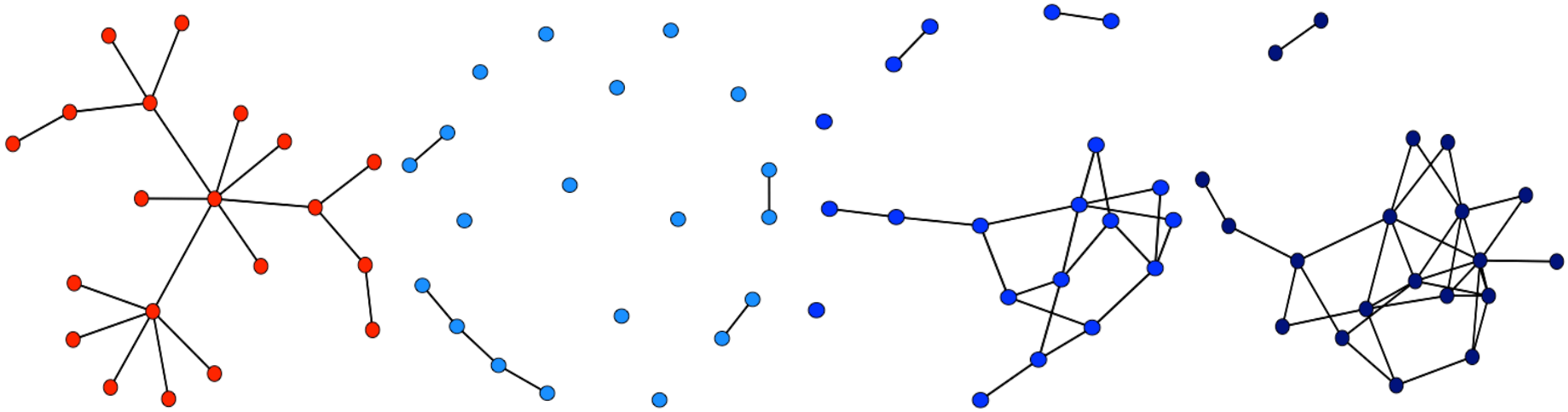
Centrality scores and their application to archaeological data



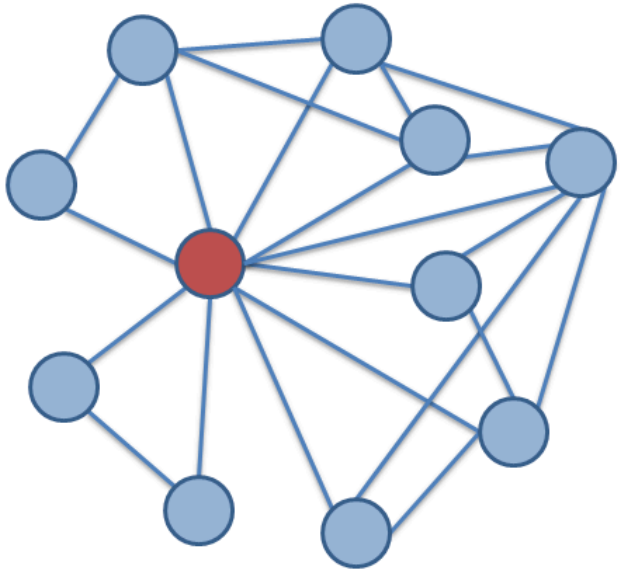
Erik Gjesfjeld, UCLA

CAA Meetings, Atlanta

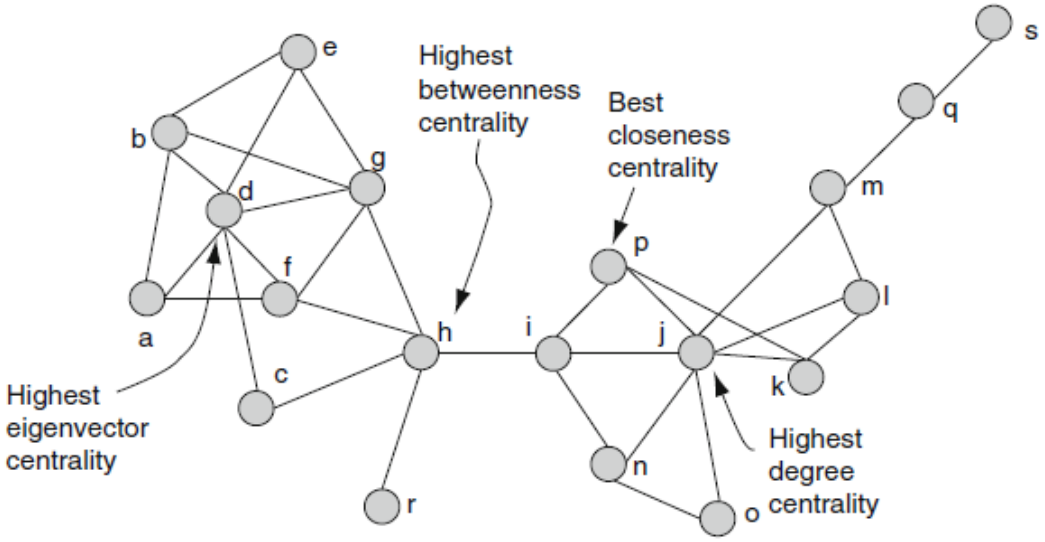
March 16, 2017

Use of Centrality

- Which node is most "important" based on their network position?



