Taking Dyads Seriously

Shahryar Minhas, Cassy Dorff, Max Gallop, Margaret Foster, Juan Tellez, Howard Liu, Michael Ward

August 30, 2018

Motivation

Much of international relations data consists of

- a set of units or nodes
- a set of measurements, y_{ij}, specific to pairs of nodes (i, j)

Sender	Receiver	Event
i	j	y_{ij}
$\vdots$	k	y_{ik}
	l	y_{il}
j	i	y_{ji}
$\vdots$	k	y_{jk}
	l	y_{jl}
k	i	y_{ki}
$\vdots$	j	y_{kj}
	l	y_{kl}
l	i	y_{li}
$\vdots$	j	y_{lj}
	k	y_{lk}

Relational data assumptions

GLM: $y_{ij} \sim \beta^T X_{ij} + e_{ij}$

Networks typically show evidence against independence of $e_{ij} : i \neq j$

Not accounting for dependence can lead to:

- biased effects estimation
- uncalibrated confidence intervals
- poor predictive performance
- inaccurate description of network phenomena

We've been hearing this concern for decades now:

Thompson & Walker (1982)	Beck et al. (1998)	Snijders (2011)
Frank & Strauss (1986)	Signorino (1999)	Erikson et al. (2014)
Kenny (1996)	Li & Loken (2002)	Aronow et al. (2015)

Outline

The approach that we will discuss here to deal with dependencies is the Additive and Multiplicative Effects (AME) model (Hoff 2015; Minhas, Hoff & Ward 2018)

- Nodal and dyadic dependencies in networks
  - Can model using the "A" in AME
- Third order dependencies
  - Can model using the "M" in AME
- Simulation
- Empirical Application

Dependencies in relational data

To start to think about dependencies that arise in relational data ...

Sender	Receiver	Event
i	j	y_{ij}
$\vdots$	k	y_{ik}
	l	y_{il}
j	i	y_{ji}
$\vdots$	k	y_{jk}
	l	y_{jl}
k	i	y_{ki}
$\vdots$	j	y_{kj}
	l	y_{kl}
l	i	y_{li}
$\vdots$	j	y_{lj}
	k	y_{lk}

$\longrightarrow$

	i	j	k	l
i	NA	y_{ij}	y_{ik}	y_{il}
j	y_{ji}	NA	y_{jk}	y_{jl}
k	y_{ki}	y_{kj}	NA	y_{kl}
l	y_{li}	y_{lj}	y_{lk}	NA

$\downarrow$

What network phenomena? Sender heterogeneity

Values across a row, say $\{y_{ij}, y_{ik}, y_{il}\}$, may be more similar to each other than other values in the adjacency matrix because each of these values has a common sender i

	i	j	k	l
i	NA	y_{ij}	y_{ik}	y_{il}
j	y_{ji}	NA	y_{jk}	y_{jl}
k	y_{ki}	y_{kj}	NA	y_{kl}
l	y_{li}	y_{lj}	y_{lk}	NA

What network phenomena? Receiver heterogeneity

Values across a column, say $\{y_{ji}, y_{ki}, y_{li}\}$, may be more similar to each other than other values in the adjacency matrix because each of these values has a common receiver i

	i	j	k	l
i	NA	y_{ij}	y_{ik}	y_{il}
j	y_{ji}	NA	y_{jk}	y_{jl}
k	y_{ki}	y_{kj}	NA	y_{kl}
l	y_{li}	y_{lj}	y_{lk}	NA

What network phenomena? Sender-Receiver Covariance

Actors who are more likely to send ties in a network may also be more likely to receive them

	i	j	k	l
i	NA	y_{ij}	y_{ik}	y_{il}
j	y_{ji}	NA	y_{jk}	y_{jl}
k	y_{ki}	y_{kj}	NA	y_{kl}
l	y_{li}	y_{lj}	y_{lk}	NA

