Supplemental Material

Modeling Asymmetric Relationships from Symmetric Networks

Arturas Rozenas
NYU
arturas.rozenas@nyu.edu

Shahryar Minhas
MSU
minhassh@msu.edu

John Ahlquist
UCSD
jahlquist@ucsd.edu

Contents

1. DETAILS ON THE P-GBME

As detailed in the paper, the partial observability generalized bilinear mixed effects (P-GBME) framework treats the observed symmetric outcome, $y_{ij} = y_{ji}$, as resulting from a joint decision taken by a pair of actors. We formalize the joint decision making process using a bivariate probit model with a standard normal link function:

$$y_{ij} = y_{ji} = \begin{cases} 1 & \text{if } z_{ij} > 0 \text{ and } z_{ji} > 0, \\ 0 & \text{else}, \end{cases} \tag{1}$$

$$\begin{pmatrix} z_{ij} \\ z_{ji} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{ij} + a_i + b_j + \boldsymbol{u}_i'\boldsymbol{v}_j \\ \mu_{ji} + a_j + b_i + \boldsymbol{u}_j'\boldsymbol{v}_i \end{pmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \right), \tag{2}$$

$$(a_i, b_i)' \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{ab}\right), \tag{3}$$

$$\boldsymbol{u}_i \sim \mathcal{N}_K(\mathbf{0}, \sigma_u^2 \boldsymbol{I}), \tag{4}$$

$$\boldsymbol{v}_i \sim \mathcal{N}_K(\mathbf{0}, \sigma_v^2 \boldsymbol{I}). \tag{5}$$

a_i and b_j represent sender and receiver random effects that account for first order dependence patterns that often arise in relational data, while $\boldsymbol{u}_i'\boldsymbol{v}_j$ captures the likelihood of a pair of actors interacting with one another based on third order dependence patterns such as transitivity, balance, and clustering. For identification purposes, we fix $\sigma^2 = 1$ and $\rho = 0$. The former is a standard restriction in probit frameworks with a binary outcome. We undertake the latter restriction because **?** shows that in this framework it is difficult to recover reliable estimates for ρ as the parameter is highly sensitive to the initial value.

The sender and receiver random effects (a_i and b_j) are drawn from a multivariate normal distribution centered at zero with a covariance matrix, Σ_{ab}, parameterized as follows:

$$\Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \tag{6}$$

The nodal effects are modeled in this way to account for the fact that in many relational datasets we often find that actors who send a lot of ties are also more likely to receive a lot of ties. Heterogeneity in the the sender and receiver effects is captured by σ_a^2 and σ_b^2, respectively, and σ_{ab} describes the covariance between these two effects.

μ represents the systematic component of actors' utilities and is expressed as a linear function of sender $^{(s)}$, receiver $^{(r)}$, and dyadic $^{(d)}$ covariates:

$$\mu_{ij} = \boldsymbol{\beta}^{(s)}\boldsymbol{x}_i^{(s)} + \boldsymbol{\beta}^{(r)}\boldsymbol{x}_j^{(r)} + \boldsymbol{\beta}^{(d)}\boldsymbol{x}_{ij}^{(d)}, \tag{7}$$

$$\mu_{ji} = \boldsymbol{\beta}^{(s)}\boldsymbol{x}_j^{(s)} + \boldsymbol{\beta}^{(r)}\boldsymbol{x}_i^{(r)} + \boldsymbol{\beta}^{(d)}\boldsymbol{x}_{ji}^{(d)}. \tag{8}$$

