Inference on the Effects of Observed Features in Latent Variable Models for Networks

Zachary Jones Matthew Denny Bruce Desmarais Hanna Wallach

April 23, 2018

Abstract

Due to the complex interdependence found in network data, networks scholars draw upon an increasingly sophisticated toolkit for building models of networks. The Latent space model (LSM) for network data combines a conventional regression model with an embedding of actors in a latent space. In the LSM, the expected relationship between two actors is dependent on observed covariates and a function of their positions in the latent space. In numerous applications, researchers have assumed that the latent spatial embedding can control for latent confounding structure. latent network structure includes unmeasured actor attributes that can be represented spatially, as well as network dependencies such as reciprocity and transitivity. There has been little research that considers the LSM's performance in adjusting for latent network structure when making inferences about covariates. We investigate the LSM's performance via a simulation study. In the presence of an unmeasured actor attribute that can be modeled perfectly using a latent space, we find that the LSM exhibits unacceptable levels of bias and Type 1 error for inferences on measured covariates. However, the prediction error of the LSM when latent network structure is present is substantially lower in most cases. We conclude that the LSM is most appropriate for exploratory or predictive modeling, unless it can be assumed that all potential confounders have been measured.

1 Introduction

Statistical models in which the outcome data (i.e., the dependent variable) consists of network data have proliferated, and grown increasingly sophisticated, over the past decade. The defining feature of network data that renders conventional statistical models—namely, regression models—inappropriate, is that ties in a network exhibit complex forms of interdependence such as reciprocity, transitivity, and homophily. If not accounted for, these complex dependencies can lead to biased estimates and errors in hypothesis testing, much in the way that omitted variable bias can affect results in conventional regression models (Ward, Siverson and Cao 2007; Kinne 2014; Cranmer and Desmarais Accepted; Hays, Kachi and Franzese 2010). We refer to network dependencies that cannot be modeled with observed covariates as *latent network structure*. A number of statistical modeling frameworks have been proposed to account for latent network structure in network data. These include, but are not limited to, the exponential random graph model (ERGM) **[UPDATE STRING CITES TO LOOK OUTSIDE POLISCI]** (e.g., Lazer, Rubineau, Chetkovich, Katz and Neblo 2010; Cranmer and Desmarais 2011; Desmarais and Cranmer 2012), the latent space model (LSM) (e.g., Ward, Siverson and Cao 2007; Ward and Hoff 2007; Kirkland 2012), and the stochastic actor oriented model (SAOM) (e.g., Berardo and Scholz 2010; Kinne 2014).

The approach to adjusting for latent network structure in the ERGM and SAOM is quite similar to adjusting for confounding covariates in regression modeling. The researcher specifies a set of dependencies that (s)he hypothesizes to be important in the generative model for the network. These dependencies are then explicitly included in a model that simultaneously represents the effects of observed covariates (Cranmer and Desmarais 2011). The LSM takes a different approach, which involves the incorporation of latent variables to model latent network structure. The LSM has an advantage over ERGM and SAOM in that researchers need not develop a precise set of hypothesized dependencies in order to model latent network structure. However, this advantage hinges upon the capacity for the LSM to differentiate the effects of latent network structure from those of the observed covariates.

Despite their growing popularity, few studies exist that investigate the performance of these models in adjusting for confounding network structure. In the current study, we focus on the LSM, examining its performance in reducing estimation and inferential errors regarding the effects of observed covariates via adjustment for confounding network structure. There are two reasons that using the LSM may lead to increased error—one may result in Type 1 inferential error, and the other in Type 2 error. First, if the latent network structure is truly correlated with the observed covariates, the unobserved structure that can be correlated with the observed variable may be attributed to the observed variable. Under this condition the latent space parameters are used to model sources of variation that cannot be attributed to the observed covariate, and inferences regarding the observed covariates remain subject to omitted variable bias, which leads to an inflated Type 1 inferential error rate. Second, the latent configurations inferred may result in a representation of the network wherein a node's position in the latent space is spuriously correlated with the observed covariates (i.e., the latent space crowds out the effects of observed covariates). Inferring latent variables that are correlated with observed covariates would lead to reduced efficiency in estimating observed covariate effects, and result in a high Type 2 error rate.

2 The Latent Space Model

The LSM, introduced by Hoff, Raftery and Handcock (2002), is used to estimate the effect of covariates in the presence of latent network structure. Here the distance function $|z_i - z_j|$ represents latent network structure as homophily with respect to latent variables. The distance function is additively combined with a regression on observed dyadic covariates, x_{ij}, to form a linear predictor for tie prediction. As with a GLM, a link function, g^{-1}, maps this linear predictor to the appropriate edge distribution, giving

$$\mathbb{E}(y_{ij}|x_{ij}) = g^{-1}(\alpha + \beta x_{ij} - |z_i - z_j|).$$

This "Euclidean" LSM is the original form proposed by Hoff, Raftery and Handcock (2002), and, as far as we can tell, the most commonly used specification in the literature.

The LSM has seen use in a variety of fields in which network data is common, particularly the social sciences. The apparent appeal of the LSM appears to be driven primarily by the LSM's

usefulness in modeling transitivity (i.e., clustering) and homophily, which are ubiquitous in social networks. In political science the LSM and variations on the form developed in Hoff, Raftery and Handcock (2002) have been used to estimate the effect of democracy on the probability of a militarized interstate dispute (Ward, Siverson and Cao 2007), the amount of portfolio investment between states (Cao and Ward 2013), and the effect of multimember districts on the probability of collaboration between state legislators in the United States (Kirkland 2012). Variations of the LSM developed for networks measured over time have been applied to the study of international trade, wherein the effects of various features of trading partners are estimated (Ward, Ahlquist and Rozenas 2013). In ecology the LSM has been used to study the sociality of elephants (Vance, Archie and Moss 2009) and orcas (Fearnbach, Durban, Ellifrit, Waite, Matkin, Lunsford, Peterson, Barlow and Wade 2014), birds (Nomano, Browning, Savage, Rollins, Griffith and Russell 2015), to discover ecological communities (Fletcher, Acevedo, Reichert, Pias and Kitchens 2011; Fletcher Jr, Revell, Reichert, Kitchens, Dixon and Austin 2013), and to study food webs (Chiu and Westveld 2011). In epidemeology it has been used to identify clusters of infected persons for later isolation (Zhang, Wang, Wang and Fang 2015) and to study patterns of interaction amongst physicians (Paul, Keating, Landon and O'Malley 2014). In marketing and business research it has been used to study inter-group trust (Dass and Kumar 2011), optimal bundling and pricing of goods and brands for retailers (Dass and Kumar 2012). It has been used to describe topic-specific patterns of interaction in e-mail communication networks (Krafft, Moore, Desmarais and Wallach 2012). It has been used to estimate the effects of education policy interventions on the structure of friendship networks among students (Sweet and Junker 2011). Lastly, in neuroscience it has been proposed as a method for modelling fMRI data (Simpson, Bowman and Laurienti 2013).

Although the LSM has often been used as an exploratory or predictive model, it has been applied in many cases for explanatory causal modeling—to reduce estimation and/or inferential error with respect to the effects of observed covariates. We do not claim that the methodological literature in which the LSM was introduced or extended has presented the LSM as capable of adjusting for unmeasured confounding structure. However, as we document below, many researchers who have applied the LSM claim that it has the capacity to adjust for confounders, and use it for this purpose. For example Ward, Siverson and Cao (2007) argue that the the LSM improves inference about the effects of democracy, international trade, and participation in international organizations on the probability of inter-state conflict.

> "The history of international disputes, and consequently the extant data on militarized interstate disputes, is replete with ... dependencies. We formally incorporate and estimate the extent of these ... dependencies in our model of the Kantian peace in order to more precisely determine the effects of the Kantian tripod on international conflict" (Ward, Siverson and Cao 2007, p. 585)

This example is representative of justifications for using the LSM that we see in other work in political science and international relations.

> "For the most part, however, most dyadic research in international relations ignores the

essential features of dyads in that they fail to satisfy the assumption of independence or, by construction, have missing data but ignore its effects: both of these bias the results in a fundamental way" (Dorff and Ward 2013, p. 2).

"This approach combines a network analysis with a standard-looking regression to permit us to access the importance of our explanatory factors without having them biased by the interdependencies in the network we are studying" (Cao and Ward 2013, p. 15).

