NME

Network Modeling for Epidemics

1

Network Data Requirements

Leveraging the principle of sufficiency

to estimate ERGMs and TERGMs

from minimal data

High level overview

- To fully parameterize the network component of EpiModel we need
  - A model for the network structure
  - A model for the dynamics of tie formation/dissolution
- How much data do we need?
  - To support a principled statistical estimate
- There turn out to be several useful “tricks”
  - Where theory helps to minimize the data burden.

3

Network structure data needs

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.

- So, what network statistics are observed in egocentric designs?

Netstats observed in egocentric designs

- 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 and End or 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

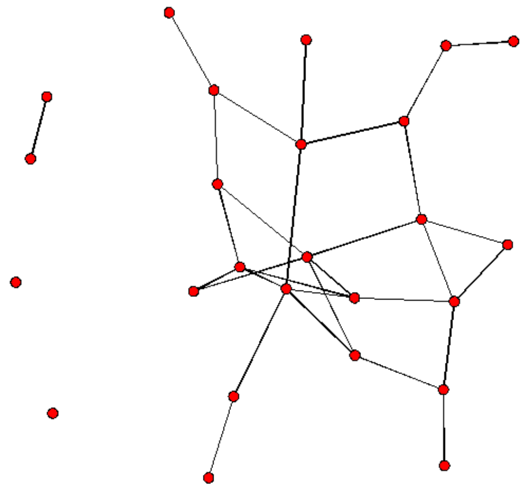
And what are the data needed for ERGMs?

- The ***g(y)*** statistics
  - are defined by the model
  - are sufficient for estimating **θ**
  - and will function as "target statistics" during estimation

- So any data source for these "target stats" can be used
  - A network census
  - An egocentric survey dataset
  - Egocentric statistics reported in the literature
  - Hypothetical statistics that you want to explore
  - Counterfactual statistics that you want to posit

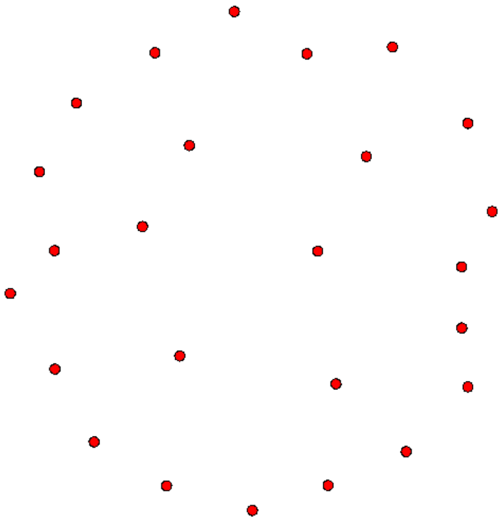
Behind the estimation curtain

Network census



=

Nodeset + target stats



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

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

More on all this is coming up

- EpiModel has the flexibility to accept many different types of data as inputs for the network model component
  - You'll get lots of practice during the labs with different data types
  - And we will be reviewing published examples

- There's just one caveat:
  - If you're not working with a network census
  - You need to pay attention to consistency and balance constraints in your target statistics
  - You'll get some practice with that too (esp in NME II)

10 Network dynamics

What data do you need to estimate the processes of tie formation and dissolution?

Now we’re talking about 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

- The only addition: data on partnership duration

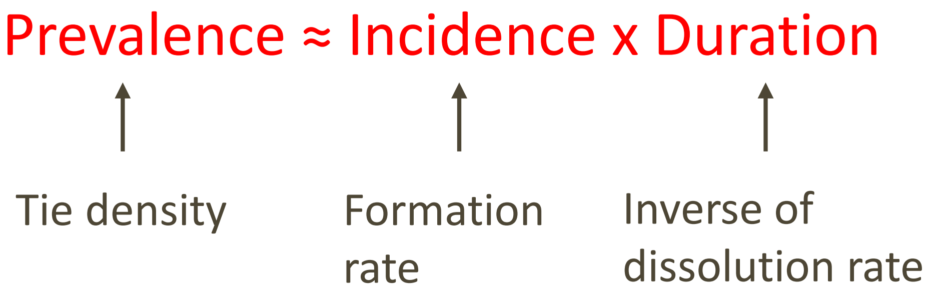
How to measure this in a survey?

- 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

How these data are used in TERGM

- Recall the approximation



Prevalence ≈ Incidence x Duration

Tie density

Formation rate

Inverse of dissolution rate

- If we know prevalence and duration, we can estimate incidence
  - Prevalence/Duration
  - or on the log scale, log(Prevalence) – log(Duration)

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, the probability of dissolution at each timestep is 1/90, and `EpiModel` will calculate (not estimate) the edges dissolution coefficient θ like this:

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

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

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

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

$$4.49 = \theta$$

Using this dissolution coefficient

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

* An offset is a term added to a linear predictor 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

