DISCUSSION PAPER SERIES

DP17638

GRAPHICAL MODEL INFERENCE WITH EXTERNAL NETWORK DATA

Jack Jewson, Li Li, Laura Battaglia, Stephen Hansen, David Rossell and Piotr Zwiernik

PUBLIC ECONOMICS AND ASSET PRICING

CEPR

ISSN 0265-8003

GRAPHICAL MODEL INFERENCE WITH EXTERNAL NETWORK DATA

Jack Jewson, Li Li, Laura Battaglia, Stephen Hansen, David Rossell and Piotr Zwiernik

Discussion Paper DP17638
Published 02 November 2022
Submitted 28 October 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

GRAPHICAL MODEL INFERENCE WITH EXTERNAL NETWORK DATA

Abstract

A frequent challenge when using graphical models in applications is that the sample size is limited relative to the number of parameters to be learned. Our motivation stems from applications where one has external data, in the form of networks between variables, that provides valuable information to help improve inference. Specifically, we depict the relation between COVID cases and social and geographical network data, and between stock market returns and economic and policy networks extracted from text data. We propose a graphical LASSO framework where likelihood penalties are guided by the external network data. We also propose a spike-and-slab prior framework that depicts how partial correlations depend on the networks, which helps interpret the fitted graphical model and its relationship to the network. We develop computational schemes and software implementations in R and probabilistic programming languages. Our applications show how incorporating network data can significantly improve interpretation, statistical accuracy, and out-of-sample prediction, in some instances using significantly sparser graphical models than would have otherwise been estimated.

JEL Classification: C11, C55

Keywords: N/A

Jack Jewson - jack.jewson@upf.edu
Universitat Pompeu Fabra and Data Science Center Barcelona

Li Li - liliecon@stu.scu.edu.cn
Sichuan University

Laura Battaglia - laura.battaglia@keble.ox.ac.uk
Department of Statistics, University of Oxford

Stephen Hansen - stephen.hansen@ucl.ac.uk
Department of Economics, University College London and CEPR

David Rossell - david.rossell@upf.edu
Universitat Pompeu Fabra and Data Science Center Barcelona

Piotr Zwiernik - piotr.zwiernik@utoronto.ca
University Of Toronto

Graphical model inference with external network data

Jack Jewson[1,2,*], Li Li[3], Laura Battaglia[2,4], Stephen Hansen[5], David Rossell[1,2], and Piotr Zwiernik[1,2,6]

[1] *Department of Business and Economics, Universitat Pompeu Fabra, Barcelona, Spain*

[2] *Data Science Center, Barcelona School of Economics, Spain*

[3] *School of Economics, Sichuan University, China*

[4] *Department of Statistics, University of Oxford, UK*

[5] *Department of Economics, University College London, UK*

[6] *Department of Statistical Sciences, University of Toronto, Canada*

[*] *Correspondence address jack.jewson@upf.edu*

October 2022

Abstract

A frequent challenge when using graphical models in applications is that the sample size is limited relative to the number of parameters to be learned. Our motivation stems from applications where one has external data, in the form of networks between variables, that provides valuable information to help improve inference. Specifically, we depict the relation between COVID-19 cases and social and geographical network data, and between stock market returns and economic and policy networks extracted from text data. We propose a graphical LASSO framework where likelihood penalties are guided by the external network data. We also propose a spike-and-slab prior framework that depicts how partial correlations depend on the networks, which helps interpret the fitted graphical model and its relationship to the network. We develop computational schemes and software implementations in R and probabilistic programming languages. Our applications show how incorporating network data can significantly improve interpretation, statistical accuracy, and out-of-sample prediction, in some instances using significantly sparser graphical models than would have otherwise been estimated.

Keywords: GLASSO; Bayesian Inference; Spike-and-Slab.

1 Introduction

Graphical models form a convenient framework to describe the dependence among random variables in an interpretable manner. Despite numerous theoretical and methodological advances, an important practical limitation is that they require the estimation of an inherently large number of parameters, which can be challenging unless the sample size is large enough. The main idea behind our work is that there are numerous applications where external data provides valuable information to help guide the graphical model selection and estimation, and hence improve their accuracy. We propose a frequentist and a Bayesian framework to exploit this complementary information.

Although our ideas can be used in many other examples, here we showcase two motivating applications where the external data comes in the form of one or more networks between variables, and we develop the corresponding methodology. First, we study the dependence between COVID-19 infection rates across U.S. counties, and whether said dependence is related to the physical distances between counties and network data measuring Facebook connections. We refer to these network data as external data, in the sense that they do not directly measure COVID-19. To illustrate how such data may be informative, Figure 1 shows the estimated partial correlations between each county pair vs. their geographical distance and the Facebook index. Counties that are highly connected on Facebook have a higher proportion of non-zero, and positive, partial correlations. A similar observation applies to the geographic network, defined such that nearby counties have a stronger network connection. As a second application, we study the dependence of stock market excess returns across companies, incorporating external information given by company similarity scores based on text data about economic and policy risk indicators. The idea is that if two companies are described by similar text data, then their dependence might be stronger. A common feature in both applications is that the sample size n is moderate relative to the number of variables p, and to the $p(p+1)/2$ parameters in the Gaussian graphical model.

To our knowledge, there are no methods to incorporate multiple network-valued external data in undirected graphical models. There has been however active research on incorporating external data in regression. For example, Stingo et al. (2010) proposed a multivariate regression of gene expression on micro-RNA, where the prior probabilities that micro-RNAs have a non-zero coefficient depend on an external biological and structural similarity scores. Along similar lines, Stingo et al. (2011) incorporated pathway information into regression models for gene expression, Quintana and Conti (2013) proposed a Bayesian variable selection framework where prior inclusion probabilities depend

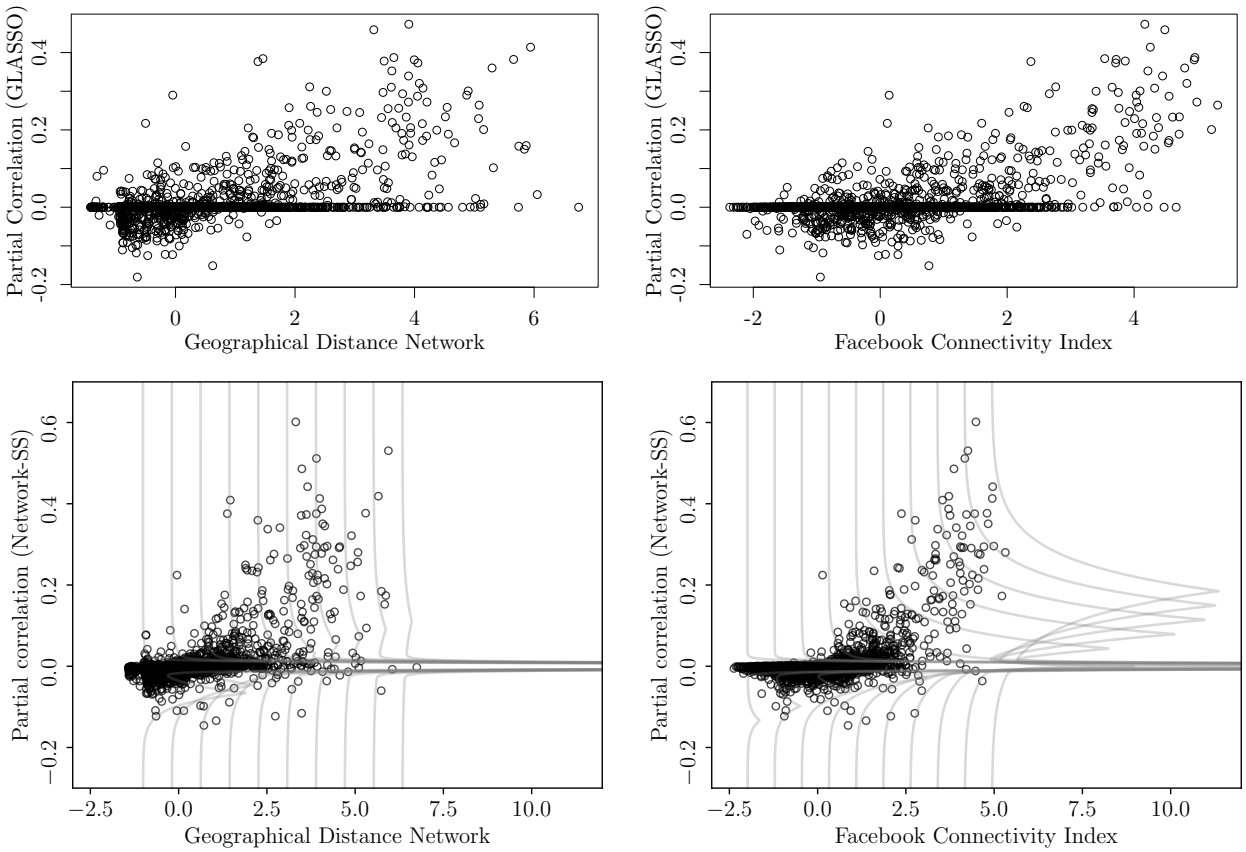


Figure 1: Residual partial correlations in COVID-19 infections (adjusted for covariates) across counties vs Geographical Distance Network defined as $1/\log(Geodistance)$ (left) and log-Facebook Connectivity Index (right). Top panel: partial correlations estimated with graphical LASSO, with penalization parameter set via BIC. Bottom panel: fitted spike-and-slab distributions and fitted partial correlations estimated with network graphical spike-and-slab LASSO.

