Standard errors for regression on relational data with exchangeable errors

Frank W. Marrs
Colorado State University
frank.marrs@colostate.edu

Collaborators

Bailey K. Fosdick
Colorado State University

Tyler H. McCormick
University of Washington

Network regression

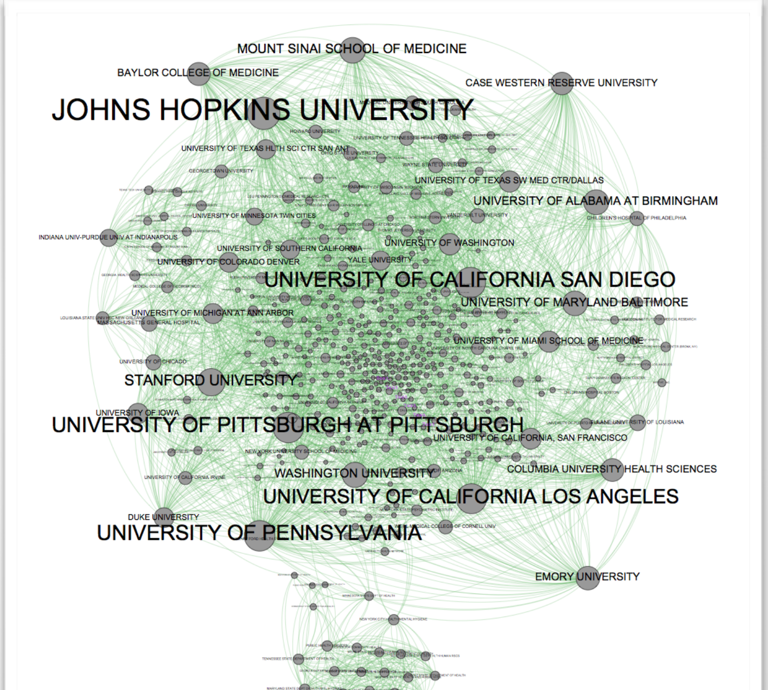
- response Y: **weighted, directed,** between actors *i* and *j*

- covariates X: individual or pairwise attributes

- Model linear relationship of covariates and response

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \xi_{ij}$$

$$Y = X\boldsymbol{\beta} + \boldsymbol{\xi} \;\in \mathbb{R}^{n(n-1)}$$



Brown 2015

Network regression

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \xi_{ij}$$

$$Y = X\boldsymbol{\beta} + \boldsymbol{\xi} \ \in \mathbb{R}^{n(n-1)}$$

- response Y: **weighted, directed,** between actors i and j
- covariates X: individual or pairwise attributes

$$\begin{array}{c|cccc|} & A & B & C & D \\ \hline A & \blacksquare & y_{AB} & y_{AC} & y_{AD} \\ B & y_{BA} & \blacksquare & y_{BC} & y_{BD} \\ C & y_{CA} & y_{CB} & \blacksquare & y_{CD} \\ D & y_{DA} & y_{DB} & y_{DC} & \blacksquare \\ \hline \end{array}$$

$$Y = \begin{bmatrix} y_{BA} \\ y_{CA} \\ \cdots \\ y_{CD} \end{bmatrix} \qquad X = \begin{bmatrix} \mathbf{x}_{BA}^T \\ \mathbf{x}_{CA}^T \\ \cdots \\ \mathbf{x}_{CD}^T \end{bmatrix}$$

Network regression

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \xi_{ij}$$

- **Goal**: inference about $\boldsymbol{\beta}$
  - point estimates $(\widehat{\boldsymbol{\beta}})$
  - confidence intervals $(\widehat{\boldsymbol{\beta}} \pm \widehat{\mathrm{se}}\{\widehat{\boldsymbol{\beta}}\})$
- ξ_{ij} highly structured error
  - i.e. ξ_{ij} and ξ_{ik} share a node, expect correlation

