DYNAMIC NETWORK MODELS

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

Underlying methods: Statnet Development Team

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

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.

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 from 0 to 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?

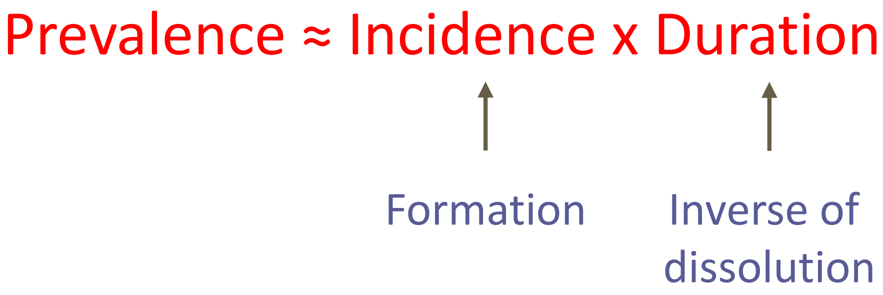
STERGMs

- Intuition: The social forces that facilitate formation of ties are often different from those that facilitate their dissolution.
- Interpretation: Because of this, we want model parameters that can be interpreted in terms of ties formed and ties dissolved. (Of course we need data that can allow us to estimate these).
- 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 applies here as well:



Prevalence ≈ Incidence x Duration

Formation

Inverse of dissolution

STERGMs

- Core idea:
  - Y is 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

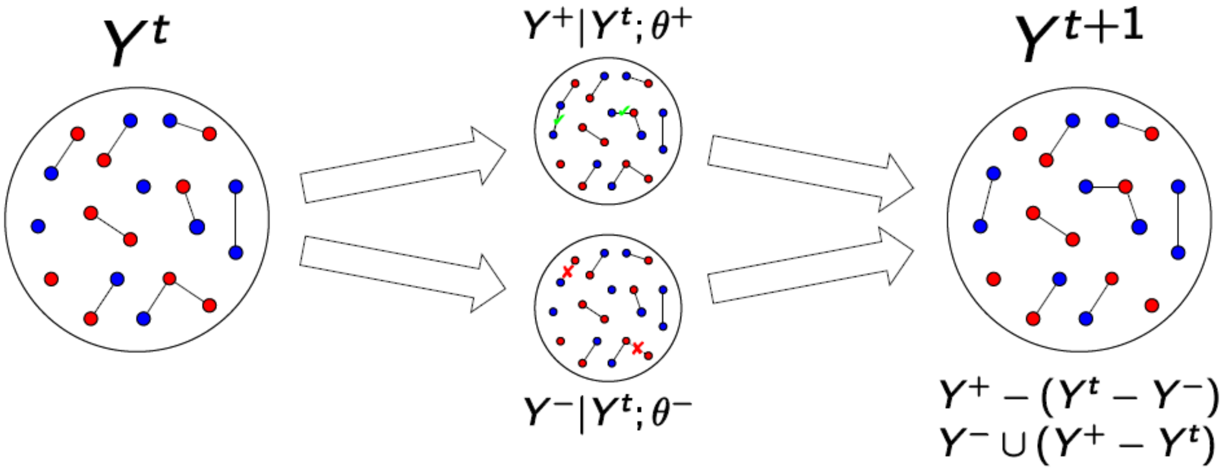
Dissolution? Or persistence?

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

- 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`

	$\boldsymbol{\theta}$ **is** $+$	$\boldsymbol{\theta}$ **is** $-$
Formation model	>50% of empty dyads have ties created during each timestep	<50% of empty dyads have ties created during each timestep
Dissolution (persistence) model	>50% existing ties preserved (fewer dissolved); longer average duration	<50% existing ties preserved (more dissolved); shorter average duration

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

STERGMs: Example of interpretation

Term = `~edges`

	$\boldsymbol{\theta}$ **is** $+$	$\boldsymbol{\theta}$ **is** $-$
Formation model	>50% of empty dyads have ties created during each timestep	<50% of empty dyads have ties created during each timestep
Dissolution (persistence) model	>50% existing ties preserved (fewer dissolved); longer average duration	<50% existing ties preserved (more dissolved); shorter average duration

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

STERGMs: Example of interpretation

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

	$\boldsymbol{\theta}$ **is** $+$	$\boldsymbol{\theta}$ **is** $-$
Formation model	ties added to actors with exactly 1 tie with relatively high probability	ties added to actors with exactly 1 tie with relatively low probability
Dissolution (persistence) model	actors with 2 ties more likely than others to have them persist	actors with 2 ties more likely than others 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+)