on meta-covariates, with applications in genomics, Cassese et al. (2014) a multivariate regression of gene expression versus copy number variations that incorporates their physical distance in the genome, and Chiang et al. (2017) a brain activity vector auto-regression that incorporates external brain information. Guha and Rodriguez (2020) proposed a Bayesian shrinkage prior to regress the mean of a univariate outcome on network-valued covariates, encouraging similar regression coefficients for covariates that are connected in the network. Chen et al. (2021) predicted disease outcomes given single nucleotide polymorphisms (SNPs), where the LASSO regularization parameter for each SNP

depends on several functional annotation categories. The main difference between these previous works and ours is that we incorporate external network data to model the (inverse) covariance rather than the mean, i.e. we use graphical models to study the dependence structure. Also, we allow the external data to inform not only prior inclusion probabilities or overall regularization but also the location and variance of non-zero parameters. This can be important in applications, where one may want to shrink estimates to values other than zero, e.g. in Figure 1 large Facebook connectivity is associated with positive partial correlations.

It should also be noted that there exist methods to incorporate external data into covariance models. Azose and Raftery (2018) proposed an estimation procedure for correlations that uses Laplace priors and can capture different prior beliefs on each correlation. Although not done by the authors, one could structure such prior beliefs using external data. Schiavon et al. (2021) proposed a factor model where variable loadings are allowed to depend on meta-covariates. Relative to these authors we focus on the graphical model, i.e. on describing partial rather than marginal correlations, and propose more flexible parametric forms to incorporate the network data.

Our main contribution is developing two frameworks to incorporate external network data into Gaussian graphical model selection and parameter estimation, and applying them to the COVID-19 and stock market motivating applications. The first framework is a hierarchical extension of the graphical LASSO (GLASSO) (Friedman et al. (2008); Yuan and Lin (2007), see also Wang (2012) for a discussion of Bayesian counterparts). The construction resembles the GLASSO priors of Khondker et al. (2013), where each precision matrix entry has a different regularization parameter, with the important difference that we allow the latter to depend on external network data. A practically-appealing feature of this first framework is that, by using tailored optimization algorithms that build on the GOLAZO algorithm of Lauritzen and Zwiernik (2020), the computational cost is similar to that of a standard GLASSO problem. A limitation however is that, as mentioned, in our applications the external data appears to be not only informative about whether a parameter is zero but also about its sign. To address this, our second framework uses a spike-and-slab prior, with the novel feature that the slab's location and variance depend on the external network data. To ensure its practical applicability we developed a software implementation in the probabilistic programming languages `Stan` (Carpenter et al., 2017) and `NumPyro` (Bingham et al., 2019; Phan et al., 2019). The latter capitalizes on efficient automatic differentiation and GPUs to help boost the computational speed. Similarly, the first framework is implemented in *R*.

The paper proceeds as follows. Section 2 discusses our motivating applications in more detail.

Section 3 reviews the GLASSO, introduces our network-adjusted extension and its Bayesian analogue. Section 4 discusses our computational strategy for learning the graphical model and hyper-parameters that depict its association with the external network data. Section 5 investigates the performance of our model in some simulations and Section 6 applies our model to our COVID-19 and stock market datasets. Both examples demonstrate that the network data is informative, the Facebook network being particularly informative about the structure of COVID-19 cases and the Economic risks network about stock-market dependence. Code to implement all of our experiments and data pre-processing is available at https://github.com/llaurabat91/graphical-models-external-networks.

2 Motivating applications

2.1 Dependence in COVID-19 infections versus Facebook and geographical networks

Studying the evolution of pandemics such as COVID-19 is of great importance for health, economic and societal reasons. The emphasis of such studies is typically either to forecast infections or to understand how expected infections are related to various factors (e.g. health measures, temperature). Here we consider a further important aspect that is often neglected: understanding how the disease co-evolves across (possibly distant) geographical units, and what factors may drive such co-evolution. For example, if several counties were expected to simultaneously exhibit higher-than-expected infection rates, health and economic authorities might need to plan resources accordingly. Further, identifying factors that are related to the co-evolution (e.g. the Facebook index) may suggest strategies to limit such coordinated growth (e.g. targeted social media campaigns).

To study COVID-19 co-evolution across U.S. counties, we downloaded weekly infection rates from CSSE (2020a) for the period 22 January 2020 to 30 November 2021 (97 weeks total) for the 99 most-populated counties. We also obtained data on covariates that may be associated with the disease's evolution, such as temperature, population density, vaccination rates and an index measuring the stringency of pandemic measures (CSSE (2020c); Bureau (2020); CSSE (2020d); CSSE (2020b)). We defined the outcome of interest as the county log-infection rates, i.e. log infections relative to the county's population. Our interest is in studying the disease co-evolution *after* accounting for factors driving the mean structure. To this end, we fitted a linear regression model that included temperature, vaccination rates, the stringency of pandemic measures, a weekly fixed effect term estimating the mean infections across all counties in that particular week, and a first-order auto-regressive term measuring

the infection rate in the previous week. See Section A.2 and the supplementary code for the data collection, pre-processing, and residual checks assessing the linearity and normality assumptions, and that higher-order auto-regressive terms are not needed.

Although the model explained most of the variance in infection rates (adjusted R^2 coefficient 0.942), certain counties were systematically above or below the model predictions. Motivated by Kuchler et al. (2021), who found that *marginal* correlations between county infections rates were related to their Facebook connectivity index, we wish to study how *partial* correlations depend on the Facebook index. Said index defines a network of counties, measuring the strength of the connection between every pair of counties. We also consider a second network defined by their geographical distance (see Section 6.1), given that nearby counties may be more likely to co-evolve. We believe that partial correlations are a more appealing measure of disease co-evolution than marginal correlations. For example, suppose that infections in County A drive those of County B, which in turn drive those of County C, then all three counties would have non-zero marginal correlation. In contrast, the partial correlation between counties A and C would be zero, suggesting there is no direct link between them.

As discussed, Gaussian graphical models are a natural strategy to estimate partial correlations. Here we develop a framework to fit Gaussian graphical models that can incorporate the external information provided by the Facebook and geographical networks, and neglect said information when not needed. The Facebook and geographical network data should help improve the graphical model estimation, e.g. regularize to a lesser degree the partial correlations for county pairs that are highly-connected in Facebook and *vice versa*. This desideratum led us to develop a network-regularized graphical LASSO framework, see Section 3.

An important observation stemming from Figure 1 is that counties that are highly connected on Facebook have a higher proportion of non-zero, and positive, partial correlations. A similar observation applies to geographic distances. Hence one wishes not only to regularize to a lesser extent county pairs with a strong Facebook connection but also to describe how the average non-zero partial correlation depends on Facebook (or geographical) connectivity. This desideratum led us to develop a network-regularized spike-and-slab framework, where the slab location and probability are driven by the network, see Section 3.

2.2 Dependence in stock market returns versus text data

A longstanding insight in finance is that the dependence structure among assets informs optimal portfolios (Markowitz, 1952). In particular, the precision matrix determines the weights across assets that

minimize a portfolio's standard deviation. Bringing optimal portfolio theory to data requires estimating high-dimensional covariance/precision matrices, which is an important barrier to its practical application (Elton and Gruber, 1973). A variety of approaches have been used to tackle the problem including, recently, GLASSO (Goto and Xu, 2015). Moreover, Senneret et al. (2016) compare several approaches for the empirical implementation of the theory and find GLASSO to be competitive with other approaches in the financial econometrics literature. However, relational data for stocks has never—to the best of our knowledge—been incorporated into the estimation of the precision matrix.