"The presence of . . . dependence implies misspecification and a high likelihood of bias in most current applications. Building on the latent space framework we model the world trading system without assuming a particular network structure or the sufficiency of particular network statistics" (Ward, Ahlquist and Rozenas 2013, p. 20).

". . . the latent space model allows for the assessment of distance between two unconnected actors while simultaneously controlling for the interdependence inherent in network data. This interdependence in latent space positions allows the model to control for common network effects like reciprocity or transitivity that would ordinarily bias results" (Kirkland 2012, p. 336).

"Ignoring third-order dependence in dyadic data and treating dyads Germany-France, France-Italy, and Germany-Italy as independent observations can cause bias in parameter estimates (Hoff 2005). The statistical literature has proposed a series of latent space to control for autocorrelation among dyadic observations (Hoff et al. 2002; Hoff 2005). Countries? unobserved characteristics are captured by latent vectors..." (Cao and Ward 2016, p. 17).

We also see examples in sociology,

"the models include latent space positions that implicitly control for structural aspects of networks, such as reciprocity and centrality, allowing for estimation of covariate effects (Hoff, Raftery, and Handcock 2002)." (Spillane, Shirrell and Sweet 2017, p. 157)

"Higher order dependencies can be taken into account by assigning each country a position in a ?social space,? based on its (unobserved) characteristics and the presence of other ties. Conditional independence between observations can then subsequently be assumed, given the countries? latent position within the social space (Hoff et al., 2002)." (Berlusconi, Aziani and Giommoni 2017, p. 107)

in ecology,

> "...can create artificially exaggerated synchrony rates regardless of motivation for signaling, which can be modeled with a term known as "transitivity" in the social network literature. A latent space model (Krivitsky et al. 2009) was used to examine the propensities for synchrony between helper males and the primary male while accounting for the variability deriving from the transitivity." (Nomano et al. 2015, p. 989)

and in the business literature.

> "This model controls for higher order team dynamics using the bilinear component $z_i' z_j$ and also allows both trustor ($x_{tor,i}$) and trusted ($x_{ted,j}$) characteristics to be investigated along with the dyadic covariates ($x_{d,i,j}$)." (Dass and Kumar 2011, p. 7)

The latent space framework offers an attractive general purpose approach to adjusting for confounding structure in network models, as, unlike the major alternative framework for network modeling (ERGM), using the latent space model does not require the researcher to specify a set of network dependencies for which to control. However, the latent variables in the LSM are free parameters that will be inferred to explain variation that is not explained by the observed covariates. In a simulation study that follows, we examine whether the LSM can be used to adjust for confounding in a simple and ideal scenario..

3 Simulation Study

We use a simulation study to evaluate the LSM's performance at inferring covariate parameters in the situation in which we have some observed covariates as well as omitted network structure which can be represented using a Euclidean latent space. We are particularly interested in whether the LSM can adjust for confounding when the confounding variable is unmeasured. In the simulation design, we study how the level of confounding affects the performance of the LSM relative to the generalized linear model (GLM). We evaluate performance in terms of bias and inferential error. Since the LSM has also been used for prediction, we also evaluate predictive performance (Ward, Ahlquist and Rozenas 2013; Fletcher et al. 2011; Fletcher Jr et al. 2013; Chiu and Westveld 2011, e.g.,).

3.1 Simulation Design

For simplicity and computational efficiency we consider a one-dimensional latent space. That is, we create a variable that can be represented using the Euclidean distance between two nodes in a one-dimensional space. We expect our results to generalize to higher dimensions and other variants of the LSM since the tendency for latent variables to be used in explaining structure that cannot be attributed covariates is universal with respect to both the dimension of the latent space and the form of the LSM. To generate an observed covariate that exhibits a controllable degree

of correlation with the latent network structure we follow a four-step process. The steps are as follows.

1. Simulate one-dimensional positions for each node (i.e., a scalar value for each node), drawing from a standard normal distribution, and calculate the Euclidean distance (i.e., absolute distance) $\mathbf{d}$ between each pair of positions. We then normalize $\mathbf{d}$ to assure it has unit variance by $\mathbf{d} := \mathbf{d}/\text{sd}(\mathbf{d})$, where $\text{sd}(\cdot)$ is the empirical standard deviation.

2. Generate a correlation coefficient (ρ) on the interval $(-1, 1)$ as $\rho = 1 - 2 \times u$, where $u \sim \text{Beta}(\eta, \eta)$. η is used to control the magnitude of the correlation coefficient. The smaller the value of η, the larger the absolute magnitude of the correlation coefficient, in expectation. This relationship is illustrated in Figure 1. When $\eta = 1$, the correlations generated exhibit the expected uniform distribution. When $\eta = 1,000,000$, each correlation coefficient is approximately 0. [1]

3. Simulate $\mathbf{x} = \frac{\rho}{\sqrt{1-\rho^2}}\mathbf{d} + \mathbf{w}$, where $\mathbf{w} \sim N(0, 1)$. This assures that the correlation between $\mathbf{x}$ and $\mathbf{d}$ is ρ.[2]

4. Finally we standardize $\mathbf{x} := \mathbf{x}/\text{sd}(\mathbf{x})$ to have unit variance to ensure that the variance of $\mathbf{x}$ relative to $\mathbf{d}$ is unrelated to ρ. This is to assure that ρ does not affect the efficiency with which the effect of $\mathbf{x}$ can be estimated through some property unrelated to the relationship between $\mathbf{x}$ and $\mathbf{d}$.

We consider three exponential family distributions from which the adjacency matrix entries are drawn: Gaussian, Bernoilli, and Poisson. We also vary the size of the network under study, considering networks with $n = 25, 50$, and 100 nodes. To vary the degree of confounding attributable to the latent positions, we consider three values of the collinearity parameter η: 1, 100, and $1,000,000$. Here where $\eta = 1$ gives a matrix where the correlation is in expectation $.5$, $\eta = 100$ results in moderate collinearity between the observed covariate and the latent distances, and $\eta = 1,000,000$ corresponds to independence between the observed covariate and the latent distances. See Figure 1 for the distribution of the absolute value of the correlation generated at each value of η.

We consider the LSM with several different priors on the coefficient for the observed covariate and the latent space. We set the prior variance of β to be either 1 or 10 and use a diffuse normal prior on the latent space.

The LSM is estimated using the canonical implementation in the `latentnet` package in `R` (Krivitsky and Handcock 2008). For each iteration of the simulation, we run 10,000 burn in iterations, followed by 1,000,000 iterations of the sampler. Every 100th post burn in iteration is saved.

[1] This methodology generalizes to higher dimensional correlation matrices as the 'c-vine' method developed by Lewandowski, Kurowicka and Joe (2009), and implemented in the R package `clusterGeneration` (**?**), which we use in our simulations.

[2] This could be generalized to more than one covariate using the conditional normal distribution derived by (Eaton 1983, pp. 116–117).

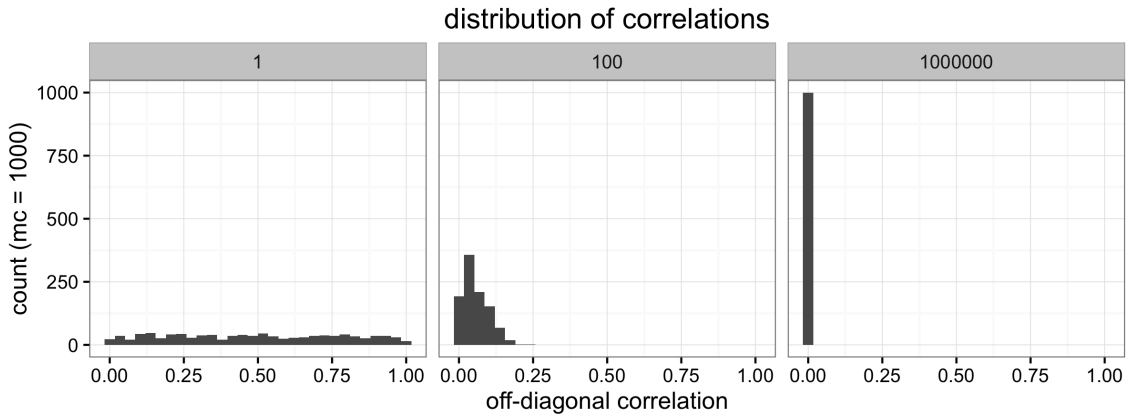

Figure 1: The distribution of the absolute value of the correlation between the observed covariate and the latent distances. The value in the plot title corresponds to the value of η.