What network phenomena? Reciprocity

Values of y_{ij} and y_{ji} may be statistically dependent

	i	j	k	l
i	NA	y_{ij}	y_{ik}	y_{il}
j	y_{ji}	NA	y_{jk}	y_{jl}
k	y_{ki}	y_{kj}	NA	y_{kl}
l	y_{li}	y_{lj}	y_{lk}	NA

Social Relations Model (The "A" in AME)

We use this model to form the additive effects portion of AME (Warner et al. 1979, Li & Loken 2002, Hoff 2005):

$$
\begin{aligned}
y_{ij} &= \color{red}{\mu + e_{ij}} \\
e_{ij} &= a_i + b_j + \epsilon_{ij} \\
\{(a_1, b_1), \ldots, (a_n, b_n)\} &\sim N(0, \Sigma_{ab}) \\
\{(\epsilon_{ij}, \epsilon_{ji}) : \; i \neq j\} &\sim N(0, \Sigma_\epsilon), \text{ where} \\
\Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} & \quad \Sigma_\epsilon = \sigma_\epsilon^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}
\end{aligned}
$$

- \- μ baseline measure of network activity
- \- e_{ij} residual variation that we will use the SRM to decompose

Social Relations Model (The "A" in AME)

$$
\begin{aligned}
y_{ij} &= \mu + e_{ij} \\
e_{ij} &= a_i + b_j + \epsilon_{ij} \\
\{(a_1, b_1), \ldots, (a_n, b_n)\} &\sim N(0, \Sigma_{ab}) \\
\{(\epsilon_{ij}, \epsilon_{ji}) : \; i \neq j\} &\sim N(0, \Sigma_{\epsilon}), \text{ where}
\end{aligned}
$$

$$
\Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \qquad \Sigma_{\epsilon} = \sigma_{\epsilon}^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}
$$

- \- row/sender effect (a_i) & column/receiver effect (b_j)
- \- Modeled jointly to account for correlation in how active an actor is in sending and receiving ties

Social Relations Model (The "A" in AME)

$$
\begin{aligned}
y_{ij} &= \mu + e_{ij} \\
e_{ij} &= a_i + b_j + \epsilon_{ij} \\
\{(a_1, b_1), \ldots, (a_n, b_n)\} &\sim N(0, \Sigma_{ab}) \\
\{(\epsilon_{ij}, \epsilon_{ji}) : \ i \neq j\} &\sim N(0, \Sigma_\epsilon), \ \text{where}
\end{aligned}
$$

$$
\Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \qquad \Sigma_\epsilon = \sigma_\epsilon^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}
$$

- \- σ_a^2 and σ_b^2 capture heterogeneity in the row and column means
- \- σ_{ab} describes the linear relationship between these two effects (i.e., whether actors who send [receive] a lot of ties also receive [send] a lot of ties)

Social Relations Model (The "A" in AME)

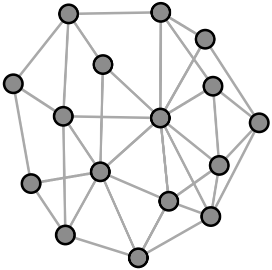
$$
\begin{aligned}
y_{ij} &= \mu + e_{ij} \\
e_{ij} &= a_i + b_j + \epsilon_{ij} \\
\{(a_1, b_1), \ldots, (a_n, b_n)\} &\sim N(0, \Sigma_{ab}) \\
\{(\epsilon_{ij}, \epsilon_{ji}) :\ i \neq j\} &\sim N(0, \Sigma_\epsilon), \text{ where}
\end{aligned}
$$

$$
\Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \qquad \Sigma_\epsilon = \sigma_\epsilon^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}
$$

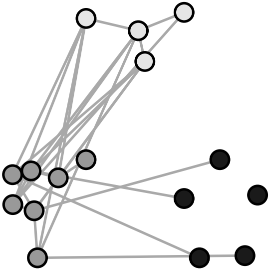
- - ϵ_{ij} captures the within dyad effect
- - Second-order dependencies are described by σ_ϵ^2
- - Reciprocity, aka within dyad correlation, represented by ρ