Linear Regression

- Recall Ordinary Least Squares

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\beta} ||Y - X\boldsymbol{\beta}||_2^2 = (X^T X)^{-1} X^T Y$$

$$Var(\widehat{\boldsymbol{\beta}}|X) = (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}$$

$$\Sigma = Var(\boldsymbol{\xi})$$

- X is *(n*^{2} *- n)* *x* *p* matrix of covariates
- Y and $\boldsymbol{\xi}$ are *(n*^{2} *- n)* vectors of relations and errors
- For inference on $\widehat{\boldsymbol{\beta}}$, need an estimate of Σ

Dyadic Clustering

- Assumes that non-overlapping pairs independent

$$
\xi = \begin{array}{c|cccc}
 & A & B & C & D \\
\hline
A & & \xi_{AB} & \xi_{AC} & \xi_{AD} \\
B & \boxed{\xi_{BA}} & & \xi_{BC} & \xi_{BD} \\
C & \xi_{CA} & \xi_{CB} & & \boxed{\xi_{CD}} \\
D & \xi_{DA} & \xi_{DB} & \xi_{DC} &
\end{array}
$$

$$
\widehat{Cov}(\xi_{BA}, \xi_{CD}) = 0
$$

Dyadic Clustering

- Model nonzero entries in Σ with residual products

$$\xi = \begin{array}{c|cccc} & A & B & C & D \\ \hline A & & \xi_{AB} & \xi_{AC} & \xi_{AD} \\ B & \xi_{BA} & & \xi_{BC} & \xi_{BD} \\ C & \xi_{CA} & \xi_{CB} & & \xi_{CD} \\ D & \xi_{DA} & \xi_{DB} & \xi_{DC} & \end{array}$$

