Social Network Analysis

ERGM Diagnostics

Robin Burke

DePaul University

Chicago, IL

Thanks to Carter Butts for some slides and illustrations

Tools

- gof model plot
  - shows how well the parameters match in simulated networks
- gof (non-model) plots
- mcmc.diagnostics
  - I prefer center=F, which keeps the graphs in terms of model variables
  - instead of z-scores
- Calculates various diagnostics on MCMC output
- Correlations and lagged correlations of model statistics
- Convergence diagnostics
- Term statistic plots

gof model plot

- Function call
  - gof(fit, GOF=~model)
- View
  - summary(gof)
  - plot(gof)
- What it does
  - collect statistics over 100 simulated networks
  - compare to observed network
  - p value compares distributions
    - high is good

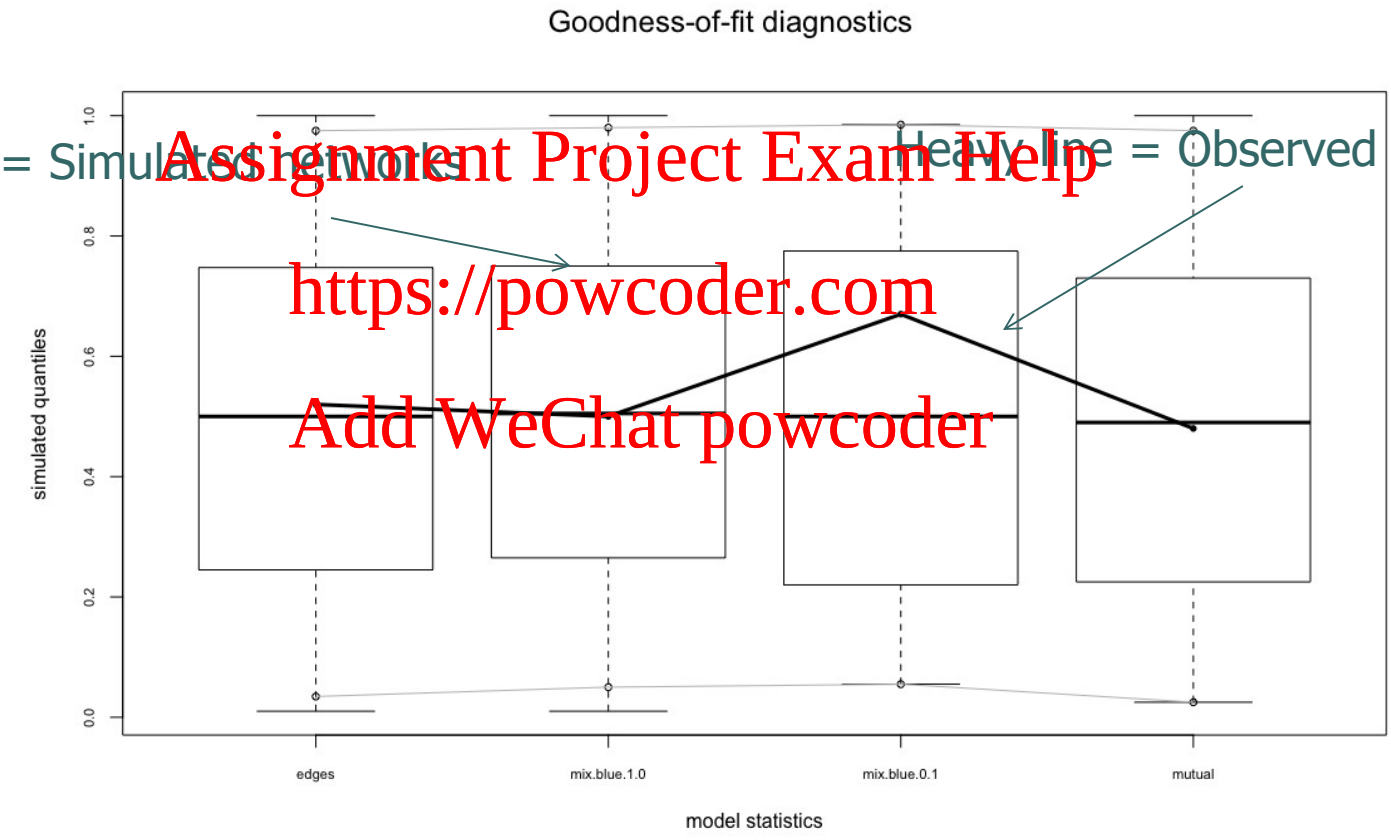
gof model summary

```
Goodness-of-fit for model statistics
                obs min   mean max MC p-value
edges           103  78 102.37 128       1.00
mix.blue.1.0      7   1   6.56  13       1.00
mix.blue.0.1      2   0   2.07   5       1.00
mutual           20  19  20.63  31       0.36
>
```