Convergence in the log probability of the model is assessed using the Geweke diagnostic in the `coda` package (Plummer, Best, Cowles and Vines 2006; Geweke et al. 1991). If the convergence criterion is satisfied, the simulation continues to the next set of arguments, otherwise the number of iterations is doubled. If the convergence criterion is still not satisfied, then the aforementioned step in the simulation is flagged for review. However, this was not necessary in any cases. At each point in the simulation's parameter space, we execute 1,000 Monte Carlo iterations. [3]

We compute the MLE estimated by iteratively reweighted least squares via the `glm` function in `R` (R Core Team 2015). For the LSM we use the posterior mode as our point estimate. In the cases where edges are Bernoulli and there is omitted network structure, we scale the estimates using the reciprocal of the bias of β when $\mathbf{x}$ and $\mathbf{d}$ are independent: $\sqrt{\frac{3.28+\beta^2 \text{Var}(\mathbf{d})}{3.29}}$, where 3.29 is the variance of a standard logistic distribution. We do this to adjust for bias in the coefficient estimate that arises due to the lack of a scale coefficient in logistic regression (See the derivation of the bias under an independent but omitted covariate in Mood (2010)).

For each combination of simulation conditions we evaluate the mean square prediction error of the model on new edges drawn condtional on $\mathbf{x}$ and $\mathbf{d}$: $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, the bias of the estimated coefficient for the measured covariate $\frac{1}{n}\sum_{i=1}^{n}(\beta - \hat{beta})$, and the Type-1 and 2 error rates. These error rates are estimated by computing coverage: $\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(\beta \in \text{HPD}(\hat{\beta}_i))$, where $\text{HPD}(\hat{beta})$ is the region of highest posterior density which covers 95% of the marginal posterior distribution and m is the number of samples from said posterior Turkkan and Pham-Gia (1993). When in the simulation $\beta = 0$, $1 - P(\text{coverage})$ gives the type-1 error rate, and $P(\text{coverage})$ givesthe type-2 error rate when $\beta = 1$. Hence we consider β to be statistically significant at the 0.05 level if the

[3] One of the primary difficulties in executing the above simulation design is the computational cost of estimating the LSM. We utilize the `BatchExperiments` R package to construct and execute our computational experiments on a Torque cluster Bischl, Lang, Mersmann, Weihs et al. (2015).

95% credible interval excludes 0.

Recent work on the LSM has explicitly addressed the potential problem of Type 2 error. The AMEN framework for latent variable modeling of networks is explicitly designed to model structure in the residuals—structure that is leftover after accounting for the observed covariates. Hoff (2015, p. 43) notes that the latent factors represent patterns in the network that, "aren't explained by the known regressors." Minhas, Hoff and Ward (2016, pp. 12–13) also describe how AMEN is designed such that the multiplicative effects, "capture higher-order dependence patterns that are left over [in the stochastic linear predictor] after accounting for any known covariate information." The approach incorporated in AMEN is effective at avoiding correlation between the observed covariates and latent factors, which would result in Type 2 error. However, there have been no methodological innovations to avoid Type 1 error in latent variable modeling with networks. Through a simulation study we find that the LSM, as implemented in both the R (R Core Team 2017) packages `latentnet` (Krivitsky and Handcock 2008) and `amen` (Hoff, Fosdick, Volfovsky and He 2015), performs poorly in terms of Type 1 error when the unmeasured network structure confounds the relationship between the observed covariate and the network.

We also repeat this simulation study using `amen` (Hoff 2015; Minhas, Hoff and Ward 2016). We omit the Poisson family, and, due to the computational cost of the simulation, we consider only networks with 25 nodes or 100 nodes. Rather than constructing an omitted variable as described above, we draw from a standard normal distribution $\mathbf{u}$, and, because we are considering symmetric networks, take the outer product of $\mathbf{u}$ with itself to construct the latent factor. We take 100,000 MCMC samples after 500 burn-in iterations, and we assess convergence using the diagnostic of Raftery and Lewis (1992).

3.2 Results

To evaluate estimation error we compute the bias. Figures 2 and 3 shows the results. In Figure 2 we can see that with a uniform distribution over the correlation between the distances and the observed covariate, all methods (except the true model wherein the distances are included via an observed covariate) show substantial bias. In most cases the LSM performs better than the GLM, sometimes by as much as 50%. However, in absolute terms, the amount of bias is large. Our results demonstrate that using the LSM will not permit the discovery of and adjustment for the true latent positions when the observed covariates and latent distances are confounded. For both the GLM and the LSM, bias decays rapidly as the degree of correlation between the distances and the measured covariate decreases. In 3 we can see that when there is no omitted network structure, the LSM exhibits bias greater than that of the GLM.

Figures 4 and 5 show the inferential error rates of the LSM with different priors for β, as well as a GLM. Type-1 error rates shown in Figure 4 for the LSM are in general somewhat lower than those of a GLM when an unmeasured covariate exists. Under uniform correlation between the omitted network structure and the observed covariate, the Type-1 error rate is often more than 10 times greater than the nominal rate of 0.05. This illustrates that the LSM cannot be considered a viable substitute to measuring the omitted structure, as the true model exhibits a Type-1 error rate

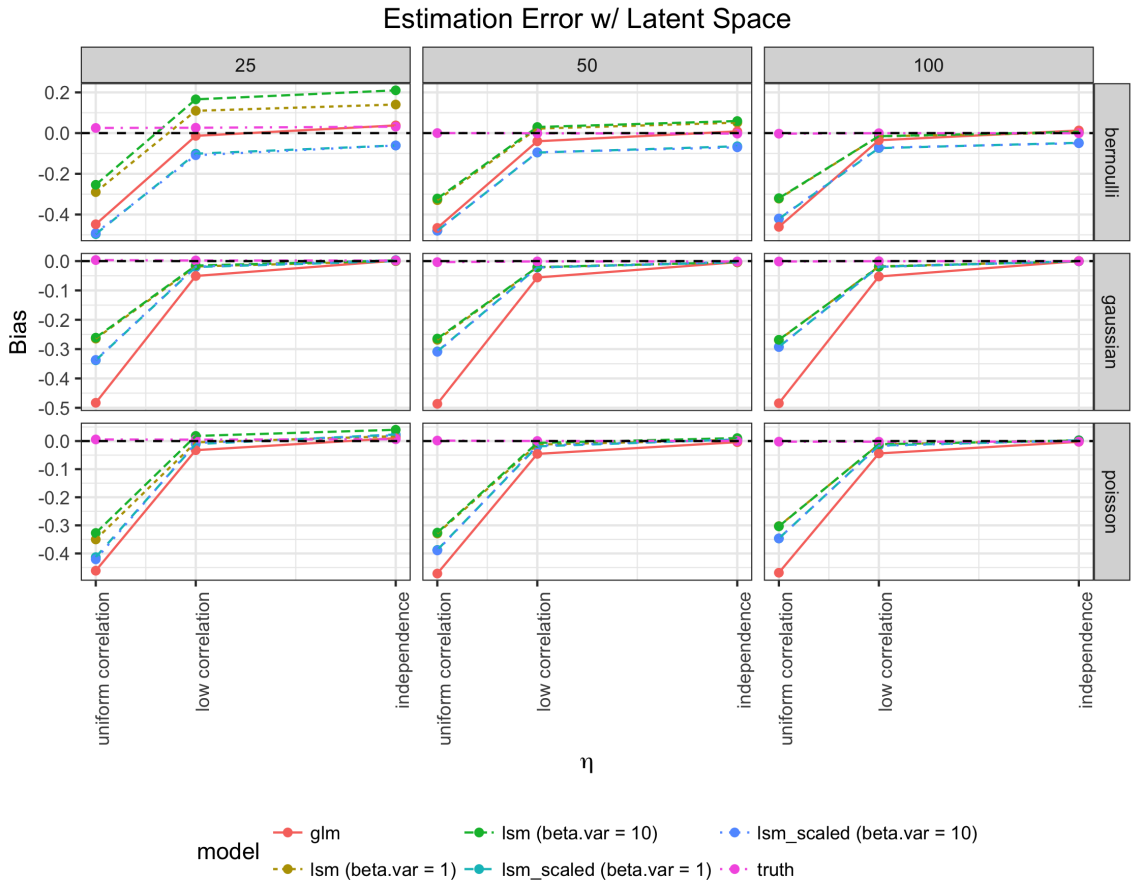

Figure 2: The bias of estimates of the effect of the observed covariate $\mathbf{x}$ when there is an ommitted variable. The x-axis gives the value of the parameter η which controls the degree of dependence between $\mathbf{x}$ and ommitted covariate. Lower values of η indicate higher levels of dependence between the observed and ommitted covariate. The y-axis gives a Monte Carlo estimate of the bias. The number of nodes are indicated in the top panels, while the distributional family of the edges is shown on the right panel. Each panel represents 4 values of η with 1,000 Monte Carlo iterations executed at each point.