The relational data we use is built from the unstructured text of the *Risk Factors* (*RF*) section of publicly traded firms' annual 10-K filings to the US Securities and Exchange Commission. *RF* texts provide an exhaustive description of sources of future earnings risk that firms face. There is an incentive for full disclosure because investors can take legal action against firms that withhold relevant information that if revealed would have prevented financial losses. A natural assumption is that firms that face similar risks have more dependent stock returns. For example, firms that both mention risks arising from a rise in oil prices will tend to co-move when oil prices move. Consistent with this intuition, Hanley and Hoberg (2019) regress the covariance of excess returns between pairs of financial firms on a measure of *RF* text overlap and show a positive relationship in the lead-up to the global financial crisis in 2008. More recently, Davis et al. (2020) show that firms with similar *RF* texts reacted similarly to the arrival of COVID-19.

To implement the idea that *RF* text overlap is useful for precision matrix estimation, we first sample 200 American firms at random from the population of firms described in Davis et al. (2020).[1] For each trading day in 2019 (252 in total), we obtain firm-level stock prices from the CRSP database, and then construct daily excess returns using the Fama-French three-factor model.[2] To measure *RF* text overlap, we first use the construction of Baker et al. (2019), who define dozens of risk exposures at the firm level that encompass both *economic* and *policy* risks. Examples of the former include 'exchange rates' and 'macroeconomic news', and of the latter include 'taxes' and 'trade policy'. For each exposure, the authors define a term set meant to capture the concept, e.g. the 'food and

[1] These firms are those for which one can retrieve a 10-K filing; are sufficiently large; and have available financial records in the Compustat database. We choose 200 firms for computational reasons, but in principle, our method is scalable to more given sufficient computing resources (e.g. multiple GPUs). Moreover, optimal portfolio allocation does not require the continuous updating of precision matrix estimates and so infrequent, but large-scale, estimation is adequate for the problem.

[2] To do so, we regress each firm's daily return on the variables contained in the daily, three Fama/French factors file downloaded from Kenneth French's Data Library website.

drug policy' risk is captured by the term set {prescription drug act, drug policy, food and drug administration, fda}. For each firm's 2019 *RF* text, we tabulate the number of appearances of terms for each category and scale it by the number of words in the text. Finally, we compute the pairwise correlation between firms based on the vectors of economic risk categories and, separately, based on the vectors of policy risk categories. This creates two separate network measures that may inform precision matrix estimation.

The particular representation of *RF* texts we adopt is not definitive and is meant for illustration purposes. Our goal is to establish that text-based relational data is useful in principle to estimate the dependence structure in equity returns, which we establish below is the case in this setting. The optimal representation of text for this task is an open question that we do not consider here. Still, as we show below, separately controlling for economic and policy risks yields important insights regarding whether government policy generates return co-movement above and beyond that generated by firm fundamentals.

See Section A.3 and the supplementary code for the data collection, pre-processing, linear model fit, and residual checks assessing our model assumptions.

3 Model

This section reviews the definition of an undirected Gaussian graphical model on p variables and proposes two new model-fitting strategies to incorporate external network data. We first set notation. Let $y_i \in \mathbb{R}^p$ be the outcomes for individuals $i = 1, \ldots, n$ (e.g. log-infection rates in p counties at week i) and $x_i \in \mathbb{R}^d$ a covariate vector (e.g. temperature, percentage of fully vaccinated individuals in week i), which might be omitted in a simplest case with no covariates. We assume that $y_i \sim \mathcal{N}_p\left(Bx_i, \Theta^{-1}\right)$ independently across $i = 1, \ldots, n$, where B is a $p \times d$ regression coefficients matrix and Θ a $p \times p$ positive-definite precision (or inverse covariance) matrix. To ensure that the independence assumption across i is tenable, one may include lagged versions of y_i into the covariates x_i. For simplicity in our applications, we start by subtracting the estimated mean $\hat{B}x_i$ from y_i, where $\hat{B}$ is the least-squares estimator, and in the sequel we assume the outcomes to have zero mean, i.e. $y_i \sim \mathcal{N}_p(0, \Theta^{-1})$.

Importantly, in our framework, one also observes external data in the form of $Q \geq 1$ networks between variables. These are denoted by $p \times p$ symmetric matrices $A^{(1)}, \ldots, A^{(Q)}$, where $a^{(q)}_{jk}$ measures strength of the connection between variables (j, k). For example, $a^{(1)}_{jk}$ may be the geographical distance between counties (j, k) and $a^{(2)}_{jk}$ their Facebook connection index.

3.1 Gaussian graphical models

A convenient property of modelling $y_i \sim \mathcal{N}_p\left(0, \Theta^{-1}\right)$ is that conditional independence statements can be drawn from the graph defined by the non-zero elements of Θ. Specifically, (y_{ij}, y_{ik}) are independent given the remaining elements in y_i if and only if $\Theta_{jk} = 0$. Hence, by determining what elements in Θ are zero one learns about conditional independence. As argued earlier, in our applications we consider using partial correlations as a measure of association. In the Gaussian model, conditional independence is equivalent to zero partial correlation. We denote negative partial correlations by

$$\rho_{jk} := \frac{\Theta_{jk}}{\sqrt{\Theta_{jj}\Theta_{kk}}} = -\text{corr}(y_{ij}, y_{ik} \mid y_{i\{1,\ldots,p\}\setminus\{j,k\}}). \tag{1}$$

A popular approach to estimate Θ, and hence the partial correlations, is the Graphical LASSO (GLASSO) (Friedman et al., 2008; Yuan and Lin, 2007). GLASSO produces an estimate of Θ that contains zeroes by maximising the Gaussian log-likelihood with a LASSO penalty. Specifically,

$$\underset{\Theta \in \mathcal{S}_+^p}{\arg\max} \quad \log\det(\Theta) - \text{tr}(S\Theta) - \lambda \sum_{i \neq j} |\Theta_{ij}|, \tag{2}$$

where $\mathcal{S}_+^p$ is the set of non-negative definite matrices, $\lambda > 0$ is a regularization parameter, $\text{tr}(\cdot)$ denotes the matrix trace and S is the empirical covariance matrix of $(y_1, \ldots, y_n)$. Alternatives to (2) include the adaptive graphical LASSO, SCAD and MCP (Fan et al., 2009; Wang et al., 2016), which were proposed to reduce bias in the estimation of large entries in Θ. We focus on (2) however due to its practical appeal of defining a concave problem for which one may establish efficient optimisation methods.

The hyperparameter λ is a critical quantity that determines the level of sparsity in the estimated partial correlations. Two popular strategies to set λ are using cross-validation (Friedman et al., 2008) and information criteria such as the Bayesian information criterion (BIC) (Schwarz, 1978). An issue with cross-validation is that it does not lead to consistent model selection even in simple linear regression models, i.e. the probability of recovering the set of non-zero parameters does not converge to 1 as $n \to \infty$. In contrast, the BIC and certain other generalised Bayesian information criteria are model selection consistent. For further details we refer the reader to Foygel and Drton (2010); Zhang et al. (2010); Wang and Zhu (2011); Fan and Tang (2013). Based on these observations, here we use information criteria. Viewing the estimate $\hat{\Theta}(\lambda)$ as a function of λ, the BIC selects the non-negative λ minimising

$$\text{BIC}(\lambda) \;=\; -2\ell_n(\hat{\Theta}(\lambda)) + \left|\mathbf{E}(\hat{\Theta}(\lambda))\right| \cdot \log n, \tag{3}$$

where $\ell_n(\hat{\Theta})$ is the Gaussian log-likelihood function and $|\mathbf{E}(\hat{\Theta}(\lambda))|$ counts the number of edges in the graph associated with $\hat{\Theta}(\lambda)$. An alternative to the BIC is the Extended BIC (EBIC) (Chen and Chen, 2008) which contains an additional penalty to the BIC for model complexity. As a sensitivity check, we provide results using the EBIC to select λ in Sections A.1.6, A.2.8 and A.3.7. In our examples the EBIC was overly conservative in selecting edges, which resulted in high mean-squared-error. Finally, we note that there are alternative approaches to choosing λ, see Kuismin and Sillanpää (2021), but they require more extensive computations that become prohibitive in our setting.

On the Bayesian side, Gan et al. (2018) proposed a spike-and-slab framework designed to bypass computational difficulties in standard conjugate frameworks, while attaining good model selection properties. We defer further discussion to Section 3.3, where we develop a network-driven extension.

