WELCOME!

NETWORK MODELS FOR HIV/STI TRANSMISSION DYNAMICS WITH EPIMODEL

HARVARD 2017

Samuel M. Jenness, Ph.D.
Steven M. Goodreau, Ph.D.
Eli S. Rosenberg, Ph.D.

First things first

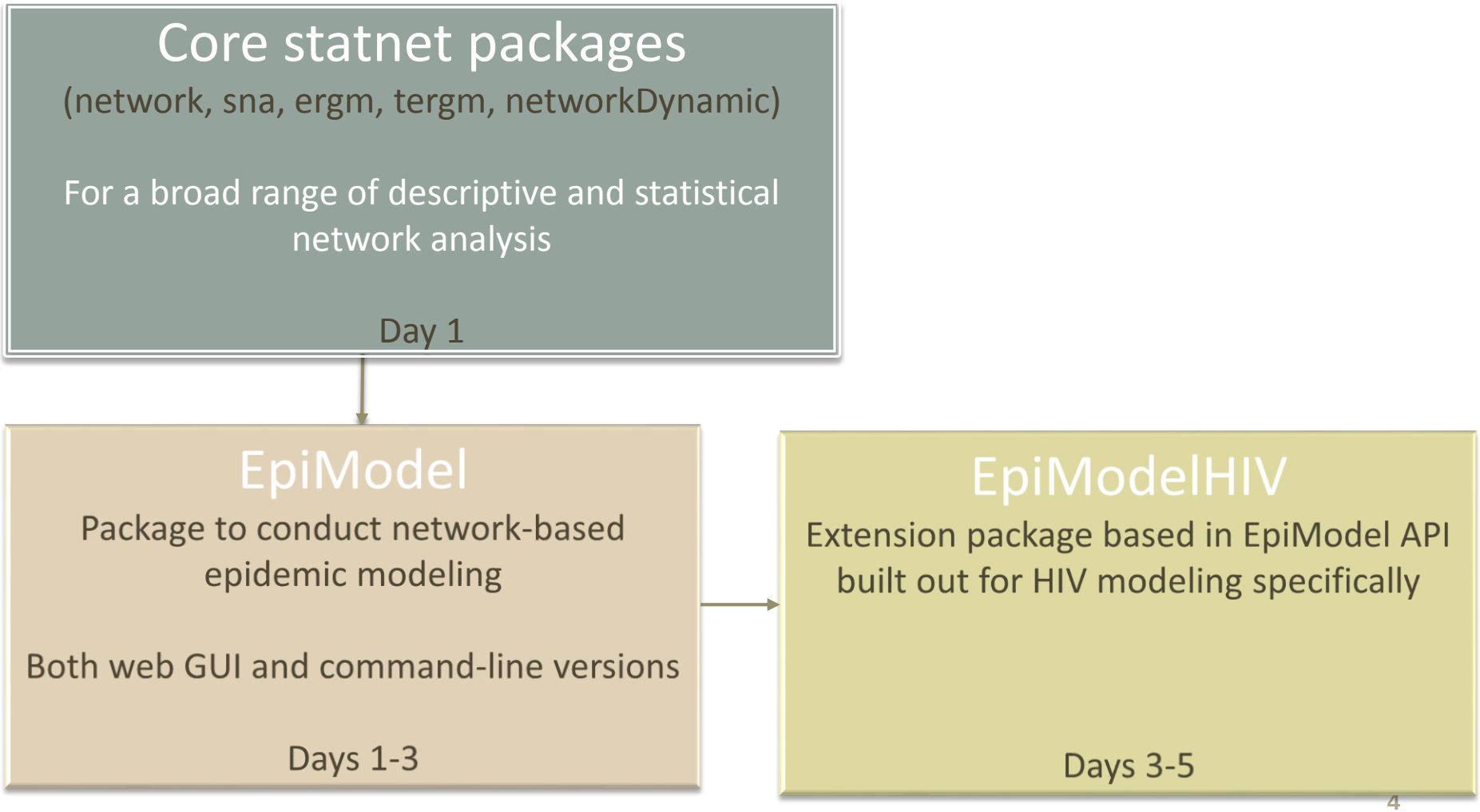
- **Materials in Days 1-3 of this workshop**
  - **based on the *Network Modeling for Epidemics* short course at UW**
  - **Co-developed with Martina Morris**
  - **Made possible by person-decades of hard work by the Statnet Development Team**
    - **Mark Handcock, Martina Morris, David Hunter, Carter Butts, Pavel Krivitsky, Steve Goodreau, Sam Jenness, Skye Bender-deMoll, and many students**

- **Materials from all days**
  - **Supported by NIH through many grants**

Overview of the week

Day	Content
1	Statistical models for static networks (ERGMs) Statistical models for dynamic networks (TERGMs)
2	Running “independent” models in EpiModel Running “dependent” models in EpiModel
3	Running “dependent” models in EpiModel (cont.) Intro to extending EpiModel Introduction to EpiModelHIV
4	Parametrization of studies on MSM and HIV/STI Interface of studies and EpiModelHIV
5	EpiModelHIV labs Calibration, validation, computation, reproducibility Open discussion

Software: based in R



Outline for today

- **Quick introduction to networks**
- **ERGM tutorial**
- **ERGM lab**
- **TERGM tutorial**
- **TERGM lab**

Intros

Why network models?

Concurrent partnerships and the spread of HIV

Martina Morris and Mirjam Kretzschmar*

Objective: To examine how concurrent partnerships amplify the rate of HIV spread, using methods that can be supported by feasible data collection.

Methods: A fully stochastic simulation is used to represent a population of individuals, the sexual partnerships that they form and dissolve over time, and the spread of an infectious disease. Sequential monogamy is compared with various levels of concurrency, holding all other features of the infection process constant. Effective summary measures of concurrency are developed that can be estimated on the basis of simple local network data.