Estimating relationship length

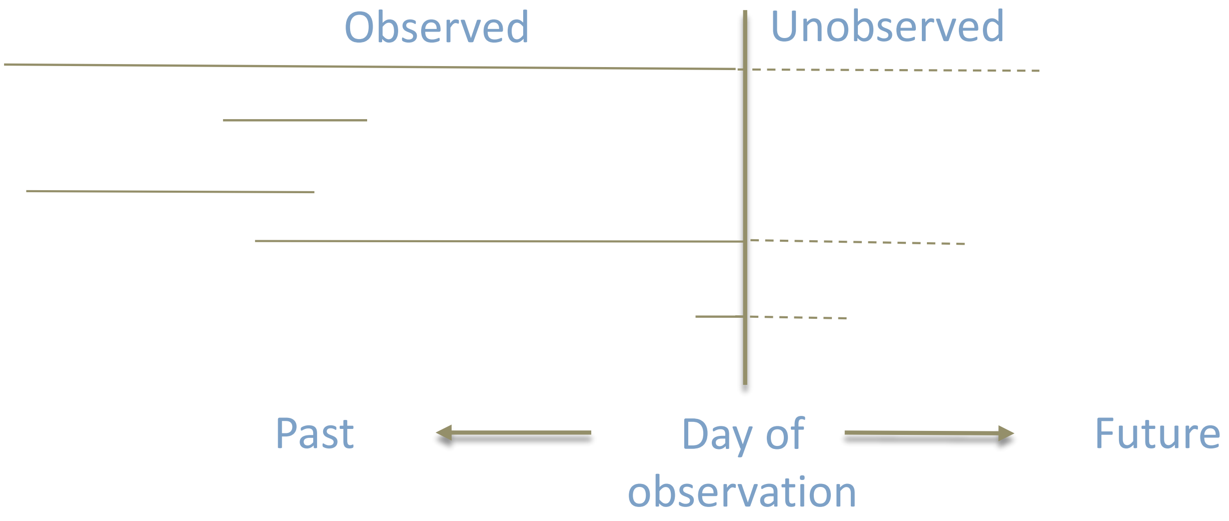
One last trick in the basket

We typically rely on retrospective data

- This is also reduces the data collection burden
- But it means we need to be careful with estimation
- The methods here come from survival analysis
  - Traditional stat, not network specific

Estimating relationship length from data

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

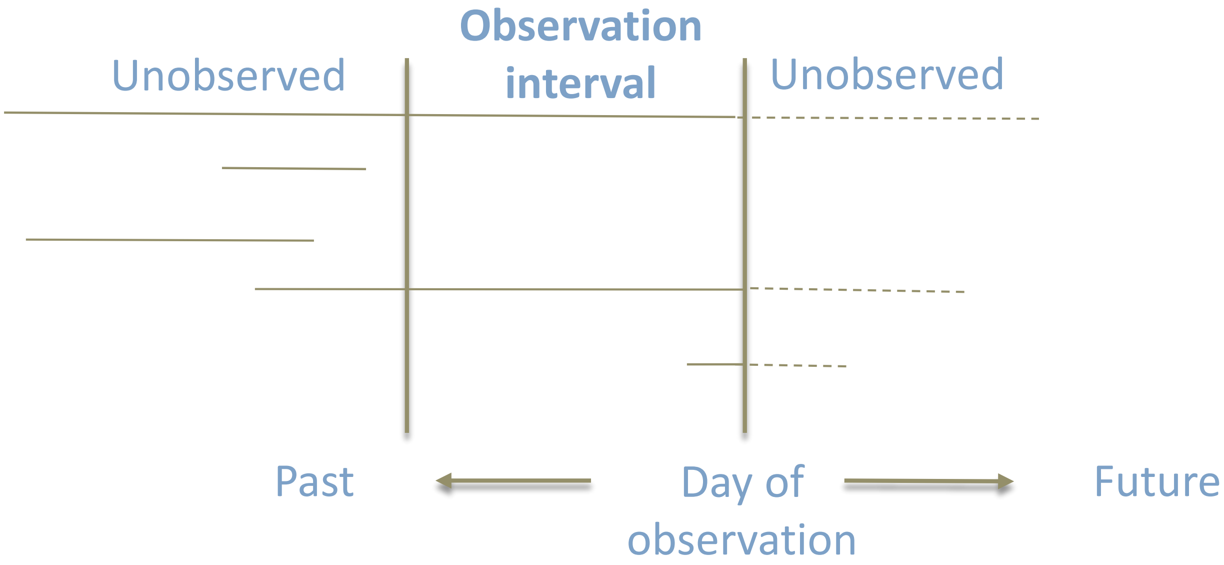


<u>Censoring</u>

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

Estimating relationship length from data

If you use only partnerships in an interval, what then …?



Any interval is more likely to capture 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

So …

- That was a lot!
- Packed into a very short presentation

- It is not essential to understand all of this in order to use EpiModel

- But, it is worth knowing how much statistical theory is there in the background working for you

In summary

- Because this is a general statistical modeling framework
  - We can leverage the principle of sufficiency
  - The assumption of form/diss separability (within time step)
  - The assumption of geometrically distributed durations

- To estimate complex temporal network models

- Very efficiently
  - Surprisingly little data needed
  - Just a single cross sectional sample of the network
  - That is *representative* of the population of interest