3.2 Network graphical LASSO

The main idea behind our network graphical LASSO framework is to estimate $\Theta \in \mathcal{S}_+^p$ maximising

$$\log\det(\Theta) - \operatorname{tr}(S\Theta) - \sum_{j\neq k} \lambda_{jk}|\Theta_{jk}|, \tag{4}$$

where the regularisation $\lambda_{jk} = \lambda_{jk}(A^{(1)}, \ldots, A^{(Q)})$ depends on the network data. That is, each Θ_{jk} gets a potentially different penalty parameter λ_{jk}, which is a function of $A^{(1)}, \ldots, A^{(Q)}$. To simplify notation, we omit the dependence on $A^{(1)}, \ldots, A^{(Q)}$ and simply use λ_{jk}, and let $A = (A^{(1)}, \ldots, A^{(Q)})$. For convenience we parameterise the penalties in terms of a scaled version of $A^{(q)}$ that is centered to have zero sample mean and unit sample variance, and which we denote by $\bar{A}^{(q)}$.

Consider first a simple example with only one external matrix $A^{(1)}$ with binary entries $a_{jk}^{(1)} \in \{0, 1\}$ for $k \neq j$. One could then use $A^{(1)}$ to structure the regularization by considering

$$\lambda_{jk} = \begin{cases} \lambda_0 & \text{if } a_{jk}^{(1)} = 0 \\ \lambda_1 & \text{if } a_{jk}^{(1)} = 1 \end{cases}$$

with $\lambda_0, \lambda_1 \geq 0$. More generally when one has multiple potentially non-binary networks, as is the case for the COVID-19 and stock market data introduced in Section 2, we consider the simple extension

$$\lambda_{jk} = \exp\left\{\beta_0 + \sum_{q=1}^{Q} \beta_q a_{jk}^{(q)}\right\} \tag{5}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_Q) \in \mathbb{R}^{Q+1}$ are regularisation hyperparameters. To set β we choose the value optimising the BIC in (3), which is now a function of β. Note that GLASSO is nested within our framework and is recovered when $\beta_1 = \cdots = \beta_Q = 0$ and $\lambda = \exp(\beta_0)$.

One could of course consider alternative parameterisations to (5), e.g. let λ_{jk} depend non-parametrically on the network data. However, (5) requires fewer hyper-parameters than a non-parametric treatment and is easy to interpret: the log-regularisation depends linearly on the networks. Further, a model-checking exercise suggested that (5) is a reasonable parameterisation for our two motivating applications. Said model-checking is best understood by adopting a Bayesian interpretation. The penalised estimator associated to (5) is equivalent to the posterior mode under independent Laplace priors (Wang, 2012) with scale parameter $1/\lambda_{jk}$, that is

$$\pi(\Theta \mid A, \beta) \propto \prod_{j>k} \frac{\lambda_{jk}}{2} \exp\{-\lambda_{jk}|\theta_{jk}|\} \, \mathrm{I}(\Theta \succ 0), \tag{6}$$

where $\mathrm{I}(\Theta \succ 0)$ is an indicator for Θ being positive-definite, λ_{jk} is as in (5) and β are now prior parameters. The Bayesian interpretation is that the λ_{jk}'s arise from a Laplace random effects distribution with parameter β. The *a priori* expected value of θ_{jk} is 0, which induces sparsity, and the prior variance is

$$\mathrm{Var}\left[\theta_{jk} \mid \beta, A\right] = \mathbb{E}\left[\theta_{jk}^2 \mid \beta, A\right] = \frac{2}{\lambda_{jk}^2}. \tag{7}$$

Therefore (5) assumes that the log-variance of the partial covariances θ_{jk} depends linearly on the network data

$$\log \mathbb{E}\left[\theta_{jk}^2 \mid \beta, A\right] = \log(2) - 2\left(\beta_0 + \beta_1 \bar{a}_{jk}^{(1)} + \ldots + \beta_Q \bar{a}_{jk}^{(Q)}\right). \tag{8}$$

Provided one has an initial estimate of the left-hand side of (8), which in our examples we derived from standard GLASSO estimates of θ_{jk}, one may check whether its relation to the network data is roughly linear. Such a check motivated taking the logarithm of the raw distances and Facebook connectivities to define our networks for the COVID-19 data. See Supplementary Sections A.2.5 and A.3.5 for further details.

3.3 Network graphical spike-and-slab LASSO

The network graphical LASSO in Section 3.2 provides sparse point estimates of partial correlations and, via its Bayesian interpretation, describes how their variance depends on the network data. In applications, however, it is also interesting to describe how the proportion of non-zero partial correlations and their mean depend on the network. For example, in the COVID-19 data both the probability that two counties are conditionally dependent and the mean partial correlation grow as their Facebook connection grows (Figure 1), and similarly for the stock market data (Figure A.12). To address this

issue, we developed a spike-and-slab framework that builds on the regression setting of Rockova and George (2014) and the graphical spike-and-slab of Gan et al. (2018). The main novelty is that both the slab prior probability and its parameters depend on network data. In particular, the slab need not be centered at zero, a feature that is novel—to our knowledge—and has some independent interest.

We parameterise Θ in terms of partial correlations ρ_{jk} in (1), which facilitates interpretation and ensures that the posterior mode is invariant to scale transformations. By scale invariance we refer to the property that the estimated partial correlations remain the same regardless of whether one applies a scale transformation to the input data or not. See Carter et al. (2021) for a detailed discussion, showing that applying a GLASSO penalty on the precision entries θ_{jk} is not scale-invariant, whereas applying it to the partial correlations ρ_{jk} is. We set a prior density $\pi(\text{diag}(\Theta), \rho) = \pi(\text{diag}(\Theta))\pi(\rho)$, where $\sqrt{\Theta_{ii}} \sim \mathcal{IG}(a, b)$ with $a = 0.01$ and $b = 0.01$ reflecting an uninformative prior on the diagonal elements of Θ, and

$$\pi(\rho) \propto \mathrm{I}(\rho \succ 0) \prod_{j>k} (1 - w_{jk})\mathrm{DE}(\rho_{jk}; 0, s_0) + w_{jk}\mathrm{DE}\left(\rho_{jk}; \eta_0^T a_{jk}, s_{jk}\right) \tag{9}$$
$$w_{jk} = \left(1 + e^{-\eta_2^T a_{jk}}\right)^{-1}, \quad s_{jk} = s_0(1 + \exp\left\{-\eta_1^T a_{jk}\right\})$$

The spike is a double-exponential with zero mean and small scale s_0 meant to capture partial correlations that are practically zero. We model the slab prior probability w_{jk} via a logistic regression on the network data $a_{jk} = (1, a_{jk}^{(1)}, \ldots, a_{jk}^{(Q)})$. Note that because of the positive-definiteness constraint the marginal prior $\pi(\rho_{jk})$ could be fairly different from the spike-and-slab in (9), and then w_{jk} could not be interpreted as the marginal slab prior probability. To address this issue we elicit prior parameters such that the positive-definiteness indicator $\mathrm{I}(\rho \succ 0)$ is satisfied with high prior probability, see below. The slab's mean $\eta_0^T a_{jk}$ depends linearly on the network data a_{jk}, and its variance s_{jk} is larger than that of the spike by a factor that also depends on a_{jk}. Finally, $\eta_0, \eta_1, \eta_2 \in \mathbb{R}^{Q+1}$ are prior hyper-parameters. We remark that, although we adopt a Bayesian treatment, one could also take $-\log \pi(\rho)$ as a likelihood penalty and devise suitable optimisation algorithms.

A popular strategy to set prior hyper-parameters such as (η_0, η_1, η_2) in (9) is an empirical Bayes framework where one maximises the marginal likelihood. Such a framework allows us to do inference on the η's themselves through the marginal posterior $\pi(\eta \mid y)$ and inference for Θ through the empirical Bayes posterior

$$\pi(\Theta|y, \hat{\eta}) = f(y|\Theta)\pi(\Theta|\hat{\eta}),$$

where $\hat{\eta}$ maximises the marginal posterior of η given the data

$$\hat{\eta} := \arg\max_{\eta} \pi(\eta \mid y) = \arg\max_{\eta} \int \pi(\Theta, \eta | y) d\Theta = \arg\max_{\eta} \int f(y|\Theta)\pi(\Theta|\eta)\pi(\eta) d\Theta.$$

One could consider using the joint posterior $\pi(\Theta, \eta|y)$ for inference on Θ and η, but we found the empirical Bayes approach to perform better in our experiments. Insofar as one of our goals is to learn the appropriate degree of sparsity from data in social science applications, this approach complements that of Giannone et al. (2021) which uses a spike-and-slab prior to infer the relevant predictors for mean outcomes in common economic and financial datasets.

