TEMPORAL EXPONENTIAL-FAMILY RANDOM GRAPH MODELING (TERGMS) WITH STATNET

Prof. Martina Morris
Prof. Steven Goodreau
The *statnet* team

Source for all things STERGM

- Pavel N. Krivitsky and Mark S. Handcock (2014). A Separable Model for Dynamic Networks. *Journal of the Royal Statistical Society, Series B*, Volume 76, Issue 1, pages 29–46.

Terminology

- The phrase “temporal ERGMs,” or TERGMs, refers to all ERGMs that are dynamic
- The specific class of TERGMs that have been implemented thus far are called “separable temporal ERGMs,” or STERGMs
- In the relevant R package, we left open the possibility that we would develop more in the future
- Thus:

	Cross-sectional	Dynamic
Name of package	ergm	tergm
Name of function in package	ergm	stergm

ERGMs: Review

Probability of observing a graph (set of relationships) y on a fixed set of nodes:

$$P(Y = y \mid \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}' \boldsymbol{g}(\boldsymbol{y}))}{k(\boldsymbol{\theta})}$$

Conditional log-odds of a tie

$$logit\left(P\left(Y_{ij} = 1 \middle| \text{rest of the graph}\right)\right) = log\left(\frac{P\left(Y_{ij} = 1 \middle| \text{rest of the graph}\right)}{P\left(Y_{ij} = 0 \middle| \text{rest of the graph}\right)}\right)$$

$$= \boldsymbol{\theta}' \boldsymbol{\partial}\left(\boldsymbol{g}(\boldsymbol{y})\right)$$

where: **g(y)** = vector of network statistics
θ = vector of model parameters
k(*θ*) = numerator summed over all possible networks on node set y
$\boldsymbol{\partial}\left(\boldsymbol{g}(\boldsymbol{y})\right)$ represents the change in **g(y)** when Y_{ij} is toggled between 0 and 1

STERGMs

- ERGMs are great for modeling cross-sectional network structure
- But they can only predict the *presence* of a tie; they are unable to separate the processes of *tie formation* and *dissolution*
- Why separate formation from dissolution?

STERGMs

- Intuition: The social forces that facilitate formation of ties are often different from those that facilitate their dissolution.
- Interpretation: Because of this, we would want model parameters to be interpreted in terms of ties formed and ties dissolved.
- Simulation: We want to be able to control cross-sectional network structure and relational durations separately in our disease simulations, matching both to data

STERGMs

- E.g. if a particular type of tie is rare in the cross-section, is that because:
  - They form infrequently?
  - They form frequently, but then dissolve frequently as well?

- The classic approximation formula from epidemiology helps us see the basic relationship among our concepts:

Prevalence ≈ Incidence x Duration

↑ Formation (Incidence)

↑ Inverse of dissolution (Duration)

STERGMs

- Core idea:
  - The y_{ij} values (ties in the network) and Y (the set of all y_{ij} values) are now indexed by time
  - Represent evolution from Y_t to Y_{t+1} as a product of two phases: one in which ties are formed and another in which they are dissolved, with each phase a draw from an ERGM.
  - Thus, two formulas: a formation formula and a dissolution formula
  - And, two corresponding sets of statistics

STERGMs

ERGM: Conditional log-odds of a tie existing

$$logit\left(P\left(Y_{ij} = 1 \middle| \text{rest of the graph}\right)\right) = \boldsymbol{\theta}'\boldsymbol{\partial}\left(\boldsymbol{g}(\boldsymbol{y})\right)$$

STERGM: Conditional log-odds of a tie *forming* (formation model)*:*

$$logit\left(P\left(Y_{ij,t+1} = 1 \middle| Y_{ij,t} = 0, \text{rest of the graph}\right)\right) = \boldsymbol{\theta}^{+\prime}\boldsymbol{\partial}\left(\boldsymbol{g}^{+}(\boldsymbol{y})\right)$$

STERGM: Conditional log-odds of a tie *persisting* (dissolution model)*:*

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

where: $\boldsymbol{g}^{+}(\boldsymbol{y})$ = vector of network statistics in the formation model
$\boldsymbol{\theta}^{+}$ = vector of parameters in the formation model
$\boldsymbol{g}^{-}(\boldsymbol{y})$ = vector of network statistics in the dissolution model
$\boldsymbol{\theta}^{-}$ = vector of parameters in the dissolution model