This formulation allows us to incorporate exogenous actor and dyad level characteristics into how actors make decisions within the partial probit framework. Following **?**, to enable a more efficient estimation, we reparameterize the model to implement hierarchical centering of the random effects (**?**):

$$z_{i,j} \approx \mathcal{N}(\boldsymbol{\beta}^{(d)}\boldsymbol{x}_{ij}^{(d)} + s_i + r_j + \boldsymbol{u}_i'\boldsymbol{v}_j), \tag{9}$$

$$s_i = \boldsymbol{\beta}^{(s)}\boldsymbol{x}_i^{(s)} + a_i, \tag{10}$$

$$r_j = \boldsymbol{\beta}^{(s)}\boldsymbol{x}_j^{(r)} + b_j. \tag{11}$$

1.1. *Parameters and Priors*

To estimate the parameters discussed in the previous section, we utilize conjugate priors and a Monte Carlo Markov Chain (MCMC) algorithm. Prior distributions for the parameters are specified as follows:[1]

- $\boldsymbol{\beta}^{(s)}$, $\boldsymbol{\beta}^{(r)}$, and $\boldsymbol{\beta}^{(d)}$ are each drawn from multivariate normals with mean zero and a covariance matrix in which the covariances are set to zero and variances to 10
- $\Sigma_{a,b} \sim$ inverse Wishart$(I_{2\times 2}, 4)$
- σ_u^2, and σ_v^2 are each drawn from an i.i.d. inverse gamma(1,1).

Starting values for each of the parameters are determined using maximum likelihood estimation.

1.2. *The MCMC algorithm*

To estimate this model a Gibbs sampler is used. This sampler follows the procedure laid out in **??** with the exception of the first step in which we extend the GBME by accounting for the possibility that seemingly symmetric events are the result of a joint decision between a pair of actors. This first step involves sampling from a truncated normal distribution, we show the full conditional distribution below.

1. Modeling partially observable outcome. Conditional on there being an observed link between i and j, and conditional on other parameters, we draw the latent variables z_{ij} and z_{ji} from the bivariate normal distribution such that both latent variables are positive:

$$\begin{pmatrix} z_{ij} \\ z_{ji} \end{pmatrix} \Bigg| \, y_{ij} = 1 \sim \mathcal{N}\left(\begin{bmatrix} \mu_{ij} + a_i + b_j + \boldsymbol{u}_i' \boldsymbol{v}_j \\ \mu_{ji} + a_j + b_i + \boldsymbol{u}_j' \boldsymbol{v}_i \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \right) \mathbb{1}\{z_{ij} > 0 \cap z_{ji} > 0\}.$$

[1] For details on the full conditional distributions of each of the parameters see **?**.

Conditional $y_{ij} = 0$ (there is no observed link between i and j), we sample the latent variables from the bivariate normal distribution where at least one of the latent variables, z_{ij} or z_{ji}, is constrained to be negative:

$$\begin{pmatrix} z_{ij} \\ z_{ji} \end{pmatrix} \Bigg| \, y_{ij} = 0 \Bigg) \sim \mathcal{N} \left(\begin{bmatrix} \mu_{ij} + a_i + b_j + \boldsymbol{u}_i' \boldsymbol{v}_j \\ \mu_{ji} + a_j + b_i + \boldsymbol{u}_j' \boldsymbol{v}_i \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \right) \mathbb{1}\{z_{ij} < 0 \cup z_{ji} < 0\}.$$

2. Additive effects

   - Sample $\boldsymbol{\beta}^{(d)}, s, r \mid \boldsymbol{\beta}^{(s)}, \boldsymbol{\beta}^{(r)}, \Sigma_{a,b}, Z, \boldsymbol{U}, \boldsymbol{V}$ (linear regression)
   - Sample $\boldsymbol{\beta}^{(s)}, \boldsymbol{\beta}^{(r)} \mid s, r, \Sigma_{a,b}$ (linear regression)
   - Sample $\Sigma_{a,b}$ from full conditional distribution

3. Multiplicative effects[2]

   - For $i = 1, \ldots, n$:
     - Sample $u_i \mid \{u_j, j \neq i\}, Z, \boldsymbol{\beta}^{(d)}, s, r, \sigma_u^2, \sigma_v^2, \boldsymbol{V}$ (linear regression)
     - Sample $v_i \mid \{v_j, j \neq i\}, Z, \boldsymbol{\beta}^{(d)}, s, r, \sigma_u^2, \sigma_v^2, \boldsymbol{U}$ (linear regression)