We next discuss our default elicitation for the prior $\pi(\rho)$. The guiding principle was to set a minimally-informative prior, so that data may suitably update prior beliefs, while encouraging sparse solutions and preserving the interpretability of (9). Briefly, we set independent Gaussian priors on (η_0, η_1, η_2). The prior on η_2 was set such that the prior mean number of edges is proportional to p, which induces sparsity, and the prior sample size can be thought of as 1, in analogy to a Beta prior in a Binomial experiment. The prior on η_1 was set such that the prior mode of the slab's scale is $10s_0$ and greater than $3s_0$ with probability 0.99, i.e. the slab captures partial correlations of a larger magnitude than the slab. Finally, the prior on η_0 was set such that the slab has zero prior mean and such that sampling entries of ρ independently from the double-exponential priors in (9) returns a positive-definite matrix with 0.95 prior probability. This ensures that $\pi(\rho)$ is similar to its unconstrained version where one drops the positive-definiteness indicator, as otherwise w_{jk} cannot be interpreted as the marginal slab probability.

To assess the impact of these default prior choices, it is useful to display the implied prior marginal distribution on the ρ_{jk}'s. Figure 2 shows that in both the COVID-19 and stock market applications most of the prior probability is contained in $\rho_{jk} \in (-0.5, 0.5)$, which seems a sensible prior interval. The prior concentrates significant mass around 0, which induces shrinkage, but also features thick tails, which favors capturing truly non-zero ρ_{jk}'s. Indeed, the corresponding posteriors (Figure 2, bottom panels) set significant mass away from zero, suggesting that the prior shrinkage towards 0 was not excessive. We refer the reader to Section A.1.3 for further details, where we also list the hyper-parameter values used in our examples. Our supplementary material contains R code to implement our prior elicitation.

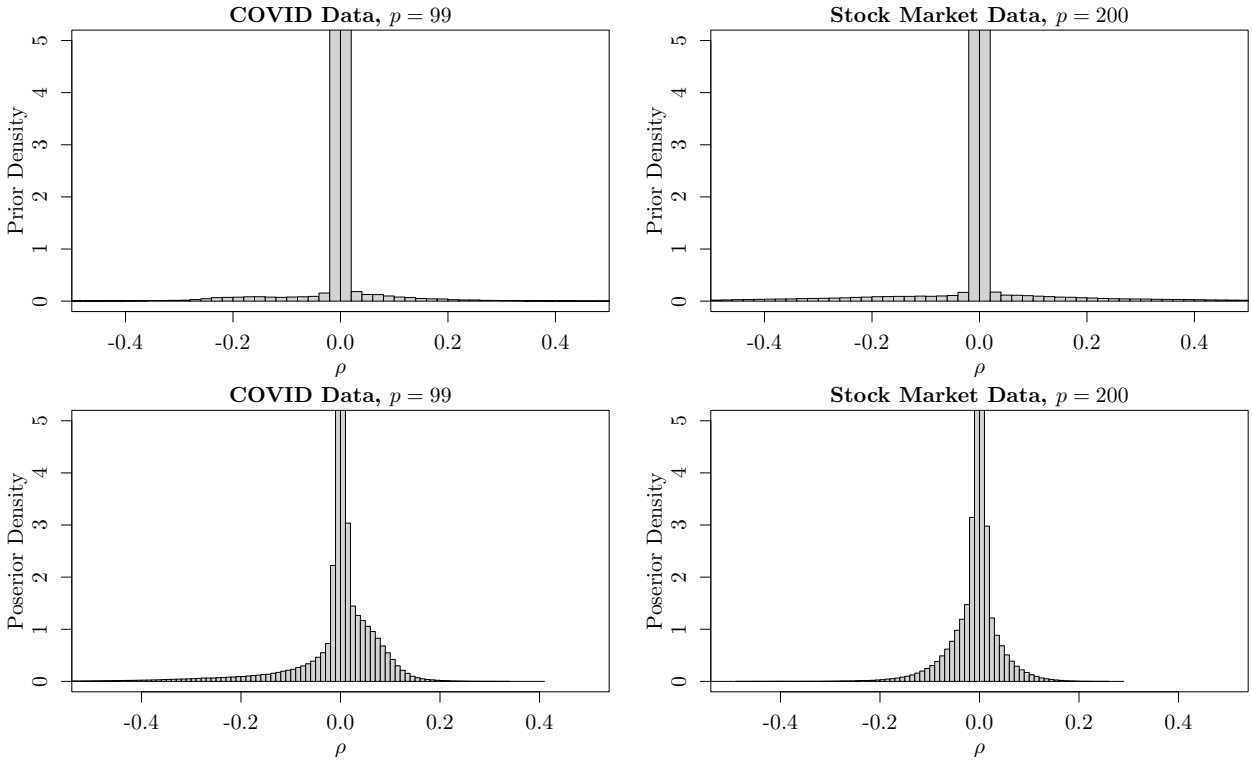


Figure 2: Elicited prior distribution and posterior distribution for ρ_{jk}, $j = 1, \ldots, p$, $k < j$ for the COVID-19 data $p = 99$ and stock market data $p = 200$.

3.4 Beyond Gaussian data

In certain applications such as our stock market example, data exhibit non-Gaussian behavior such as thick tails and asymmetries, even after taking logarithmic or similar transforms (see the normality checks in Section A.3.4). To address this issue in this application we used a non-paranormal model, which can accommodate said departures from normality. The distribution of $y_i = (y_{i1}, \ldots, y_{ip})$ is non-paranormal if there exist strictly increasing functions $f_j : \mathbb{R} \to \mathbb{R}$ for $j = 1, \ldots, p$ such that the vector $f(y_i) := (f_1(y_{i1}), \ldots, f_p(y_{ip}))$ is Gaussian. Such a non-paranormal model may be estimated by first obtaining an estimate $\hat{f}$ from the data, for which we used the R package `huge`, and subsequently applying our methodology to the transformed data $\hat{f}(y_i)$.

An interesting property of the non-paranormal family is that the graphical model can be interpreted as in the Gaussian case. The partial correlation between the transformed $f_j(y_{ij})$ and $f_k(y_{ik})$ is zero if and only if (y_{ij}, y_{ik}) are conditionally independent. Partial correlations retain an interest-

ing interpretation in the trans-elliptical family: zero partial correlation $\rho_{jk} = 0$ indicates that y_{ij} is linearly independent with any transformation of y_{ik} (Rossell and Zwiernik, 2021).

4 Computation and inference

4.1 Network GLASSO

We first describe how to optimise (4) for a fixed β, and subsequently how to estimate $\hat{\beta}$. The main idea is that, since $\lambda_{jk} = \lambda_{jk}(A, \beta)$ are fixed for a fixed β, the network GLASSO objective in (4) is a special case of the GOLAZO class of models in Lauritzen and Zwiernik (2020).

Motivated by the desire to penalise positive and negative partial correlations differently, GOLAZO algorithms consider Gaussian graphical models with likelihood penalties of the form

$$\sum_{j=1}^{p} \sum_{k \neq j} \max\left\{L_{jk}\rho_{jk}, U_{jk}\rho_{jk}\right\}, \tag{10}$$

where $-\infty \leq L_{jk} \leq 0 \leq U_{jk} \leq \infty$ are fixed. Noting that $\lambda|x| = \max\{-\lambda x, \lambda x\}$ for positive λ, we see that the penalty in (4) is in the form of (10) with $L_{jk} = -\lambda_{jk}$ and $U_{jk} = \lambda_{jk}$. Penalising the Gaussian log-likelihood with (10) leads to a convex problem that can be efficiently solved using a block-coordinate ascent algorithm similar to that proposed for GLASSO in (Banerjee et al., 2008). An R package is also provided for the GOLAZO optimisation - `https://github.com/pzwiernik/golazo`.

Analogously to (18), we set β by minimising the BIC

$$\hat{\beta}_{\text{BIC}} := \underset{\beta \in \mathbb{R}^{Q+1}}{\arg\min}\ \text{BIC}\left(\lambda(A, \beta)\right) \tag{11}$$

The function $\text{BIC}(\lambda(A, \beta))$ in (11) is a non-concave function of β that exhibits discontinuities, hence the use of gradient optimisation methods is challenging. Instead, we use a grid-search optimisation strategy, akin to that used to set the regularisation parameter in standard GLASSO. The efficiency of the GOLAZO optimisation combined with the implementation details outlined next make this feasible.