- How do we measure centrality?
  - Degree centrality
  - Betweenness centrality
  - Eigenvector centrality
  - And many more…



http://www.scottbot.net/HIAL/index.html@p=6526.html

https://ultrabpm.wordpress.com/2013/03/25/social-network-analysis-part-two/

Abuse of Centrality

- Methodological challenges (Brughmans, Collar and Coward 2016)

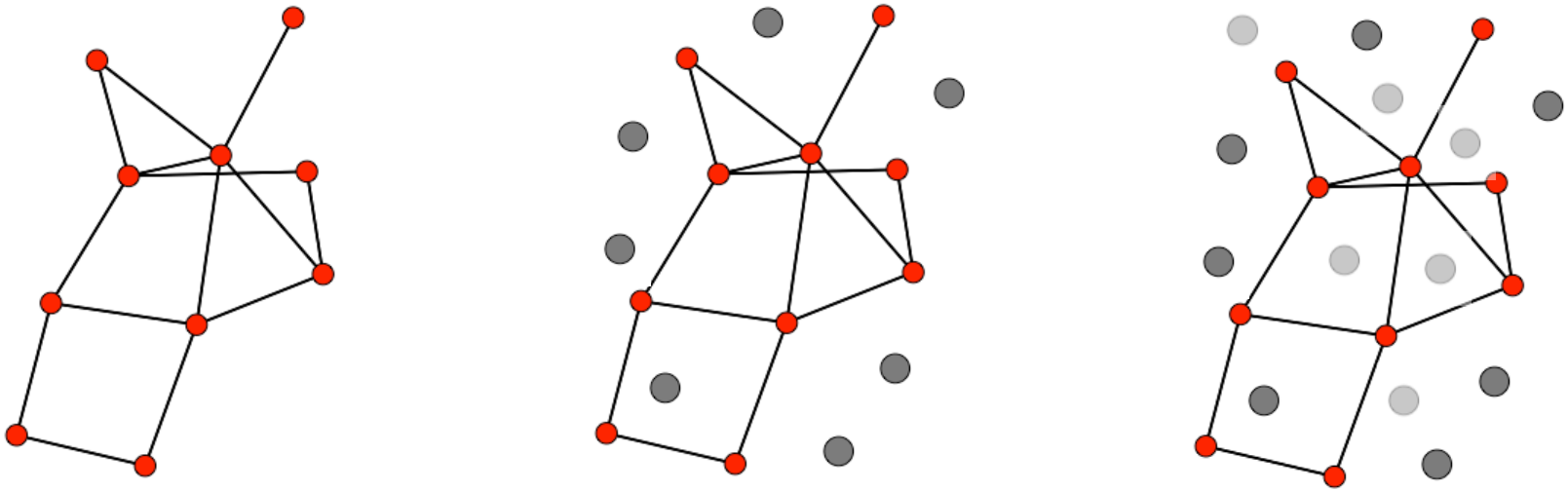
- "Lack of knowledge or confidence to critique or tailor the mathematics behind network techniques"

- "Mechanized application of network methods"

- *"Applying formal network perspectives to inappropriate data"*

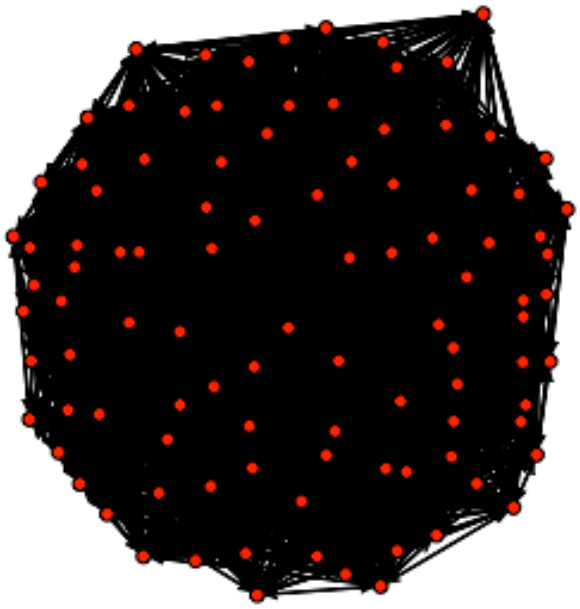
Missing Data

- Network models are always an incomplete or simplified version of a “real” network



- How sensitive are centrality scores from sampled network to changes in network size and density?

