NME
2021

Network Modeling for Epidemics

DAY 2:
STATISTICAL METHODS FOR NETWORK ANALYSIS

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

Supported by the US National Institutes of Health

Today we will cover

Three classes of statistical models for networks

- Simple null models
- Generative models for static networks

morning

- Generative models for dynamic networks

afternoon

Morning session outline

- Session 1: Basics of null hypothesis testing in networks
  - Testing single statistics

Tea Break

- Session 2: ERGMs part 1, joint estimation for multiple terms
  - The complete ERGM workflow: descriptives to model assessment

Group Lab: Fit ERGMs to faux.mesa.high

- Session 3: ERGMs part 2
  - Improving model specifications (for dyad-dependent terms)
  - Estimating ERGMs from egocentrically sampled data

Note

- We'll cover a lot of ground here
  - some of the material and vocabulary may be unfamiliar

- Don't worry if you don't understand everything
  - Focus on getting the big picture, not the details
  - EpiModel puts a lot of this behind the curtain
    - So you don't have to deal with it, for the most part
    - But … these details do matter when you have a problematic model
    - That's why we give you this overview

- And – *don't be afraid to ask questions*

Statistical Testing: Basics

How do you know if your network is significantly different than a simple random graph?

Description vs. Inference in statistics

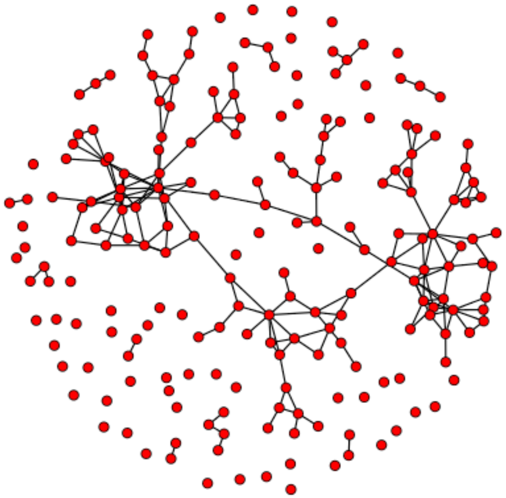
- So far we have been using descriptive statistics to explore our network data
  - Density
  - Degree and geodesic distributions
  - Mixing matrices
  - Component size distributions

- Next, we might want to compare these statistics to what we would expect by chance
  - What do we mean “by chance”?
  - Is there a natural “null hypothesis test” in this context?

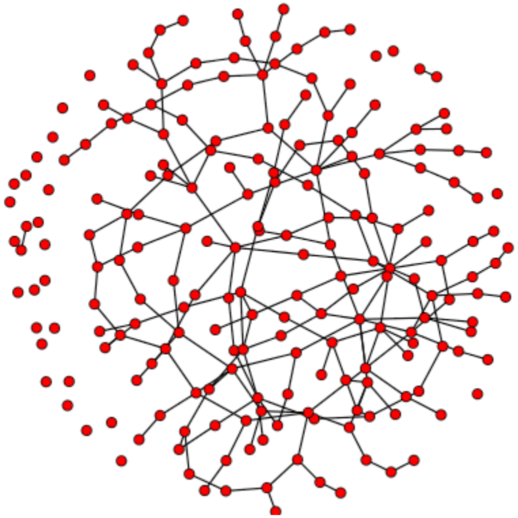
Recap

- Does the structure of our social network differ from a simple random graph?

faux.mesa.high network



Simple random graph with the same tie probability



- What are some structural differences you can see?

Consider triangles

- 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 for a specific network?

A basic statistical test for triangles

- 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?

For that you need a null hypothesis of some sort

What is the natural null hypothesis?

- It turns out there’s more than one …

- But they all get *used* the same way when constructing a statistical test.
  - To create a sampling distribution consistent with the null
  - And compare your observed value to that distribution

Null probability distribution (1)

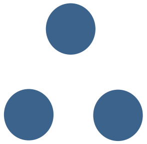
Unconditional: For a network this size (size = # nodes)

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

- Where does the number of triangles in your network lie in this distribution?
  - Top 5%?
  - Bottom 5%
  - Near the middle?