First we initialise $\hat{\beta}_0$ (the first entry in $\hat{\beta}$), such that $\hat{\lambda} = \exp(\hat{\beta}_0)$, where $\hat{\lambda}$ maximises the BIC in (18) over a univariate grid. Assuming that all variables in y_i are standardised to unit sample variance, the grid search is facilitated by an analytic upper bound that

$$\hat{\beta}_0 \leq \log\left(\max_{k \neq j}\{|R_{jk}|\}\right) \tag{12}$$

where R is the empirical correlation matrix, see Appendix A.1.1 for the proof. Second, we conduct a grid search on the whole vector β, with the first entry being centered around $\hat{\beta}_0$. The grid search is again facilitated by the existence of bounds, specifically, Appendix A.1.1 shows that one may restrict attention to β such that

$$\max_{j \neq k} \lambda_{jk} = \max_{j \neq k} e^{\beta_0 + \sum_{q=1}^{Q} \beta_q \bar{a}_{jk}^{(q)}} \leq 1 - |R_{jk}|$$

since increasing λ_{jk} beyond this bound will not change $\hat{\Theta}$. Within the grid search, we also use the solution obtained for a particular β as a warm start for subsequent values of β.

4.2 Spike-and-slab

The full parameter of interest is $(\text{diag}(\Theta), \rho, \eta)$, where $\eta = (\eta_0, \eta_1, \eta_2)$ are the hyper-parameters in (9). To approximate their posterior distribution $\pi(\text{diag}(\Theta), \rho, \eta \mid y)$ given the data y we used Hamiltonian Monte Carlo (see Neal (2011) for a review). Specifically, we developed an `R` implementation using the `Stan` software (Carpenter et al., 2017), as well as a Python implementation using the `NumPyro` package (Phan et al., 2019). Sections A.1.2 and A.1.5 describe further implementation details and our code provides both implementations. The purpose of the `R` version is to make our methods available to the ample `R` community, whereas `NumPyro` provides significant computational savings by using improvements in automatic differentiation and enabling the use of GPUs. The savings were substantial, Section A.4 demonstrates that greater than an order of magnitude speed up was possible even in simple experimental settings.

The output of both implementations are N posterior samples $(\text{diag}(\Theta^{(i)}), \rho^{(i)}, \eta^{(i)})$ for $i = 1, \ldots, N$ that can be used to approximate the posterior distribution or suitable summaries such as the marginal posterior mean and standard deviation of any parameter. Of particular interest to us is to estimate the posterior probability for the presence of an edge between any two nodes (j, k), i.e. that the partial correlation ρ_{jk} was generated by the slab in (9). We next discuss how to estimate said posterior probability using the posterior samples.

To ease notation re-write the prior as

$$\pi(\rho_{jk} \mid \eta) = (1 - w_{jk}(\eta))\pi_0(\rho_{jk} \mid \eta) + w_{jk}(\eta)\pi_1(\rho_{jk} \mid \eta) \tag{13}$$

where $\pi_0(\rho_{jk} \mid \eta)$ is the spike prior density, $\pi_1(\rho_{jk} \mid \eta)$ the slab prior density, and $w_{jk}(\eta)$ the slab prior probability. The idea is that any ρ_{jk} generated by the spike takes a near-zero value, i.e. the partial correlation is either truly zero or small enough to be practically irrelevant. Let $z_{jk} = 1$

indicate that ρ_{jk} was generated from the slab and $z_{jk} = 0$ that it was generated from the spike, i.e. $P(z_{jk} = 1 \mid \eta) = w_{jk}$. A measure of evidence in favor of the presence of the edge is the posterior probability

$$P(z_{jk} = 1 \mid y) = \int P(z_{jk} = 1 \mid \rho_{jk}, \eta)\pi(\rho_{jk}, \eta \mid y)d\rho_{jk}d\eta, \tag{14}$$

where from Bayes rule

$$P(z_{jk} = 1 \mid \rho_{jk}, \eta) = \frac{w_{jk}(\eta)\pi_1(\rho_{jk} \mid \eta)}{(1 - w_{jk}(\eta))\pi_0(\rho_{jk} \mid \eta) + w_{jk}(\eta)\pi_1(\rho_{jk} \mid \eta)}. \tag{15}$$

Given B posterior samples from $\pi(\rho, \eta \mid y)$, (14) may be easily estimated by

$$\hat{P}(z_{jk} = 1 \mid y) = \frac{1}{N}\sum_{I=1}^{N} P(z_{jk} = 1 \mid \rho_{jk}^{(I)}, \eta^{(I)}) \tag{16}$$

Our decision rule is to include edge (j, k) whenever $\hat{P}(z_{jk} = 1 \mid y) \geq t$ for some threshold $t \in [0, 1]$. In problems where the goal is to estimate Θ it is customary to use $t = 0.5$, see Barbieri and Berger (2004). In contrast, in structural learning where one seeks to control a Bayesian version of the false discovery rate (the posterior expected false discovery proportion) below some given level α, Müller et al. (2004) showed that the optimal threshold maximising statistical power is to set the largest t such that

$$\frac{1}{|D|}\sum_{(j,k)\in D} \hat{P}(z_{jk} = 0 \mid y) \leq \alpha$$

where D is the set of included edges. In particular, setting $t = 1 - \alpha$ ensures that the posterior expected false discovery proportion is below α.

4.3 Empirical Bayes

The empirical Bayes estimate $\hat{\eta}$ discussed in Section 3.3 requires marginalizing the joint posterior $\pi(\Theta, \eta \mid y)$. This is possible given N posterior samples $(\Theta^{(i)}, \eta^{(i)})$ for $i = 1, \ldots, N$ from the latter, since then by definition $\eta^{(i)}$ are samples from $\pi(\eta \mid y)$. Then one may obtain $\hat{\eta}$ by maximising a kernel density estimate of $\pi(\eta \mid y)$, for example. Given that the accuracy of kernel density estimators degrades as dimensionality grows, in our examples when $\dim(\eta) > 2$ we instead obtain marginal mode estimators $\hat{\eta}_j = \arg\max_{\eta_j} \pi(\eta_j \mid y)$.

5 Simulation study

We conducted a simulation study as proof of principle that, by incorporating external network data, one may improve graphical model inference. To this end, we compared standard GLASSO with the network GLASSO of Section 3.2 and the network graphical spike-and-slab of Section 3.3 in several settings. As discussed in Section 4, the network GLASSO hyper-parameters β are set via the BIC, and the spike-and-slab hyper-parameters η using empirical Bayes. We considered a setting where there is a single binary network A with entries $a_{jk} \in \{0, 1\}$ and considered $p \in \{10, 50\}$ and sample sizes $n \in \{100, 200, 500\}$. We then generated 50 independent datasets where $y_i \sim \mathcal{N}(0, \Theta^{-1})$ independently across $i = 1, \ldots, n$. We set the data-generating Θ to have unit diagonal and most non-zero entries along the main tri-diagonal (Θ_{jk} where $|j-k| = 1$). Specifically, a proportion of 0.95 of the tri-diagonal entries were set to non-zero values uniformly spaced in $[0.2, 0.5]$. Regarding entries outside the main tri-diagonal (i.e. Θ_{jk} where $|j - k| > 1$), a proportion of $0.5/p$ were set to non-zero values uniformly spaced in $[-0.1, 0.1]$ (i.e. the number of edges in the graphical model grows linearly with p).

The primary purpose of the study is two-fold. First, to show that if the network data provide no valuable information about the graphical model structure, then our methods perform similarly to standard GLASSO where the network data is not used. Second, to show that as the network becomes more informative, the inference provided by our methods improves gradually. To measure the degree to which the network data $a_{jk} \in \{0, 1\}$ is informative we count the proportion of overlaps where $a_{jk} = \mathrm{I}(\Theta_{jk} \neq 0)$, i.e. the presence/absence of an edge in the network A matches that of an edge in Θ. We considered the following settings:

1. Independent network: The tri-diagonal elements of A are set such that half of them are 1 and half of them 0, equally for the elements outside the main tri-diagonal, half of these are 1 and half of these are 0. This led to a 0.533 and 0.502 proportion of edges that agree between A and $\mathrm{I}(\Theta \neq 0)$ for $p = 10$ and 50 respectively.

2. Mildly informative network: The tri-diagonal elements of A are set such that the proportion $a_{jk} = 1$ is 0.75, alternatively for the elements outside the main tri-diagonal the proportion of $a_{jk} = 1$ is 0.25. This led to a 0.778 and 0.747 proportion of edges that agree between A and $\mathrm{I}(\Theta \neq 0)$ for $p = 10$ and 50 respectively.