Third Order Dependencies

Homophily



Stochastic Equivalence



To account for these patterns we can build on what we have so far and find an expression for γ:

$$y_{ij} \approx \beta^T X_{ij} + a_i + b_j + \gamma(u_i, v_j)$$

Latent Factor Model: The "M" in AME

Each node i has an unknown latent factor:

$$\mathbf{u}_i, \mathbf{v}_j \in \mathbb{R}^k \;\; i, j \in \{1, \ldots, n\}$$

The probability of a tie from i to j depends on their latent factors (Hoff 2008, Hoff 2015, Minhas, Hoff & Ward 2018):

$$\begin{aligned} \gamma(\mathbf{u}_i, \mathbf{v}_j) &= \mathbf{u}_i^T D \mathbf{v}_j \\ &= \sum_{k \in K} d_k u_{ik} v_{jk} \\ & D \text{ is a } K \times K \text{ diagonal matrix} \end{aligned}$$

Can account for both stochastic equivalence and homophily

AME

The full AME framework enables us to not only account for the types of dependencies that often arise in dyadic data but also estimate them in a fully Bayesian framework:

$$
\begin{aligned}
y_{ij} \;=\; & f(\theta_{ij}), \text{ where} \\
\theta_{ij} \;=\; & \boldsymbol{\beta}_d^\top \mathbf{X}_{ij} + \boldsymbol{\beta}_s^\top \mathbf{X}_i + \boldsymbol{\beta}_r^\top \mathbf{X}_j && \text{(Exogenous parameters)} \\
& + a_i + b_j + \epsilon_{ij} && \text{(SRRM parameters)} \\
& + \mathbf{u}_i^\top \mathbf{D} \mathbf{v}_j && \text{(LFM parameters)}
\end{aligned} \tag{1}
$$

Simulation

Assume that the true data-generating process for a particular Y is given by:

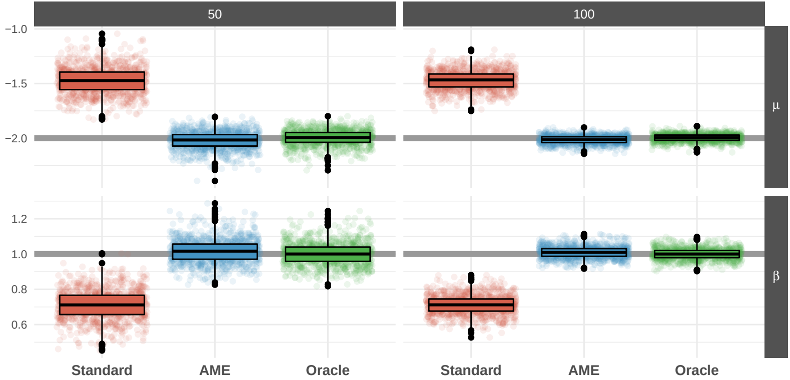
$$y_{i,j} \sim \mu + \beta x_{i,j} + \gamma w_{i,j} + \epsilon_{i,j} \tag{2}$$

We compare inference for μ and β—the latter parameter would be of primary concern for applied scholars—using three models:

- the "standard" international relations approach estimated through a typical generalized linear model;
- the AME approach outlined in the previous section with a unidimensional latent factor space ($K = 1$);
- and an "oracle" regression model that assumes we have measured all sources of dependencies and thus includes both $x_{i,j}$ and $w_{i,j}$.