$$\widehat{Cov}(\xi_{BA}, \xi_{AC}) = e_{BA} e_{AC}$$

$$e_{AB} := y_{AB} - \mathbf{x}_{ij}^T \widehat{\boldsymbol{\beta}}$$

Dyadic Clustering

$$\widehat{\Sigma}_{DC} =$$

$$n(n-1) \times n(n-1)$$

	BA	CA	DA	AB	CB	DB	AC	BC	DC	AD	BD	CD
BA	$e_{BA}e_{BA}$	$e_{CA}e_{BA}$	$e_{DA}e_{BA}$	$e_{AB}e_{BA}$	$e_{CB}e_{BA}$	$e_{DB}e_{BA}$	$e_{AC}e_{BA}$	$e_{BC}e_{BA}$		$e_{AD}e_{BA}$	$e_{BD}e_{BA}$	
CA	$e_{BA}e_{CA}$	$e_{CA}e_{CA}$	$e_{DA}e_{CA}$	$e_{AB}e_{CA}$	$e_{CB}e_{CA}$		$e_{AC}e_{CA}$	$e_{BC}e_{CA}$	$e_{DC}e_{CA}$	$e_{AD}e_{CA}$		$e_{CD}e_{CD}$
DA	$e_{BA}e_{DA}$	$e_{CA}e_{DA}$	$e_{DA}e_{DA}$	$e_{AB}e_{DA}$		$e_{DB}e_{DA}$	$e_{AC}e_{DA}$		$e_{DC}e_{DA}$	$e_{AD}e_{DA}$	$e_{BD}e_{DA}$	$e_{CD}e_{DA}$
AB	$e_{BA}e_{AB}$	$e_{CA}e_{AB}$	$e_{DA}e_{AB}$	$e_{AB}e_{AB}$	$e_{CB}e_{AB}$	$e_{DB}e_{AB}$	$e_{AC}e_{AB}$	$e_{BC}e_{AB}$		$e_{AD}e_{AB}$	$e_{BD}e_{AB}$	
CB	$e_{BA}e_{CB}$	$e_{CA}e_{CB}$		$e_{AB}e_{CB}$	$e_{CB}e_{CB}$	$e_{DB}e_{CB}$	$e_{AC}e_{CB}$	$e_{BC}e_{CB}$	$e_{DC}e_{CB}$		$e_{BD}e_{CB}$	$e_{CD}e_{CB}$
DB	$e_{BA}e_{DB}$		$e_{DA}e_{DB}$	$e_{AB}e_{DB}$	$e_{CB}e_{DB}$	$e_{DB}e_{DB}$		$e_{BC}e_{DB}$	$e_{DC}e_{DB}$	$e_{AD}e_{DB}$	$e_{BD}e_{DB}$	$e_{CD}e_{DB}$
AC	$e_{BA}e_{AC}$	$e_{CA}e_{AC}$	$e_{DA}e_{AC}$	$e_{AB}e_{AC}$	$e_{CB}e_{AC}$		$e_{AC}e_{AC}$	$e_{BC}e_{AC}$	$e_{DC}e_{AC}$	$e_{AD}e_{AC}$		$e_{CD}e_{AC}$
BC	$e_{BA}e_{BC}$	$e_{CA}e_{BC}$		$e_{AB}e_{BC}$	$e_{CB}e_{BC}$	$e_{DB}e_{BC}$	$e_{AC}e_{BC}$	$e_{BC}e_{BC}$	$e_{DC}e_{BC}$		$e_{BD}e_{BC}$	$e_{CD}e_{BC}$
DC		$e_{CA}e_{DC}$	$e_{DA}e_{DC}$		$e_{CB}e_{DC}$	$e_{DB}e_{DC}$	$e_{AC}e_{DC}$	$e_{BC}e_{DC}$	$e_{DC}e_{DC}$	$e_{AD}e_{DC}$	$e_{BD}e_{DC}$	$e_{CD}e_{DC}$
AD	$e_{BA}e_{AD}$	$e_{CA}e_{AD}$	$e_{DA}e_{AD}$	$e_{AB}e_{AD}$		$e_{DB}e_{AD}$	$e_{AC}e_{AD}$		$e_{DC}e_{AD}$	$e_{AD}e_{AD}$	$e_{BD}e_{AD}$	$e_{CD}e_{AD}$
BD	$e_{BA}e_{BD}$		$e_{DA}e_{BD}$	$e_{AB}e_{BD}$	$e_{CB}e_{BD}$	$e_{DB}e_{BD}$		$e_{BC}e_{BD}$	$e_{DC}e_{BD}$	$e_{AD}e_{BD}$	$e_{BD}e_{BD}$	$e_{CD}e_{BD}$
CD		$e_{CA}e_{CD}$	$e_{DA}e_{CD}$		$e_{CB}e_{CD}$	$e_{DB}e_{CD}$	$e_{AC}e_{CD}$	$e_{BC}e_{CD}$	$e_{DC}e_{CD}$	$e_{AD}e_{CD}$	$e_{BD}e_{CD}$	$e_{CD}e_{CD}$

Dyadic Clustering

- **Issues:**
  - More estimates than data points, $O(n^3) > O(n^2)$
  - No sharing of information
  - Singular with probability 1
- Can we add a reasonable assumption to improve the estimate?

Exchangeability

- Many network models are exchangeable: e.g. latent space, stochastic block, etc.
- **Intuition**: Ordering of rows/columns uninformative
- $\boldsymbol{\xi}$ *jointly exchangeable* if, for any permutation $\pi(.)$,

$$\mathbb{P}\left(\{\xi_{ij} : i \neq j, 1 \leq i, j \leq n\}\right) = \mathbb{P}\left(\{\xi_{\pi(i)\pi(j)} : i \neq j, 1 \leq i, j \leq n\}\right)$$

(akin to homogenous variance assumption)

