SFU
SIMON FRASER
UNIVERSITY

Bayesian tools for Aggregated Network Data

Owen G. Ward
Department of Statistics and Actuarial Science, Simon Fraser University

SSC 2025

SFU

- Sociologists often want to estimate the size of hidden subpopulations in a population/social network, individual degree of nodes in network

Motivation

- Sociologists often want to estimate the size of hidden subpopulations in a population/social network, individual degree of nodes in network
- Standard procedure to do this is to collect **A**ggregated **R**elational **D**ata, many Bayesian models for such data

SFU

- Sociologists often want to estimate the size of hidden subpopulations in a population/social network, individual degree of nodes in network
- Standard procedure to do this is to collect **A**ggregated **R**elational **D**ata, many Bayesian models for such data
- Unclear how to choose an appropriate model, compare potential models

Outline

- A brief review of Statistical Network Analysis

Outline

SFU

- A brief review of Statistical Network Analysis
- Aggregated Relational Data (ARD) for Networks

Outline

- A brief review of Statistical Network Analysis
- Aggregated Relational Data (ARD) for Networks
- Bayesian Models for ARD

Outline

SFU

- A brief review of Statistical Network Analysis
- Aggregated Relational Data (ARD) for Networks
- Bayesian Models for ARD
- Model checking/comparison for ARD models

Outline

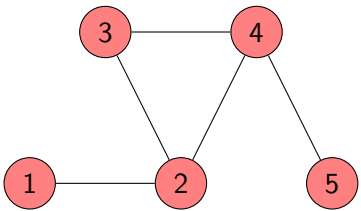
SFU

- A brief review of Statistical Network Analysis
- Aggregated Relational Data (ARD) for Networks
- Bayesian Models for ARD
- Model checking/comparison for ARD models
  - This is ongoing work, feedback welcome!

Overview of network data

- Observe (binary) relational data about pairs of nodes i and j in a network
- Relationship might correspond to friendship, social media interactions, citations, etc
- Represent this using an $N \times N$ adjacency matrix A for a network of N nodes

- Observe (binary) relational data about pairs of nodes i and j in a network
- Relationship might correspond to friendship, social media interactions, citations, etc
- Represent this using an $N \times N$ adjacency matrix A for a network of N nodes



A simple (undirected) network

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

The corresponding (symmetric) adjacency matrix

SFU

- Suppose you wanted to estimate how many people were in a certain subpopulation in your population, N_u
- Could be a hard to reach population such as homeless, drug users, etc
- If you knew the degree $d_i = \sum_j A_{ij}$ for sample of size n, then for each person ask how many people they knew in the subpopulation, y_{iu}
- An estimate for N_u is

$$\widehat{N}_u = \frac{N}{n} \sum_{i=1}^{n} \frac{y_{iu}}{d_i},$$

known as the Network Scale Up Estimate (NSUM) (Bernard et al., 1991)

Computing this

$$\widehat{N}_u = \frac{N}{n} \sum_{i=1}^{n} \frac{y_{iu}}{d_i},$$

- Requires knowing the degree d_i, hard/expensive to estimate

SFU

$$\widehat{N}_u = \frac{N}{n} \sum_{i=1}^{n} \frac{y_{iu}}{d_i},$$

- Requires knowing the degree d_i, hard/expensive to estimate
- Sociologists devised better ways to collect data to solve this

Computing this

$$\widehat{N}_u = \frac{N}{n} \sum_{i=1}^{n} \frac{y_{iu}}{d_i},$$

- Requires knowing the degree d_i, hard/expensive to estimate
- Sociologists devised better ways to collect data to solve this
- Ask people "how many X do you know?" about *multiple* subpopulations in the population, some of interest, *some we know the size of already*

SFU

$$\widehat{N}_u = \frac{N}{n} \sum_{i=1}^{n} \frac{y_{iu}}{d_i},$$

- Requires knowing the degree d_i, hard/expensive to estimate
- Sociologists devised better ways to collect data to solve this
- Ask people "how many X do you know?" about *multiple* subpopulations in the population, some of interest, *some we know the size of already*
- Gives **Aggregated Relational Data**