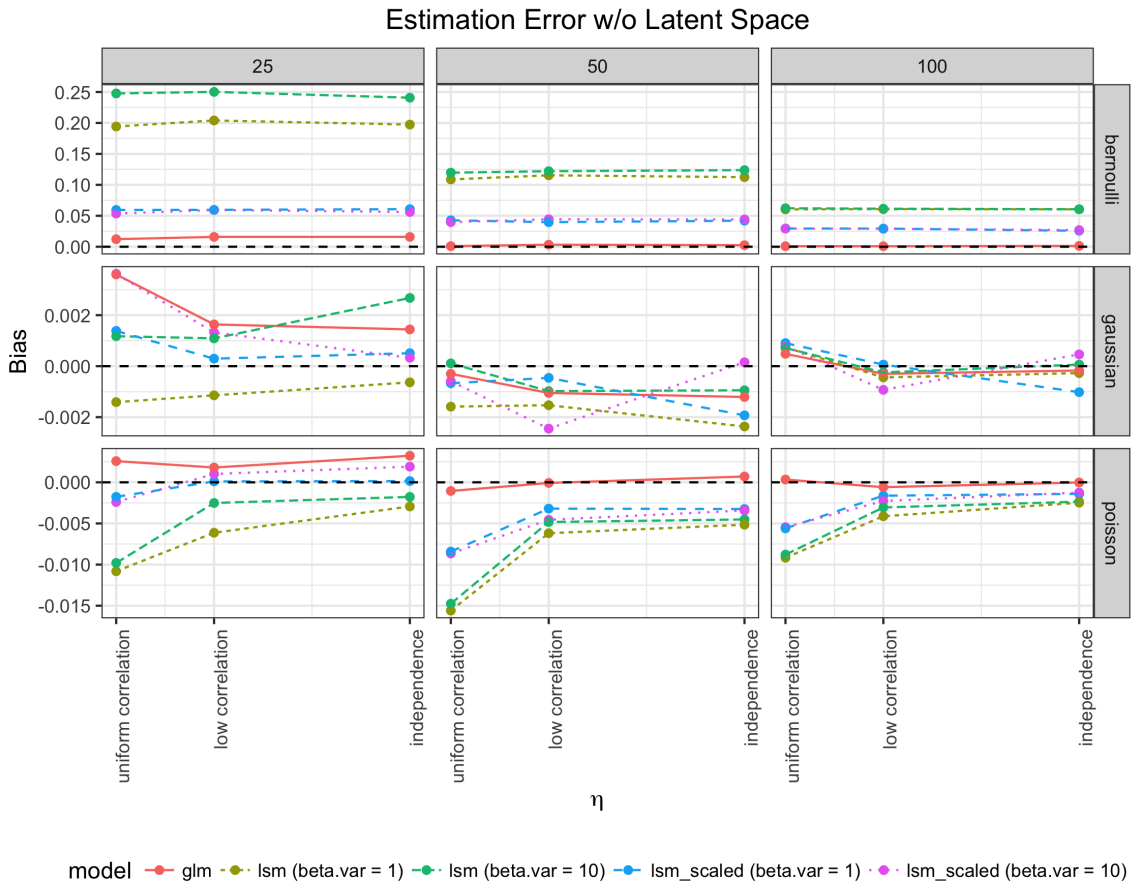

Figure 3: The bias of estimates of the effect of the observed covariate $\mathbf{x}$ when there is no ommitted variable.

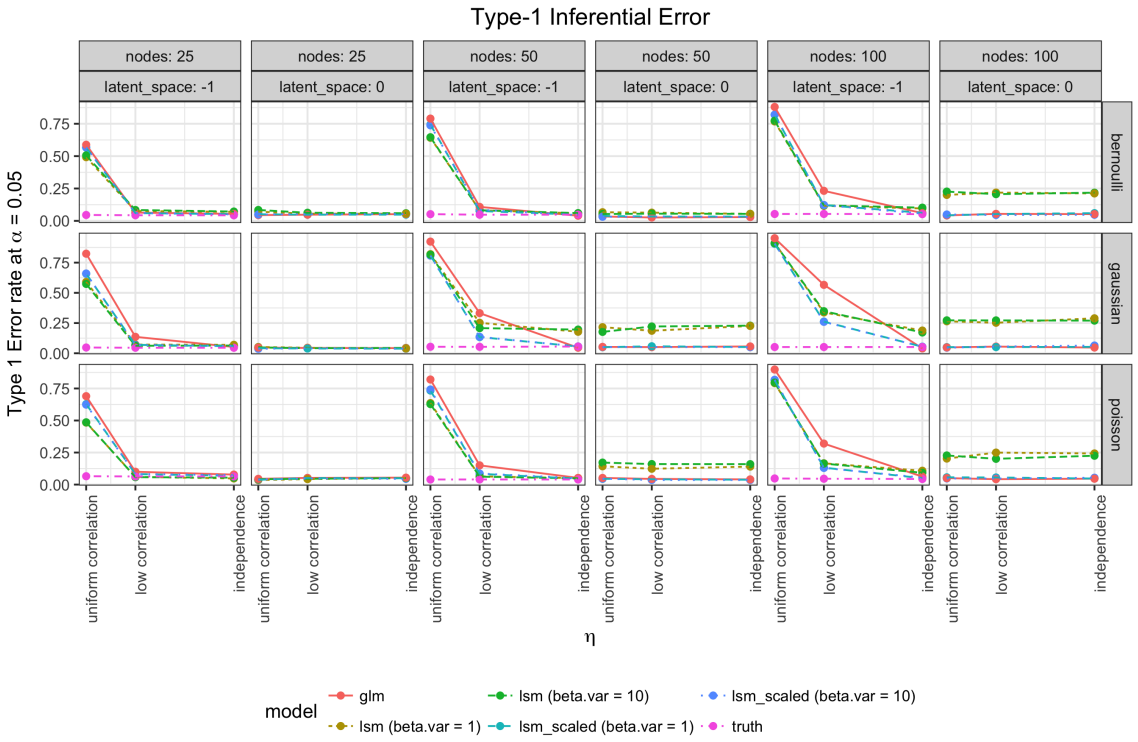

Figure 4: Monte Carlo estimates of the Type-1 error regarding the effect of the observed covariate $\mathbf{x}$ are shown on the y-axis. Here, $\beta = 0$ and the error rate shown in each panel in that row gives 1 minus the probability of a 95% confidence region (for the LSM) or interval (for the GLM) including 0, giving the probability that a true null hypothesis of $\beta = 0$ is falsely rejected.

that matches the nominal rate. Type-2 error rates, shown in Figure 5 are also substantially above their nominal rate when there is a uniform distribution on the strength of confounding, though the inflation is not as bad as that of the Type-1 rates. These results indicate that, in the presence of unobserved confounders that can be represented through the LSM, the LSM does not adequately correct the inferential errors that would arise due to omitted structure among the dyads in the network.