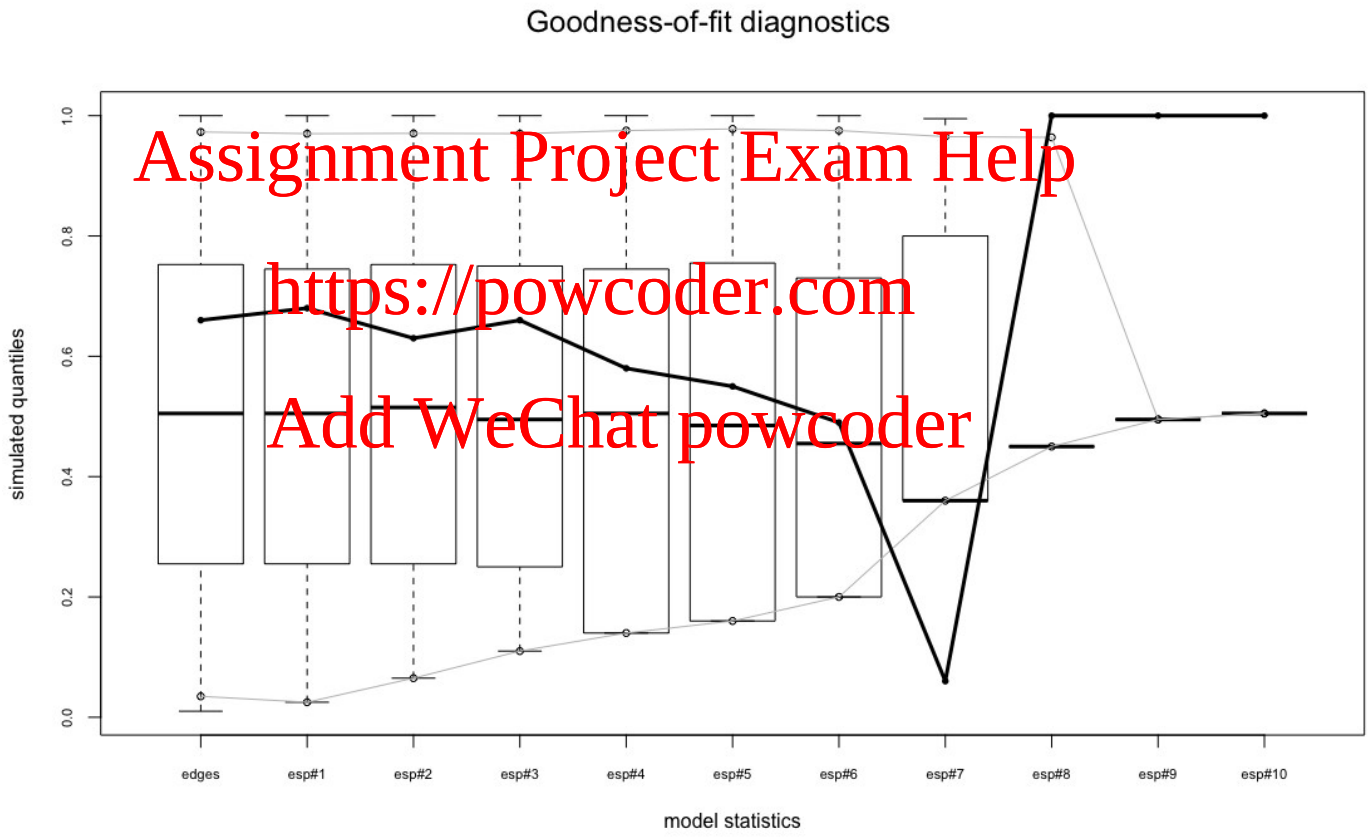
gof model plot

Boxplots = Simulated networks

Heavy line = Observed network



Bad example



What to learn

- Do the simulated networks match the observed one?
  - in terms of the model parameters
- If not
  - possibly not enough computation
    - longer burn-in interval
    - larger sample size
  - possibly a bad model