3. Strongly informative network: The tri-diagonal elements of A are set such that the proportion $a_{jk} = 1$ is 0.85, alternatively for the elements outside the main tri-diagonal, the proportion of

$a_{jk} = 1$ is 0.15. This led to a 0.867 and 0.844 proportion of edges that agree between A and $\mathrm{I}(\Theta \neq 0)$ for $p = 10$ and 50 respectively.

Code to reproduce our simulations is available in the GitHub repository. For each setting, we report the mean squared estimation error (MSE), the false discovery rate (FDR), and the false negative rate (Benjamini and Hochberg, 1995). The FDR is the expected proportion of false positive edges among the edges estimated to be present, a measure of type I error, whereas the FNR is the expected proportion of false negative edges among those not reported to be present, which measures statistical power. Under the GLASSO methods, an edge is declared if the corresponding estimate of ρ_{jk} was non-zero (rounded to 5 decimal places). For the spike-and-slab model an edge is declared when the posterior probability that ρ_{jk} arises from slab (15) is above 0.95.

Table 1 summarises the results. For all sample sizes, the network GLASSO significantly reduced the MSE when the network data were mildly or strongly informative ($A_{0.75}$ and $A_{0.85}$), whereas it attained a similar MSE to standard GLASSO in the uninformative network setting (A_{ind}). The FDR was significantly above the usually accepted level of 0.05, which is to be expected, since LASSO-based point estimates are geared towards prediction rather than statistical inference. Regarding the spike-and-slab formulations, they consistently achieved an FDR below 0.05 and a small FNR, and in large p situations a further improvement of the MSE compared with the network-GLASSO methods. Adding network data improved the spike-and-slab MSE and FNR, particularly when p was large relative to n. The FDR did not noticeably improve, but it was already near-zero when not using the network data. These findings suggest that the spike-and-slab formulations tend to attain better inference than the GLASSO counterparts. However the latter may be more appealing in settings with pressing computational demands. For example, in the $p = 10$, $n = 100$, $A_{.85}$ setting GOLAZO took 9 seconds to run, whereas the `NumPyro` spike-and-slab implementation took 47 seconds (and `Stan` 8.95 minutes), see Section A.4 for further details.

6 Results

6.1 COVID-19 infection rates

We illustrate the findings afforded by our methodology in the COVID-19 application. Recall that the outcomes are log-infection rates for the $p = 99$ most populous US counties during $n = 97$ weeks and that a regression model was fit to account for various factors such as temperature and vaccination

Table 1: Simulation results under non, mildly and strongly informative networks A_{ind}, $A_{0.75}$ and $A_{0.85}$. For SS and network SS models edges declared when posterior probability > 0.95.

		$p = 10$			$p = 50$		
	n	MSE	FDR	FNR	MSE	FDR	FNR
GLASSO	100	0.350	0.370	0.098	3.505	0.442	0.292
Network GLASSO, $A_{ind.}$	100	0.354	0.340	0.122	3.623	0.392	0.306
Network GLASSO, $A_{0.75}$	100	0.291	0.258	0.093	2.847	0.421	0.251
Network GLASSO, $A_{0.85}$	100	**0.170**	0.174	0.120	2.246	0.426	0.223
SS	100	0.222	**0.000**	0.086	1.611	**0.000**	0.023
Network SS, $A_{ind.}$	100	0.237	0.003	0.082	1.631	0.004	0.025
Network SS, $A_{0.75}$	100	0.234	0.007	0.073	1.462	0.005	0.023
Network SS, $A_{0.85}$	100	0.189	0.047	**0.060**	**1.280**	0.002	**0.022**
GLASSO	200	0.184	0.416	0.022	1.794	0.476	0.181
Network GLASSO, $A_{ind.}$	200	0.201	0.378	0.040	1.871	0.439	0.189
Network GLASSO, $A_{0.75}$	200	0.161	0.309	0.022	1.515	0.412	0.181
Network GLASSO, $A_{0.85}$	200	0.096	0.204	0.098	1.241	0.388	0.173
SS	200	0.109	**0.000**	0.056	0.672	0.002	0.017
Network SS, $A_{ind.}$	200	0.127	0.007	0.053	0.671	**0.002**	0.017
Network SS, $A_{0.75}$	200	0.114	0.007	0.048	0.597	0.003	0.015
Network SS, $A_{0.85}$	200	**0.091**	0.023	**0.041**	**0.527**	0.002	**0.015**
GLASSO	500	0.082	0.367	0.002	0.825	0.410	0.032
Network GLASSO, $A_{ind.}$	500	0.085	0.315	0.007	0.766	0.443	0.035
Network GLASSO, $A_{0.75}$	500	0.066	0.270	0.000	0.604	0.419	0.031
Network GLASSO, $A_{0.85}$	500	0.045	0.195	0.008	0.512	0.386	0.027
SS	500	**0.030**	0.000	0.023	0.198	0.002	**0.009**
Network SS, $A_{ind.}$	500	0.034	**0.000**	0.023	0.201	**0.001**	0.010
Network SS, $A_{0.75}$	500	0.032	0.002	**0.018**	0.193	0.001	0.009
Network SS, $A_{0.85}$	500	0.033	0.008	0.022	**0.183**	0.001	0.009

rate, fixed time and county effects, and serial correlation (see Section 2). See also Section A.2 for a description of the data pre-processing. The two networks are a geographical network A_1 where $a_{jk}^{(1)}$

Table 2: Four models for the COVID-19 data. A_1 and A_2: networks defined by $1/\log(Geodistance)$ and $\log(Facebook)$. BIC values account for the extra hyper-parameters in the network GLASSO models. 10-fold: 10-fold cross-validated log-likelihood

Method	BIC	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	Edges	10-fold
GLASSO	7066.313	-1.711			646	273.61
Network GLASSO- A_1	5555.422	-0.132	-1.211		245	285.00
Network GLASSO- A_2	5561.311	0.263		-1.368	217	286.77
Network GLASSO- A_1 & A_2	**5525.378**	-0.056	-0.333	-1.000	235	**287.22**

is the reciprocal of the log-geographic distance between counties (j, k) (hence larger values indicate smaller distance), and a Facebook network A_2 where $a_{jk}^{(2)}$ is the log-Facebook connection index between (j, k). We note that Pearson's correlation between A_1 and A_2 is 0.746 and therefore there is a large overlap in the information introduced by both networks.

As a first exercise, we analyze the data using four strategies: a standard GLASSO using no network information, our network GLASSO using only the geographical or only the Facebook networks, and using both networks. Table 2 shows a summary comparing the four models. The model attaining the best BIC value includes the two networks, suggesting that they both carry relevant information to help learn the graphical model. The estimated coefficients for both networks $(\hat{\beta}_1, \hat{\beta}_2)$ were negative, i.e. counties that are close geographically or highly-connected at Facebook are regularised less. The larger coefficient $\hat{\beta}_2$ in the joint model suggests that the effect of the Facebook network is greater. Interestingly, the three network-regularised solutions were significantly sparser relative to the 628 edges detected by GLASSO.

Despite these solutions being sparser, they included some edges that were not included by GLASSO. Figure 3a shows edges that were only selected when adding the geographical network A_1, which largely correspond to counties that are close to each other. Figure 3b shows an analogous plot when using the Facebook network A_2, interestingly there are connections between faraway counties in the west, north-east and south-east. Figure A.7 further portrays the estimated graphical model when using both networks.