	θ is +	θ is −
Formation model	ties added to actors with exactly 1 tie with relatively high probability	ties added to actors with exactly 1 tie with relatively low probability
Dissolution (persistence) model	actors with 2 ties more likely than others to have them persist	actors with 2 ties more likely than others 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+)

	θ is $+$	θ is $-$
Formation model	ties added to actors with exactly 1 tie with relatively high probability	ties added to actors with exactly 1 tie with relatively low probability
Dissolution (persistence) model	actors with 2 ties more likely than others to have them persist	actors with 2 ties more likely than others to have them dissolve

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

Why 2, and not 2+ in the interpretation of dissolution ?

DATA

Data types

- Network censuses
  - Rare in most forms of network epidemiology
  - (Nearly?) non-existent in HIV/STI research

- Egocentric network data
  - Enroll population sample (“egos”)
  - Ask them the usual questions about themselves
  - Ask them non-identifying information about their partners (“alters”)
    - Number
    - Timing
    - Alter characteristics
    - Relational characteristics (e.g. type, act frequency, condom use)
  - Optional: ask about alter-alter ties
  - Optional: ask about perceptions of alters’ alters more generally

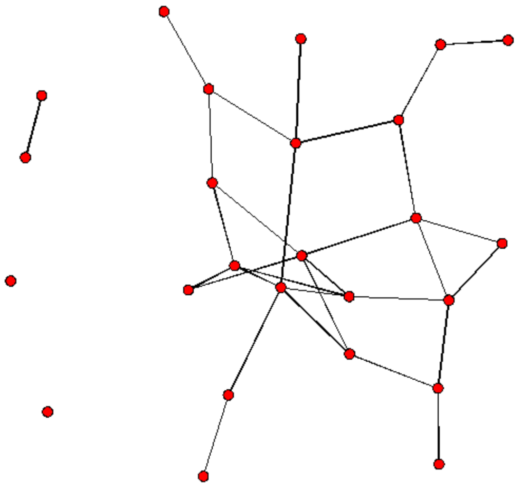
Egocentric data in ERGMs and STERGMs

- Mechanically, these can be handled in the software quite easily.
- E.g. in the ERGM session, we supplied:
  - Model formula
  - A network containing:
    - nodes with their attributes
    - the relations among those nodes

- But alternatively, one can pass:
  - Model formula
  - A network containing:
    - nodes with their attributes
  - The sufficient statistics for the terms in the model formula for that set of nodes (“target stats”)

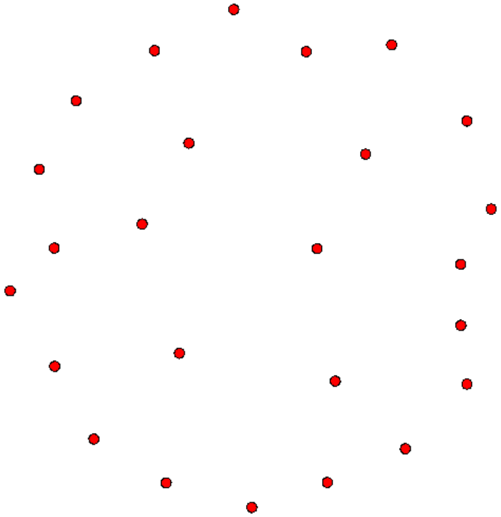
Egocentric data in ERGMs and STERGMs

Option 1:



```
net~edges+triangle
```

Option 2:



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

Egocentric data in ERGMs and STERGMs

- Why does this work?
  - (ST)ERGMs are based in exponential family theory
  - one of the properties of MLEs for exponential families is that exp(model sufficient stats) = observed sufficient stats.
  - any graph with the same observed sufficient stats has the same probability under the model
  - so we just iterate our way to finding the coefficients that generate exp(model sufficient stats) = observed sufficient statistics