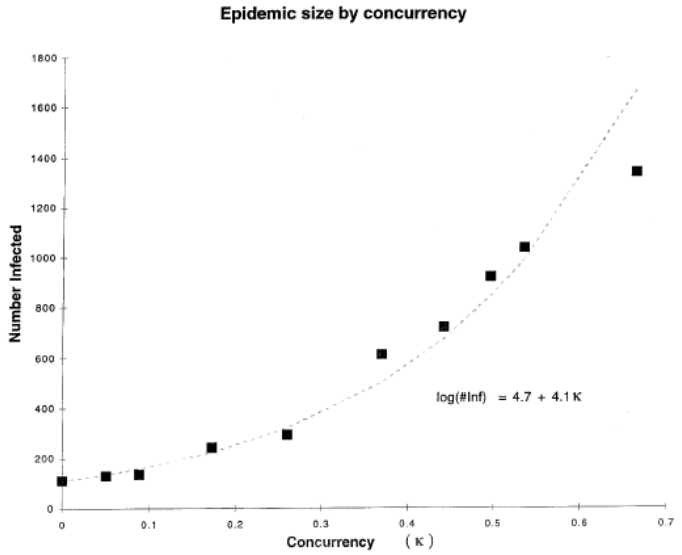


Fig. 3. The mean final size of the epidemic as a function of concurrency. Each observation represents the mean of 100 runs under the same value for the concurrency index κ. The full distribution of epidemic size under each scenario is shown in Fig. 2.

Why network models?

- Concurrency microsimulation

```
if (!require("devtools")) install.packages("devtools")
devtools::install_github("statnet/concurrency.sim")
library(concurrency.sim)
concweb()
```

Concurrency in DCMs

The Role of Acute and Early HIV Infection in the Spread of HIV-1 in Lilongwe, Malawi: Implications for "Test and Treat" and Other Transmission Prevention Strategies

Kimberly A. Powers, **Azra C. Ghani**, **William C. Miller**, **Irving F. Hoffman**, **Audrey E. Pettifor**, **Gift Kamanga**, **Francis E.A. Martinson**, and **Myron S. Cohen**

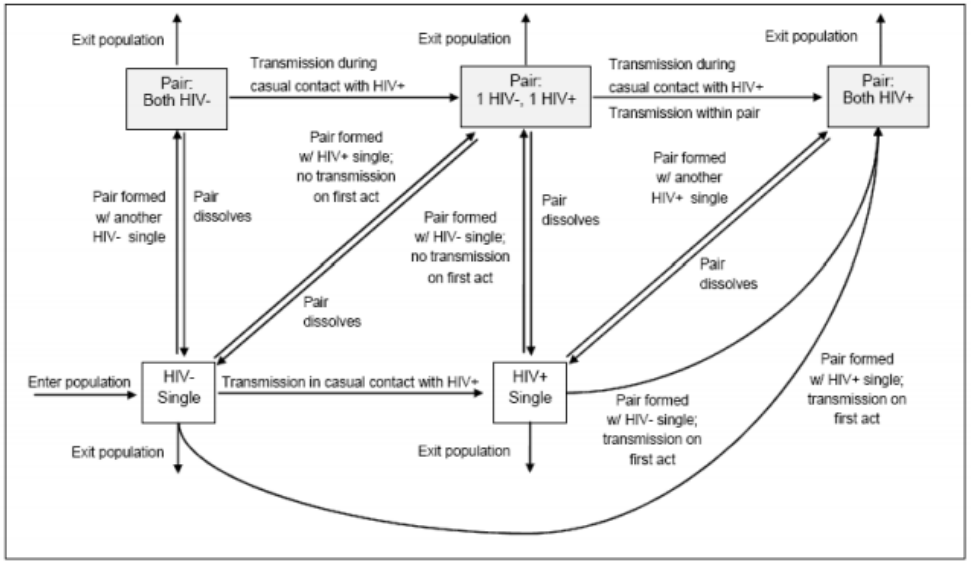


Figure 2. Simplified diagram of model structure
Unshaded boxes represent single (unpaired) individuals; shaded boxes represent steady partnerships. As detailed by the accompanying labels, arrows represent flows from one compartment to another via demographic processes (entering & exiting the population), partnership formation and dissolution, or HIV transmission. For ease of illustration, the diagram does not illustrate the two separate risk groups or the multiple stages of infection.