1.3. Simulation Exercise

To test the capabilities of the P-GBME framework in representing the data generating process for a partially observable outcome we conduct a simulation exercise. In each simulation, we randomly construct a directed network from a pair of dyadic covariates, nodal covariates, and the random effects structure detailed in the previous section. The regression parameters for the dyadic covariates are set at 1 and -1/2, and the parameters for the nodal covariates are set at 0 and 1/2. At this stage, the network simulated from this data generating process is directed. We modify the simulated network so that a link between a dyad only appears in the network if

[2] See **?** for further details on how multiplicative effects are estimated in a directed context within the GBME framework.

both the i, j dyad and the j, i dyad both have a link in the simulated network, thus making the network appear undirected.

Next, we examine whether the P-GBME model can recover the data generating process underlying the partially observed simulated network. We run the P-GBME model in every simulation for 20,000 iterations with a 10,000 burn-in period. We repeat this simulation process 100 times.

With the simulation results our first step is to examine whether the P-GBME accurately recovers the regression parameter estimates for the dyadic and nodal covariates. To test whether this is the case we calculate the mean regression parameter estimate from the MCMC results for each simulation, and we summarize these results in Figure **??**. For each parameter we indicate its true value by a colored horizontal line and summarize the distribution of the mean regression values estimated from the P-GBME using a boxplot. Given that for each of the parameters the true value almost exactly crosses the median value indicated in the box plot, this simulation shows that the P-GBME is quite effective in estimating the true parameter values underlying a partially observed outcome.

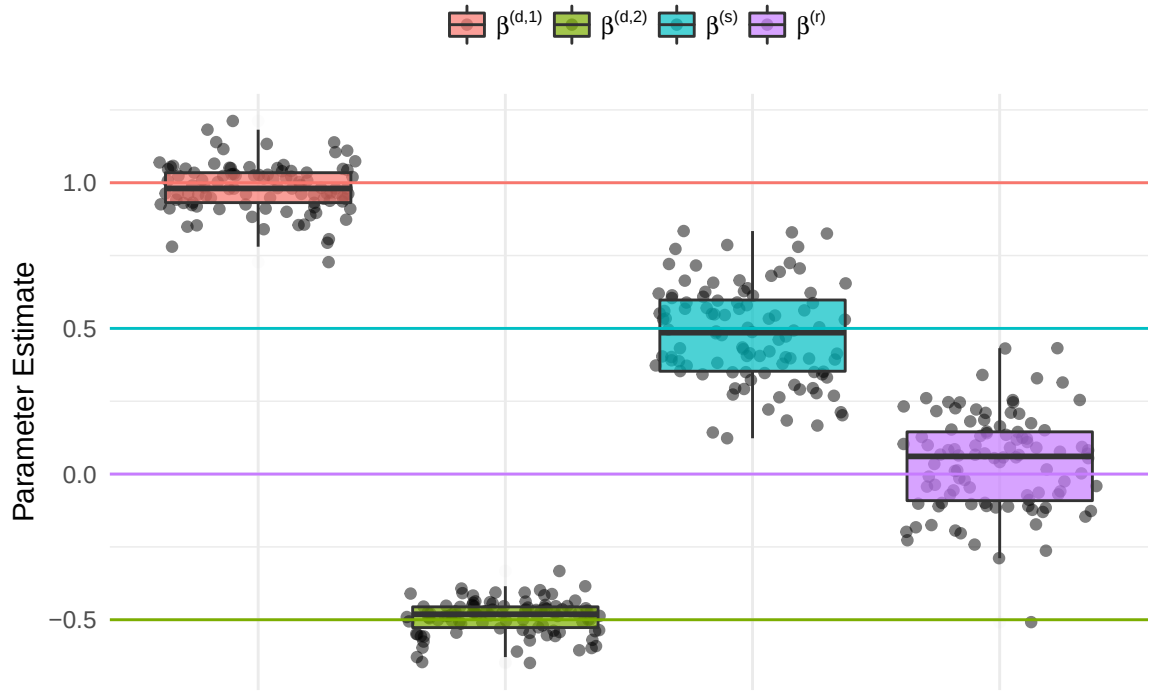

Figure 1: Boxplot of mean value of regression parameters estimated using the P-GBME across 100 simulations. Horizontal colored lines indicate the true parameter values.

