NME 2024

Network Modeling for Epidemics

1

Network Data Requirements

Leveraging the principle of sufficiency to estimate ERGMs from egocentric samples

What is “sufficiency” ?

- Intuitively:
  - you can only estimate what you observe (obvious)
  - but if you observe it, you can estimate it (less obvious)

- Formally:
  - A principle in statistical theory
  - That defines what you need to observe in data
  - In order to estimate the parameters in your model
    - The data “sufficient” for estimation

Example: from simple linear regression

- The OLS regression coefficient is related to the data as:

$$\hat{\beta} = \frac{Cov(X,Y)}{Var(X)}$$

- I only need to observe these 2 sets of summary statistics
  - $Cov(X,Y)$ and $Var(X)$
- In order to estimate β
- They are "sufficient"
  - I don't need to have the original data from the individual observations
  - Just these two aggregate summary values

This is *very* helpful for network models

Because it reduces the burden of data collection

Network data: Three main types (review)

- Network census
  - Data on every node and every link

Often infeasible in practice

- Adaptively sampled networks
  - Link tracing designs (e.g., snowball or RDS)

Challenging to collect, and the statistical methods for analysis are very limited

- Egocentrically sampled networks
  - Enroll population sample ("egos")
  - Ask them the usual questions about themselves
  - Ask them non-identifying information about their partners ("alters")
    - Timing (start and end of partnership)
    - Alter characteristics (sex, age, race, etc.)
    - Relational characteristics (type, cohabitation, etc.)
    - Pair-specific behaviors (act frequency, condom use, etc.)
  - Optional: ask about alter-alter ties
  - Optional: ask about perceptions of alters' alters more generally

Feasible, statistically supported and general

"partnership module"

Partnership modules

- These can be very short, or very long
  - DHS AIDS-related module had 6-8 questions – asked in over 25 countries around the world

    (example quex is linked below this slideset in the web book)
  - A Ugandan study had a sexual network module with ~70 questions – it was almost like a conversation with the respondent

- Module informs both network and epi modeling parameters
  - E.g., frequency of acts within partnerships, etc.

What is observed in the egocentric design?

- Degree
  - Mean degree, which sets density
  - Degree distributions
- Nodal attribute heterogeneity
  - Heterogeneity in degree
  - Mixing by nodal attributes
- Triads
  - Only if the alter-alter matrix data are collected
- Timing
  - Start/End, Duration of both active and completed partnerships

Much of the global structure of a network is set by these local properties

We can use what we observe to estimate the ERGM coefficients

Egocentric data in ERGMs

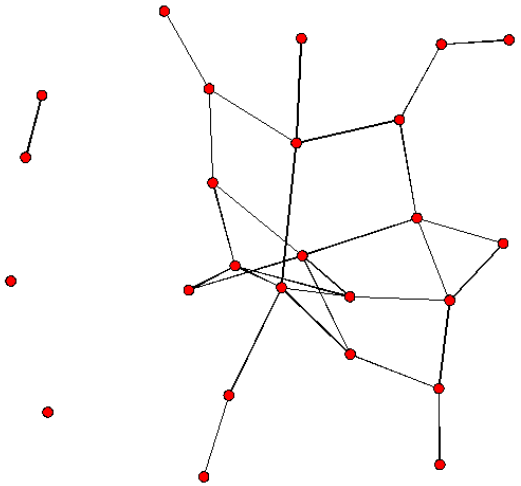
- These can be handled in the software quite easily.

- Recall with faux.mesa.high above, we fit the ergm by providing:
  - A model formula
  - A complete network containing:
    - nodes with their attributes
    - the relations among those nodes

- But alternatively, one can pass:
  - A model formula
  - An set of nodes with their attributes
  - The sufficient statistics for the terms in the model formula
    - Calculated from the observed data, and scaled if desired
    - These are called “target stats” in ergm

