SFU

SIMON FRASER
UNIVERSITY

Scalable Bayesian computation for networks utilising Aggregated Relational Data

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

ICSDS 2025

Real data motivation

- 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, community structure present

Outline

- The model we want to fit to this form of network data

Outline

- The model we want to fit to this form of network data
- Aggregated Relational Data

Outline

- The model we want to fit to this form of network data
- Aggregated Relational Data
- Fitting this network model without the network

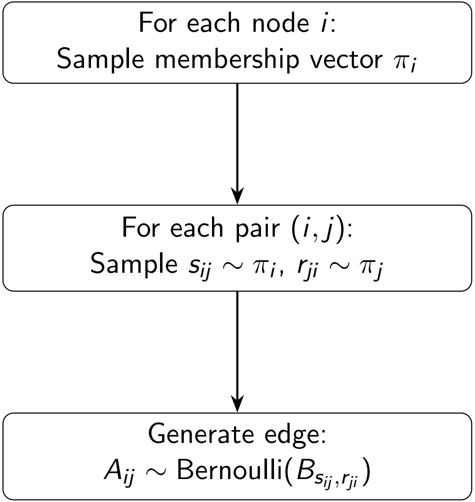
Outline

- The model we want to fit to this form of network data
- Aggregated Relational Data
- Fitting this network model without the network
- Simulations, real data results

A model for latent structure

- Want to identify community structure in citation network
- Suitable model is the Mixed-Membership Stochastic Block Model (Airoldi et al., 2008), overlapping community structure
- Airoldi et al. (2008) originally fit MMSBM using a Variational EM approach
  - Assume variational families for latent parameters
  - Optimize over parameters of these families to minimize distance from true posterior
- Fit this to a network with hundreds of nodes

Data generating procedure



For each node i:
Sample membership vector π_i

For each pair (i, j):
Sample $s_{ij} \sim \pi_i$, $r_{ji} \sim \pi_j$

Generate edge:
$A_{ij} \sim \text{Bernoulli}(B_{s_{ij}, r_{ji}})$

- κ latent communities in network
- π_i a κ dimensional membership vector, affiliation of node to each of the communities
- Nodes given community assignment for each interaction
- Probability of an edge corresponding entry from $\kappa \times \kappa$ community edge probability matrix B

Scaling this model

- Hoffman et al. (2013) proposed Stochastic Variational Inference to scale Variational methods to large datasets

Scaling this model

- Hoffman et al. (2013) proposed Stochastic Variational Inference to scale Variational methods to large datasets
  - Use noisy (unbiased) updates of the gradient

Scaling this model

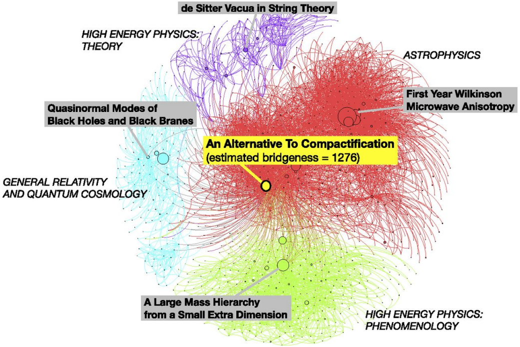
- Hoffman et al. (2013) proposed Stochastic Variational Inference to scale Variational methods to large datasets
  - Use noisy (unbiased) updates of the gradient
- Gopalan and Blei (2013) construct unbiased estimates for MMSBM, hundreds of thousands of nodes

Scaling this model

- Hoffman et al. (2013) proposed Stochastic Variational Inference to scale Variational methods to large datasets
    - Use noisy (unbiased) updates of the gradient
- Gopalan and Blei (2013) construct unbiased estimates for MMSBM, hundreds of thousands of nodes

Scaling this model

- Hoffman et al. (2013) proposed Stochastic Variational Inference to scale Variational methods to large datasets
    - Use noisy (unbiased) updates of the gradient
- Gopalan and Blei (2013) construct unbiased estimates for MMSBM, hundreds of thousands of nodes



Gopalan and Blei, 2013

Fitting Network Models without Network Data

- Can we fit this model without having the underlying adjacency matrix A?

Fitting Network Models without Network Data

- Can we fit this model without having the underlying adjacency matrix A?
- Can be difficult/expensive to collect A

Fitting Network Models without Network Data

- Can we fit this model without having the underlying adjacency matrix A?
- Can be difficult/expensive to collect A
- Sociologists have considered this problem to estimate population level social network information, use **Aggregated Relational Data**

An alternative approach

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

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

- 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)

Hard to compute network statistics

- Suppose A was everyone in Canada, wanted to estimate d_i, degree of person i (How many people does node i know)?