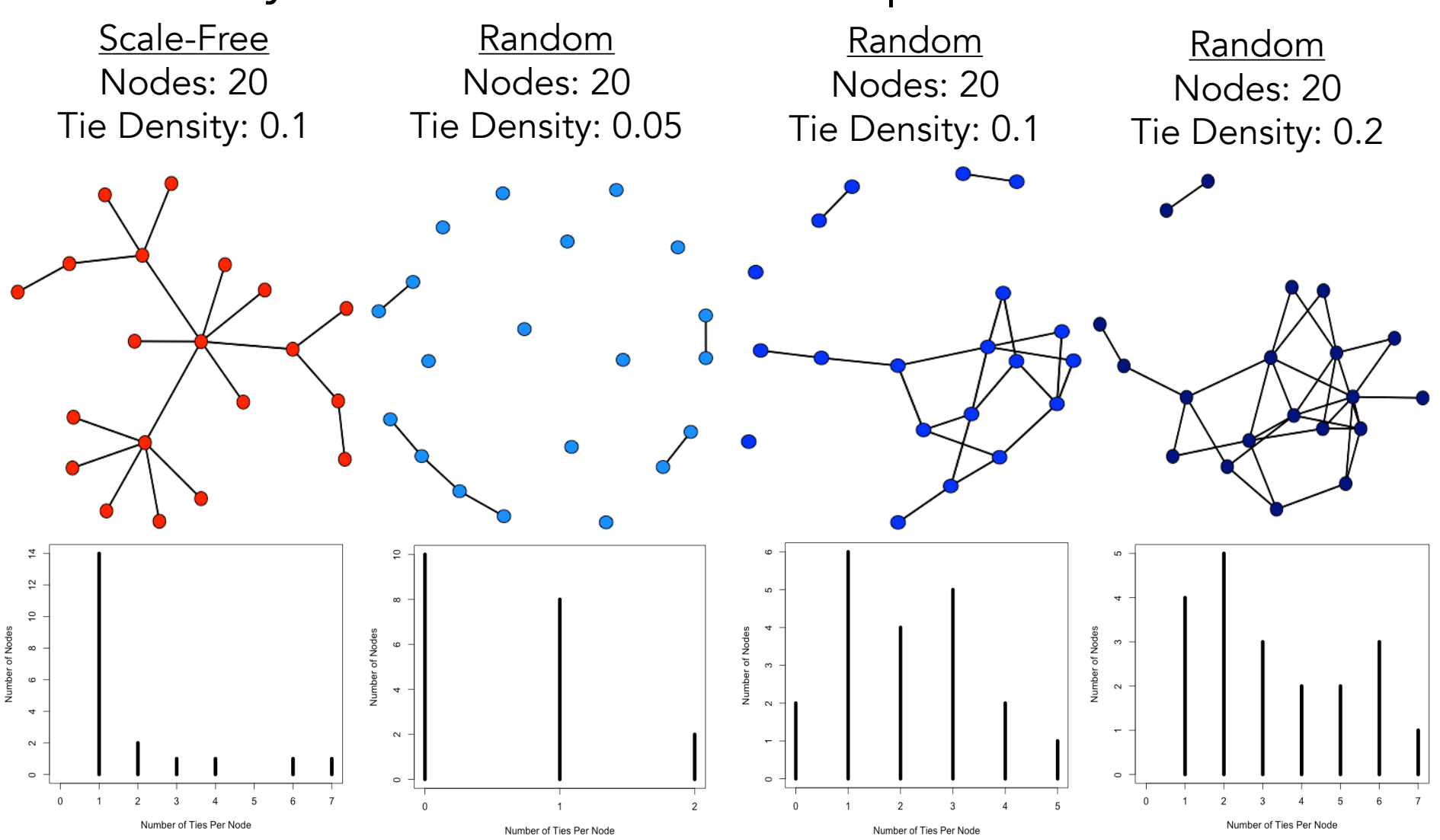
- Stability of centrality scores
  - Bolland 1988; Galaskiewicz 1991; Marsden 1993; Costenbader and Valente 2003; Borgatti et al. 2006; Kossinets 2006; Mills et al. 2013; Gjesfjeld 2015, Peeples et al. 2016
- Tendencies of previous studies
  - Use empirical data
  - Cyclical networks
  - Large and dense networks
    - 100 nodes, 0.5 density



- What about networks that tend to be small and sparse?
  - Networks from compositional data
  - Historic trade network
  - Most real-world networks

What does this research add?