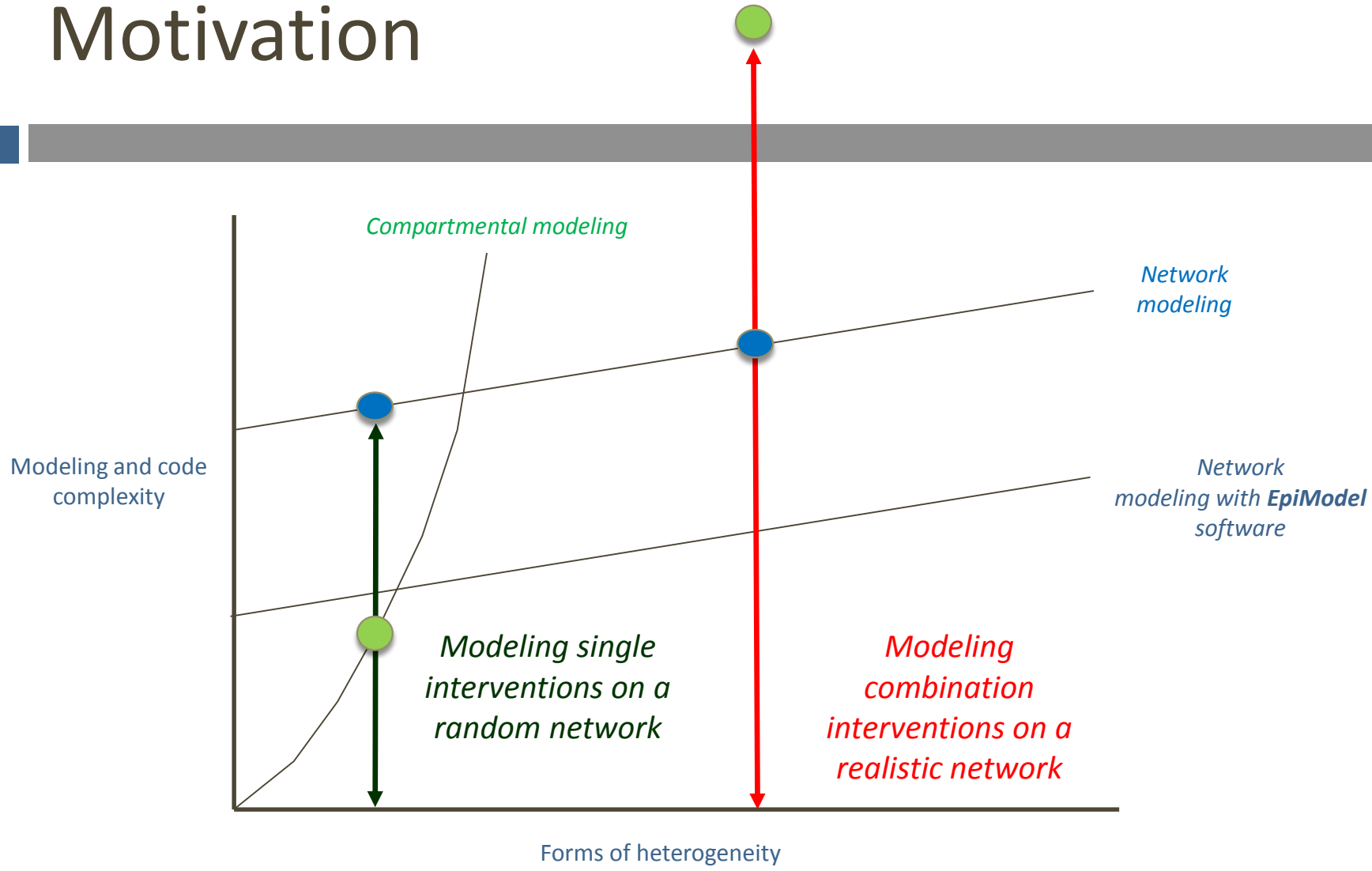
Networks

- This DCM approach leaves out:
  - Long-term concurrency
  - More than one person in the partnership having other partners
  - The outside partner having other partners
  - Any network effects beyond this (e.g. other kinds of mixing)
- Full network models can allow for all this
- Network models (and other agent-based models) also make it easier to have:
  - Many forms of heterogeneity
  - Heterogeneity that is continuous
  - Transition probabilities that are not memoryless
  - Interactions among all of these

Networks

- Network models allow for:
  - Estimation and simulation of complex network structures
  - Simulation of sexual acts (or other contacts) within the relationships in those structures
  - Simulation of infection on those networks
  - Simulation of an arbitrary number of other interacting processes
  - Feedback among all these

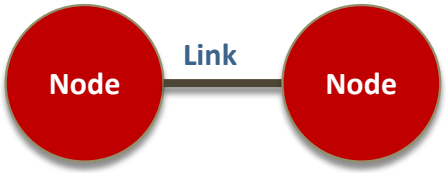
Motivation


Compartmental modeling
Network modeling
Modeling and code complexity
Network modeling with EpiModel software
Modeling single interventions on a random network
Modeling combination interventions on a realistic network
Forms of heterogeneity

Terminology

- **Node**: the entity of interest
  - for us, nodes represent people; also called actors or vertices

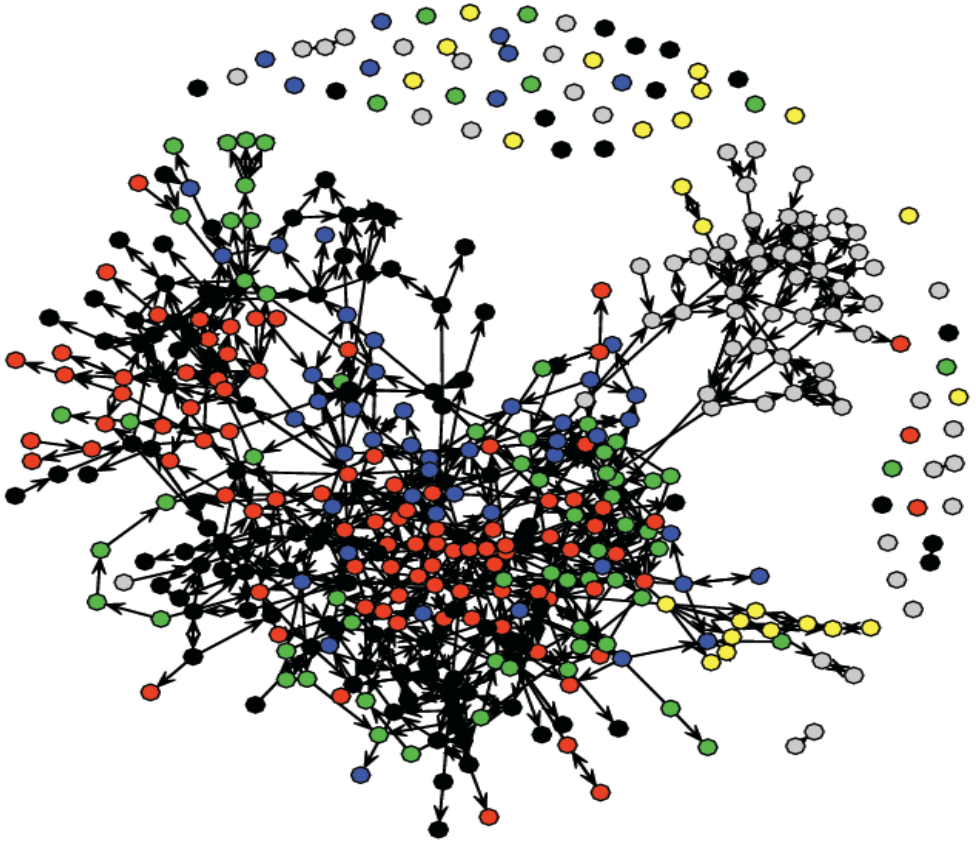
- **Link**: the relationship of interest
  - also called tie, edge, line, relationship, partnership



- **Network**: a collection of nodes and links
  - also called a graph

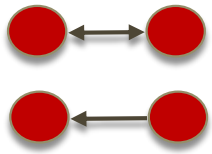
Nodes, links and networks

Beyond the pretty pictures, there are many different attributes of nodes, links and networks that have implications for the structures we can observe, and what we want to model