Null probability distribution (1)

For example: Take a network with 3 nodes

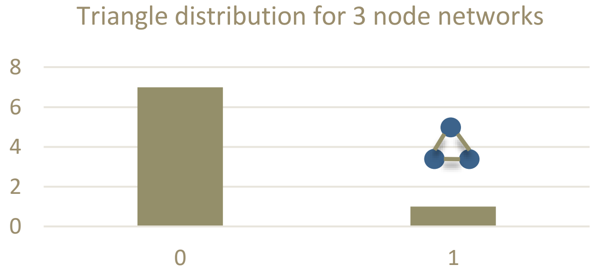


- How many dyads are there?

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = 3$$

- How many different networks on these dyads?
  - Every dyad has 2 possible values, and there are 3 dyads
  - So the number of possible networks is: $2^3 = 8$

- What is the distribution of triangle counts?
  - 7 networks have 0 triangles
  - 1 network has 1 triangle


Triangle distribution for 3 node networks

- So if your network has 1 triangle, what do you think?

Null probability distribution (1)

One problem with the unconditional null distribution

- enumerate *all possible networks* for a fixed number of nodes

this is not so easy with larger networks

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 = $2^{190} \approx 10^{57}$

We can solve this problem by sampling from the space of networks.

Null probability distribution (1)

More important question for the unconditional null distribution

- Do you really care about comparing your network to networks with zero ties?
- Or all possible ties?
- Or does it make more sense to compare your network to other networks with the same number of ties?

Conditional on density,
does your network have more or less triangles than expected?

Null probability distribution (2)

Condition on the *density*, the number of nodes and links

This is the Conditional Uniform Graph test (CUG)

- enumerate all possible networks for a fixed number of nodes **and links**,
- count the number of triangles in each network,
- construct the frequency distribution of the counts
- compare the value in your network

This also reduces the sample space

but it's still a lot of graphs… $\binom{\binom{n}{2}}{e} = \binom{n}{2}!\,/e!\,(\binom{n}{2} - e)!$

so we will still need to sample from this space in practice

The CUG is implemented as a permutation test

- Since full enumeration is typically not possible

- We sample the enumeration space by permutation
  - Randomly choose a tied dyad, and a dyad without a tie
  - *Permute* the tie and the non-tie
    - This preserves the exact density in the network
  - Count the number of triangles in the new network
  - Repeat until you have the desired sample size

- Permutation tests are often used in statistics
  - When the distribution of a sample statistic is not known

Null probability distribution (3)

Condition on the *probability of a tie*

This is the Bernoulli Random Graph test (BRG)

- Similar to the CUG, but treats density as a random variable
- Implemented via Markov Chain Monte-Carlo (MCMC)
  - Randomly choose a dyad
  - Flip a coin with probability(tie) = density of the network
    - This will not preserve the exact density for each network, but will preserve it *on average*
  - Repeat many times, then count the number of triangles in the final network
  - Repeat until you have a sample of the desired size

Null models in statnetWeb

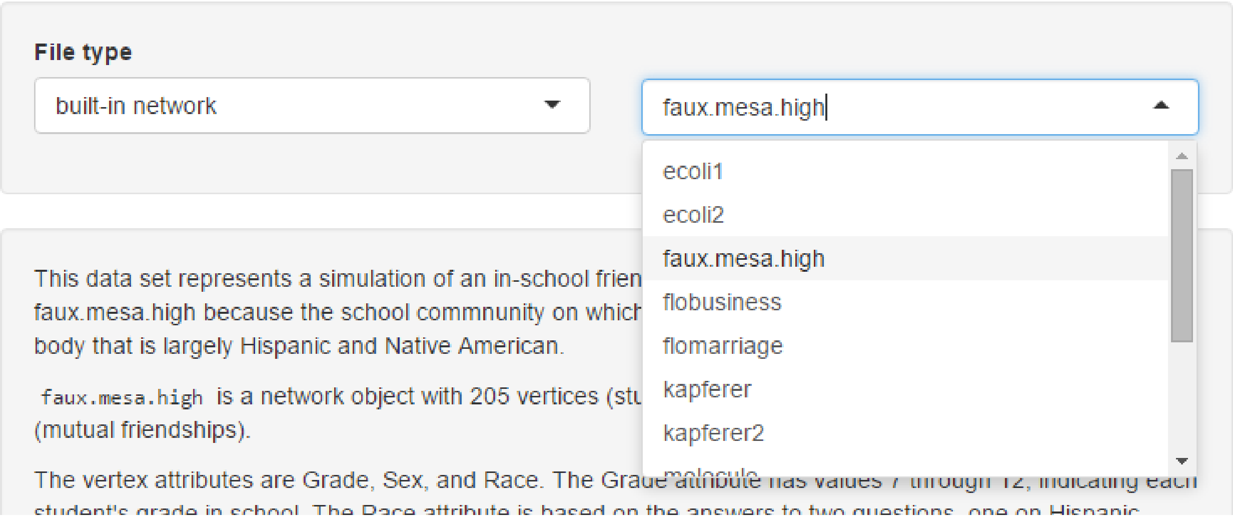
- *Select* a summary measure for the observed data
- *Compare* it to the distribution simulated from a null model
- In statnetWeb:
  - We can plot null distribution overlays on degree and geodesic distributions
  - And plot the CUG and BRG distributions for selected network summary measures