STERGMs

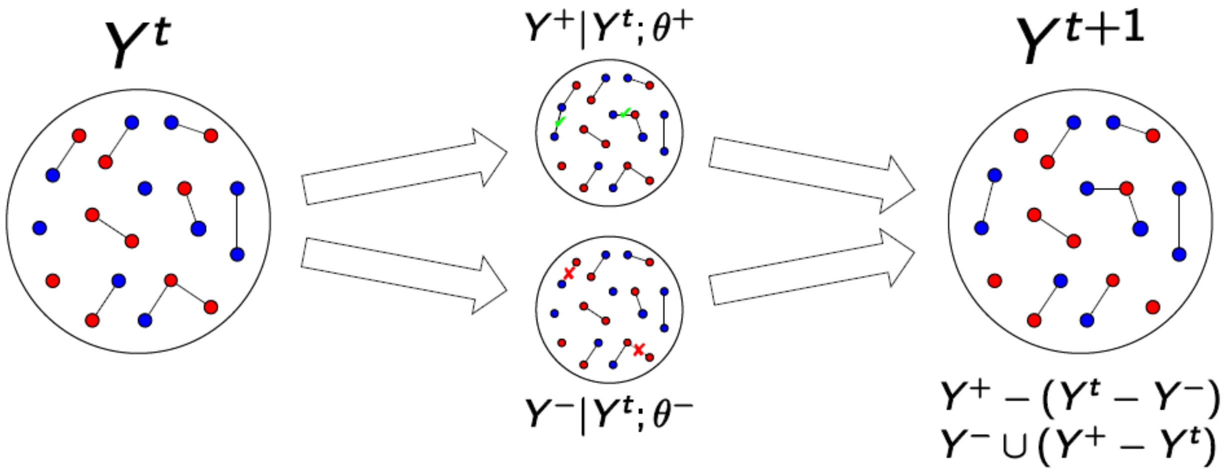
Dissolution? Or persistence?

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

- The model is expressed as log odds of tie equaling 1 given it equaled 1 at the last time step
- This is done to make it consistent with the formation model, so all the math works out nicely
- But it implies that the model, and thus the coefficients, should be interpreted in terms of effects on relational persistence
- That said, people tend to thing in terms of relational formation and dissolution, since relational dissolution is a more salient event than relational persistence
- Thus, we often use the language of dissolution

STERGMs

- During simulation, two processes occur separately within a time step:



- Y^+ = network in the formation process after evolution
- Y^- = network in the dissolution process after evolution
- This is the origin of the "S" in STERGM

STERGMs

- The statistical theory in Krivitsky and Handcock 2014:
  - demonstrates a given combination of formation and dissolution model will converge to a stable equilibrium, i.e.:

Prevalence ≈ Incidence x Duration

- This and other work in press provide the statistical theory for methods for estimating the two models, given certain kinds of data

STERGMs: Example of interpretation

Term = `~edges`

	$\boldsymbol{\theta}$ ↗	$\boldsymbol{\theta}$ ↘
Formation model	more new ties created each time step	fewer new ties created each time step
Dissolution (persistence) model	more existing ties pre-served (fewer dissolved); longer average duration	fewer existing ties pre-served (more dissolved); shorter average duration

What combo do you think is most common in empirical networks?

STERGMs: Example of interpretation

Term = `~edges`

	$\boldsymbol{\theta} \nearrow$	$\boldsymbol{\theta} \searrow$
Formation model	more new ties created each time step	fewer new ties created each time step
Dissolution (persistence) model	more existing ties pre-served (fewer dissolved); longer average duration	fewer existing ties pre-served (more dissolved); shorter average duration

What combo do you think is most common in empirical social networks?

STERGMs: Example of interpretation

Term = `~concurrent` (# of nodes with degree 2+)

	$\theta \nearrow$	$\theta \searrow$
Formation model	more ties added to actors with exactly 1 tie	fewer ties added to actors with 1 tie
Dissolution (persistence) model	actors with 2 ties more likely to have them be preserved	actors with 2 ties more likely to have them dissolve

What combo do you think is most common in empirical sexual networks?

STERGMs: Example of interpretation

Term = `~concurrent` (# of nodes with degree 2+)

	$\boldsymbol{\theta} \nearrow$	$\boldsymbol{\theta} \searrow$
Formation model	more ties added to actors with exactly 1 tie	fewer ties added to actors with 1 tie
Dissolution (persistence) model	actors with 2 ties more likely to have them be preserved	actors with 2 ties more likely to have them dissolve

What combo do you think is most common in empirical sexual networks?

STERGMs: Data sources

- 1. Multiple cross-sections of complete network data
  - easy to work with
  - but rare-to-non-existent in some fields

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

STERGMs: nodal dynamics

- All of the statistical theory presented so far regards networks with
  - Dynamic relationships, but still
  - Static actors

- I.e. no births and deaths, no changing of nodal attributes

- The statistical theory of STERGM can handle nodal dynamics during simulation, with a few added tweaks
  - Most important is an offset term to deal with changing population size
  - Without it, density is preserved as population size changes
  - With it, mean degree is preserved as population size changes

STERGMs: nodal dynamics

- For more info, see:

  Pavel N. Krivitsky, Mark S. Handcock, and Martina Morris (January 2011). **Adjusting for Network Size and Composition Effects in Exponential-Family Random Graph Models**. *Statistical Methodology*, 8(4): 319–339

- And for more help with using STERGMs to simulate dynamic networks along with changing nodes and attributes:

  - Take our intensive summer workshop on network modeling for epidemic diffusion
  - Explore the online materials for the workshop (on the statnet webpage)
  - Try the EpiModel package

To the tutorial.....