Network statistics in ERGMs

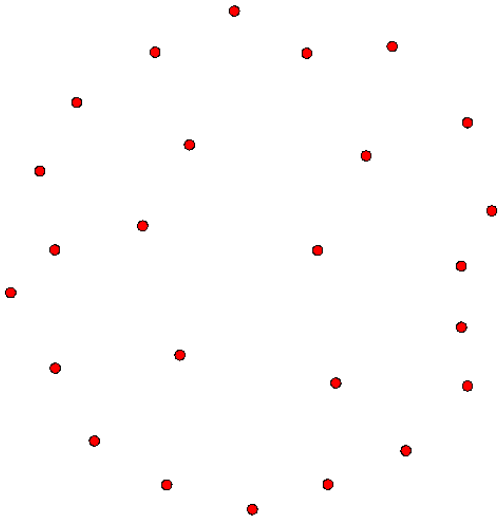
Option 1: network census



```
net ~ edges+degree(1)
```

ergm automatically calculates net stats from the data

Option 2: pass nodeset and targets



```
net ~ edges+degree(1)
target.stats = c(40, 7)
```

targets can come from any data set (or chosen as counterfactual)

We'll be using this extensively this week

- EpiModel is designed to work with both
  - Complete network data (census)
  - Egocentric data with target stat specifications
- You'll get lots of practice during the labs with target stats
- And we will be reviewing published examples

What about data for TERGMs?

- Recall: Temporal network data study designs
  - Panel data of network census (Discrete time)
  - Event history of network census (Continuous time)
  - Egocentric sample with retrospective information on duration

- It turns out the same principles hold for estimating TERGMs
  - Because this is just 2 ERGMs

- So we can use an egocentric sample with duration info

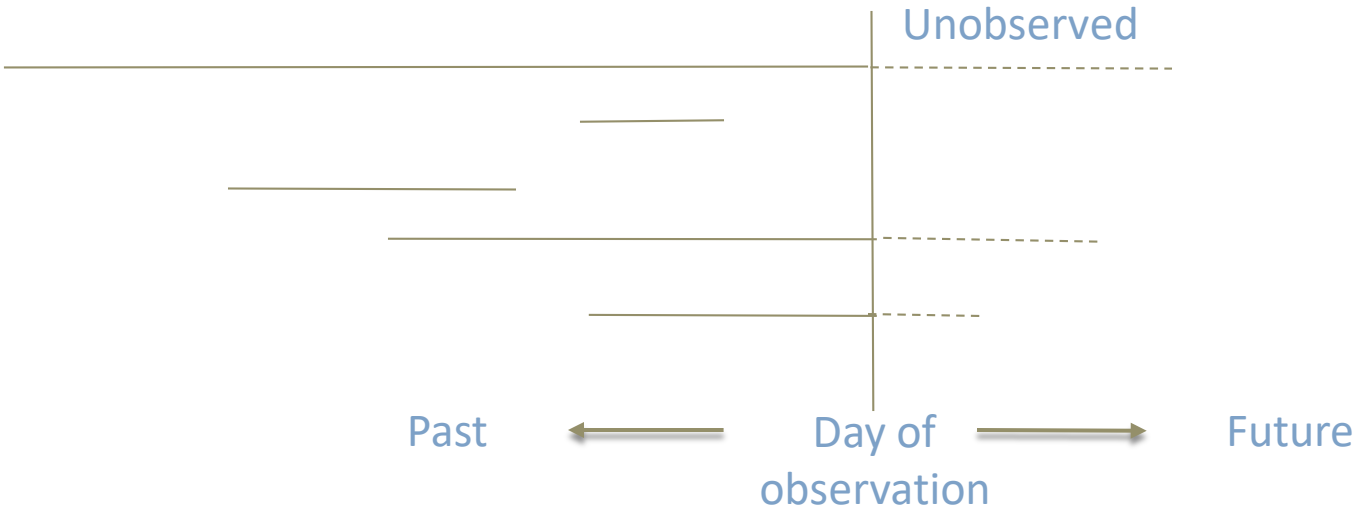
How to instrument this

- In the partnership module question set
  - Ask when a partnership started
  - Ask whether it is currently ongoing
    - if no: ask how long it lasted (or when it ended)
  - Ask what kind of relationship this is (if there are identifiable types)

- From this we can estimate
  - Mean duration of relationships
  - Heterogeneity in durations
    - By nodal attributes
    - By relationship type

Estimating relationship length from data

If you use all of the partnerships, what issue does this raise?