Exchangeability

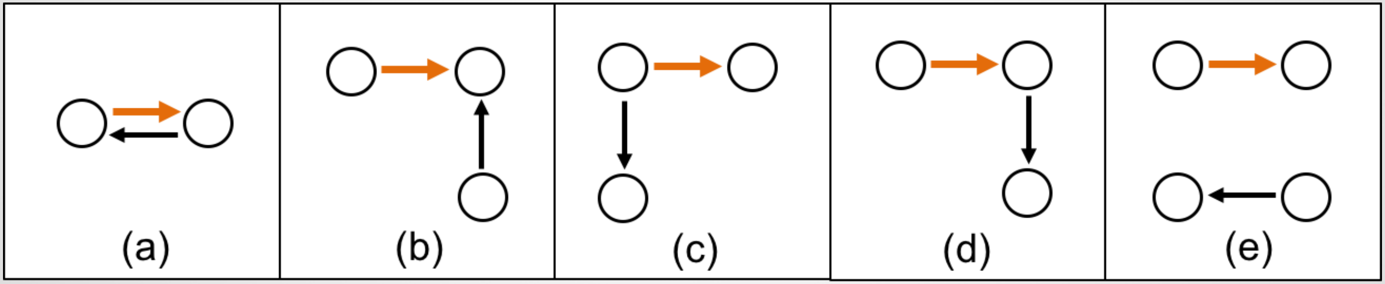
$\pi(\{A, B, C, D\})$ = Swap *B* and *D*

$$
\begin{array}{c|cccc}
 & A & B & C & D \\
\hline
A & & \xi_{AB} & \xi_{AC} & \xi_{AD} \\
B & \xi_{BA} & & \xi_{BC} & \xi_{BD} \\
C & \xi_{CA} & \xi_{CB} & & \xi_{CD} \\
D & \xi_{DA} & \xi_{DB} & \xi_{DC} &
\end{array}
\quad \overset{\mathbb{P}}{=} \quad
\begin{array}{c|cccc}
 & A & D & C & B \\
\hline
A & & \xi_{AD} & \xi_{AC} & \xi_{AB} \\
D & \xi_{DA} & & \xi_{DC} & \xi_{DB} \\
C & \xi_{CA} & \xi_{CD} & & \xi_{CB} \\
B & \xi_{BA} & \xi_{BD} & \xi_{BC} &
\end{array}
$$

Original $\boldsymbol{\xi}$ Permuted $\boldsymbol{\xi}$

Exchangeability

- **Major contribution**: Covariance matrix of jointly exchangeable vector $\boldsymbol{\xi}$ has 5 unique covariances and 1 variance, regardless of n
- We explicitly define this matrix for the first time



Exchangeability

	A	B	C	D
A		ξ_{AB}	ξ_{AC}	ξ_{AD}
B	ξ_{BA}		ξ_{BC}	ξ_{BD}
C	ξ_{CA}	ξ_{CB}		ξ_{CD}
D	ξ_{DA}	ξ_{DB}	ξ_{DC}	

	ξ_{BA}	ξ_{CA}	ξ_{DA}	ξ_{AB}	ξ_{CB}	ξ_{DB}	ξ_{AC}	ξ_{BC}	ξ_{DC}	ξ_{AD}	ξ_{BD}	ξ_{CD}
ξ_{BA}	σ^2	ϕ_b	ϕ_b	ϕ_a	ϕ_d	ϕ_d	ϕ_d	ϕ_c		ϕ_d	ϕ_c	
ξ_{CA}	ϕ_b	σ^2	ϕ_b	ϕ_d	ϕ_c		ϕ_a	ϕ_d	ϕ_d	ϕ_d		ϕ_c
ξ_{DA}	ϕ_b	ϕ_b	σ^2	ϕ_d		ϕ_c	ϕ_d		ϕ_c	ϕ_a	ϕ_d	ϕ_d
ξ_{AB}	ϕ_a	ϕ_d	ϕ_d	σ^2	ϕ_b	ϕ_b	ϕ_c	ϕ_d		ϕ_c	ϕ_d	
ξ_{CB}	ϕ_d	ϕ_c		ϕ_b	σ^2	ϕ_b	ϕ_d	ϕ_a	ϕ_d		ϕ_d	ϕ_c
ξ_{DB}	ϕ_d		ϕ_c	ϕ_b	ϕ_b	σ^2		ϕ_d	ϕ_c	ϕ_d	ϕ_a	ϕ_d
ξ_{AC}	ϕ_d	ϕ_a	ϕ_d	ϕ_c	ϕ_d		σ^2	ϕ_b	ϕ_b	ϕ_c		ϕ_d
ξ_{BC}	ϕ_c	ϕ_d		ϕ_d	ϕ_a	ϕ_d	ϕ_b	σ^2	ϕ_b		ϕ_c	ϕ_d
ξ_{DC}		ϕ_d	ϕ_c		ϕ_d	ϕ_c	ϕ_b	ϕ_b	σ^2	ϕ_d	ϕ_d	ϕ_a
ξ_{AD}	ϕ_d	ϕ_d	ϕ_a	ϕ_c		ϕ_d	ϕ_c		ϕ_d	σ^2	ϕ_b	ϕ_b
ξ_{BD}	ϕ_c		ϕ_d	ϕ_d	ϕ_d	ϕ_a		ϕ_c	ϕ_d	ϕ_b	σ^2	ϕ_b
ξ_{CD}		ϕ_c	ϕ_d		ϕ_c	ϕ_d	ϕ_d	ϕ_d	ϕ_a	ϕ_b	ϕ_b	σ^2