The last property we consider in comparing the GLM and LSM is generalization error—the performance exhibited by the models in predicting new data that was drawn from the same model that generated the data used for estimation (i.e., out of sample predictive performance). We estimate generalization error as the expected prediction error on new edges generated using a fixed set of latent positions for the nodes and a fixed observed covariate. Generalization error results are shown in Figure 6. In nearly all cases the LSM substantially outperforms the GLM, especially when the omitted covariate is not highly collinear with the observed covariate. In nearly all cases it performs

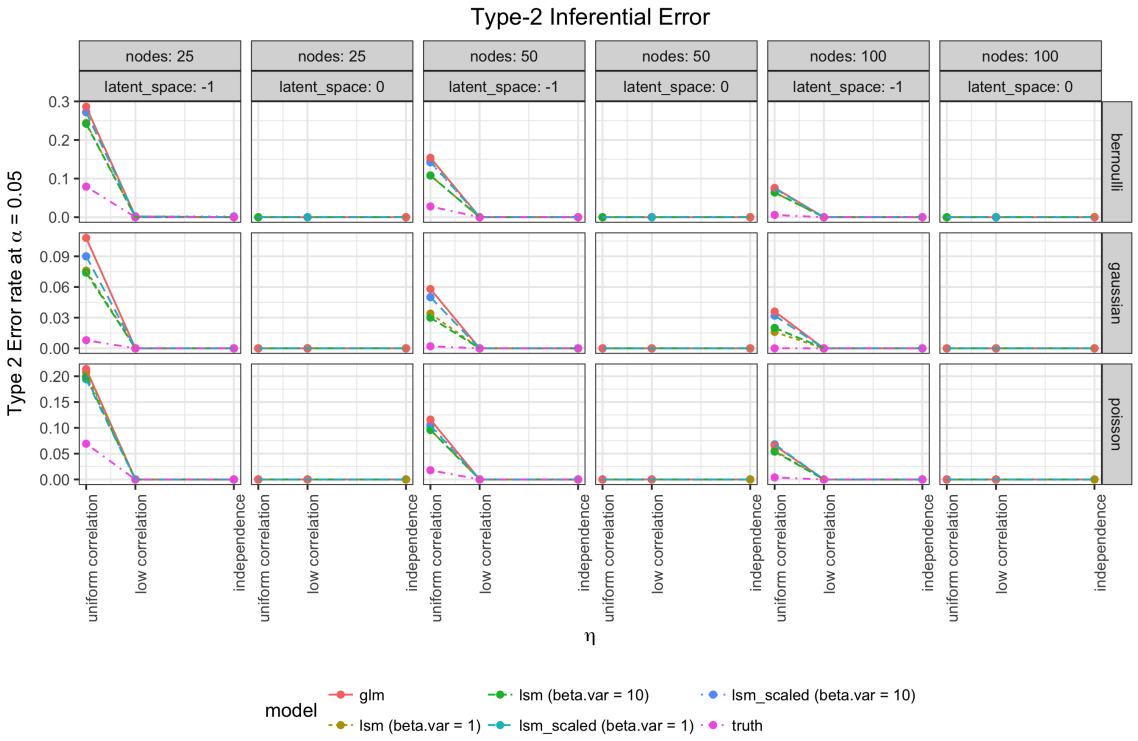

Figure 5: Monte Carlo estimates of the Type-2 error regarding the effect of the observed covariate $\mathbf{x}$ are shown on the $y - axis$. Here, $\beta = 1$, and the error rate shown in each panel in each row gives the probability of the probability/confidence intervals covering 0, giving the probability of accepting a false null hypothesis.

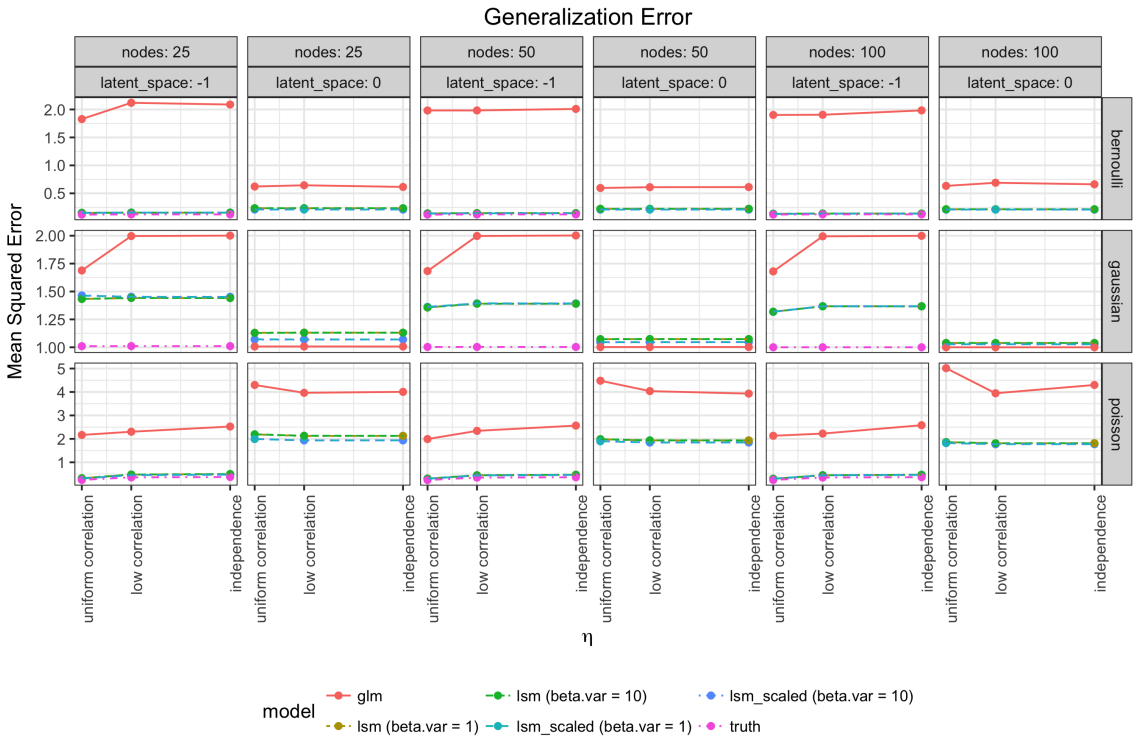

Figure 6: The x-axis shows a Monte Carlo estimate of the mean square error for edge values drawn from the appropriate distribution with the observed covariate $\mathbf{x}$ and the latent positions $\mathbf{z}$ fixed. In the Bernoulli case the Brier score is computed and in the Poisson and normal cases the mean-square-error.

as well as the true model.

In our repetition of this simulation with `amen` we find similar results, with the exception of type-2 error rates, which, as expected, are substantially lower for `amen`. However, type-1 error rates remain elevated, as is bias in the case where there is a latent dependence structure that is correlated with the observed covariate. `amen` does perform better in the case where it is used despite the absence of such a latent dependence structure though, relative to `latentnet`, which is to be expected given that it outperforms `latentnet` in terms of type-2 error. Like the LSM estimated with `latentnet`, `amen` performs well in terms of generalization error. The results of our simulations with `amen` are available in the Online Appendix.

4 Conclusion

Based on our simulation study, we conclude that the primary advantage of using the LSM—relative to a dyadic regression model in which the network is fit using observed covariates only—is that the latent space provides an efficient complement to observed covariates when it comes to fitting and predicting the network; `amen` increases this advantage. Considering both (1) that the LSM does not substantially reduce bias or inferential error relative to the GLM, and (2) that the LSM exhibits strong predictive performance, we conclude that the latent positions are used to explain the systematic variation that cannot be explained through the observed covariates. This is of considerable use in research focused on developing a predictive model, or fitting and exploring latent positions. However, since the latent positions are inferred to complement the observed covariates—explaining residual variation—the LSM is ill-suited to adjust for confounding network structure, as adjusting for confounding would require that the latent positions explain variation that could otherwise be attributed to the observed covariates.

The key recommendation arising from our results is that researchers not use the LSM as an alternative to measuring potential confounding variables and/or explicitly modeling the endogenous network structure. The LSM cannot control for unmeasured confounders. This property is, perhaps, not surprising. Given that all priors are centered at zero, the parameter regions that exhibit high posterior probability will be those that explain the observed network (i.e., high likelihood) using minimal departures from zero (i.e., high prior probability). Since explaining the observed network using a covariate requires moving one parameter value away from zero, whereas explaining the network using a latent dimension requires moving many latent coordinates away from zero, the LSM exhibits an inherent preference to explain network structure using observed covariates over latent coordinates. One way to address this issue would be to use a highly diffuse prior for the latent positions, but as we discuss above, that is not possible with the LSM. If the LSM is to be used as a tool for drawing inferences regarding observed covariate effects while adjusting for confounding network structure, further development is needed to enable the LSM to avoid bias and inferential error under confounding.