To statnetWeb

For some simple null hypothesis tests

Getting started

- Open statnetWeb and load the faux.mesa.high network
  - `library(statnetWeb); run_sw();`



File type

built-in network

faux.mesa.high

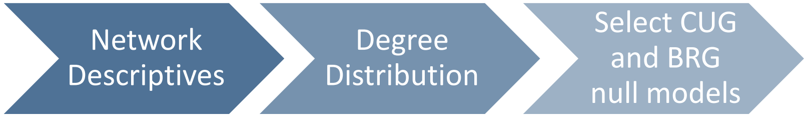
This data set represents a simulation of an in-school frien
faux.mesa.high because the school commnunity on which
body that is largely Hispanic and Native American.

`faux.mesa.high` is a network object with 205 vertices (stu
(mutual friendships).

The vertex attributes are Grade, Sex, and Race. The Grade attribute has values 7 through 12, indicating each
student's grade in school. The Race attribute is based on the answers to two questions, one on Hispanic

In statnetWeb: Degree distribution

Compare the degree distribution in faux.mesa.high to what we would expect by chance

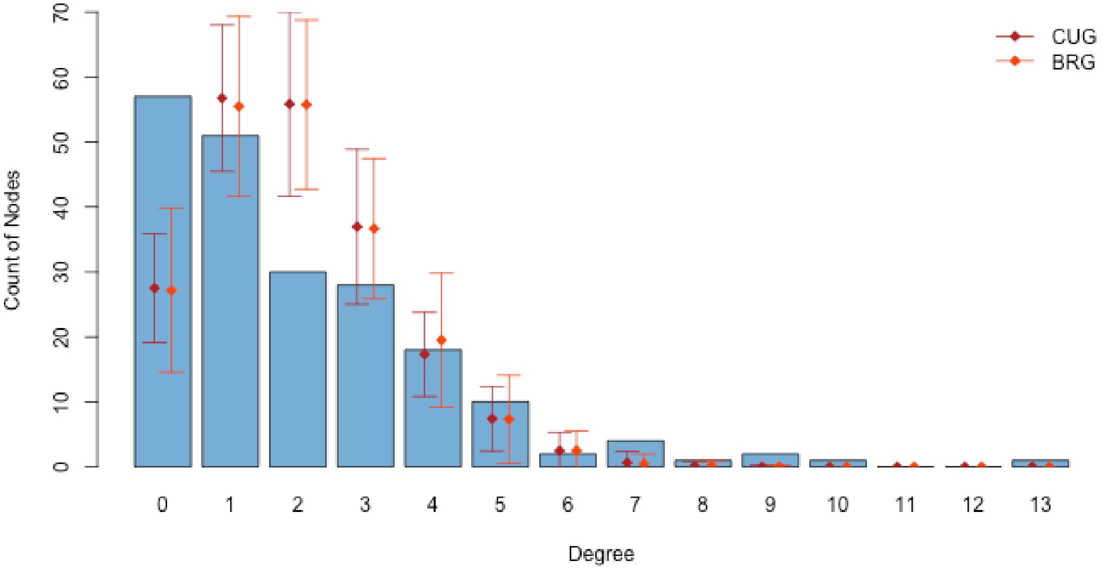


Network Descriptives

Degree Distribution

Select CUG and BRG null models

Overlays the mean and 95% confidence intervals from 100 simulations



What do you see now?