- A systematic approach to evaluating stability of centrality scores across smaller, sparser networks



Bootstrap Approach

1. Construct series of networks with varying sizes for each network category
   - Scale-free, random, random-half, random-double
   - Network sizes of 10, 20, 40, 60, 80 and 100 nodes (24 networks)

2. Resample each of the networks 1000 times based on sampling fraction
   - 1000 networks sampled at 90%
   - 1000 networks sampled at 80%….
   - ….1000 networks sampled at 10%

3. Calculate degree and betweenness centrality score for each of the 1000 networks

4. Correlate centrality scores of each sampled network with original network

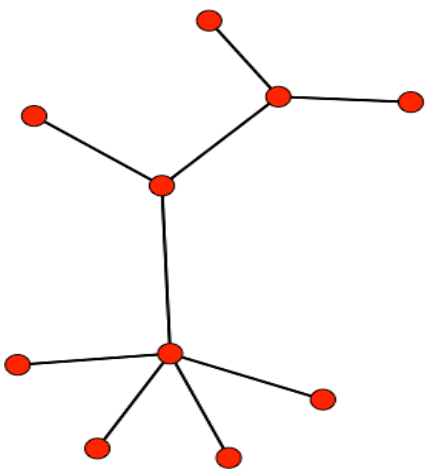
5. Take mean value of the 1000 correlations

