

Bayesian tools for Aggregated Network Data

Owen G. Ward

Department of Statistics and Actuarial Science, Simon Fraser University

WNAR 2025

- Interested in identifying community structure in academic papers using citation networks (paper i cites paper j), represent as a network with entries A_{ij}
- This is a network of more than 2 million nodes, from more than 4k journals
- Want to identify latent structure of these papers, what sort of community structure is present



- Suitable model is the Mixed-Membership Stochastic Block Model (Airoldi et al., 2008), overlapping community structure
- κ underlying latent communities, each node has some affiliation to each community
- The probability of an edge between any two nodes depends on their community assignment for that interaction (nodes can be in different communities for different interactions), probability matrix B
- Gopalan and Blei (2013) fit this model using Stochastic Variational Inference

- **Aggregate Relational Data (ARD)** a way to summarize network information
- Rather than observing adjacency matrix A_{ij} instead observe aggregate counts y_{ik} where

y_{ik} = How many people does node i know in subpopulation k

- Corresponds to summing across certain columns of network A

$$y_{ik} = \sum_{j \in G_k} A_{ij}$$

- Widely used to estimate size of hard to reach subpopulations (Laga, Bao, and Niu, 2021)

- ARD used to estimate some existing network models (Breza et al., 2023)
- Can we use ARD to fit the MMSBM for this large citation network?
- Key requirement to do this in general is a way to construct appropriate subpopulations

- ARD used to estimate some existing network models (Breza et al., 2023)
- Can we use ARD to fit the MMSBM for this large citation network?
- Key requirement to do this in general is a way to construct appropriate subpopulations
 - For citation network natural subpopulation corresponding to journals!

- ARD used to estimate some existing network models (Breza et al., 2023)
- Can we use ARD to fit the MMSBM for this large citation network?
- Key requirement to do this in general is a way to construct appropriate subpopulations
 - For citation network natural subpopulation corresponding to journals!
 - In general unclear how best to do this

- ARD used to estimate some existing network models (Breza et al., 2023)
- Can we use ARD to fit the MMSBM for this large citation network?
- Key requirement to do this in general is a way to construct appropriate subpopulations
 - For citation network natural subpopulation corresponding to journals!
 - In general unclear how best to do this
 - Subpopulations not the same as communities we want to estimate


- y_{ik} number of edges between node i and subpopulation k
 - How many times does paper i cite a paper in journal k
- Under the MMSBM model

$$P(A_{ij} = 1 | \pi_i, \pi_j) = \pi_i^T B \pi_j$$


- If knew membership vectors π_i, π_j could approximate $y_{ik} \sim \text{Poisson}(\lambda_{ik})$,

$$\lambda_{ik} = \sum_{j \in G_k} P(A_{ij} = 1 | \pi_i, \pi_j) = \sum_{j \in G_k} \pi_i^T B \pi_j,$$

Member of
subpopulation k



- Can't estimate π_j for nodes in G_k , only have y_{ik}
- Instead introduce a distribution P_k over the membership vectors $\pi_j \in G_k$
- P_k is a distribution for node in subpopulation k **across the latent communities**
- Then approximate

Size of the subpopulation 

$$\frac{1}{N_k} \sum_{j \in G_k} \pi_i^T B \pi_j \approx \mathbb{E}_{\pi_j \sim P_k} (\pi_i^T B \pi_j)$$

- This gives

$$\lambda_{ik} \approx N_k \mathbb{E}_{\pi_j \sim P_k} (\pi_i^T B \pi_j)$$

- Take P_k to be Dirichlet distribution then

$$\mathbb{E}_{\pi_j \sim P_k} (\pi_i^T B \pi_j) = \pi_i^T B \eta_k,$$

with η_k the subpopulation mean vector for P_k

Final model

$$y_{ik} | \pi_i, \eta_k, B \sim \text{Poisson}(N_k \pi_i^T B \eta_k)$$

for each node i and subpopulation k

- N_k size of the subpopulation
- π_i affiliation of the node to the communities
- B community connection probabilities
- η_k mean of distribution of subpopulation over the communities

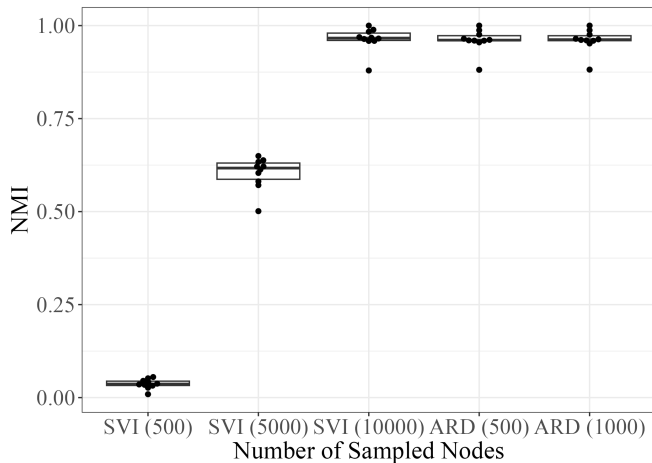
- Fit this using Variational Inference (Blei et al., 2017)
- Choose mean field distribution q for parameters, estimate the parameters of these families to minimize

$$\mathcal{L} = \mathbb{E}_q \log p(Y, \pi, B, \eta) - \mathbb{E}_q \log q(\pi, B, \eta)$$

- Take samples from ARD, use these minibatches to update parameters

Gopalan and Blei (2013) comparable method for large network data

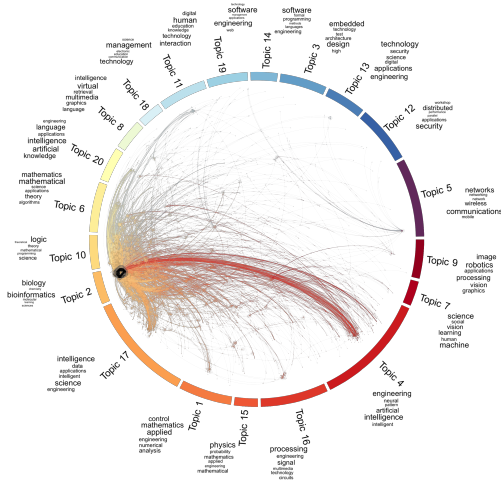
- How much of the network do they need to observe to recover communities?
- What about if we use ARD (sample) instead?
- Predictive performance, computational time, robustness?



Community Recovery (SVI = Gopalan and Blei (2013))

- Better recover communities and B for simulated data
- Our procedure performs better for sparser networks
- Faster for the same amount of data

- Fit with 20 communities
- Show papers in journal along with papers cited by those papers
- Fitting Gopalan and Blei (2013) gives communities which are hard to interpret



Many future directions to consider

- For the citation example natural choice of subpopulations (journals)
- Unclear how to best choose the subpopulations for (human) networks
- Limited work on model checking for network models in general
 - Can we develop tools for models for ARD?
 - Starting with existing Bayesian model checking techniques

- Used ARD to estimate underlying network community structure
- Allows us to scale this problem to massive networks
- More work needed to see when we don't need complete network data

- Used ARD to estimate underlying network community structure
- Allows us to scale this problem to massive networks
- More work needed to see when we don't need complete network data



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*.