References

Berardo, Ramiro and John T. Scholz. 2010. "Self-Organizing Policy Networks: Risk, Partner Selection, and Cooperation in Estuaries." *American Journal of Political Science* 54(3):632–649.

Berlusconi, Giulia, Alberto Aziani and Luca Giommoni. 2017. "The determinants of heroin flows in Europe: A latent space approach." *Social Networks* 51:104–117.

Bischl, Bernd, Michel Lang, Olaf Mersmann, Claus Weihs et al. 2015. "BatchJobs and BatchExperiments: Abstraction mechanisms for using R in batch environments." *Journal of Statistical Software* 64(1):1–25.

Cao, Xun and Hugh Ward. 2016. "Transnational Climate Governance Networks and Domestic Regulatory Action." *International Interactions* (just-accepted).

Cao, Xun and Michael D Ward. 2013. "Do democracies attract portfolio investment?" *International Interactions, forthcoming* .

Chiu, Grace S and Anton H Westveld. 2011. "A unifying approach for food webs, phylogeny, social networks, and statistics." *Proceedings of the National Academy of Sciences* 108(38):15881–15886.

Cranmer, Skyler J. and Bruce A. Desmarais. 2011. "Inferential Network Analysis with Exponential Random Graph Models." *Political Analysis* 19(1):66–86.

Cranmer, Skyler J. and Bruce A. Desmarais. Accepted. "A Critique of Dyadic Design." *International Studies Quarterly* .

Dass, Mayukh and Piyush Kumar. 2011. "The impact of economic and social orientation on trust within teams." *Journal of Business & Economics Research (JBER)* 9(2).

Dass, Mayukh and Piyush Kumar. 2012. "Assessing category vulnerability across retail product assortments." *International Journal of Retail & Distribution Management* 40(1):64–81.

Desmarais, Bruce A. and Skyler J. Cranmer. 2012. "Micro-Level Interpretation of Exponential Random Graph Models with Application to Estuary Networks." *Policy Studies Journal* 40(3):402–434.

Dorff, Cassy and Michael D Ward. 2013. "Networks, dyads, and the social relations model." *Political Science Research and Methods* 1(02):159–178.

Eaton, Morris L. 1983. *Multivariate statistics: a vector space approach*. New York: Wiley.

Fearnbach, Holly, John W Durban, David K Ellifrit, Janice M Waite, Craig O Matkin, Chris R Lunsford, Megan J Peterson, Jay Barlow and Paul R Wade. 2014. "Spatial and social connectivity of fish-eating "Resident" killer whales (Orcinus orca) in the northern North Pacific." *Marine biology* 161(2):459–472.

Fletcher Jr, Robert J, Andre Revell, Brian E Reichert, Wiley M Kitchens, Jeremy D Dixon and James D Austin. 2013. "Network modularity reveals critical scales for connectivity in ecology and evolution." *Nature communications* 4.

Fletcher, Robert J, Miguel A Acevedo, Brian E Reichert, Kyle E Pias and Wiley M Kitchens. 2011. "Social network models predict movement and connectivity in ecological landscapes." *Proceedings of the National Academy of Sciences* 108(48):19282–19287.

Geweke, John et al. 1991. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Vol. 196 Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Hays, Jude C, Aya Kachi and Robert J Franzese. 2010. "A spatial model incorporating dynamic, endogenous network interdependence: A political science application." *Statistical Methodology* 7(3):406–428.

Hoff, Peter, Bailey Fosdick, Alex Volfovsky and Yanjun He. 2015. *amen: Additive and Multiplicative Effects Models for Networks and Relational Data*. R package version 1.1.
URL: *http://CRAN.R-project.org/package=amen*

Hoff, Peter D. 2015. "Dyadic data analysis with amen." *arXiv preprint arXiv:1506.08237* .

Hoff, Peter D, Adrian E Raftery and Mark S Handcock. 2002. "Latent space approaches to social network analysis." *Journal of the american Statistical association* 97(460):1090–1098.

Kinne, Brandon J. 2014. "Dependent diplomacy: Signaling, strategy, and prestige in the diplomatic network." *International Studies Quarterly* 58(2):247–259.

Kirkland, Justin H. 2012. "Multimember Districts' Effect on Collaboration between US State Legislators." *Legislative Studies Quarterly* 37(3):329–353.

Krafft, Peter, Juston Moore, Bruce Desmarais and Hanna M Wallach. 2012. Topic-partitioned multinetwork embeddings. In *Advances in Neural Information Processing Systems*. pp. 2807–2815.

Krivitsky, Pavel N. and Mark S. Handcock. 2008. "Fitting position latent cluster models for social networks with latentnet." *Journal of Statistical Software* 24(5).

Lazer, David, Brian Rubineau, Carol Chetkovich, Nancy Katz and Michael Neblo. 2010. "The coevolution of networks and political attitudes." *Political Communication* 27(3):248–274.

Lewandowski, Daniel, Dorota Kurowicka and Harry Joe. 2009. "Generating random correlation matrices based on vines and extended onion method." *Journal of multivariate analysis* 100(9):1989–2001.

Minhas, Shahryar, Peter D Hoff and Michael D Ward. 2016. "Inferential Approaches for Network Analyses: AMEN for Latent Factor Models." *arXiv preprint arXiv:1611.00460* .

Mood, Carina. 2010. "Logistic regression: Why we cannot do what we think we can do, and what we can do about it." *European Sociological Review* 26(1):67–82.

Nomano, Fumiaki Y, Lucy E Browning, James L Savage, Lee A Rollins, Simon C Griffith and Andrew F Russell. 2015. "Unrelated helpers neither signal contributions nor suffer retribution in chestnut-crowed babblers." *Behavioral Ecology* 26(4):986–995.

Paul, Sudeshna, Nancy L Keating, Bruce E Landon and A James O'Malley. 2014. "Results from using a new dyadic-dependence model to analyze sociocentric physician networks." *Social Science & Medicine* 117:67–75.

Plummer, Martyn, Nicky Best, Kate Cowles and Karen Vines. 2006. "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News* 6(1):7–11.
URL: *http://CRAN.R-project.org/doc/Rnews/*

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org/*

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org/*

Raftery, Adrian E and Steven M Lewis. 1992. "[Practical Markov Chain Monte Carlo]: comment: one long run with diagnostics: implementation strategies for Markov Chain Monte Carlo." *Statistical science* 7(4):493–497.

Simpson, Sean L, F DuBois Bowman and Paul J Laurienti. 2013. "Analyzing complex functional brain networks: fusing statistics and network science to understand the brain." *Statistics surveys* 7:1.

Spillane, James P, Matthew Shirrell and Tracy M Sweet. 2017. "The elephant in the schoolhouse: The role of propinquity in school staff interactions about teaching." *Sociology of Education* 90(2):149–171.

Sweet, Tracy Morrison and Brian Junker. 2011. "Modeling intervention effects on social networks in education research." *Educational Evaluation and Policy Analysis* 30:203–235.

Turkkan, Noyan and T Pham-Gia. 1993. "Computation of the highest posterior density interval in Bayesian analysis." *Journal of statistical computation and simulation* 44(3-4):243–250.

Vance, Eric A, Elizabeth A Archie and Cynthia J Moss. 2009. "Social networks in African elephants." *Computational and mathematical organization theory* 15(4):273–293.

Ward, Michael D, John S Ahlquist and Arturas Rozenas. 2013. "Gravity's rainbow: a dynamic latent space model for the world trade network." *Network Science* 1(01):95–118.

Ward, Michael D and Peter D Hoff. 2007. "Persistent patterns of international commerce." *Journal of Peace Research* 44(2):157–175.

Ward, Michael D, Randolph M Siverson and Xun Cao. 2007. "Disputes, democracies, and dependencies: A reexamination of the Kantian peace." *American Journal of Political Science* 51(3):583–601.

Zhang, Zhaoyang, Honggang Wang, Chonggang Wang and Hua Fang. 2015. "Cluster-based Epidemic Control Through Smartphone-based Body Area Networks." *Parallel and Distributed Systems, IEEE Transactions on* 26(3):681–690.