Link properties

- Directed (e.g., likes)
  - Mutual
  - Asymmetric



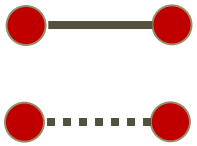
Nodes are now classified as senders and/or receivers

- Undirected (e.g., has sex with)

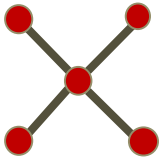
- Binary (0,1 on or off only)
- Signed and/or Valued (… -2, -1, 0, 1, 2 …)

Configurations

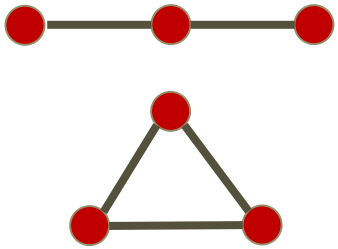
Dyads



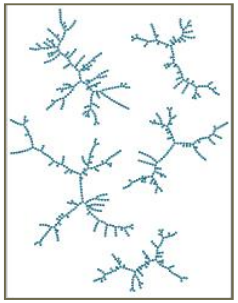
Stars



Triples & Triangles



Components



Any collection of nodes and links can be defined as a configuration

Levels of measurement

As we look at ways of describing network data, keep in mind the different levels of measurement

- Node level: *relational characteristics of a node*
  - Examples: degree
- Edge level: *relational attributes of edges*
  - Examples: duration
- Network level: *overall structural attributes and distributions*
  - Examples: number of edges, density

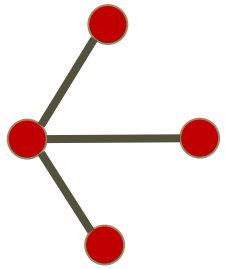
E.g. Cycles

- Paths that lead back to the origin node
  - Cycle length k = number of lines in the cycle, “k-cycles”
  - Triangles are 3-cycles
- Node level measure: Number of cycles a node is a member of
- Edge level measure: Number of cycles an edge is a member of
- Network level measure: The “cycle census” is a property of the network

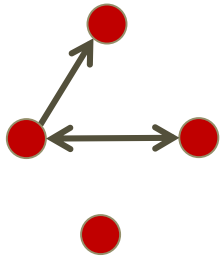
Common network level measures

- Density: Fraction of all dyads that have an edge

- Isolate count: Number of nodes without any edges



Nodes: 4
Isolates: 0
Dyads: 4*3/2 = 6
Density: 3/6 = .5



Nodes: 4
Isolates: 1
Dyads: 4*3 = 12
Density: 3/12 = .25

Types of networks

- Simplest form: 1-mode, undirected, binary ties, single relation
- 2-mode (aka *Bipartite*)
  - Two different types of nodes
  - Ties only allowed between groups

    Examples: Online network groups and persons (an Affiliation network)

    Heterosexual sex network
- Directed
  - Allows for a distinct set of in-ties and out-ties, and mutual
- Multiplex
  - More than one type of link possible

    Example: Sexual partnerships and needle sharing

Representing network data

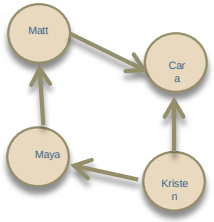
- Sociomatrix
  - (aka adjacency matrix)
  - simple but inefficient for large sparse nets (order n^2)

	Matt	Cara	Kristen	Maya
Matt	0	1	0	0
Cara	1	0	0	1
Kristen	0	1	0	1
Maya	1	0	0	0

- Edgelist

Matt	Cara
Cara	Matt
Cara	Maya
Kristen	Cara
Kristen	Maya
Maya	Matt

- Graph



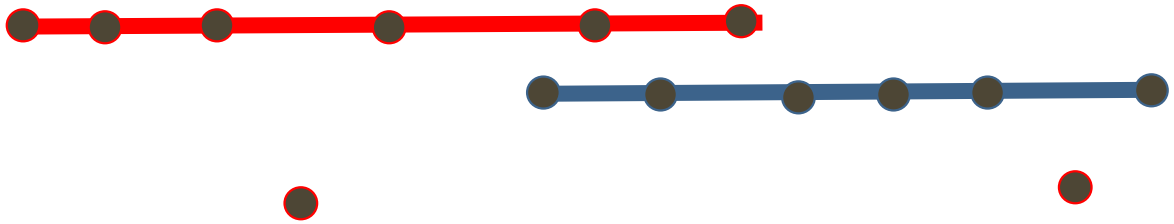
Sources of network data

- Census
- Egocentric
- Partial
  - Link tracing
  - RDS
  - Etc.

- Have no fear! Our tools work with egocentric data

Partnerships/relationships and acts

- By Day 3 you will see our fully developed EpiModelHIV code
  - Dynamic network of main partnerships (incl. individual sex acts)
  - Dynamic network of casual partnerships (incl. individual sex acts)
  - Set of one-time contacts at each time point
  - Interaction among them



Statistical Model: Basic idea

- We want to model the probability of relational ties as a function of:
  - Nodal attributes (that influence degree and mixing)
  - The propensity for certain "configurations" (like triangles)
- The dyads may be dependent
  - Note that nodal attribute effects do not induce dyad dependence
- So we model the joint distribution directly

Statistical Model: Testing

- What affects the probability that two people will become friends?

Geographical proximity	dyadic
Similar age (and other attributes)	dyadic
Main effect on age	individual
Not already having two many friends	> dyadic
Having friends in common	> dyadic

Statistical Model: Testing

- Suppose kids have a tendency to become friends with their friends' friends
  - And this is the only generative process occurring.

- Presumably, this would mean that you would observe more triangles than expected by chance in the graph.
  - How would you test this in an empirical dataset?

A basic statistical test approach

- Begin by counting the # triangles in your network
  - Say this is “T”, your test statistic
- Then determine the probability of observing T or more triangles in this network …
- And see if it is less than 5%

But … how do you determine that probability?

What is this probability distribution?