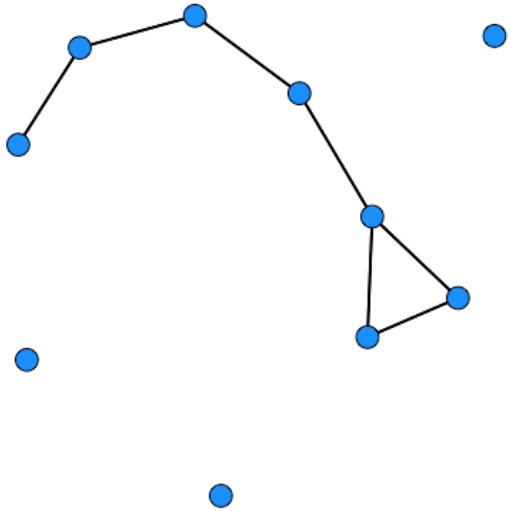
Scale-Free
Nodes: 10
Tie Density: 0.2

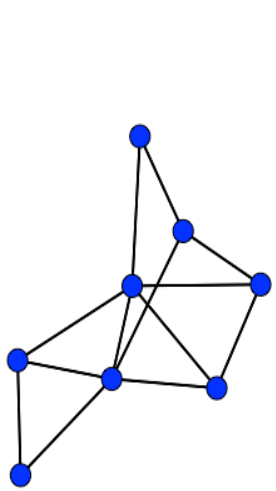
Random
Nodes: 10
Tie Density: 0.1

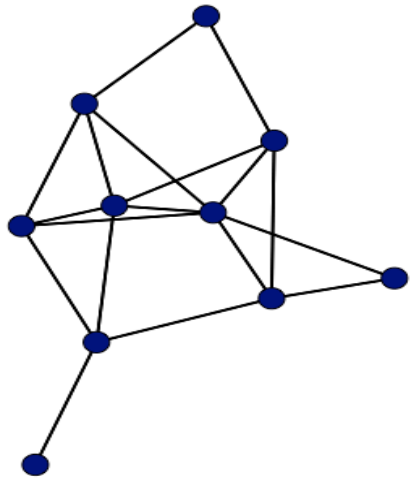
Random
Nodes: 10
Tie Density: 0.2

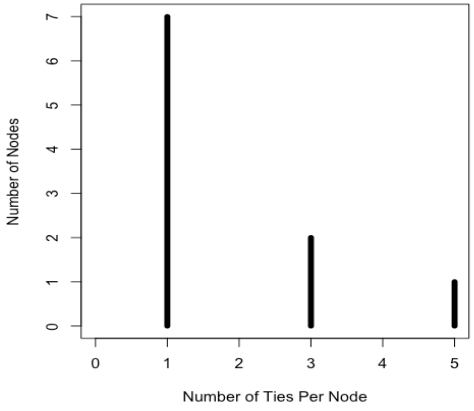
Random
Nodes: 10
Tie Density: 0.4

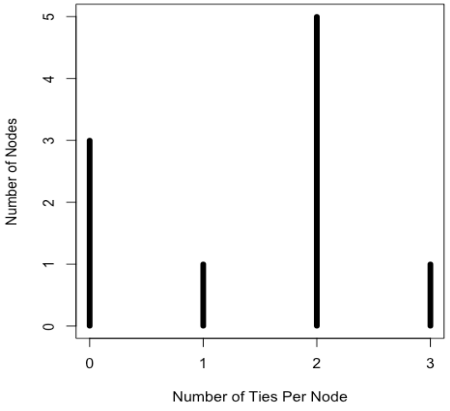


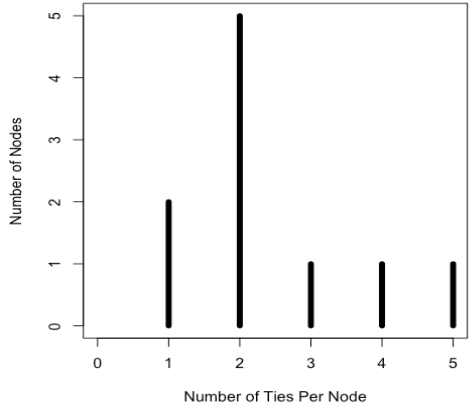