Exchangeable estimator

- Maintain independence assumption from DC

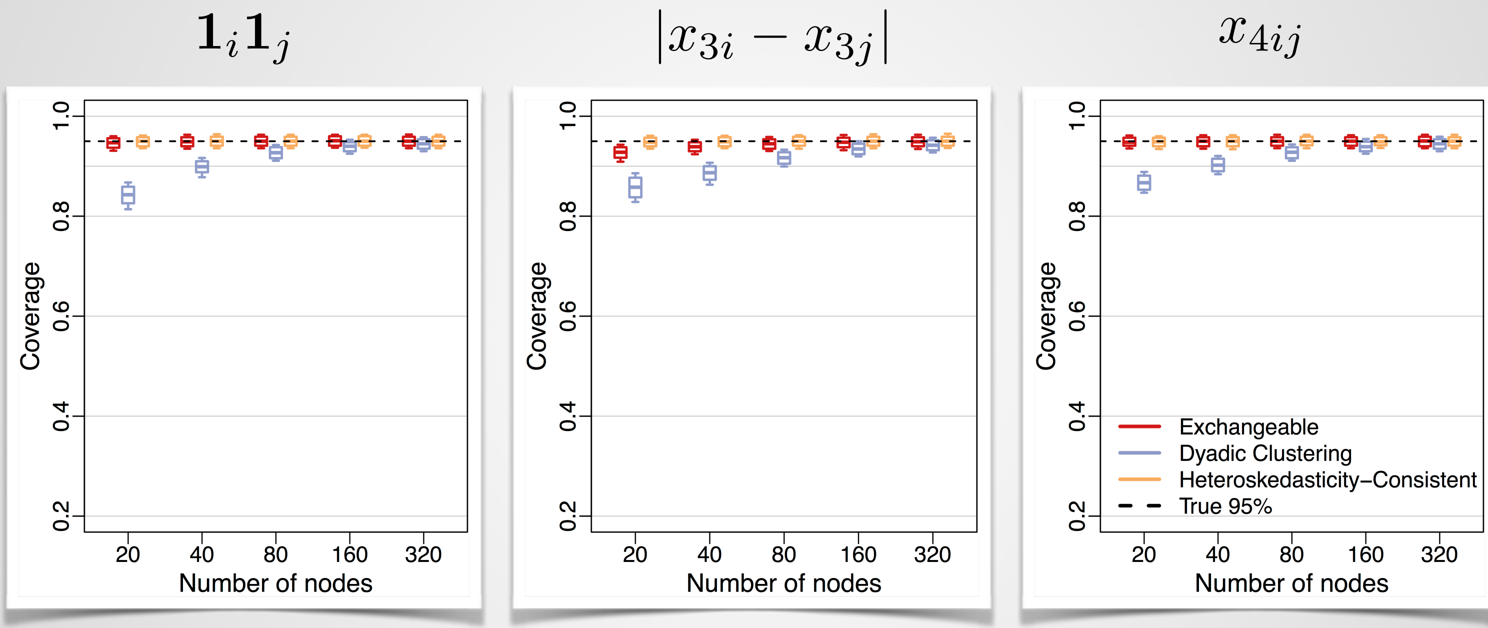
$$Cov(\xi_{ij}, \xi_{kl}) = 0 \text{ when } \{i,j\} \cap \{k,l\} = \varnothing$$