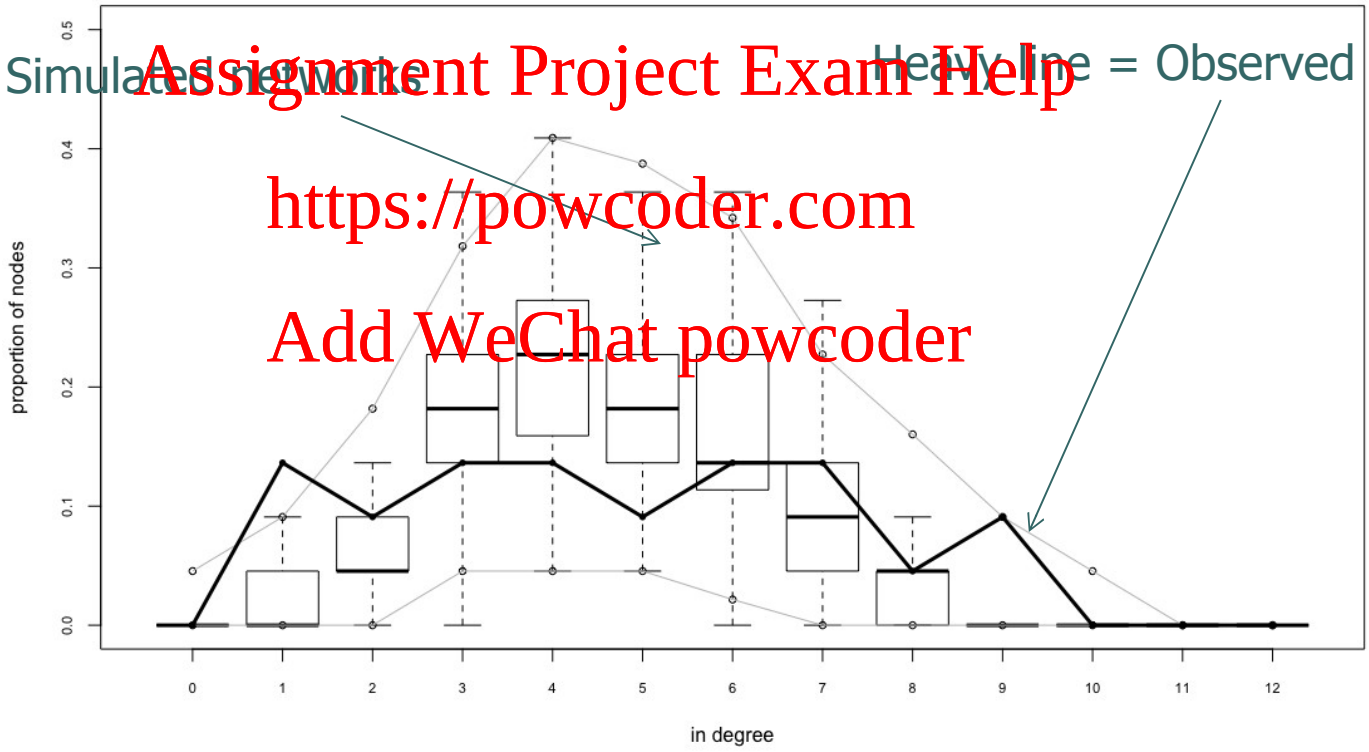
gof external

- Function call
  - gof(fit)
- View
  - summary(gof)
  - plot(gof)

Plot (in-degree)

Boxplot = Simulated networks

Heavy line = Observed network



Metrics

- Degree distribution
- Edgewise shared partner
- Minimum geodesic distance
- These are hard to fit
  - so if the model matches
    - good sign
- Degree distribution characterizes a network strongly
  - better if this matches well

What to learn

- How do the simulated networks match the original one
  - Especially useful if you didn't fit these metrics
- May need to add terms to correct
  - here maybe idegree 1 and 2
- Also computation or model selection problems

Markov chain diagnostics

- Many pieces of information here
- mcmc.diagnostics(fit, center=F)
  - otherwise values converted to z-scores

Empirical distributions

- Shows the distributions of model terms
  - not quite as useful as gof model
  - similar information

Cross-correlations

- Cross-correlations close to 1 or -1 can be bad
  - indicate that the model has terms that are co-linear
- Sometimes unavoidable
  - should always investigate
- What to learn
  - should terms be dropped from the model?