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

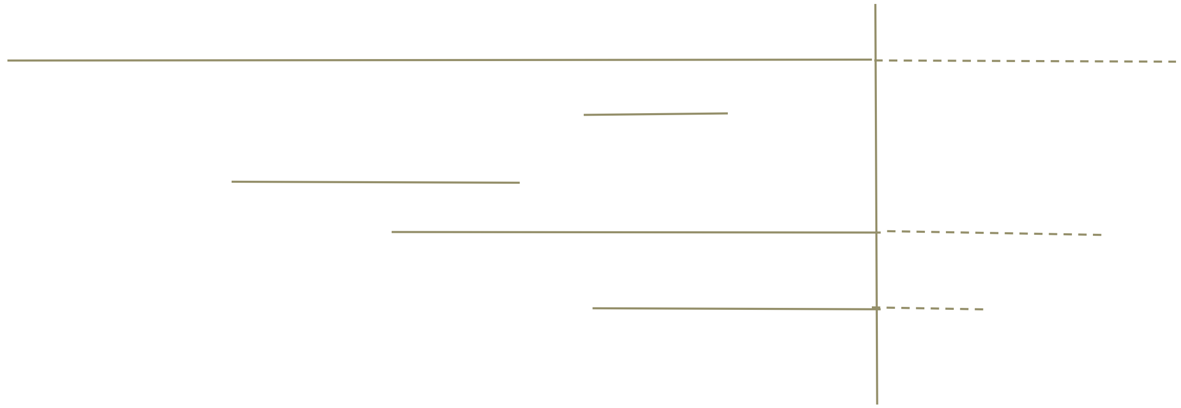
One cross-section + duration info

- 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

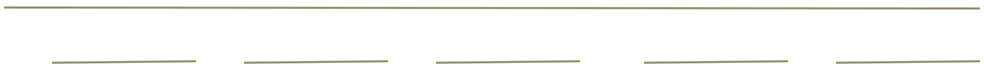
- Issues?



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:
  - 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 uncensored mean duration of relationships
  - Why?!?

One cross-section + duration info

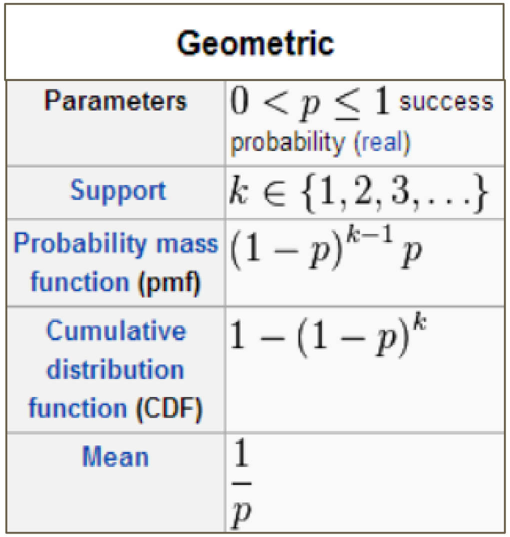
- Exponential/geometric durations suggests a memoryless processes – one in which the future does not depend on the past
- Imagine a fair, 6-sided die:

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 rolling 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 sides on the die?

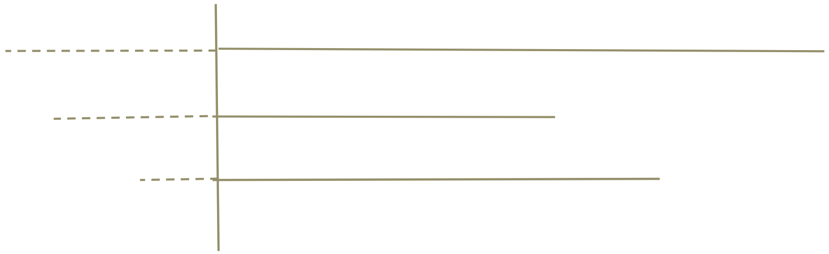
One cross-section + duration info

- 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!

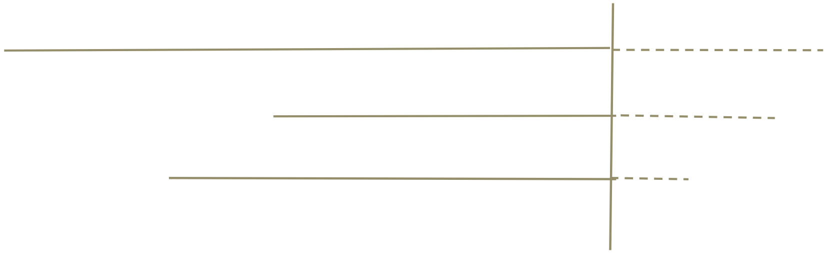
One cross-section + duration info

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.