Aggregated Relational Data

SFU

- Ask for some aggregate count information instead
  - Surveys consist of questions of the form "How many X's do you know?"
  - X is some subpopulation in the network
  - The choice of these subpopulations is important

SFU

- Ask for some aggregate count information instead
  - Surveys consist of questions of the form "How many X's do you know?"
  - X is some subpopulation in the network
  - The choice of these subpopulations is important

"For the purposes of this study, the definition of knowing someone is that you know them and they know you by sight or by name, that you could contact them, that they live within the United States, and that there has been some contact (either in person, by telephone or mail) in the past 2 years."

SFU

- Get this for n nodes about K subpopulations giving an $n \times K$ **ARD** matrix Y
- Y_{ik} the number of people node i knows in subpopulation k
- People can more reliably recall these quantities than individuals (still issues)
- Ask about known subpopulations (e.g. people named Alice, Pilots)

SFU

- Get this for n nodes about K subpopulations giving an $n \times K$ **ARD** matrix Y
- Y_{ik} the number of people node i knows in subpopulation k
- People can more reliably recall these quantities than individuals (still issues)
- Ask about known subpopulations (e.g. people named Alice, Pilots)

$$\widehat{N}_u = \frac{N}{n} \sum_{i=1}^{n} \frac{y_{iu}}{\hat{d}_i},$$

Estimate d_i using known subpopulations

SFU

- Estimating hard to reach populations, such as the number of people killed in an earthquake (Laga et al., 2021)
- Estimating personal network size (McCormick et al., 2010)
- Replicate more expensive economic experiments without need to collect entire network (Breza, Chandrasekhar, McCormick, et al., 2020)
- Most recent models for this data have been Bayesian

Bayesian ARD models

SFU

- Bayesian approaches for modeling this data popular (in statistics)
- Joint estimation of the degrees and subpopulation sizes, allow more variation in responses
- Share information across the data, provide a more principled way of estimating uncertainty

SFU

- First Bayesian ARD model was Zheng et al. (2006)
- $n \times K$ ARD matrix Y, Y_{ik} is the number of people person i knows in subpopulation k
- Model this with

$$y_{ik} \sim \text{Poisson}(e^{\alpha_i + \beta_k + \gamma_{ik}})$$

- First Bayesian ARD model was Zheng et al. (2006)
- $n \times K$ ARD matrix Y, Y_{ik} is the number of people person i knows in subpopulation k
- Model this with

$$y_{ik} \sim \text{Poisson}(e^{\alpha_i+\beta_k+\gamma_{ik}})$$

  - $d_i = \exp(\alpha_i)$ is the expected degree of person i

SFU

- First Bayesian ARD model was Zheng et al. (2006)
- $n \times K$ ARD matrix Y, Y_{ik} is the number of people person i knows in subpopulation k
- Model this with

$$y_{ik} \sim \text{Poisson}(e^{\alpha_i+\beta_k+\gamma_{ik}})$$

  - $d_i = \exp(\alpha_i)$ is the expected degree of person i
  - $b_k = \exp(\beta_k)$ is the relative size of subpopulation k in the population

SFU

- First Bayesian ARD model was Zheng et al. (2006)
- $n \times K$ ARD matrix Y, Y_{ik} is the number of people person i knows in subpopulation k
- Model this with

$$y_{ik} \sim \text{Poisson}(e^{\alpha_i+\beta_k+\gamma_{ik}})$$

  - $d_i = \exp(\alpha_i)$ is the expected degree of person i
  - $b_k = \exp(\beta_k)$ is the relative size of subpopulation k in the population
  - $g_{ik} = \exp(\gamma_{ik})$ captures variation due to interaction between node and subpopulation, related to overdispersion