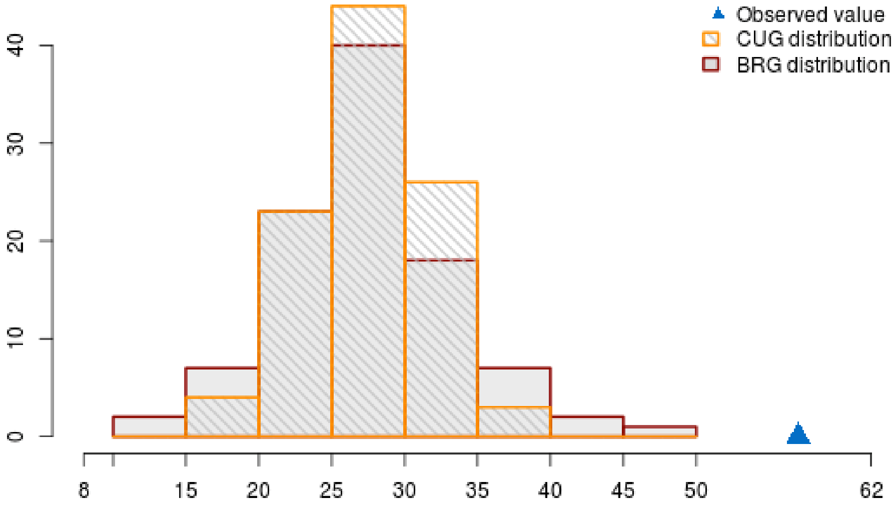
Test for the number of isolates

Compare the number of isolates in faux.mesa.high to what we would expect by chance

Network Descriptives

More

Null model tests



How likely is the number of isolates we observed in our network, under the null model?