Auto-correlations

- Correlation over time
  - final values
  - compared to earlier in the Markov evolution
  - Lag = # of mixing steps
- If autocorrelation is close to 1
  - the graphs are very similar to earlier ones
  - the Markov chain is still close to starting point
- What to learn
  - possibly more burn-in time is needed

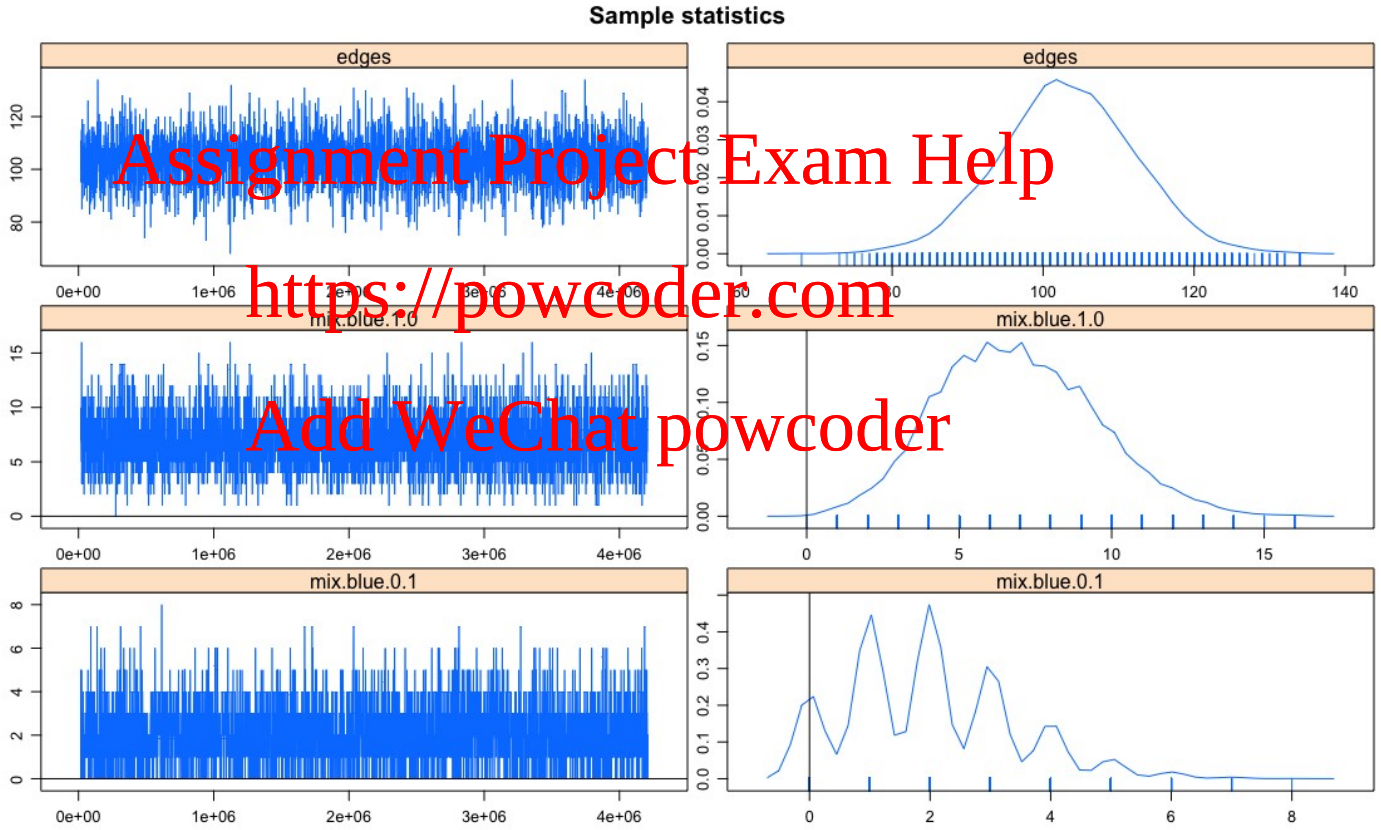
Geweke statistics

- Similar to autocorrelation
- Compare the last 50% of the samples
  - with the first 10%
- p values
  - How likely are these to be drawn from the same distribution?
  - High = bad
    - because you want the distribution at the beginning and end to be different
- What to learn
  - more burn-in time

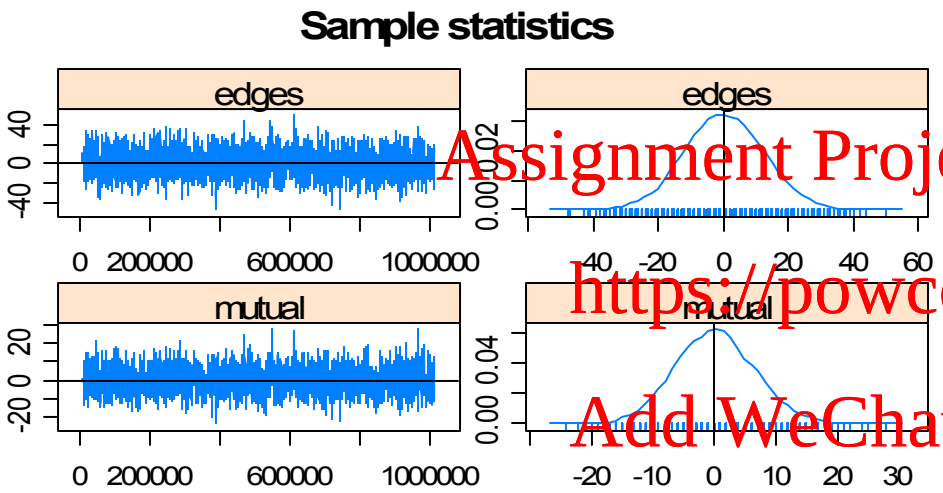
Sample statistics plots

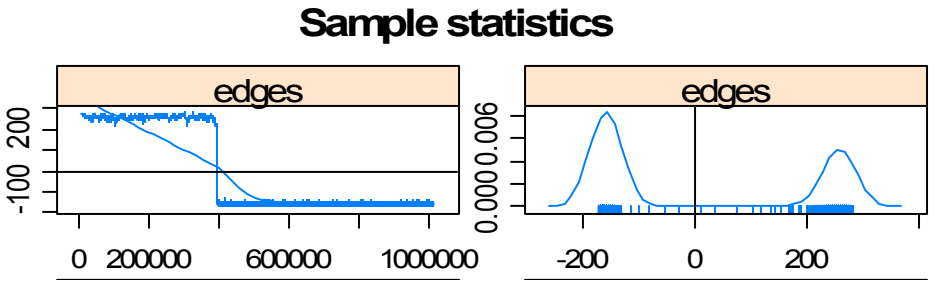
- Show how the term statistics vary over the sampled networks

Example



Good vs bad fits





What to learn

- If the plots are “normal” looking
  - then the chain converged in a “continuous” fashion
- If not
  - the results are not to be trusted
- Possibilities
  - bad model
  - additional burn-in needed

How to use

- First look at gof model
  - if this is bad, then your model fit doesn't match what you were trying to fit
  - probably your model is bad
    - Or more fitting is required.
- Next look at gof external
  - esp. degree distribution
  - decide if you can live with it
  - see where the model has problems
- Then look at diagnostics
  - do you believe the results?
  - was there enough mixing / sampling?

#1 Advice: Start simple

- Use a simple set of terms
  - consistent with your hypothesis
- Don't throw in more complex terms
  - just because you can
- Example in lab

Lab

- Room Daley 505
- Start at 7:30
- Otherwise the recording won't work!

Next week

- Distributed graph computation
- Readings:
  - GraphX
  - Pregel