Other Bayesian Models

The *i*th respondent's likelihood to know someone in subpopulation *k* depends…

Zheng, Salganik, and Gelman (2006)

$$Y_{ik} \sim \text{Poisson}(d_i b_k g_{ik})$$
$$g_{ik} \sim \text{Gamma}(1, 1/(\omega_k - 1))$$

or equivalently,

$$Y_{ik} \sim \begin{matrix}\text{Overdispersed} \\ \text{Neg. Binomial}\end{matrix}\left(\begin{matrix}\text{mean} = d_i b_k, \\ \text{overdispersion} = \omega_k\end{matrix}\right)$$

*…**and** may vary for certain combinations of individuals and subpopulations…*

…based on predefined discrete demographic categories (e.g., gender, age)

McCormick, Salganik, and Zheng (2010)

$$Y_{ik} \sim \begin{matrix}\text{Overdispersed} \\ \text{Neg. Binomial}\end{matrix}\left(d_i \sum_{a=1}^{A} \frac{N_{ak}}{N_a} m(e,a)\right)$$

g_{ik}: discrete mixing matrix

*…**or** smoothly, via kernels for continuous demographic variables (e.g., age)*

Sahai et al. (2019)

$$Y_{ik} \sim \begin{matrix}\text{Overdispersed} \\ \text{Neg. Binomial}\end{matrix} \; d_i \sum_{g_j} \rho_{g_i g_j} \left(\sum_a \frac{N_{g_j a} k}{N_{g_j a}}\right) \times N(a_i | \mu_{kg_j}, \lambda_{g_i g_j}(a_i) + \sigma^2_{kg_j})$$

g_{ik}: smooth kernels

*…**or** based on (average) proximity between individuals and subpopulations in an unobserved latent social space*

McCormick & Zheng (2015)

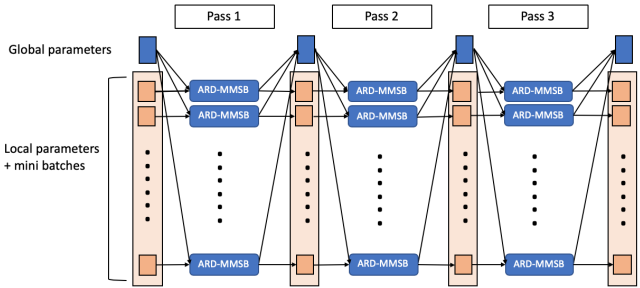
$$Y_{ik} \sim \text{Poisson}(d_i b_k \kappa(\zeta, \eta_k, \theta_{(z_i, v_k)}))$$

g_{ik}: latent space

SFU

- These appear to be straightforward models, can be fit with software such as Stan
- In practice lots of model fitting issues which need to be considered (such as scaling to known populations)
- More complex models for large ARD require other inference methods, such as Variational Inference (Jones et al., 2025)

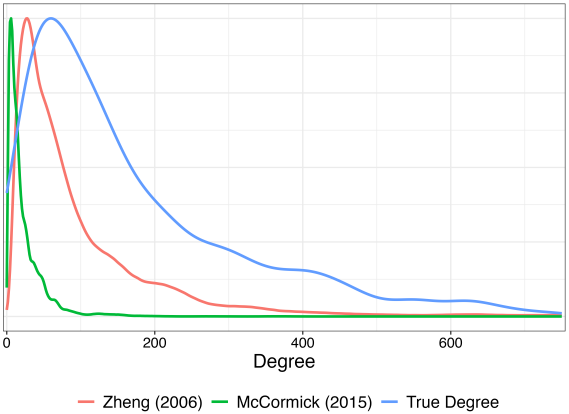
- Traditional ARD small, can fit these models locally (Stan)
- ARD has also been used for complete networks, requires scalable computation
- Tools like approximate inference needed (Jones et al., 2025)



Scaling inference for massive ARD (Jones et al., 2025)