- RShiny app that shows this actually works

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: 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:

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

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

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

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

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

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

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)
```

Balance

- The idea that the number of contacts group A has with group B must equal the number that group B has with group A

Balance: network models

- E.g. if you are building a purely heterosexual model
  - In the real world, in any population:

    # of relationships/acts that females have with males =

    # of relationships/acts males have with females

  - But this may not be exactly true in the data
    - bias (sex ratio of sample does not equal empirical sex ratio, female sex workers are under-sampled)
    - misreporting (e.g. females may under-report)

  - Nevertheless, one needs to be explicit about balance in the target statistics

STERGMs: balance

- E.g.
  - Both population and sample have a 1:1 sex ratio
  - males report mean degree of 0.74, females report 0.68
  - you must choose whether to use 0.68, 0.74, 0.71, or something else
  - Do so when build network and set target stats

- Note: once estimation is done, and simulation begins then balance will happen automatically forever, even when we introduce vital dynamics
- This is because the target stats have been converted into parameters based in log-odds
- This is true no matter the nature of complexity of the nodal dynamics

To EpiModel…..

STERGMs – dependence across time steps

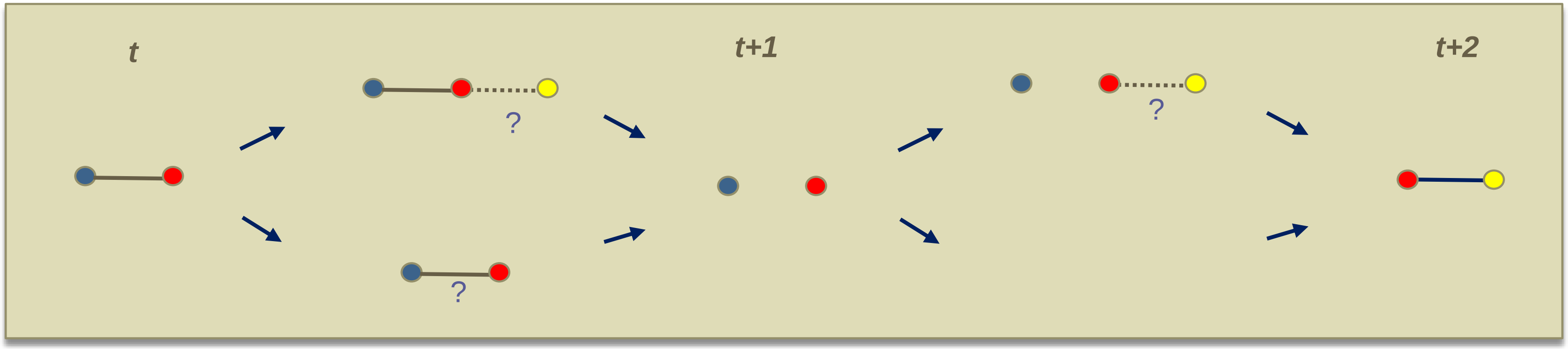
- The “separable” part of STERGMS means that within a time step, formation and dissolution are independent
- But this does not mean that they must be independent across time steps
- Imagine this model:
  - formation = ~edges+degree(2:10)
  - dissolution = ~edges
  - with increasingly negative parameters on the degree terms.
  - i.e. there is some underlying tendency for relational formation to occur, which is considerably reduced with each pre-existing tie that the two actors involved are already in.
- In other words, there is a strong prohibition against being in multiple simultaneous romantic relationships.
- However, dissolution is fully independent---all existing relationships have the same underlying dissolution probability at every time step.

STERGMs – dependence across time steps

- Imagine that Chris and Pat are in a relationship at time *t*.
- During the step between t and t+1, whether they break up does not depend on when either of them acquires a new partner, and vice versa.
- Let us assume that they do *not* break up during this time.
- Now, during the time period between t+1 and t+2:
  - whether or not they each form new partnership is dependent on whether they are still together at time t+1,
  - and that in turn depends on whether they broke up between t and t+1.

STERGMs – dependence across time steps

- Imagine that Chris and Pat are in a relationship at time *t*.