Online Appendix

Below we replicate the simulation study results figures using the results from our simulation with the recently developed `amen` package. Note that these simulations exclude (1) the Poisson family, which is not available in the `amen` package, and (2) the 50-node networks, which we left out in the interest of lessening the computational burden of the simulation.

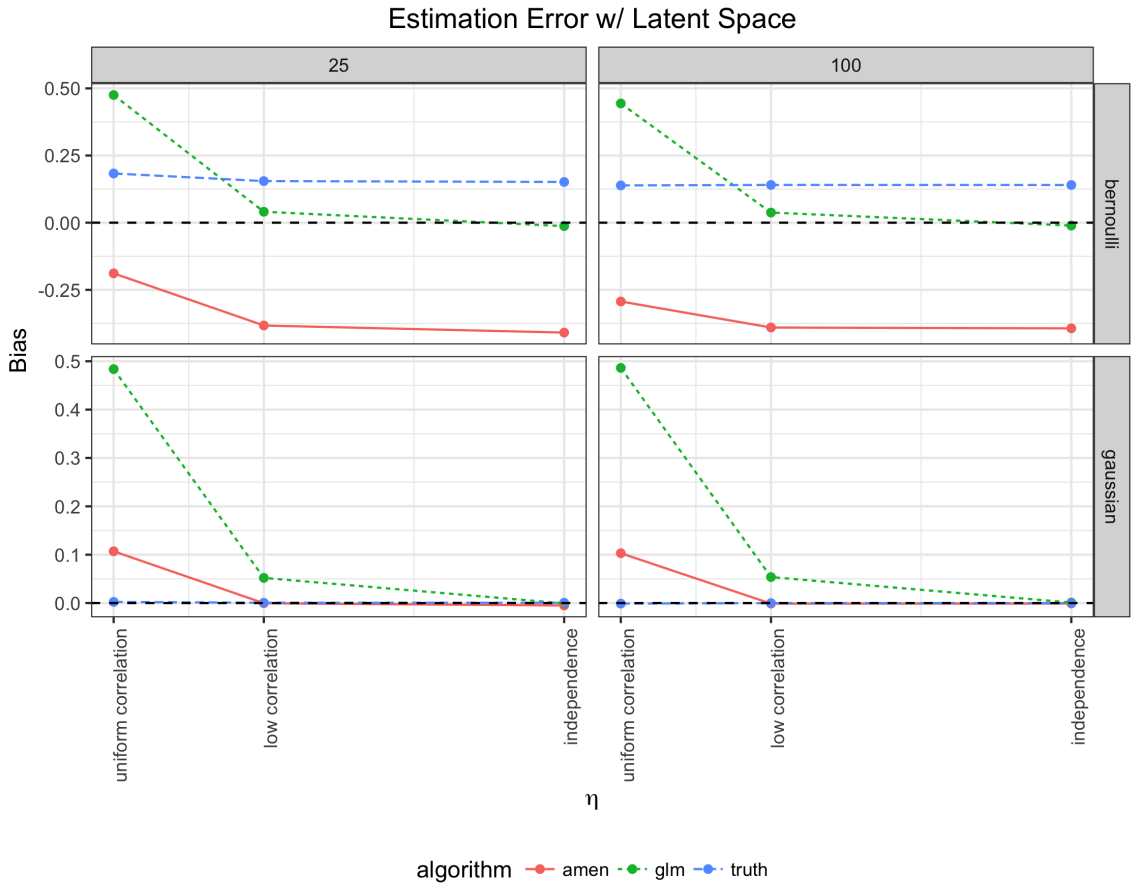

Figure 7: The bias of estimates of the effect of the observed covariate $\mathbf{x}$ when there is an ommitted variable. The x-axis gives the value of the parameter η which controls the degree of dependence between $\mathbf{x}$ and ommitted covariate. Lower values of η indicate higher levels of dependence between the observed and ommitted covariate. The y-axis gives a Monte Carlo estimate of the bias. The number of nodes are indicated in the top panels, while the distributional family of the edges is shown on the right panel. Each panel represents 4 values of η with 1,000 Monte Carlo iterations executed at each point.

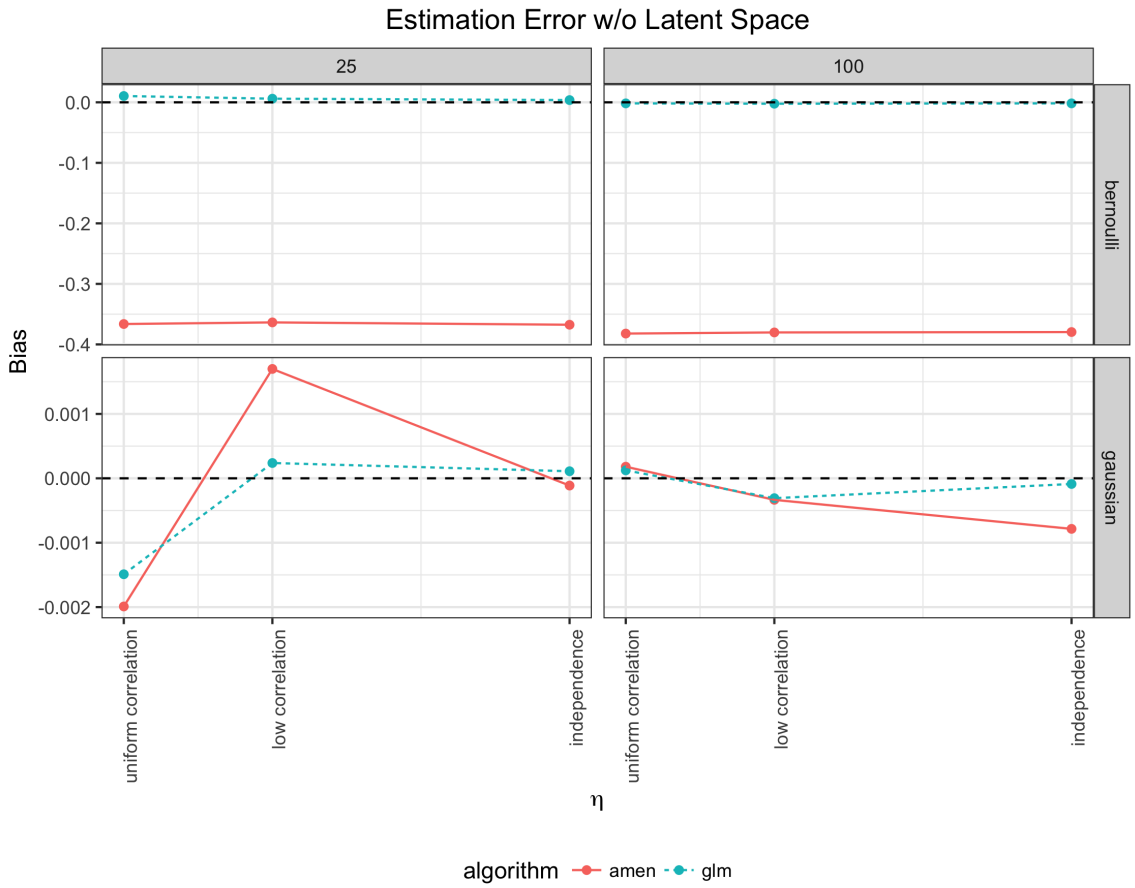

Figure 8: The bias of estimates of the effect of the observed covariate $\mathbf{x}$ when there is no ommitted variable.