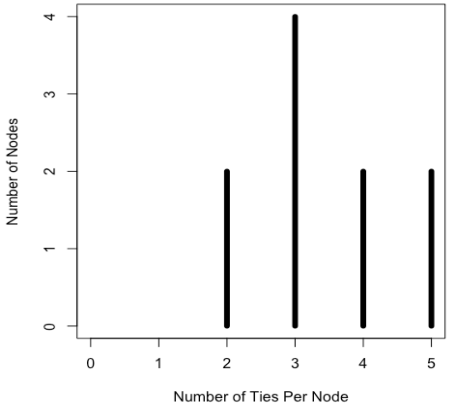


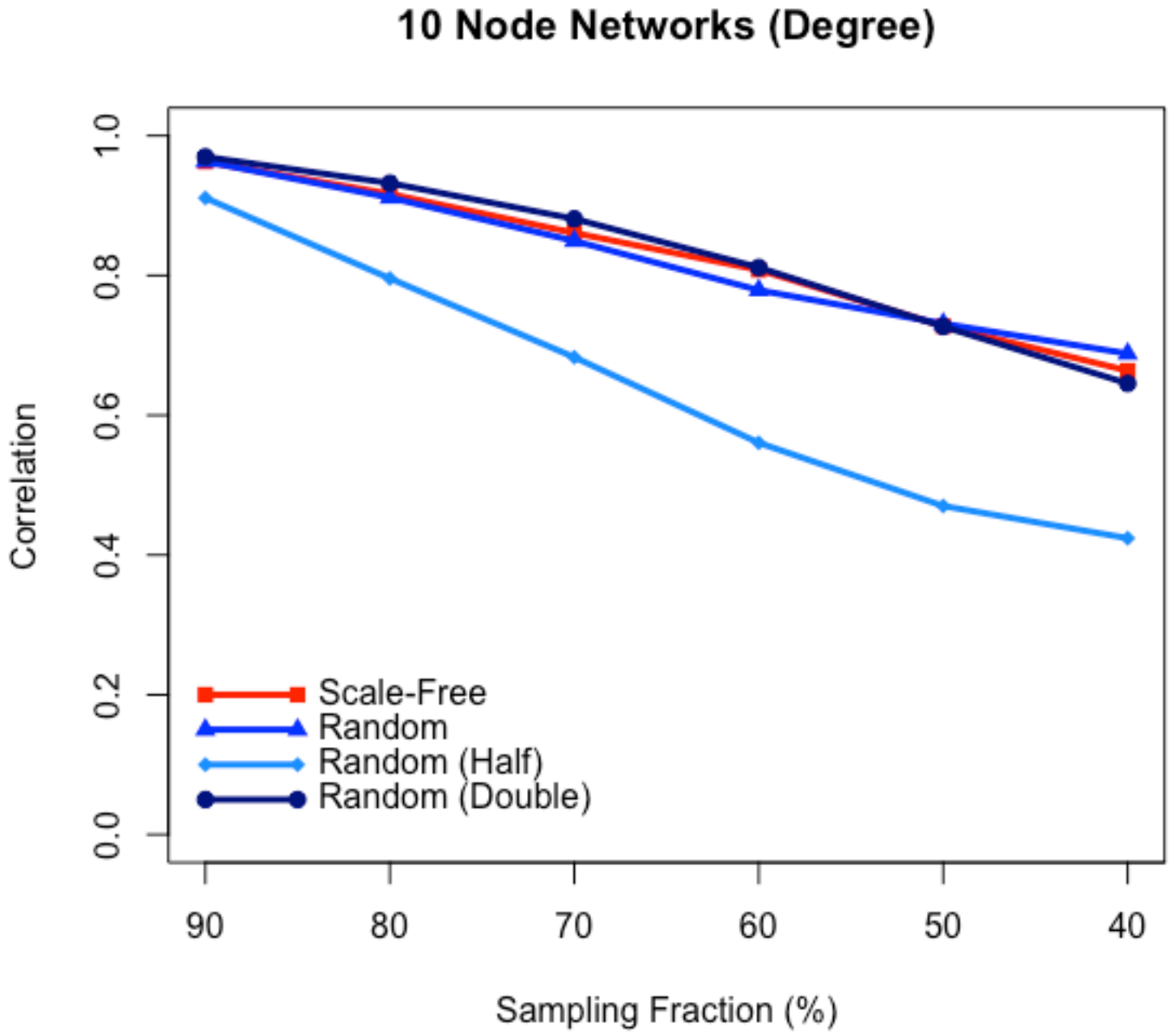




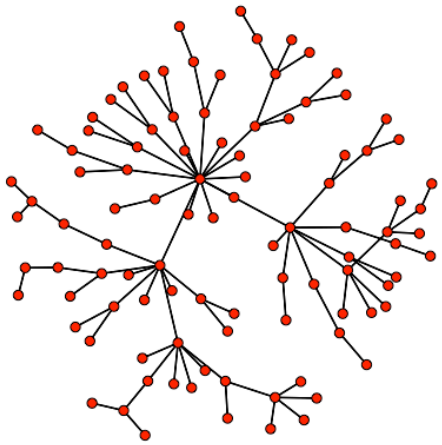


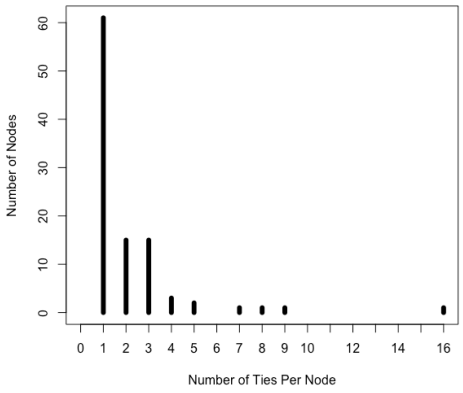




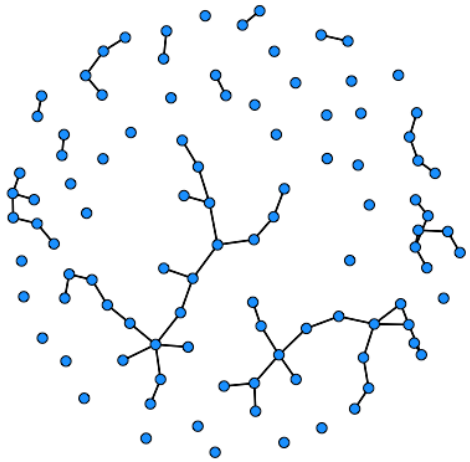

10 Node Networks (Degree)
Correlation
1.0
0.8
0.6
0.4
0.2
0.0
90
80
70
60
50
40
Sampling Fraction (%)
Scale-Free
Random
Random (Half)
Random (Double)

