from canacSBM. For the community and attribute value associated with a sample node selected for illustration, we observe its community homophily index is preserved.


Estimated Number of Communities


Mean Node Degree


Baseline Community Homophily Index for an example node

## Open Directions

Time-varying networks? Overlapping communities?
Missing attribute values? Multi-level covariates?

### References

[1] James Coleman. 1958. Relational analysis: The study of social organizations with survey methods. *Human organization* 17, 4 (1958), 28-36.

[2] KM Altenburger. Measurement Issues with Homophily: The Impact of Group Structures. *Working Paper* (2017).

[3] Fan Chung and Linyuan Lu. 2002. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics* 6, 2 (2002), 125-145.

[4] KM Altenburger and J Ugander. Bias and Variance in the Social Structure of gender. *Working Paper* (2017).

[5] Brian Karrer and MEJ Newman. 2011. Stochastic blockmodels and community structure in networks. *Physical Review E* 83, 1 (2011).