We also examine the proportion of times that the true value falls within the 95% credible interval of the estimated regression parameter. In over ninety percent of the simulations, the true value falls within the 95% credible interval of each of the estimated regression parameters from the P-GBME. Specifically, for $\beta^{(d,1)}$ the coverage rate is 0.88, for $\beta^{(d,2)}$ 0.94, for $\beta^{(s)}$ 0.91, and for $\beta^{(r)}$ 0.90.

2. Estimation and Application

2.1. Data

Table **??** provides a description for each of the variables used in the analysis.

Table 1: Variables in the analysis

Variable	Level	Definition	Source
BIT	dyadic	1 if i & j Signed a BIT by year t	UNCTAD[3]
UDS (median)	dyadic	$\lvert \text{UDS}_{it} - \text{UDS}_{jt} \rvert$	Pemstein et al. (2010)
Law & Order	dyadic	$\lvert \text{LO}_{it} - \text{LO}_{jt} \rvert$	ICRG[4]
Log(GDP per capita)	dyadic	$\lvert \text{Log(GDPcap)}_{it} - \text{Log(GDPcap)}_{jt} \rvert$	WDI
OECD	dyadic	1 if i& j Both OECD members by year t	OECD
Distance	dyadic	Minimum distance between i&j	Gleditsch & Ward (2001) [5]
FDI/GDP	node	Net FDI inflow as % GDP in year t	WDI
ICSID Disputes	node	Cumulative number of disputes by year t	ICSID[6]
GDP per capita growth	node	Level of GDP per capita growth by year t	WDI
PTAs	node	Cumulative number of PTAs signed by year t	DESTA[7]

A shortcoming of the existing GBME framework is its inability to account for applications where there is missingness in the set of exogenous covariates used in the model. For our application, a number of the nodal covariates had varying levels of missingness. Additionally, most of the dyadic covariates that we construct from nodal variables, such as the unified democracy scores, also have varying levels of missingness. The table below shows how much missingness we had for the variables included in our analysis:

Table 2: Missingness among variables used in the analysis

Variable	Proportion of Cases Missing
Law & Order	16.6%
FDI/GDP	2.8%
GDP per capita	1.4%
GDP per capita growth	1.4%
Unified Democracy Scores (UDS)	0.7%
ICSID Disputes	0%
PTAs	0%
OECD	0%

In general, the level of missingness is not high. The only exception here is with the Law & Order variable from the ICRG dataset, for this variable we had approximately 17% of country-year observations missing from 1990 to 2012. The only true dyadic variable we include in our analysis is a calculation of the minimum distance between countries, and this variable has no missingness. Additionally, our dependent variable measuring whether or not two countries had signed a BIT by year t

also has no missingness.

A number of works have noted the issues that can arise when simply using list-wise deletion,[8] thus before running the P-GBME sampler we impute missingness among the covariates used in our model with a Bayesian, semi-parametric copula imputation scheme.[9] We generate 1,000 imputed datasets from this imputation scheme and save the last 500 for use in the P-GBME MCMC sampler.

To account for missingness within the P-GBME, at the beginning of every iteration of the MCMC for model, we draw a randomly sampled imputed dataset from the posterior of the Copula, calculate the parameters associated with the P-GBME using the imputed dataset, and repeat this process for every iteration of the sampler for the model. This approach directly incorporates imputation uncertainty into our posterior distributions of the P-GBME parameters without having to run and combine separate models.

