NME 2021

Network Modeling for Epidemics

DYNAMIC NETWORK MODELS: DATA

NME WORKSHOP

Martina Morris, Ph.D.
Steven M. Goodreau, Ph.D.
Samuel M. Jenness, Ph.D.

STERGMs: Data sources

- 1. Multiple cross-sections of complete network data
  - easy to work with
  - but rare-to-non-existent in infectious disease epi

- 2. One snapshot of a cross-sectional network (census, egocentric, or otherwise), plus information on relational durations
  - much more common
  - but introduces some statistical issues

Egocentric data in ERGMs and STERGMs

Repeating previous session:

- This approach entails constructing an ‘artificial population’
  - Simulated population size does not need to match sample size (or even size of target population)
  - Don’t need to be able to ID whether different respondent’s partners are the same as one another, or are also respondents
  - We will go through simple examples in the tutorial today, delve back into this tomorrow

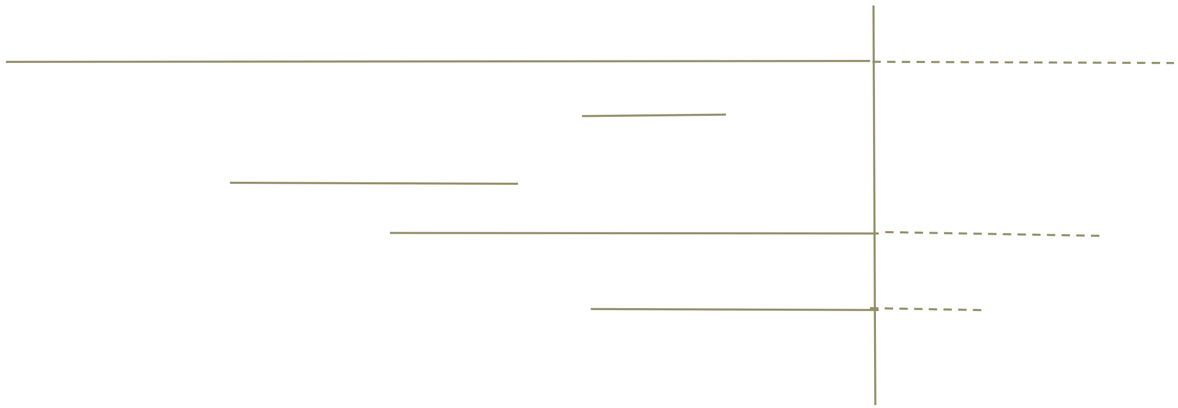
One cross-section + duration info

- Typically takes the form of
  - asking respondents about individual relationships (either with or without identifiers).
  - Often this is the *n* most recent, or all over some time period, or some combination (e.g. up to 3 in the last year)
  - asking whether the relationship is currently ongoing
  - if it's ongoing: asking how long it has been going on (or when it started)
  - if it's over: asking how long it lasted (or when it started and when it ended)

- From this we want to estimate
  - the mean duration of relationships
  - perhaps additional information about the variation in those durations (overall, across categories of respondents, etc.)

One cross-section + duration info

- Issues?



1. Ongoing durations are right-censored
    - can use Kaplan-Meyer or other techniques to deal with

One cross-section + duration info

- Issues?

2. Relationships are subject to length bias in their probability of being observed
   - This can also be adjusted for statistically
   - However, complex hybrid inclusion rules (e.g. most recent 3, as long as ongoing at some point in the last year) can make this complicated

One cross-section + duration info

- In practice (and for examples in this course), we sometimes rely on an elegant approximation based on the properties we just witnessed:
  - If relation lengths are approximately exponential/geometric, then the effects of length bias and right-censoring cancel out
  - The mean amount of time that the **ongoing** relationships have lasted until the day of interview (relationship age) is an unbiased estimator of the uncensored mean duration of relationships
  - Yes, it’s true.

One cross-section + duration info

- If you have something approximating a memoryless process for relational duration, then an unbiased estimator for relationship length is to:
  - ask people about how long their ongoing relationships have lasted up until the present
  - take the mean of that number across respondents.

One cross-section + duration info

- In practice, we find that the geometric distribution doesn't often capture the distribution of relational durations overall.
- But, if you divide the relationships into 2+ types, it can do a reasonable job within type
- Especially if you remove any 1-time contacts and model them separately (for populations where they are common)
- In our applied models (and in EpiModelHIV) we have three types
- Remember: most commonly used versions of DCMs model pretty much everything as a memoryless process, so approximating one aspect of our model that way is well within common practice

One cross-section + duration info

- When we pass our data into EpiModel as cross-sectional structure + durations, the algorithm is going to:
  - Calculate the dissolution coefficients first using data on duration
  - Then estimate the formation model condition on the dissolution model, using data on cross-sectional network structure

	Prevalence ≈	Incidence x	Duration
Data we have	Cross-sectional structure		Duration
Processes to model		Formation	Dissolution

One cross-section + duration info

- Mostly this will happen behind the scenes, but to get a flavor:

$$logit\left(P\left(Y_{ij,t+1}=1 \middle| Y_{ij,t}=1, \text{rest of the graph}\right)\right)=\boldsymbol{\theta}^{-\prime}\boldsymbol{\partial}\left(\boldsymbol{g}^{-}(\boldsymbol{y})\right)$$

- For the `~edges` model, with mean duration = 90 time steps:

$$ln\left(\frac{P(\text{tie persists})}{P(\text{tie dissolves})}\right)=\boldsymbol{\theta}^{-\prime}\boldsymbol{\partial}\left(\boldsymbol{g}^{-}(\boldsymbol{y})\right)$$

$$ln\left(\frac{P(\text{tie persists})}{P(\text{tie dissolves})}\right)=\boldsymbol{\theta}$$

$$ln\left(\frac{P(\text{tie persists})}{1/90}\right)=\boldsymbol{\theta}$$

$$ln\left(\frac{1-1/90}{1/90}\right)=\boldsymbol{\theta}$$

$$ln\left(\frac{90-1}{1}\right)=\boldsymbol{\theta}$$

$$ln(90-1)=\boldsymbol{\theta}$$

$$4.49=\boldsymbol{\theta}$$

One cross-section + duration info

- So dissolution can be solved analytically
- Then we want to condition the formation model on the dissolution model
- In R, the standard notation for indicating the parameters of a model that are to be fixed and conditioned on, rather than estimated, is with:

```
~offset(FixedParameter)
```