Unconditional: Start with a network this size (size = # nodes)

- enumerate all possible networks for a fixed number of nodes,
- count the number of triangles in each network, and
- construct the frequency distribution of the counts.

for 4 nodes:	# of dyads is 4*3/2 = 6 # of possible networks = $2^6 = 64$
for 10 nodes:	# of dyads is 10*9/2 = 45 # of possible networks = $2^{45} \approx$ 35 trillion
for 20 nodes:	# of dyads is 20*19/2 = 190 # of possible networks = $2^{190} \approx 10^{57}$

Limitations of this approach

- If we are ***only*** interested in whether the triangle counts are different than expected given the density of the graph
  - One can use graph theory

- But what if we want to understand the underlying generative process, e.g., homophily vs. transitivity?
  - And we want to quantify the impact of each process on our network?
  - This requires a more general statistical *modeling* framework

Exponential Random Graph Model (ERGM)

Probability of observing a graph (set of relationships) $\mathbf{y}$ on a fixed set of nodes:

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

where: $\mathbf{g(y)}$ = vector of network statistics

$\boldsymbol{\theta}$ = vector of model parameters

$\mathbf{k(\theta)}$ = numerator summed over all possible networks on node set $\mathbf{y}$

- Exponential family model
- Well understood statistical properties Besag (1974), Frank (1986)
- Very general and flexible

Exponential Random Graph Model (ERGM)

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

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

If you're not familiar with this kind of compact vector notation, the numerator is just:

$$\exp(\boldsymbol{\theta}_1 x_1 + \boldsymbol{\theta}_2 x_2 + \cdots + \boldsymbol{\theta}_3 x_3)$$

Kind of like a linear model, but a bit different (watch out for this later)

The conditional probability of a tie

The probability of the graph

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

can be re-expressed as

The conditional probability of the 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 $\boldsymbol{\partial}\left(\boldsymbol{g}(\boldsymbol{y})\right)$ represents the change in $\boldsymbol{g}(\boldsymbol{y})$ when Y_{ij} is toggled between 0 and 1

ERGM specification: $\boldsymbol{\theta' g(y)}$

The $\boldsymbol{g(y)}$ terms in the model represent "network statistics"

- These are counts of network configurations, for example:
  1. Edges: $\sum y_{ij}$
  2. Within-group ties: $\sum y_{ij} I(i \in C, j \in C)$
  3. 2-stars: $\sum y_{ij} y_{ik}$
  4. 3-cycles: $\sum y_{ij} y_{ik} y_{jk}$

- A key distinction in the types of terms:
  - Dyad independent (1 & 2 are examples)
  - Dyad dependent (3 & 4 are examples)

ERGM specification: $\boldsymbol{\theta' g(y)}$

Model specification involves:

1. Choosing the set of network statistics $\boldsymbol{g(y)}$
   - From minimal : # of edges
   - To saturated: one term for every dyad in the network

2. Choosing "homogeneity constraints" on the parameter θ, for example, with edges:
   - all homogeneous
   - group specific (e.g., sex or age specific)
   - dyad specific

Model Estimation

- For many models there is no “closed form” or analytic solution for the estimated coefficients (as there is in OLS)
- Instead, we rely on a defining property of Maximum Likelihood Estimates (MLEs)
  - At the MLE of the coefficients, the expected values of the statistics under the model equal the observed statistics
- And we find these MLEs using an iterative search algorithm
  - The algorithm is called “Markov Chain Monte Carlo” (MCMC)
  - Start with some initial θ values, simulate networks from those values, compare the mean statistics to the observed, update the values of θ
  - Repeat until the (expected – observed) < epsilon

Model estimation

- What does it mean to “simulate networks from those values”?
  - Pick a dyad at random
  - Toss a coin to set the tie status
    - The probability of the tie is determined by the model
  - Repeat (many many many times)
- This produces a Markov Chain of networks
  - Sample from this chain, every 1000th element (say)
- Calculate the mean of the model statistics from this sample
  - And compare the this mean to the observed network statistics

Computationally intensive estimation

- Has been key to statistical estimation of complex (i.e., realistic and interesting) models for dependent data
  - And to the emergence of the field of "data science"
- In most cases, it works really well
  - And there is lots of mathematical theory proving it has good convergence properties
- … but, it can run into trouble
  - especially if the model you're trying to fit is not a good one for the observed network

Model Degeneracy

- Models with dyad dependent terms can behave differently than we expect
  - They look simple
  - But they represent effects that cascade through a network via a chain of dependence
- Homogeneous triangle and k-star terms turn out to be some of the worst offenders

Model Degeneracy

- Technical Definition:

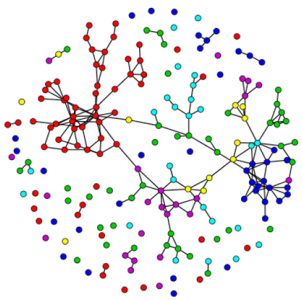
When a model places almost all probability on a small number of uninteresting graphs

- Most common "uninteresting" graphs:
  - Complete (all links exist)
  - Empty

- Model degeneracy = misspecification

Testing goodness of fit

- Traditional GOF stats can be used
  - AIC, BIC in the model summary
- We also take another approach
  - We are interested in how well we fit aggregate properties of the network structure that we did **not** include as model terms
  - This helps to identify *what* the model gets wrong
  - We use 3 “higher order” statistics:
    - Degree distribution
    - Shared partner distribution (non-parametric) (local clustering)
    - Geodesic distance distribution (global clustering)



DATA

MODEL

ESTIMATED COEFFICIENTS

SIMULATED DATA (draws from the prob. dist.)