2.2. Estimation details

In our application we estimate the P-GBME separately for each year from 1990 to 2012 using the prior distributions and MCMC algorithm described above. For each year, we ran the P-GBME MCMC sampler for 300,000 iterations, discarding the first 150,000 iterations as burn-in. We thinned the chain by saving only every 100^{th} value.

The following trace plot describes MCMC convergence for all parameters in the 2012 P-GBME model.

[8] See, for example, **?**.

[9] See **??** for details on this imputation scheme and how it differs from other approaches frequently utilized in political science.

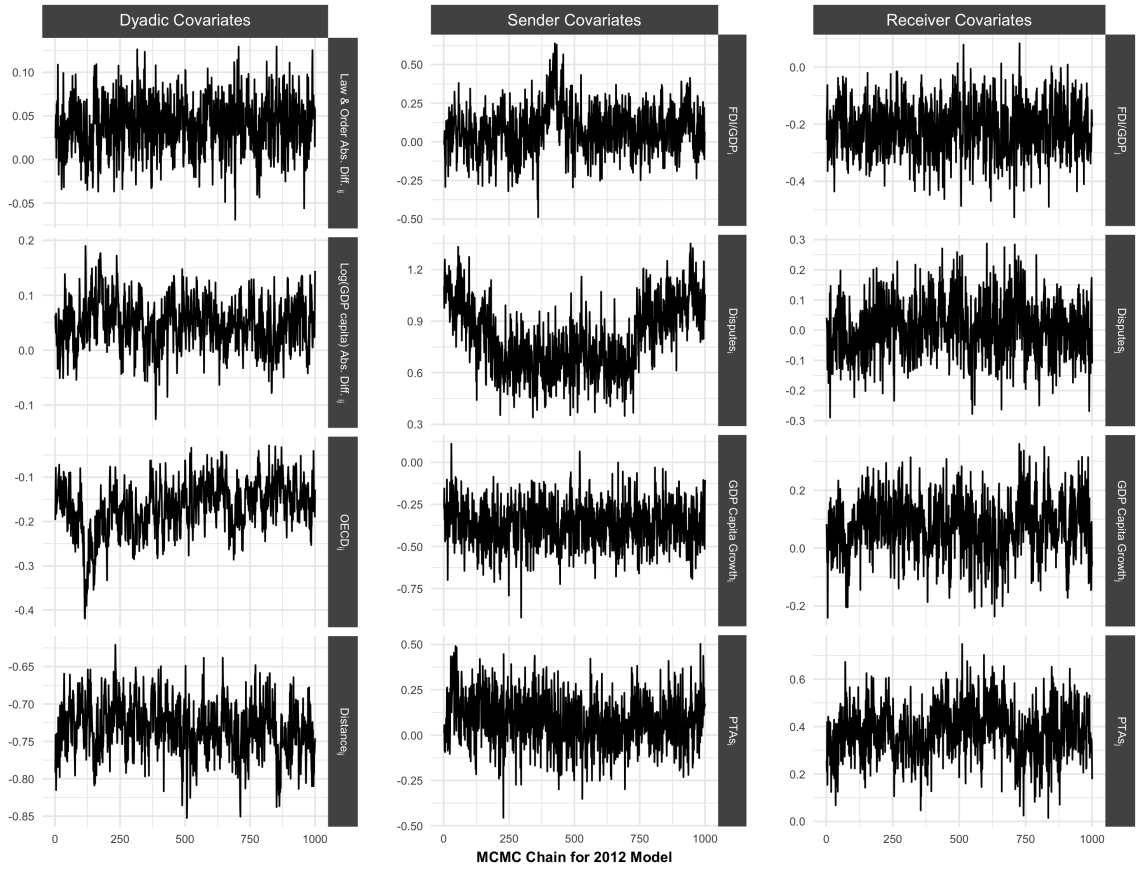

Figure 2: Traceplot for 2012 P-GBME results.