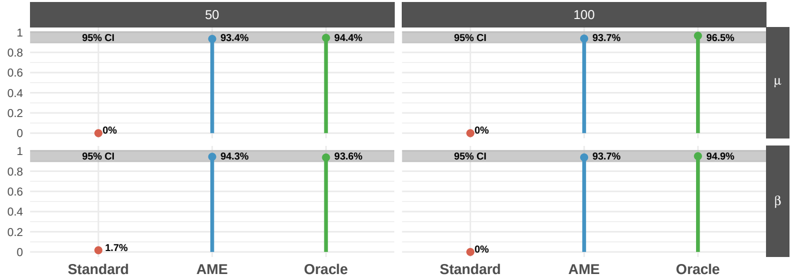
Bias Comparison

We first compare the performance of the models in terms of how well they estimate the true values of μ and β:



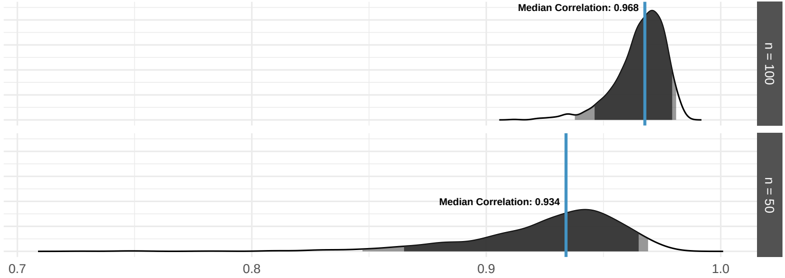
Coverage Comparison

Next, we estimate the 95% confidence interval for the three models in each of the simulations and estimate the proportion of times that the true value fell within those intervals:



What about the missing variable?

Unobserved dependencies should be captured through the multiplicative effects portion of the model, $\mathbf{U}^\top \mathbf{D} \mathbf{V}$:



Empirical Applications

Descriptive information about the replicated studies:

	Model	Date Range	N. Actors	Dyads Type
Reiter & Stam (2003)	Logit	1945–1995	193	Directed
Weeks (2012)	Logit	1946–1999	197	Directed
Gibler (2017)	Logit	1816–2008	193	Undirected

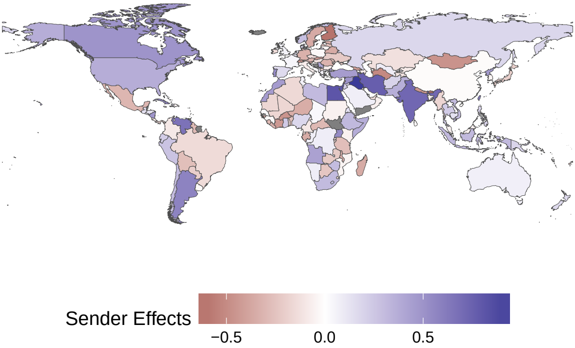
Predictive Comparison

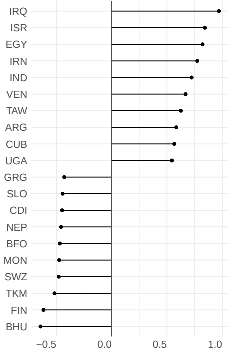
	Model	AUC (ROC)	AUC (PR)
Reiter & Stam (2003)	AME	0.96	0.15
	GLM	0.92	0.08
Weeks (2012)	AME	0.97	0.15
	GLM	0.64	0.00
Gibler (2017)	AME	0.91	0.08
	GLM	0.52	0.00

Results robust?

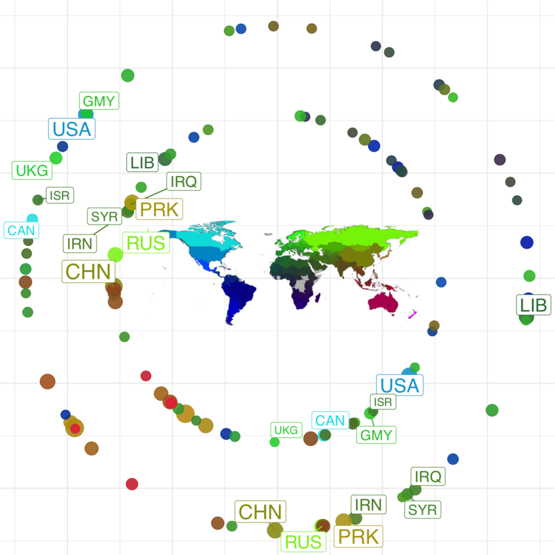
Study	Central Finding	Confirmed after accounting for dependencies?
Reiter & Stam (2003)	Personalist Regimes Attack Democracies, Not Vice Versa	Confirmed
Weeks (2012)	Bosses, Juntas, and Strongmen are more Aggressive, Machines are Not	Unconfirmed
Gibler (2017)	Power Parity at Time of Entry to International System Increases Conflict	Unconfirmed