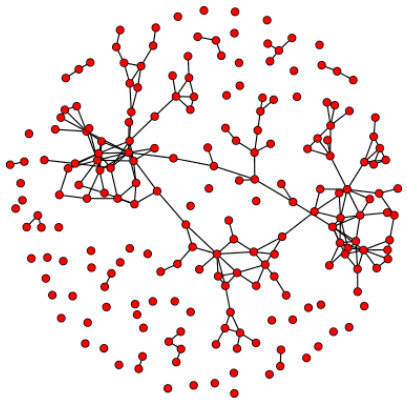
HIGHER ORDER GRAPH STATISTICS OF DATA

HIGHER ORDER GRAPH STATISTICS OF SIMULATED DATA

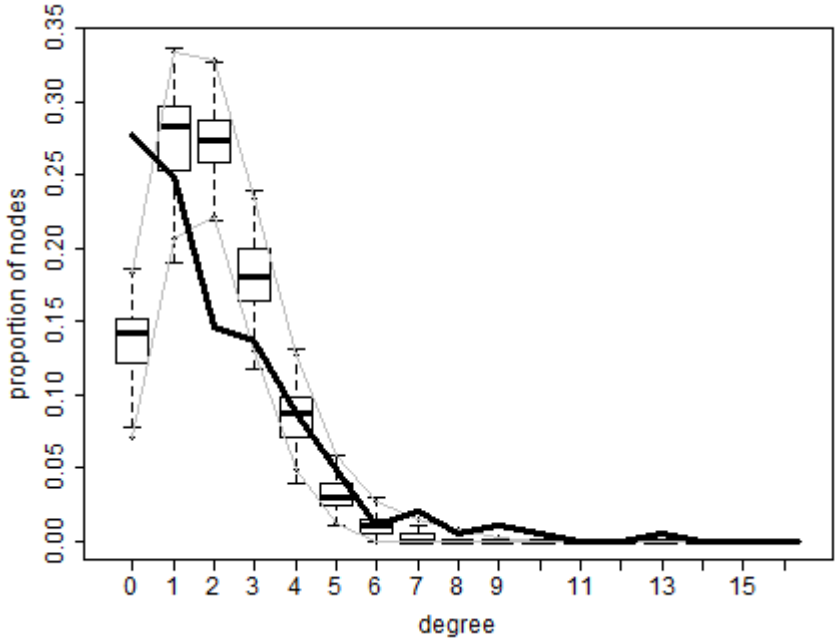
GOODNESS OF FIT OF MODEL TO DATA

Goodness of fit measure 1: degree distribution

Data:

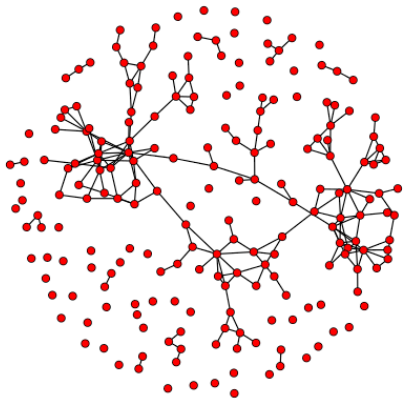


Model: Bernoulli
(i.e. edges only)

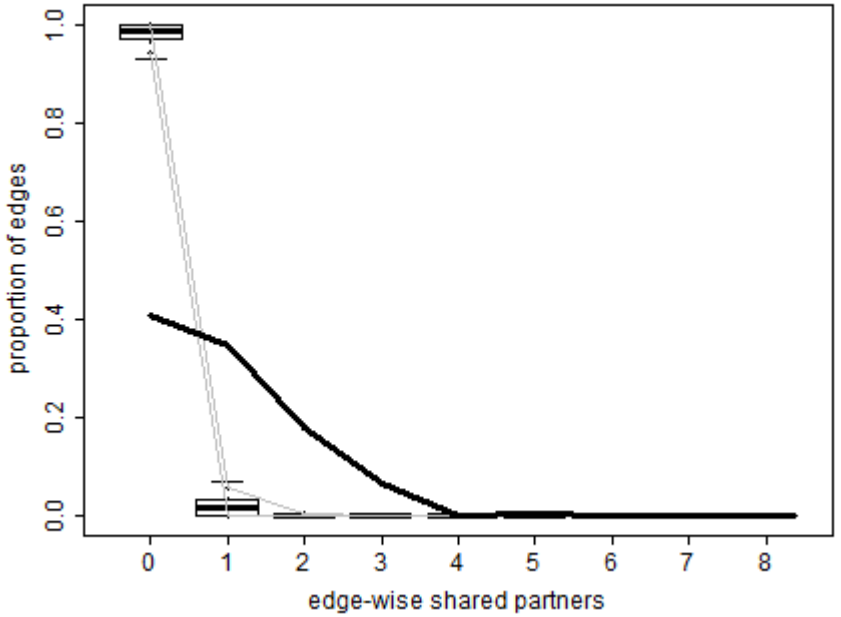


Goodness of fit measure 2: ESP distribution
(local clustering)

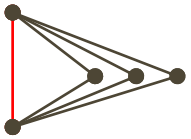
Data:



Model: Bernoulli
(i.e. edges only)