2.3. Regression Parameters

Figure **??** displays the posterior mean and the 90% (thicker) and 95% (thinner) credible intervals (CIs). The first column contains the dyadic covariates; the second, sender-level covariates; and, the third, receiver covariates. In each panel, we show the parameter estimates for that variable from 1990 to 2012. The dotted horizontal line is 0 and the thicker grey line is the posterior mean, pooling the posterior draws across all years.

The P-GBME recovers directed sender- and receiver-effects for node-level covariates from an observed undirected network, something that other approaches, including the GBME, are unable to do. Our estimation in this application indicates substantial instability in these estimates over time, both relative to a baseline of 0 and relative to the pooled posterior mean. This instability is consistent with substantive arguments that the incentives to sign BITs have changed over time (**?**).

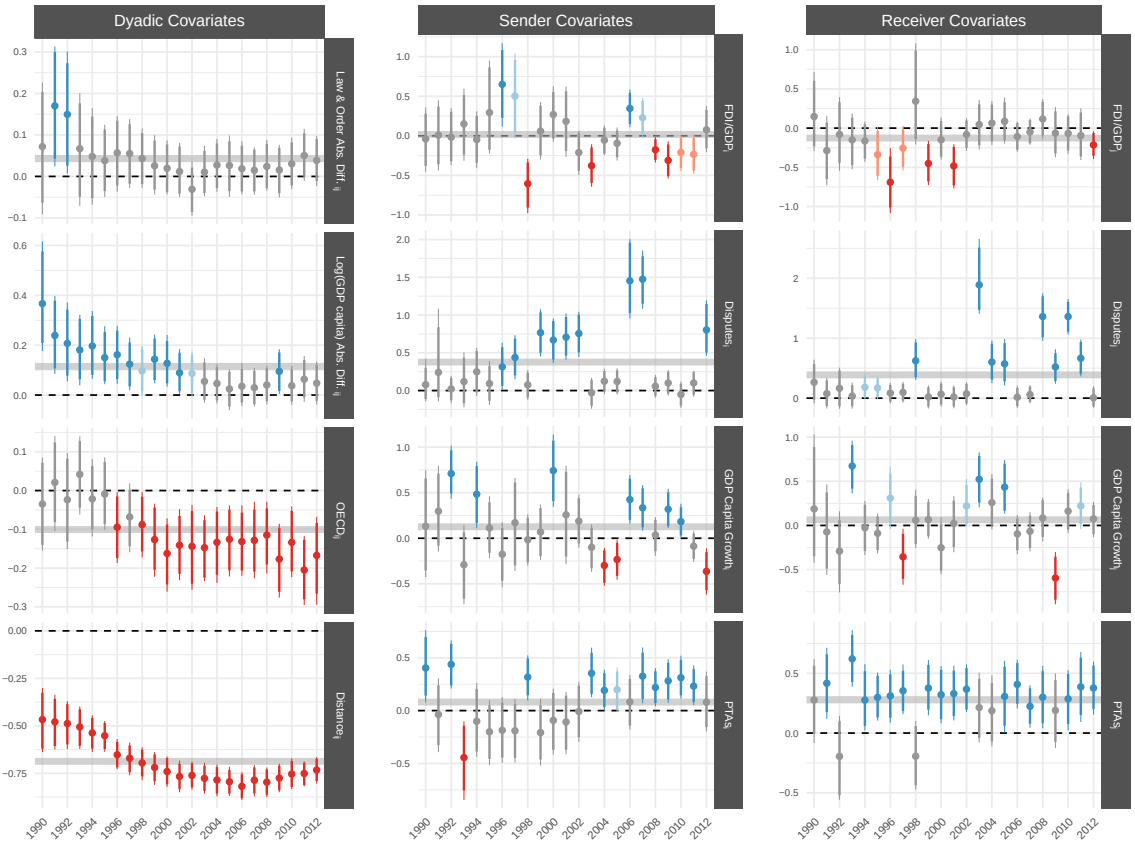

Figure 3: Left-most plot shows results by year for the dyadic parameters, next shows parameter results for sender covariates, and right-most plot show results for receiver covariates. Points in each of the plots represents the average effect for the parameter and the width the 90 and 95% credible intervals. The grey bar in the panels represents the average effect of the parameter across all years. Dark shades of blue and red indicate that the 95% CI does not contain 0 and lighter shades implies that the 0 is not in the 90% credible interval. Parameters in grey are are ones where 0 is inside the 90% credible interval.