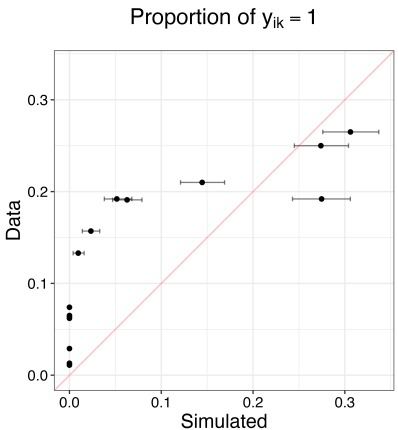
Comparing Bayesian ARD models

- Limited existing work on how to evaluate the results of fitted ARD models, particularly to real data
- For simulated data, can compare to true values such as true node degrees, true subpopulation sizes



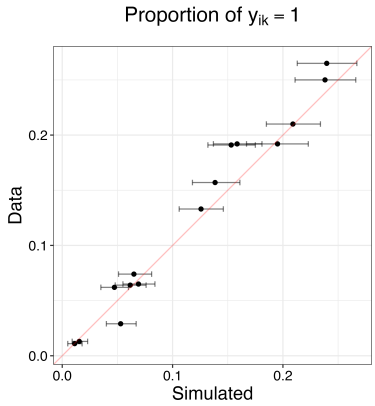
Estimated degree distributions for simulated data

- Only existing model checking technique for real data is to do Posterior Predictive Checking (PPC) (Zheng et al., 2006)
- For each set of posterior draws, simulate ARD and compare to truth
- Can look at proportion of $y_{ik} = 1$ for each subpopulation in simulated, compared to true Y

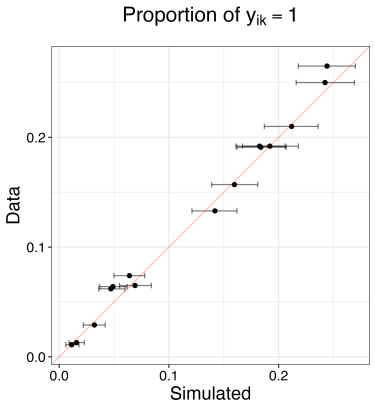


PPC for clearly incorrect model.

Works well when the model is clearly wrong, but less useful for comparing/choosing between two reasonable models

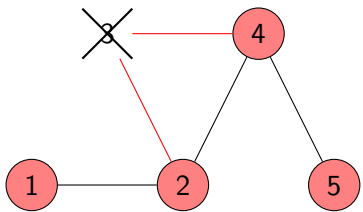


PPC for Zheng et al. (2006)



PPC for McCormick et al. (2015)

- Cross validation natural way to try compare models
- This is hard to do for traditional network data, not clear how to best preserve network structure if you sample on edges/nodes
- Important recent work (Li et al., 2020; Chakrabarty et al., 2025)



Removing a node impacts other nodes

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Edges removed if you remove a node

SFU

- Reasonable to apply CV to ARD directly, as not using the underlying network to fit model
- Initial simulations indicate it works well for recovering true model with simulated data

Model	elpd_diff	se_diff
McCormick(2015)	0.000e+00	0.000e+00
Zheng(2006)	-4.9550e+03	1.00e+02
Null Model	-1.127526e+05	3.8539e+03

CV for Expected Log Predictive Density (Vehtari et al., 2017)

SFU

- Recent work has examined how/if underlying network models such as Stochastic Block Model/Latent Space Model can be estimated from ARD (Breza, Chandrasekhar, Lubold, et al., 2023)
- Of interest to examine connection between CV for ARD and underlying network models
- What tradeoffs to doing network model selection using ARD, how best to construct this ARD?

SFU

- Bayesian ARD models widely used
- Limited tools for model checking/comparison
- Can tools for ARD be extended to full network models?

Jones, Timothy, Owen G Ward, Yiran Jiang, John Paisley, and Tian Zheng (2025). "Scalable Community Detection in Massive Networks using Aggregated Relational Data". In: *Statistica Sinica*.

Ward, Owen G, Anna L Smith, and Tian Zheng (2025). "Bayesian Models for Aggregated Relational Data: A Unified Perspective". In: *In preparation*.