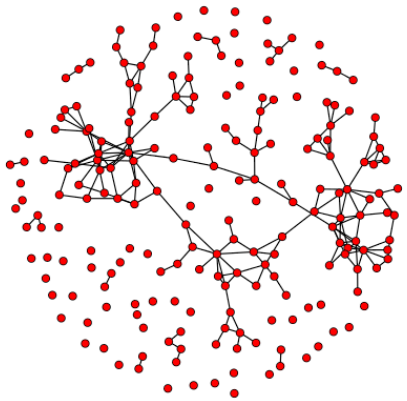


This edge has an ESP value of 3 →

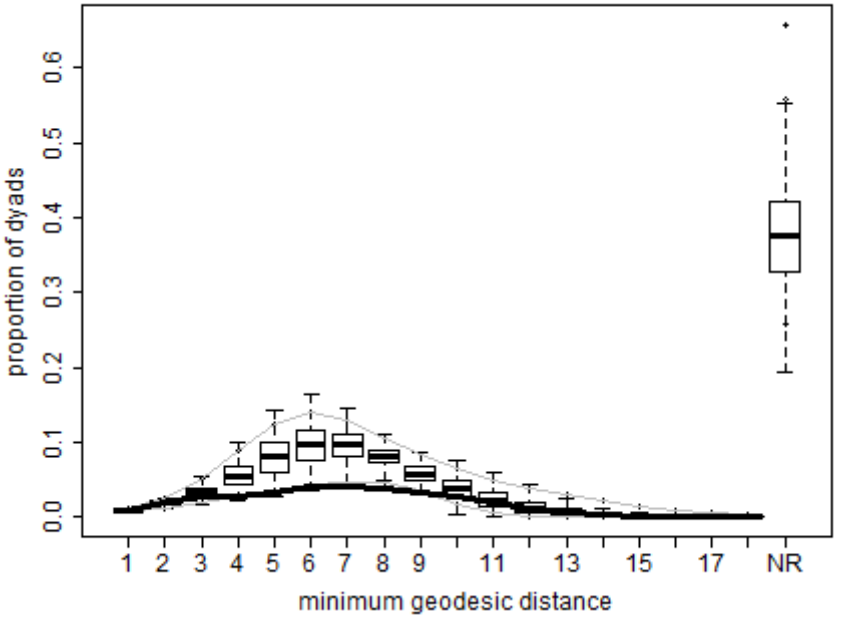


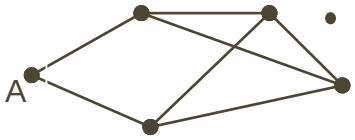
Goodness of fit measure 3: geodesic distribution
(global clustering)

Data:



Model: Bernoulli
(i.e. edges only)

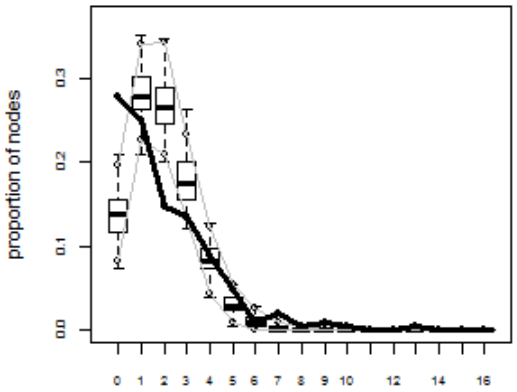




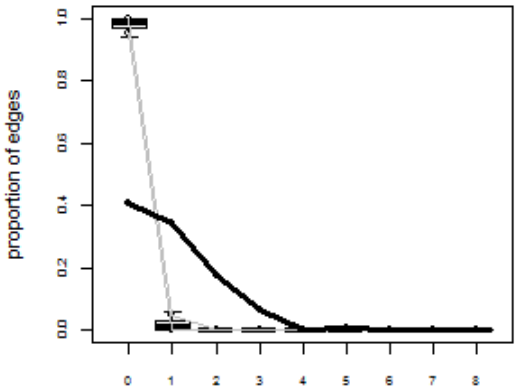
A/B have geodesic 2
A/C have geodesic ∞

Goodness of fit measures assembled

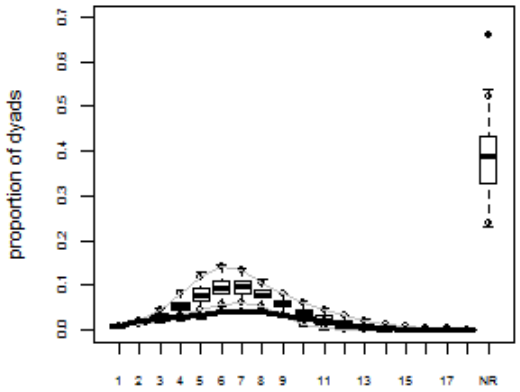
faux.mesa.high ~ edges



degree



edgewise shared partner



geodesic

Summary: Not a good fit to any of the aggregate structural properties observed

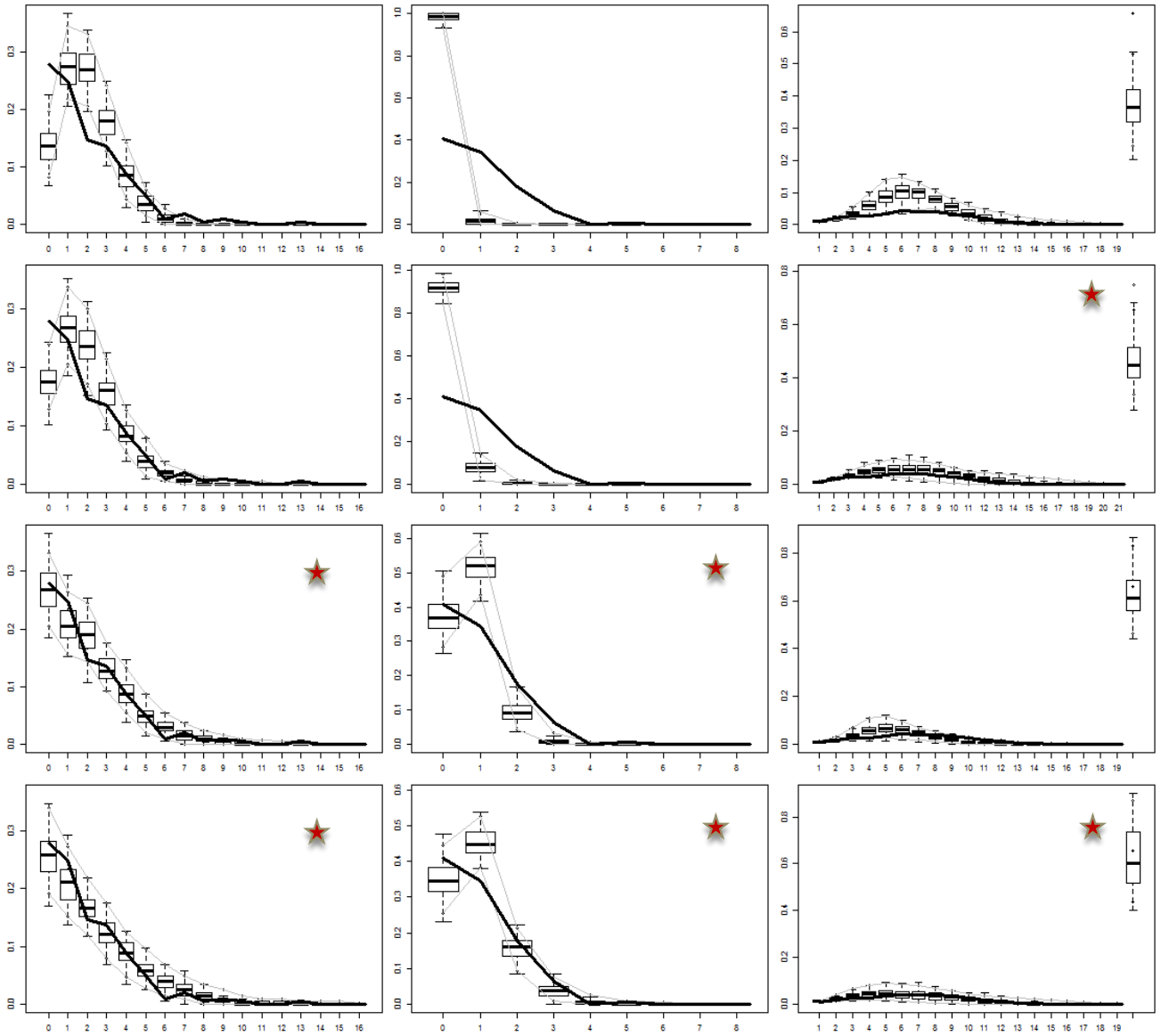
All 4 models: degree | edgewise shared partner | geodesic