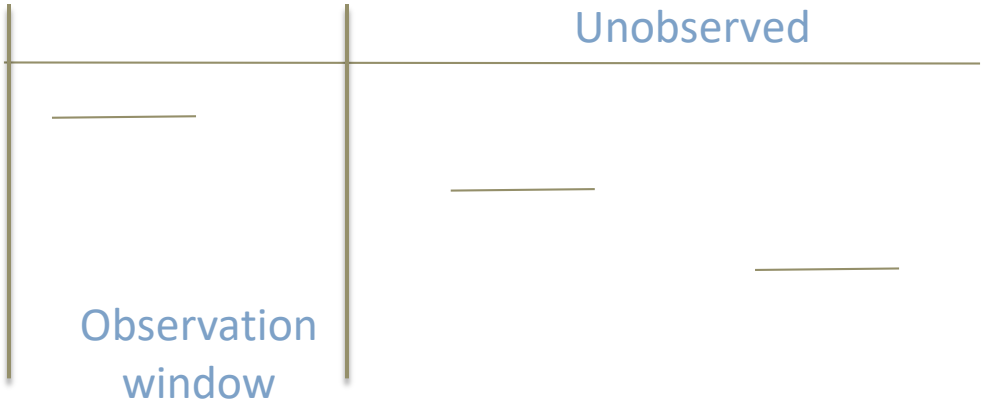


Censoring

- Ongoing durations are right-censored
- Can use Kaplan-Meyer or other techniques to deal with this

Estimating relationship length from data

- And if your data look like this, what issue does this raise?



Any one interval is more likely to pick up the longer partnerships, so your estimate of average duration will be too high

Length-biased sampling

- This can also be adjusted for statistically
- However, complex hybrid inclusion rules (e.g. most recent 3 + ongoing at some point in the last year) can make this complicated

The simple solution

If relation lengths are approximately exponential/geometric

- The average time that the **ongoing** relationships have lasted on the day of observation (relationship *age*) is an unbiased estimator of the uncensored mean duration of relationships
- The effects of length bias and right-censoring cancel out
- Surprising, amazing, and incredibly useful here

How these data are used in TERGM

- Recall the approximation from yesterday

Prevalence ≈ Incidence x Duration

Tie density | Formation rate | Inverse of dissolution rate

- If we know prevalence and duration, we can estimate incidence

Data: One cross-section + duration

When we pass data into `EpiModel` as cross-sectional structure + durations, the package will:

- Calculate the dissolution *coefficient(s)* first using data on tie age
- Then estimate the formation model conditioning on the dissolution model, using data on cross-sectional network structure

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

Calculating the dissolution coefficient

- Example: For the `~edges` dissolution model, $\partial\left(g^{-}(y)\right)$ always =1
- So if we observe mean tie age = 90 time steps, `EpiModel` will calculate (not estimate) the edges dissolution coefficient θ like this:

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

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

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

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

$$4.49=\theta$$

Fixing the dissolution coefficient

- Once the dissolution coefficient is calculated
- We tell EpiModel to treat it as an "offset"*
  - EpiModel will then fit the formation ERGM to the cross-sectional data on prevalent ties, and subtract this offset from the edges coefficient
  - This transforms the edges coefficient from a prevalence rate (density) to an incidence rate (formation)
  - The rest of the terms will capture the observed structural patterns
- In R, the standard notation is: `~offset(edges)`

* An offset is a term to be added to a linear predictor, such as in a generalised linear model, with known coefficient 1 rather than an estimated coefficient.

Capturing heterogeneity in duration

There are 3 types of heterogeneity we can represent in EpiModel

- Overall variance in the distribution of duration
  - These are stochastic models, so they produce variability in duration even for a homogeneous population (the variance of the geometric distribution)
- Heterogeneity by group (nodal attribute)
  - Add these terms to the dissolution model
- Heterogeneity by relationship type (tie attribute)
  - Separate network models for each type of data
    - But ties in one network can influence dynamics in another
  - Overlay these networks in the simulation model

In summary

- Because this is a general statistical modeling framework
- We can leverage the principle of sufficiency
- To estimate complex temporal network models
- Very efficiently
  - Surprisingly little data needed
  - Just a single cross sectional sample that is *representative* of the population of interest