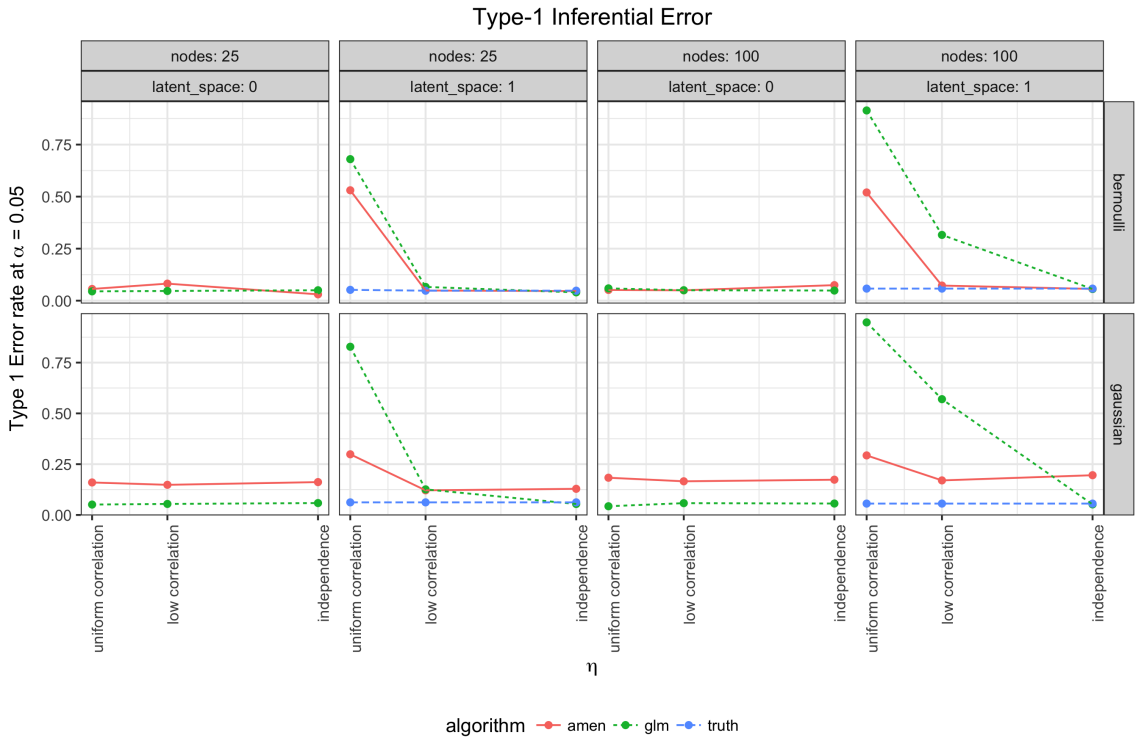

Figure 9: Monte Carlo estimates of the Type-1 error regarding the effect of the observed covariate $\mathbf{x}$ are shown on the y-axis. Here, $\beta = 0$ and the error rate shown in each panel in that row gives 1 minus the probability of a 95% confidence region (for the LSM) or interval (for the GLM) including 0, giving the probability that a true null hypothesis of $\beta = 0$ is falsely rejected.

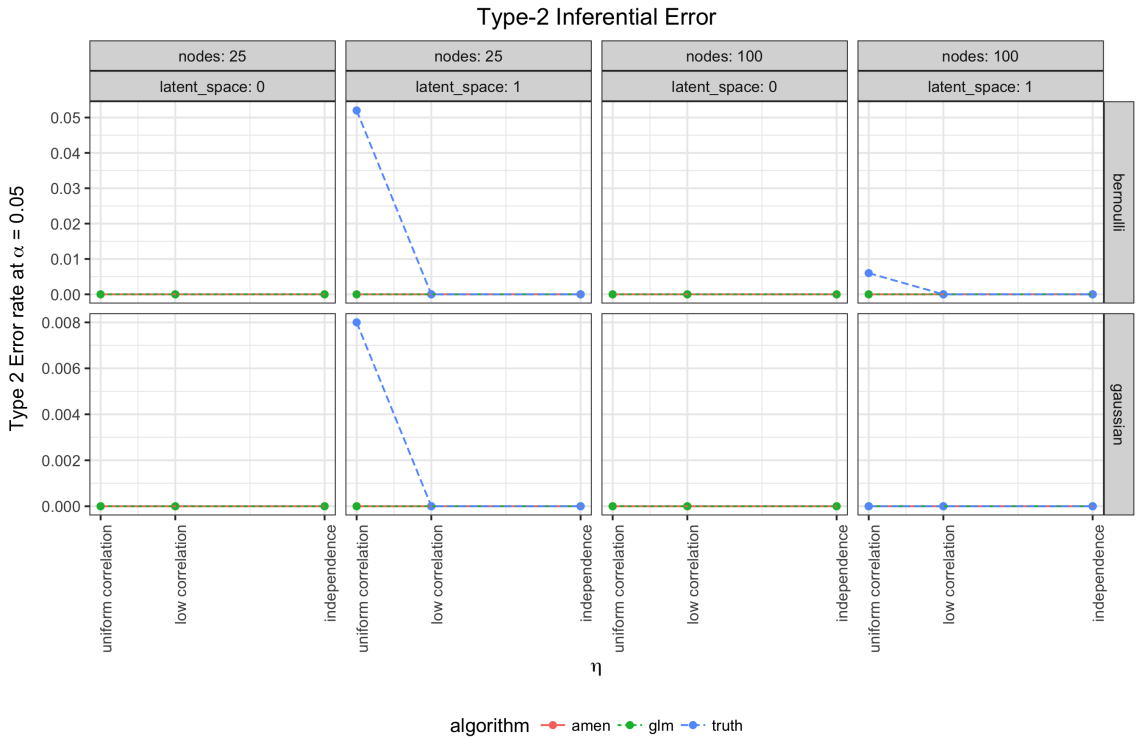

Figure 10: Monte Carlo estimates of the Type-2 error regarding the effect of the observed covariate $\mathbf{x}$ are shown on the $y - axis$. Here, $\beta = 1$, and the error rate shown in each panel in each row gives the probability of the probability/confidence intervals covering 0, giving the probability of accepting a false null hypothesis.

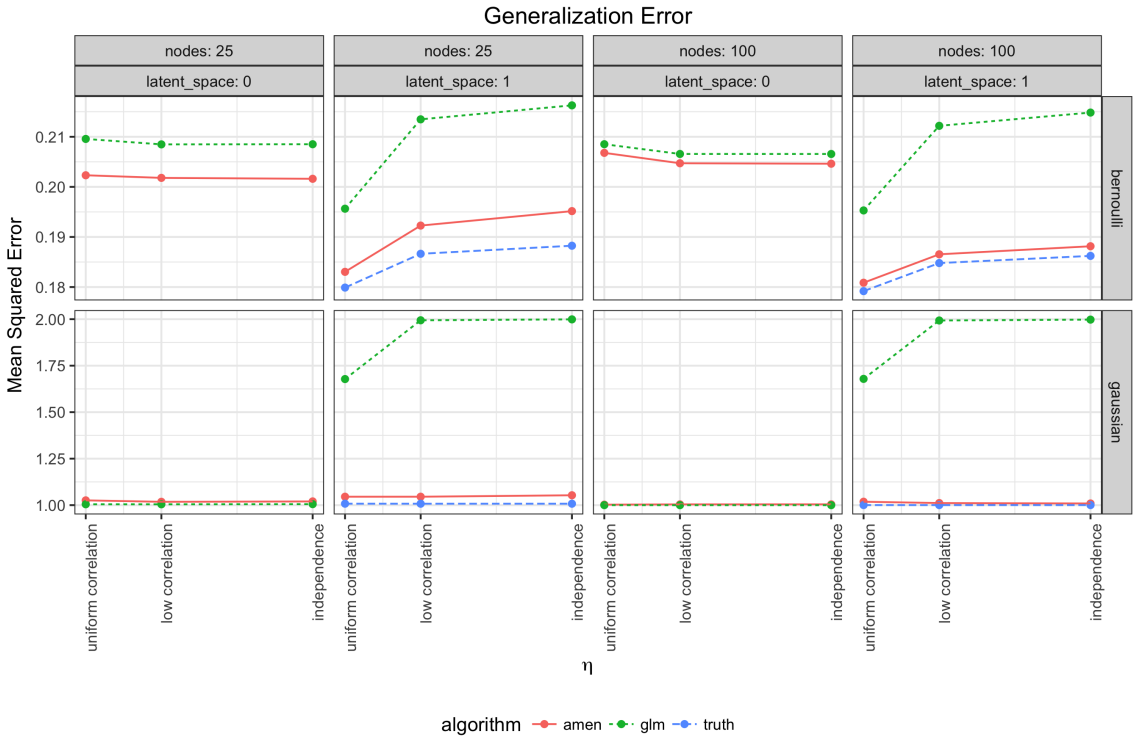

Figure 11: The x-axis shows a Monte Carlo estimate of the mean square error for edge values drawn from the appropriate distribution with the observed covariate $\mathbf{x}$ and the latent positions $\mathbf{z}$ fixed. In the Bernoulli case the Brier score is computed and in the normal case the mean-square-error.