- Pool across all relations to estimate 5 nonzero terms in $\widehat{\Sigma}_E$
- Estimate $\widehat{\sigma}^2$, $\widehat{\phi}_i$ with mean of products of OLS residuals
- Projection of $\widehat{\Sigma}_{DC}$ onto subspace of exchangeable covariance matrices

Exchangeable estimator

- Adds assumption of joint exchangeability to DC estimator
- Shares information: should see reduced variability
- Should see improved performance when assumption is reasonable
    - Covariates explain all variability except for exchangeable structure
    - Heterogeneities small relative to variability across 5 parameters

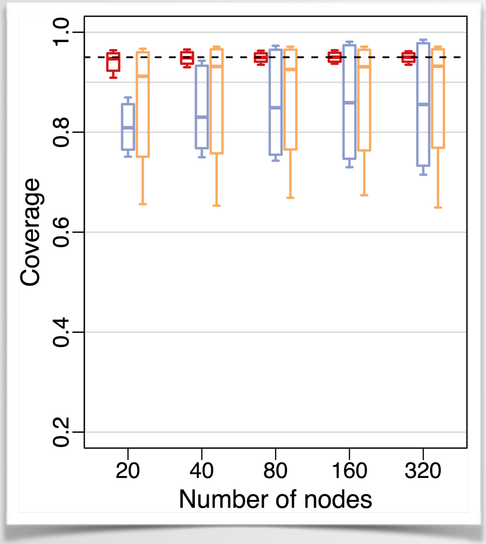
IID Errors

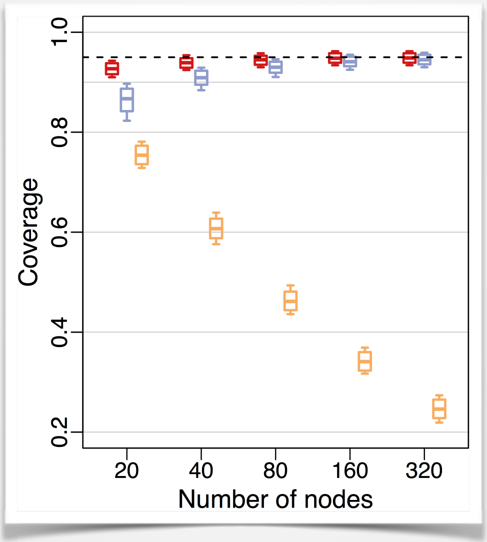


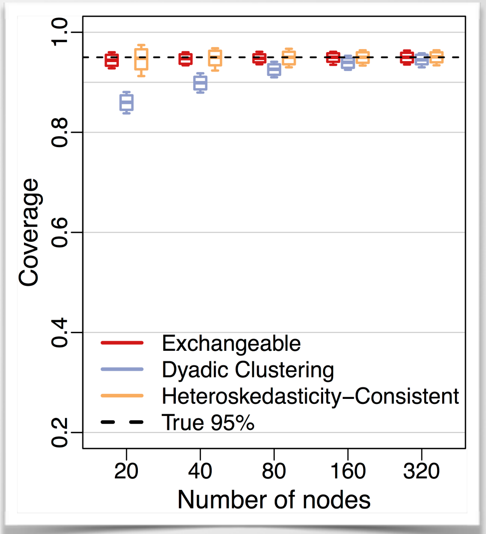
Probability true coefficient pertaining to each covariate is in 95% confidence interval