Dyadic covariates tend to be more stable across years in this application. But they, too, show that the BIT formation process has changed over time. Economic and political "distance", for example, has become less important as the network evolved and more lower-income countries have signed BITs with each other. Geographic distance, on the other hand, continues to be strongly related to the formation of BITs.

P-GBME covariate estimates shed light on the evolving processes producing the observed BIT network in the 1990-2012 period. The changing values of covariate parameters over time also indicates that common practice of pooling dyads and assuming the existence of temporally stable parameters may be dangerous.

2.4. Choosing dimension of the multiplicative effects, K

One of the parameters that users are able to set within the P-GBME to account for third order dependence patterns is K – see the MCMC algorithm section above for more details on this parameter and its relation to the model. In the results reported in the paper, we set K = 2. To understand whether or not a higher value of K is necessary users of this approach can compare the in-sample fit of the model with varying values for K. In our application exercise, we varied K from 1 to 3 to settle on an appropriate value of K that can represent the data generating process of the network. Results are shown in Table **??** below.

Table 3: In-sample performance results from running the P-GBME on data from 2012 with varying values of K.

	AUC (ROC)	AUC (PR)
K=1	0.86	0.60
K=2	0.88	0.65
K=3	0.89	0.66

As you can see after K=2, the subsequent in-sample performance improvement notably declines. There is a slight increase in performance from K=2 to K=3, however, every time one increases K we are also adding $2 * n$ more parameters to the P-GBME model. Adding this many more parameters can easily lead one to overfit the data in an out-of-sample context.

References

Gelfand, Alan E., Sujit K. Sahu and Bradley P. Carlin. 1995. "Efficient parametrisations for normal linear mixed models." *Biometrika* 82(3):479–488.

Gleditsch, Kristian S. and Michael D. Ward. 2001. "Measuring Space: A Minimum Distance Database and Applications to International Studies." *Journal of Peace Research* 38(6):749–768.

Hoff, Peter. 2009. "Multiplicative Latent Factor Models for Description and Prediction of Social Networks." *Computational and Mathematical Organization Theory* 15(4):261–272.

Hoff, Peter D. 2005. "Bilinear Mixed-Effects Models for Dyadic Data." *Journal of the American Statistical Association* 100(4690):286–295.

Hoff, Peter D. 2007. "sbgcop: Semiparametric Bayesian Gaussian copula estimation and imputation." R Package.

Hollenbach, Florian M., Iavor Bojinov, Shahryar Minhas, Nils W. Metternich, Michael D. Ward and Alexander Volfovsky. 2016. "Principled Imputation Made Simple: Multiple Imputation Using Gaussian Copulas." *Working Paper* .

Jandhyala, Srividya, Witold J. Henisz and Edward D. Mansfield. 2011. "Three waves of BITs: The global diffusion of foreign investment policy." *Journal of Conflict Resolution* 55(6):1047–1073.

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1):49–69.

Pemstein, Daniel, Steven A. Meserve and James Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18(4):426–449.

Rajbhandari, Ashish. 2014. Identification and MCMC Estimation of Bivariate Probit Models with Partial Observability. In *Bayesian Inference in the Social Sciences*, ed. Ivan Jeliazkov and Xin-She Yang. Wiley.

Weidmann, Nils B. and Kristian Skrede Gleditsch. 2015. *cshapes: CShapes Dataset and Utilities*. $\mathcal{R}$ package version 0.5-1.
URL: *https://CRAN.R-project.org/package=cshapes*