(reference slides follow)

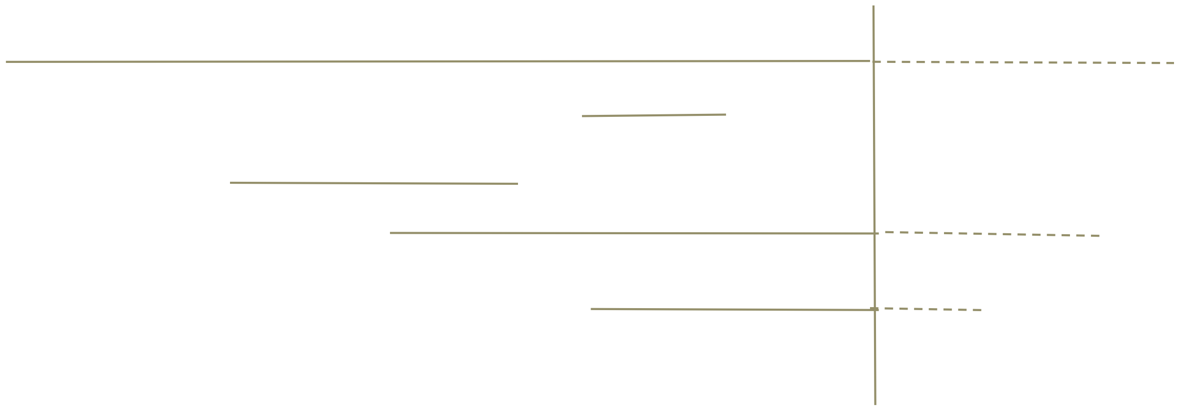
One cross-section + duration info

- In some domains, often 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:
  - If relation lengths are approximately exponential/geometric (a big if!), 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 mean duration of relationships
  - Why?!?

One cross-section + duration info

- Exponential/geometric durations suggests a memoryless processes – one in which the future does not depend on the past
- Imagine a fair, 6-sided die:

1/6
- What is the probability I will get a 1 on my next toss?

1/6
- What is the probability I will get a 1 on my next toss given that my previous 1 was five tosses ago?

6
- On average, how many tosses will I need before I get my first 1?

6
- On average, how many more tosses will I need before I get my next 1, given that my previous 1 was 8 tosses ago?

Geometric	
Parameters	$0 < p \leq 1$ success probability (real)
Support	$k \in \{1, 2, 3, \dots\}$
Probability mass function (pmf)	$(1-p)^{k-1}\,p$
Cumulative distribution function (CDF)	$1-(1-p)^k$
Mean	$\frac{1}{p}$

One cross-section + duration info

- Now, let's imagine this fairly bizarre scenario:
  - You arrive in a room where there are 100 people who have each been flipping one die; they pause when you arrive.
  - You don't know how many sides those dice have, but you know they all have the same number.
  - You are not allowed to ask any information about what they've flipped in the past.
  - The only information people will give you is: how many flips after your arrival does it take until they get their first 1?
  - You are allowed to stay until all of the 100 people get their first 1, and they can inform you of the result.
- Given the information provided you, how will you estimate the number of sizes on the die?

One cross-section + duration info

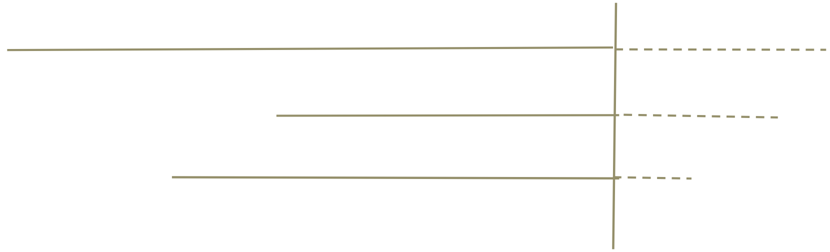
- Simple: when everyone tells you how many flips it takes from your arrival until their first 1, just take the mean of those numbers. Call it *m*.
- Your best guess for the probability of getting a 1 per flip is 1/*m*.
- And your best guess for the number of sides is the reciprocal of the probability of any one outcome per flip, which is 1/1/m, which just equals *m* again.
- Voila!

One cross-section + duration info

Retrospective relationship surveys are like this, but in reverse:

Dice:

Relationships:



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)
- Remember: DCMs model pretty much everything as a memoryless process, so approximating one aspect of our model that way is well within common practice