What else did we learn? Reiter & Stam (2003)

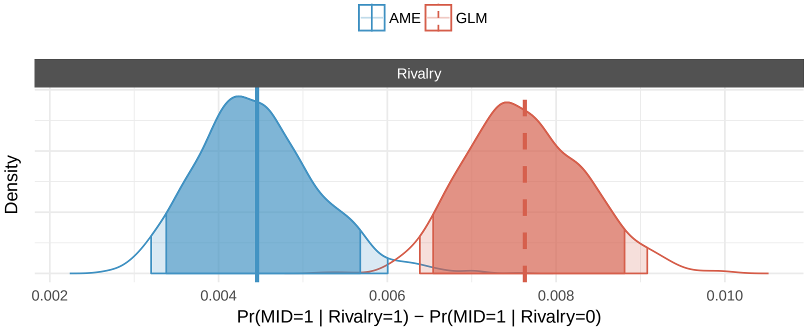




What else did we learn? Weeks (2012)



What else did we learn? Gibler (2017)

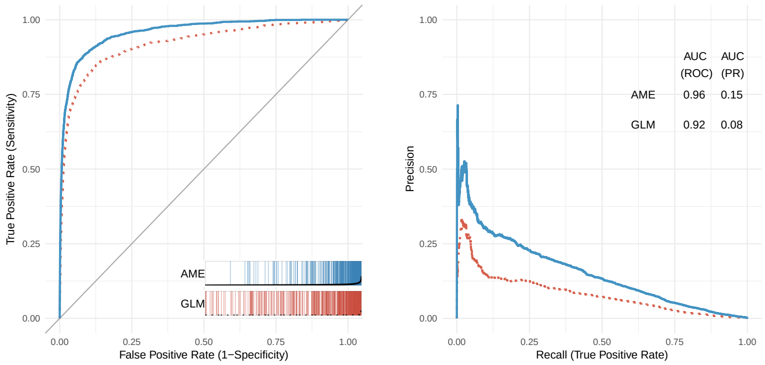


Next steps

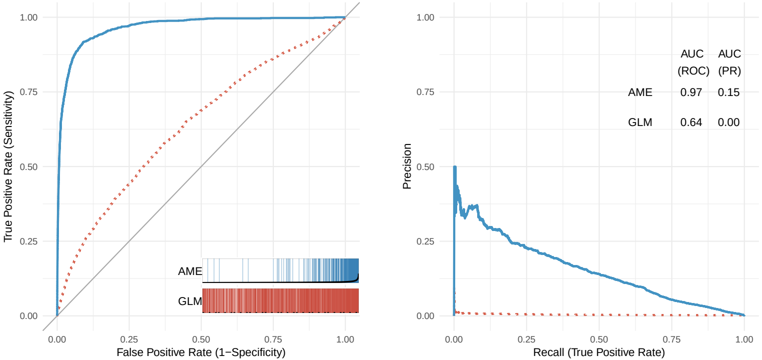
- Dealing with time varying dependence structure
- Parsing out formation, persistence and dissolution
- Many other extensions (bipartite, endogenous multilayer networks, ...)
- Computational issues ... Bayesian models take time to converge

Thanks.

Predictive Comparison: Reiter & Stam (2003)



Predictive Comparison: Weeks (2012)



Predictive Comparison: Gibler (2017)