Hard to compute network statistics

- Suppose A was everyone in Canada, wanted to estimate d_i, degree of person i (How many people does node i know)?
- Maybe you want to estimate how many people in A have HIV (group which are hard to reach directly), N_u

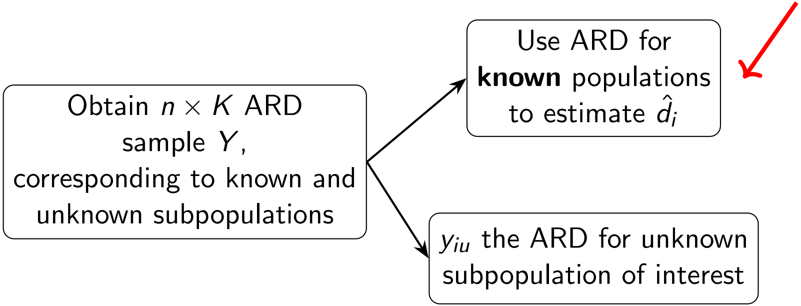
Hard to compute network statistics

- Suppose A was everyone in Canada, wanted to estimate d_i, degree of person i (How many people does node i know)?
- Maybe you want to estimate how many people in A have HIV (group which are hard to reach directly), N_u
- Want to estimate these without access to A

Using ARD

Obtain $n \times K$ ARD sample Y, corresponding to known and unknown subpopulations

Using ARD



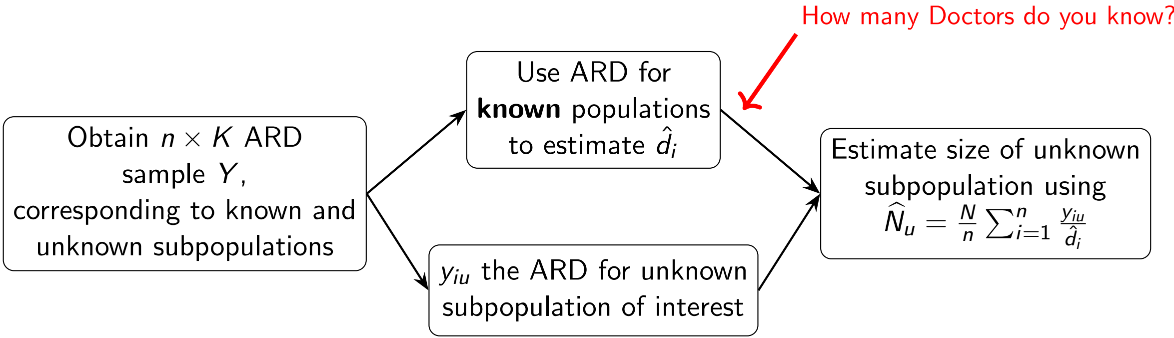
Obtain $n \times K$ ARD sample Y, corresponding to known and unknown subpopulations

Use ARD for **known** populations to estimate $\hat{d}_i$

How many Doctors do you know?

y_{iu} the ARD for unknown subpopulation of interest

Using ARD



How many Doctors do you know?

Obtain $n \times K$ ARD sample Y, corresponding to known and unknown subpopulations

Use ARD for **known** populations to estimate $\hat{d}_i$

y_{iu} the ARD for unknown subpopulation of interest

Estimate size of unknown subpopulation using $\widehat{N}_u = \frac{N}{n} \sum_{i=1}^{n} \frac{y_{iu}}{\hat{d}_i}$

Using ARD

- 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

Using ARD

- 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!

Using ARD

- 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

Using ARD

- 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 ($\kappa \neq K$)

ARD for Mixed-Membership SBM

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

ARD for Mixed-Membership SBM

- 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

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

Integrate out
sender and receiver effect

ARD for Mixed-Membership SBM

- 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

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

Integrate out
sender and receiver effect

- 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

Additional details

- 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**

Additional details

- 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

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

More details

- This gives

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

More details

- This gives

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

- 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

MMSBM for ARD

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

MMSBM for ARD

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

Fitting this model

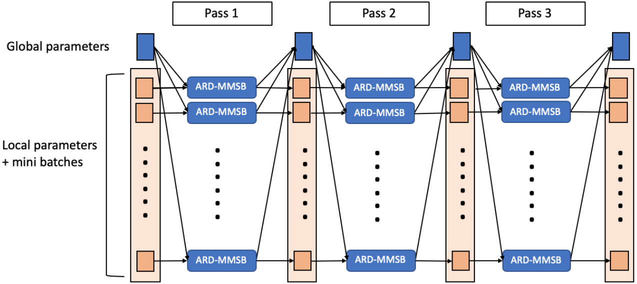
- Fit this using Variational Inference (Blei et al., 2017)
- Choose mean field distribution q for model parameters, estimate the parameters of q to minimize
$$\mathcal{L} = \mathbb{E}_q \log p(Y, \pi, B, \eta) - \mathbb{E}_q \log q(\pi, B, \eta)$$
- Take samples from ARD, use these minibatches to update parameters