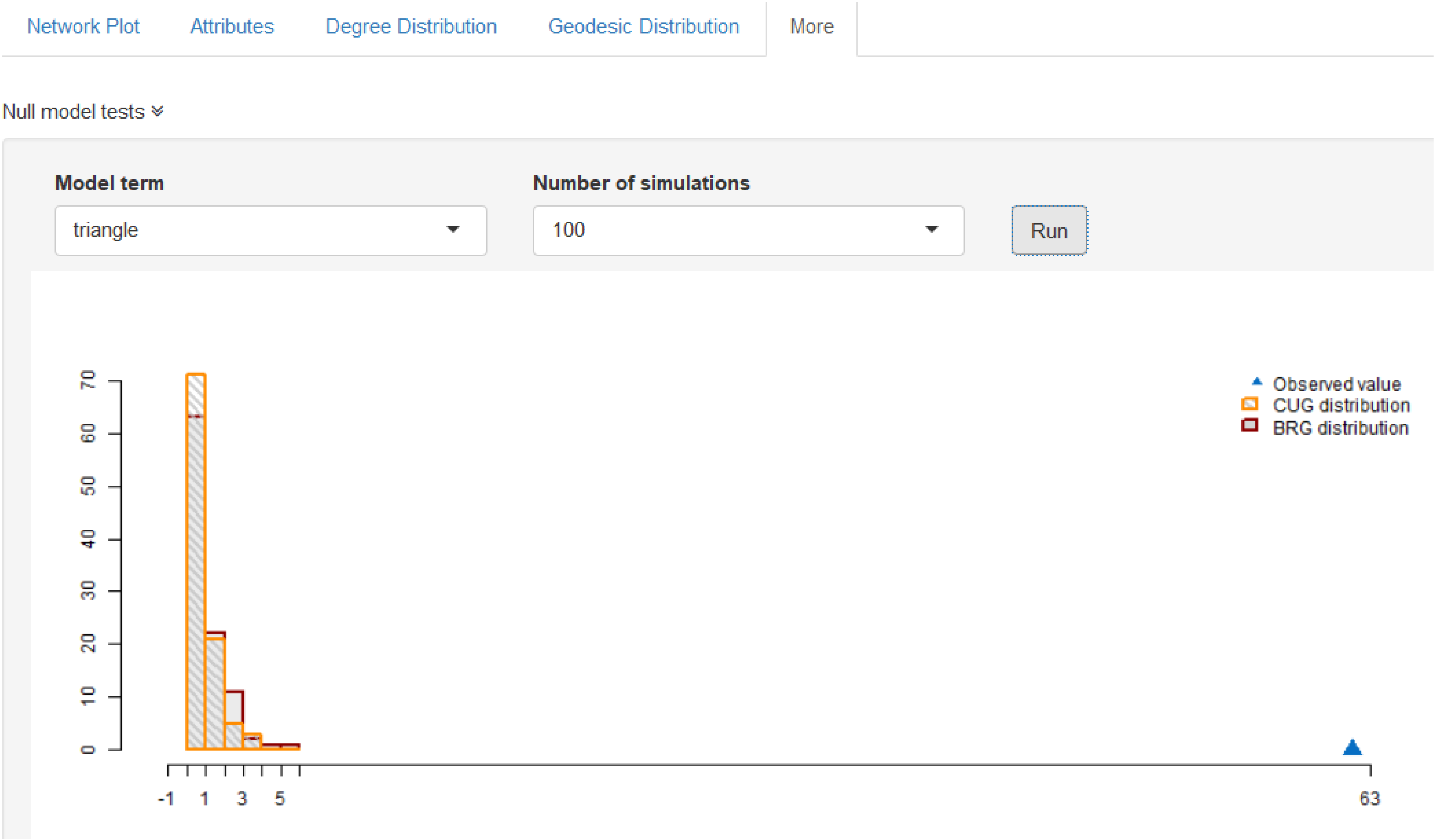
CUG test for triangles

- Are there more triangles in the observed network?

- Choose the triangle term from the dropdown menu and run 100 simulations to see how our network compares to the two null models

  - “CUG” and “BRG”

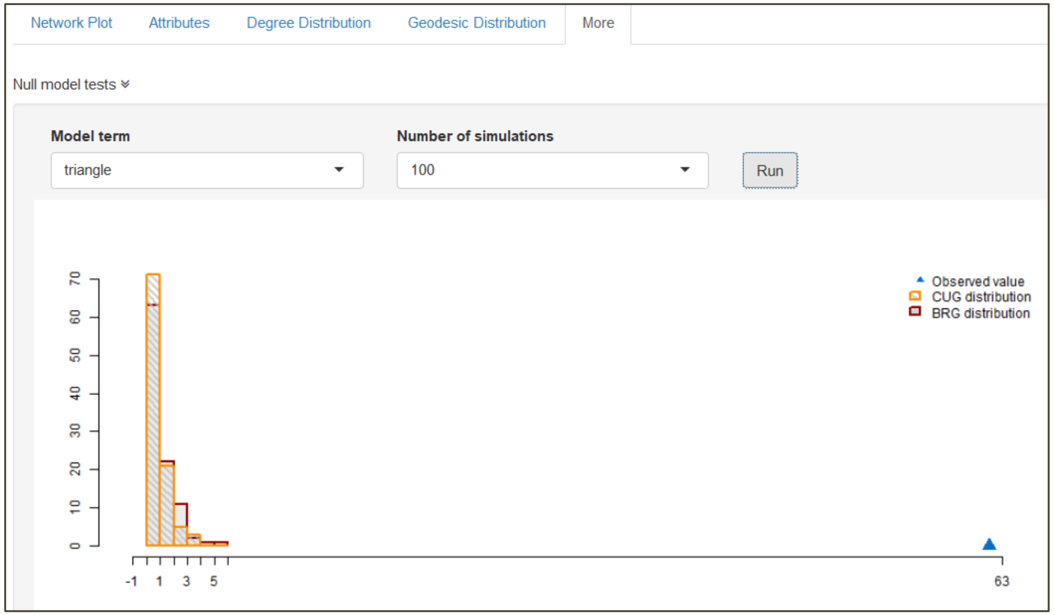
Indeed…



Model term: triangle

Number of simulations: 100

Yes the observed triangle count is high



Model term: triangle

Number of simulations: 100

- But why?

… a simple null hypothesis test doesn’t provide any insight about that.

Moving on to ERGMs