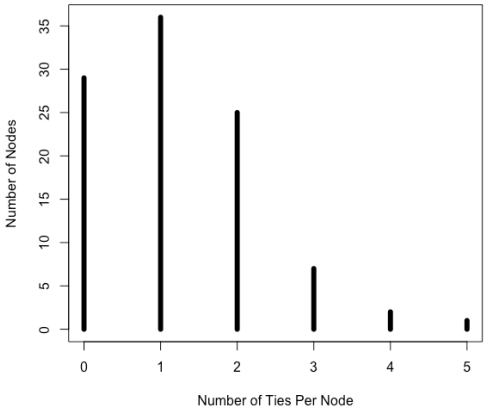
Scale-Free
Nodes: 100
Tie Density: 0.02



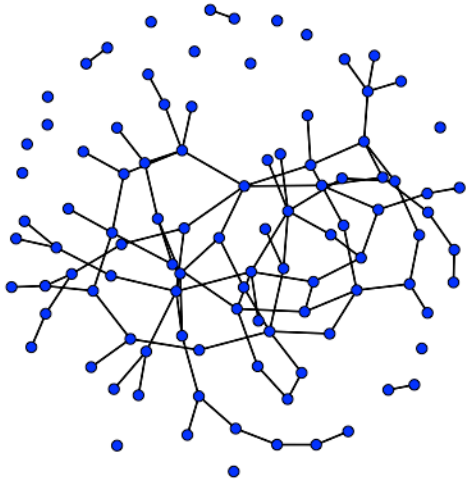


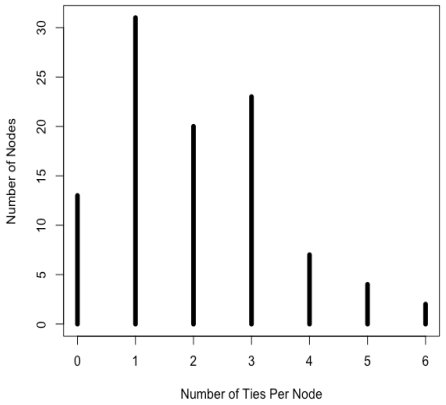
Random
Nodes: 100
Tie Density: 0.01



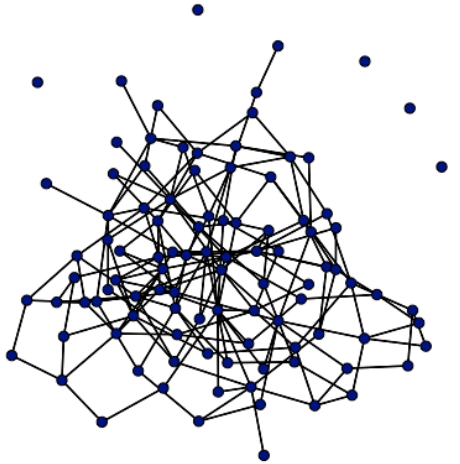


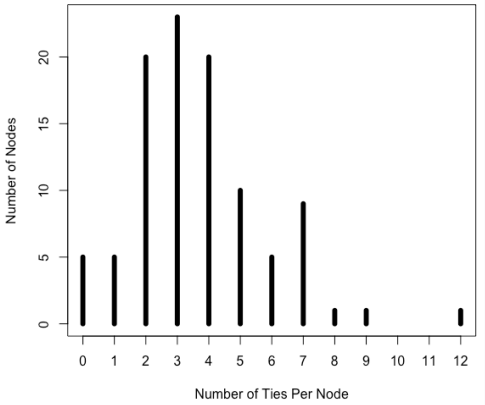
Random
Nodes: 100
Tie Density: 0.02

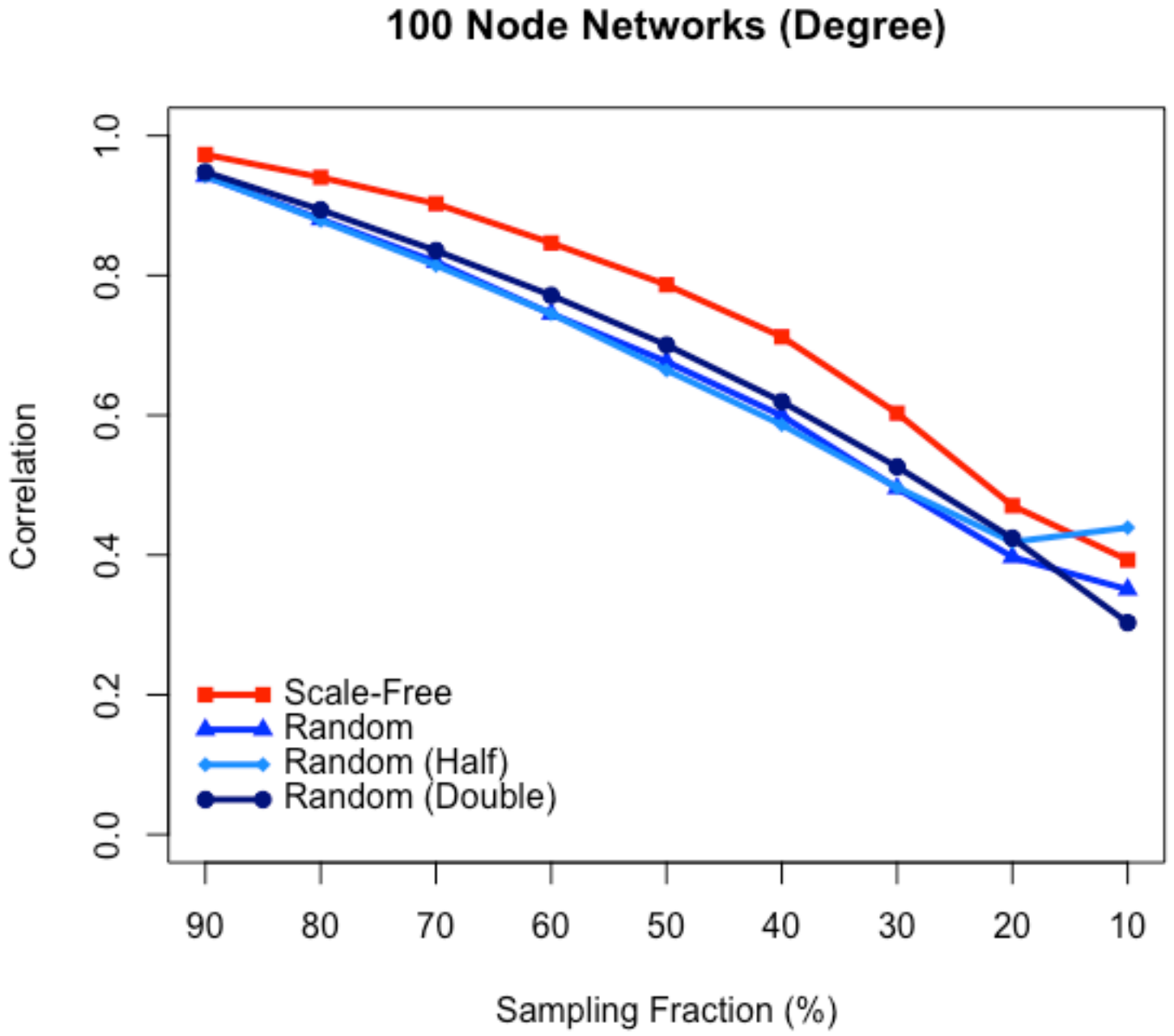




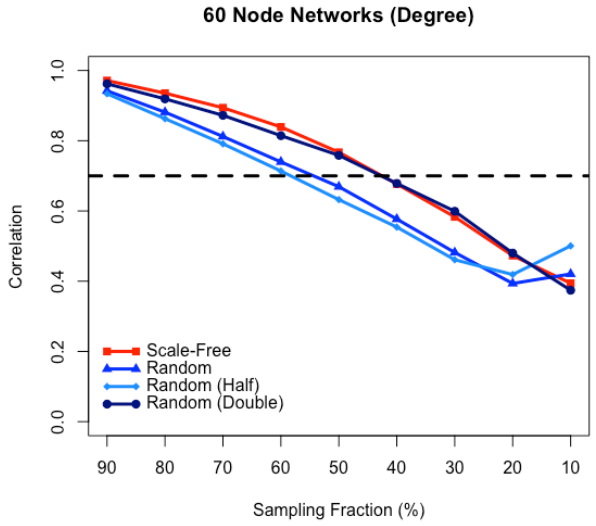
Random
Nodes: 100
Tie Density: 0.04

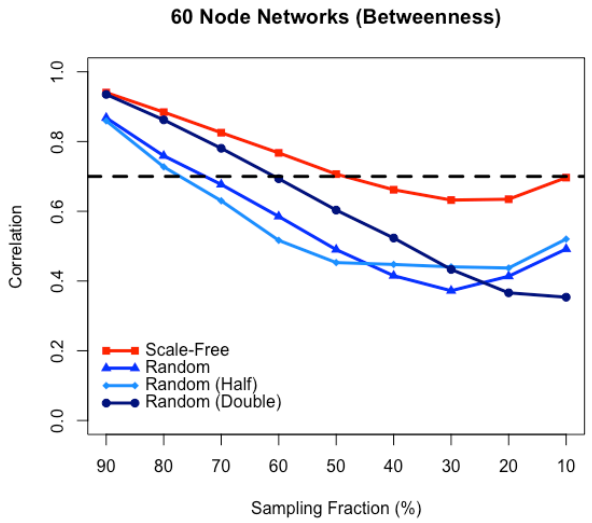





100 Node Networks (Degree)
Correlation
1.0
0.8
0.6
0.4
0.2
0.0
90
80
70
60
50
40
30
20
10
Sampling Fraction (%)
Scale-Free
Random
Random (Half)
Random (Double)

Degree (0.7)	10	20	40	60	80	100	Average
Scale-Fee	47%	42%	49%	42%	42%	40%	**43%**
Random	44%	57%	45%	52%	57%	54%	**52%**
Random - Half	72%	58%	55%	59%	55%	54%	**58%**
Random - Double	48%	62%	42%	42%	49%	50%	**48%**





Betweenness (0.7)	10	20	40	60	80	100	Average
Scale-Fee	35%	20%	52%	50%	50%	39%	**35%**
Random	58%	71%	52%	72%	80%	75%	**58%**
Random - Half	73%	80%	77%	78%	75%	78%	**65%**
Random - Double	50%	71%	63%	60%	64%	62%	**53%**

General Thoughts

- Centrality measures in small and sparse networks can be robust to sampling
  - Dependent on the distributions of ties

- For a 0.7 correlation (strong positive):
  - Scale-free networks require ~40% of "real" nodes
  - Random networks require ~50-60% of "real" nodes

- Network analysis is actually quite robust to missing data, but stability of centrality scores must be done on a case-by-case basis

Questions?

- R code and data for analyses in this presentation can be found at:
  https://github.com/erikgjes/CAA_2017

- Slides can be found at:
  http://erikgjesfjeld.net/publications