Exchangeable Errors

$\mathbf{1}_i\mathbf{1}_j$ $|x_{3i} - x_{3j}|$ x_{4ij}

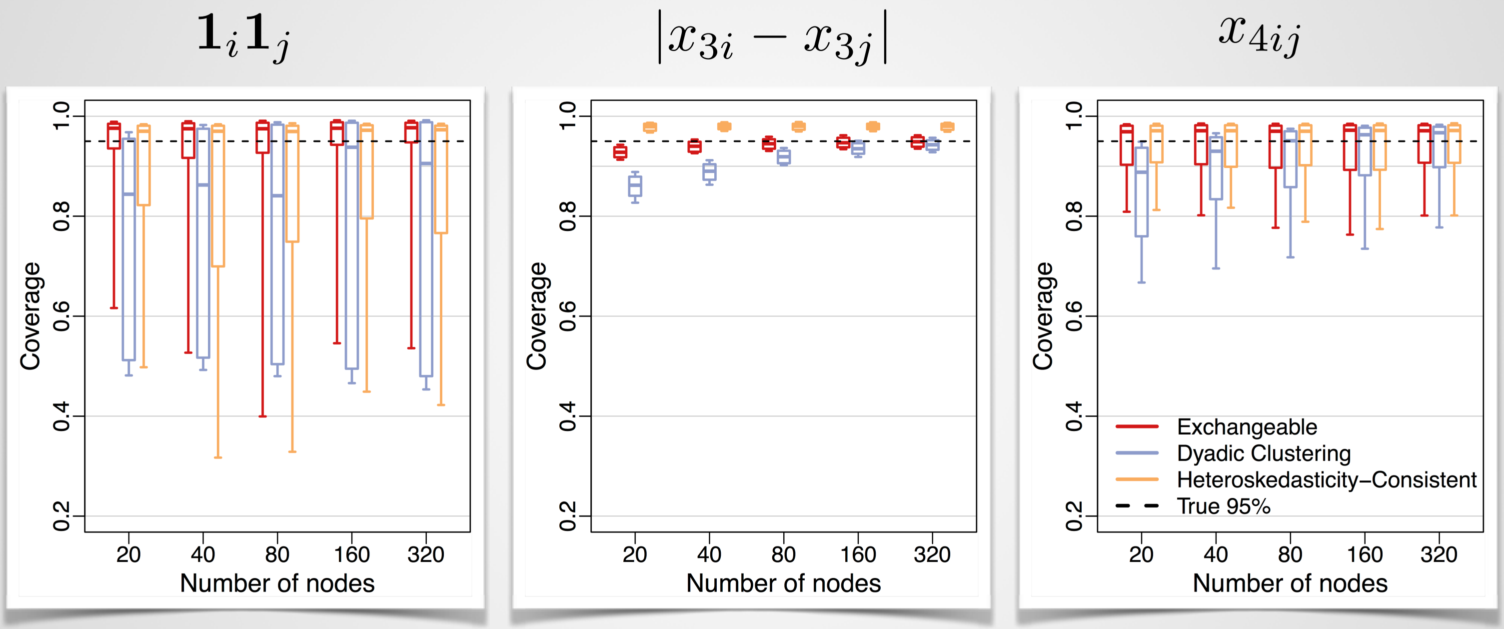






Probability true coefficient pertaining to each covariate is in 95% confidence interval

Nonexchangeable Errors



Probability true coefficient pertaining to each covariate is in 95% confidence interval

Summary

- Dyadic clustering approach may be noisy
- Many common network models are jointly exchangeable
- Exchangeable error covariance matrix has 6 unique terms
  - One of which we assume is zero
- Estimates of $\mathrm{se}\{\widehat{\beta}\}$ based on exchangeable error structure perform well
  - exchangeable structure
  - robust to non-exchangeable structure

Thank you!

Frank Marrs

Colorado State University

frank.marrs@colostate.edu

http://www.stat.colostate.edu/~marrs

Marrs, F.W., McCormick, T.H., and Fosdick, B.K. (2017) "Standard errors for regression on relational data with exchangeable errors", arXiv:1701.05530. [Preprint]