Model: Edges
AIC: 2288

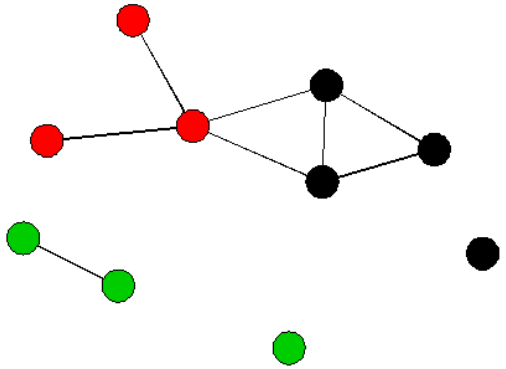
Model:
Edges + Attributes
AIC: 1809

Model:
Edges + GWESP
AIC: 1999

Model: Edges + Attributes + GWESP
AIC: 1648



Common statistics in ergms



undirected network of 10 nodes, including nodal attribute "color", with values:

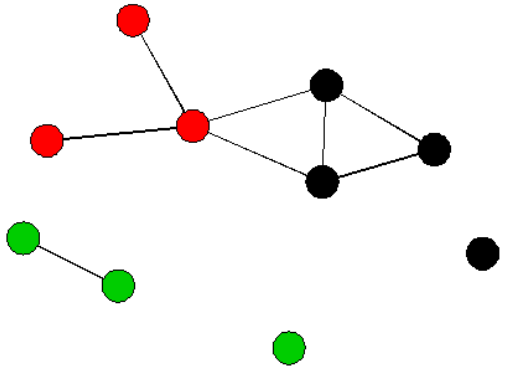
1=black,

2=red,

3=green

Term	Formula	Unit	Value(s)
`~edges`	# of edges	edges	8
`~nodefactor("color")`	Sum of degrees for nodes of each color	nodes/edges*	[8,] 6, 2
`~nodefactor("color", base=2)`	Sum of degrees for nodes of each color	nodes/edges*	8, [6,] 2
`~nodematch("color")`	# of edges between nodes of same color	edges	6
`~nodematch("color", diff = TRUE)`	# of edges between nodes of same color, for each color	edges	3, 2, 1

Common statistics in ergms



undirected network of 10 nodes, including nodal attribute "color", with values:

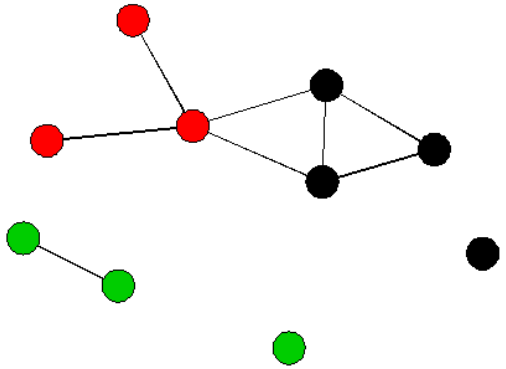
1=black,

2=red,

3=green

Term	Formula	Unit	Value(s)
`~nodemix("color", base=1)`	# of edges between nodes of each color combo	edges	[3,] 2, 2, 0, 0, 1
`~degree(0)`	# of nodes of degree 0	nodes	2
`~degree(2:5)`	# of nodes of degrees 2, 3, 4, 5 each	nodes	1, 2, 1, 0
`~concurrent`	# of nodes of at least degree 2	nodes	4

Common statistics in ergms



undirected network of 10 nodes, including nodal attribute “color”, with values:

1=black,

2=red,

3=green

Term	Formula	Unit	Value(s)
`~triangle`	# of triangles (beware!)	triangles	2
`~gwesp(0)`	# of edges in at least one triangle	edges	5

Summary

- Network structure influences transmission dynamics
- Statistical models for networks (ERGMs) provide a way to
  - estimate and evaluate hypotheses
  - about the generative processes that lead to the structures we observe
- And the fully specified models can also be used to simulate networks.
  - The expected values of the model statistics from the simulated networks will match the statistics in the observed network
- Of course, the networks we want to simulate need to be dynamic (and that's where we'll go this afternoon…..)

Selected References

Dietz K, Hadeler K: Epidemiological models for sexually transmitted diseases**. *Journal of Mathematical Biology 1988, 26(1):1-25.***

Watts CH, May RM: **The Influence of Concurrent Partnerships on the Dynamics of HIV/AIDS. *Math Biosc 1992, 108:89-104.***

Morris M, Kretzschmar M: **Measuring Concurrency. *Connections 1994, 17(1):31-34.***

Morris, M., S. Goodreau and J Moody (2007). Sexual Networks, Concurrency, and STD/HIV. Sexually Transmitted Diseases. K. K. Holmes, P. F. Sparling, W. E. Stammet al. New York, McGraw-Hill**: 109-125.**

Goodreau, S., et al. (2009). "Birds of a Feather, or Friend of a Friend? Using Statistical Network Analysis to Investigate Adolescent Social Networks." Demography **46(1): 103–125.**

Journal of Statistical Software (v42) 2008 – Eight papers on ERGMs and statnet