- During the step between *t* and *t+1*, whether they acquire a new partner does not depend on whether they break up and vice versa.
- Let us assume that they do break up during this time.
- Now, during the time period between t+1 and t+2:
  - whether or not they each form new partnership is dependent on whether they are still together are time t+1,
  - and that in turn depends on whether they broke up between t and t+1.

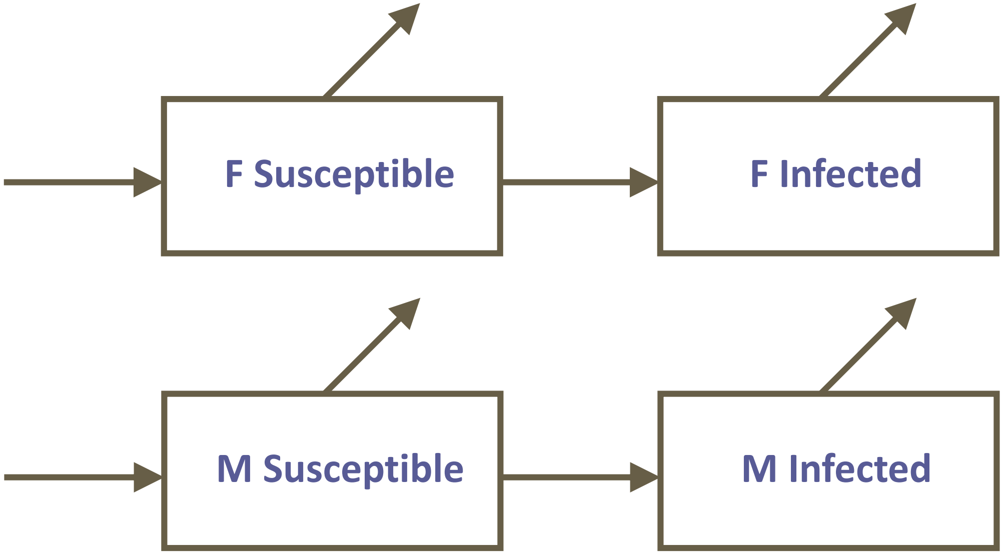


STERGMs – dependence across time steps

- Implication: formation and dissolution can be dependent, but that dependence occurs in subsequent time steps, not simultaneously.
- Note: time step is arbitrary, and left to the user to define. One reason to select a smaller time interval is that it makes this assumption more justifiable.
- I.e. with a time step of 1 month, then if I start a new relationship today, the earliest I can break up with my first partner as a *direct result* of that new partnership is in one month.
- If my time step is a day, then it is in 1 day
- The latter is likely much more reasonable.
- Tradeoff: shorter time interval means longer computation time for both model estimation and simulation
- At the limit, this can in practice approximate a continuous-time model---the only issue is computational limitations.`

Appendix

Balance: DCMs: SI model w/ 2 groups



Balance: DCMs: SI model w/ 2 groups

E.g. two sexes with purely heterosexual contact

Parameterization w/ 1 group

t = time

$s(t)$ = number of susceptible people at time t

$i(t)$ = number of infected people at time t

α = act rate per unit time

τ = prob. of transmission given S-I act

One form of parameterization w/ 2 groups

t = time

$s_f(t)$ = number of susceptible females at time t

$s_m(t)$ = number of susceptible males at time t

$i_f(t)$ = number of infected females at time t

$i_m(t)$ = number of infected males at time t

α = act rate per unit time

τ_{mf} = prob. of transmission given female S – male I act

τ_{fm} = prob. of transmission given male S – female I act

Let’s imagine we’ve added in births and deaths as well, and infected people have a higher mortality rate than others

Balance: DCMs: SI model w/ 2 groups

- Incidence in 1-group model $= s(t)\alpha\frac{i(t)}{n(t)}\tau$

- Incidence in 2-group model
  - female incidence $= s_f(t)\ \alpha\ \frac{i_m(t)}{n_m(t)}\ \tau_{mf}$
  - male incidence $= s_m(t)\ \alpha\ \frac{i_f(t)}{n_f(t)}\ \tau_{fm}$???

- Anyone see a potential issue that this introduces?

Balance: DCMs: SI model w/ 2 groups

- We assumed:
  - Pure across-group mixing, and
  - One act rate (α) for the whole population

- How many acts in total do females have with males at time t? $n_f(t)\ \alpha$
- How many acts in total do males have with females at time t? $n_m(t)\ \alpha$
- These quantities must be equal, since every time a woman has an act with a man, a man has an act with a woman
- But what if $n_f(t)$ doesn't equal $n_m(t)$?
- Would that happen in our model?

Balance: DCMs: SI model w/ 2 groups

- Behavioral Balance

Option 1: females drive things: α_f is fixed, and $\alpha_m(t) = \alpha_f \frac{n_f(t)}{n_m(t)}$

Option 2: males drive things: α_m is fixed, and $\alpha_f(t) = \alpha_m \frac{n_m(t)}{n_f(t)}$

Option 3: meet somewhere in the middle

- Balance is crucial - do it early and often!