Optimizing this quantity

- Have (not too ugly) closed form updates for B, the probability matrix
- Don't have conjugate updates for variational families of π, η
- Using Non-Conjugate Variational Message Passing (Knowles and Minka, 2011) gives closed form updates for π, η, need to adjust the step size
- Some other computational tricks required to make this work

Multiple passes through the data

- Each sample contains small number of rows/columns of Y
- Estimate local and global parameters with this information
- Average global parameters with these estimates and refit



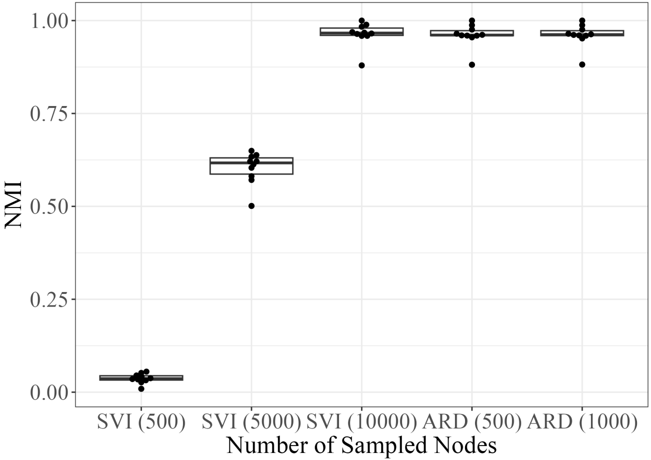
Iterating through the data

Evaluating this model

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?

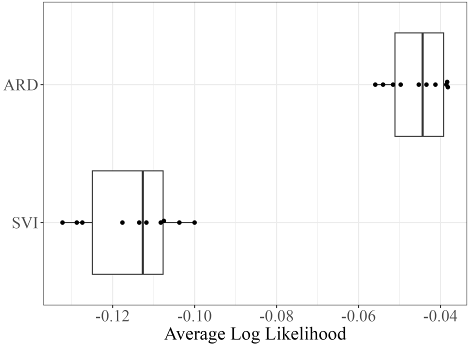
Simulation studies



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

- Better recover communities and *B* for simulated data
- Our procedure performs better for sparser networks

More simulations

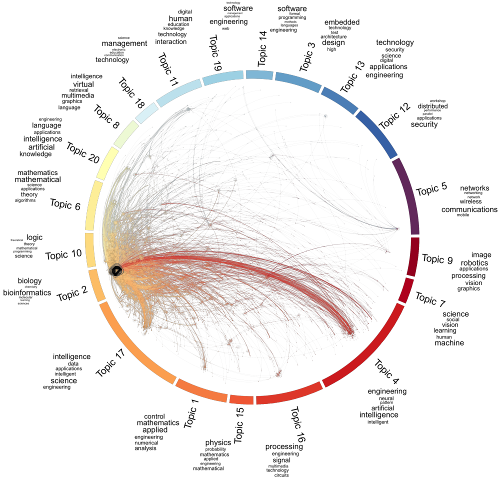


Predictive Log-likelihood using a subgraph

Our method faster to run than Gopalan and Blei (2013) observing the same amount of information

Real data

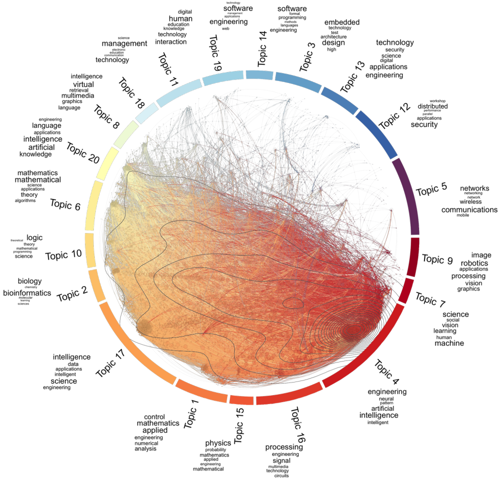
- Fit with 20 communities
- Show papers in journal along with papers cited by those papers
- Fitting Gopalan and Blei (2013) gives communities which are hard to interpret



Inferred Memberships for Journal Bioinformatics

Real data

- Capture the interdisciplinary nature of this journal
- Gopalan and Blei (2013) can better predict individual links



Memberships for Journal of Machine Learning Research

Next steps

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

Summary, Contributions

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

Summary, Contributions

- Used ARD to estimate underlying network community structure
- Allows us to scale this problem to large 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*.

Thank you