To further assess the relative performance of the four models, we undertook a 10-fold cross-validation exercise where we assessed the log-likelihood (as a measure of predictive accuracy) in an out-of-sample fashion. The models incorporating network information also performed much better

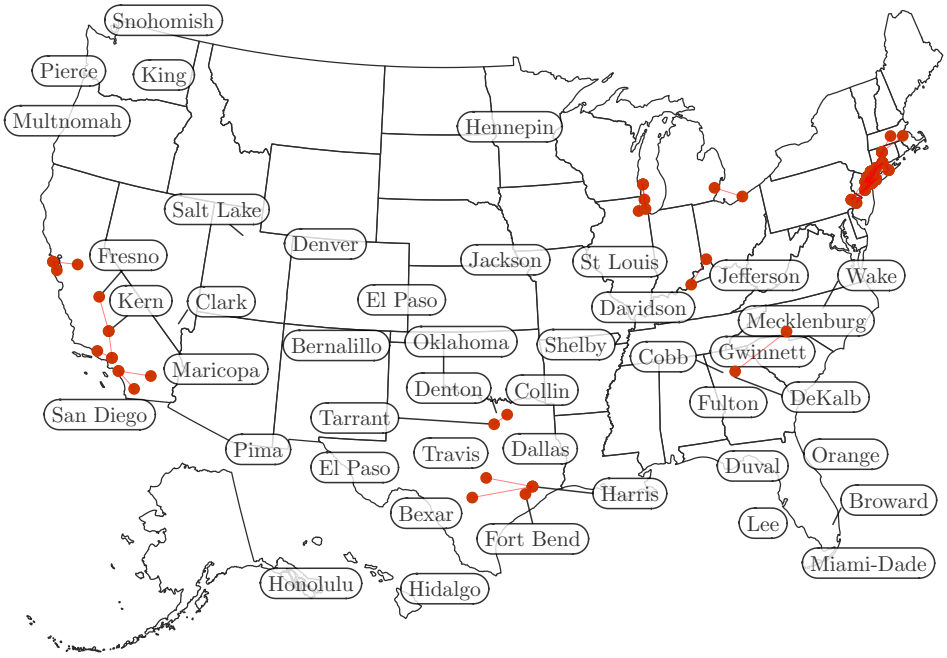
than standard GLASSO according to this predictive criterion.

Finally, we applied our spike-and-slab framework to obtain further insights on how the proportion of edge connections, as well as the mean partial correlation, depend on the two networks. As discussed earlier, the bottom panels in Figure 1 display the fitted spike-and-slab distribution as a function of both networks. Table 3 presents the corresponding (empirical Bayes) hyper-parameter estimates, and Figure A.6 display the estimated prior slab mean and prior slab probability as functions of the geographic and Facebook network connections. The mean of the non-zero partial correlations ($-\rho_{jk}$), i.e. the slab location parameter, was estimated to increase with both networks (negative $\hat{\eta}_{01}$, $\hat{\eta}_{02}$). Their variance (slab dispersion parameter) was also estimated to increase with both networks (negative $\hat{\eta}_{11}$, $\hat{\eta}_{12}$). The estimated coefficients $(\hat{\eta}_{21}, \hat{\eta}_{22})$ parameterising the slab probability (i.e. non-zero partial correlation) had opposite signs. Larger Facebook connectivity was associated with a larger probability of the corresponding ρ_{jk} being non-zero, while under the Geographical distance the opposite was observed. However the credibility interval for the latter (η_{21}) included 0, i.e. while there was strong evidence that the geographical distance affects the mean and variance of the partial correlations, this was not the case for the probability of a non-zero partial correlation (after accounting for the Facebook index). Indeed, Figure A.6 illustrates that the probability of a non-zero partial correlation increases from near-zero to near-one as the Facebook index grows, whereas that is not the case for the geographical index. These results illustrate the greater flexibility provided by the network spike-and-slab models to portray the relation between the network data and the partial correlations. For completeness, Table A.3 summarises the selected graphical model under a 0.5 and 0.95 posterior probability threshold for declaring an edge.

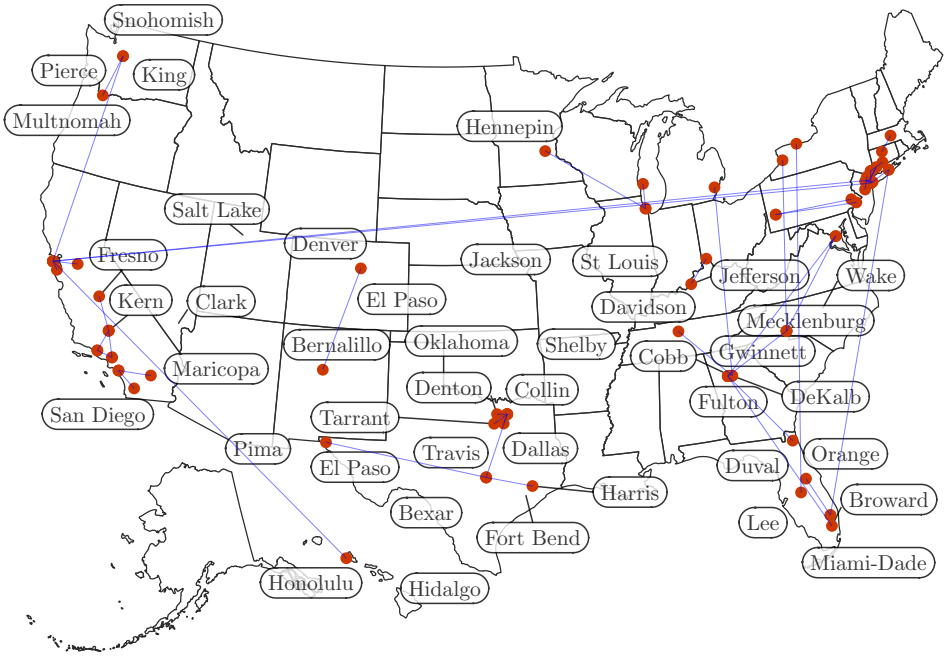
6.2 Stock market excess returns

We apply our methodology to study the log-daily excess returns (calculated by the Fama-French three-factor model) of $p = 200$ US companies. The two networks are an economic risks network A_1 where $a_{jk}^{(1)}$ is the Pearson's correlation between vectors of $\log(1 + count)$ of terms related to economic risks (Baker et al., 2019) in company j and k's 10-K filings, and A_2 the equivalent but for those terms classified as policy risks. Pearson's correlation between the two networks was 0.301, suggesting that they provide largely different information. See Section A.3 for a description of the data pre-processing.

We firstly analyze the data using the standard GLASSO, our network GLASSO using only the Economic network, only the Policy network, and finally using both networks. Table 4 compares these four models. The model including both networks attained the highest BIC and their estimated



(a) Edges identified by Network GLASSO - A_1 (geographical network) but not by GLASSO



(b) Edges identified by Network GLASSO - A_2 (Facebook network) but not by GLASSO

Figure 3: Edges identified by Network GLASSO but not by standard GLASSO. The coordinates of Honolulu (Hawaii) have been adjusted from (-164.44361, 23.87280) to (-158.2019740, 21.4613654) for presentation

Table 3: Network spike-and-slab empirical Bayes (marginal MAP) estimates and 95% posterior intervals for COVID-19 data. A_1 and A_2: networks defined by $1/log(Geodistance)$ and $log(Facebook)$.

	Intercept	A_1	A_2
η_0 (slab location)	0.046	-0.016	-0.047
95% interval	(0.027, 0.049)	(-0.023, -0.006)	(-0.051, -0.021)
η_1 (slab dispersion)	-2.325	-0.183	-0.177
95% interval	(-2.423, -1.691)	(-0.323, -0.02)	(-0.509, -0.01)
η_2 (slab probability)	-1.781	-0.187	1.154
95% interval	(-1.961, -0.232)	(-0.349, 0.191)	(0.718, 1.379)

Table 4: Four models for the stock market data. A_1 is the Economic network, A_2 the Policy network. BIC values account for the extra hyper-parameters in the network GLASSO models. 10-fold is the 10-fold cross-validation log-likelihood

Method	BIC	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	Edges	10-fold
GLASSO	47505.870	-1.474			631	-6857.00
Network GLASSO- A_1	46927.615	-1.289	-0.368		664	-6834.45
Network GLASSO- A_2	47012.765	-1.105		-0.474	604	**-6828.51**
Network GLASSO- A_1 & A_2	**46853.544**	-1.222	-0.306	-0.167	654	-6831.10

parameters $(\hat{\beta}_1, \hat{\beta}_2)$ are both negative. That is, partial correlations between companies that have large connections in the network are more likely to be non-zero, and are hence less regularised. The estimated graphical model when using both networks is slightly denser than under standard GLASSO.

To further assess the four models we evaluated their out-of-sample log-likelihood using 10-fold cross-validation. The models incorporating network information perform better than standard GLASSO in this prediction exercise.

To gain further insights into the relation between partial correlations and the network data, we applied our spike-and-slab framework. Figure A.12 shows estimated spike-and-slab distributions for the partial correlations as a function of both networks, and Table 5 the corresponding hyper-parameter estimates. Further, Figure A.14 displays the estimated effect of both networks on the slab location (mean non-zero partial correlation) and its probability. Interestingly, the policy network had a large

Table 5: Network spike-and-slab empirical Bayes (marginal MAP) estimates and 95% posterior credible intervals for the stock market data. A_1 is the Economic network, A_2 the Policy network.

	Intercept	A_1	A_2
η_0 (slab location)	-0.001	-0.014	-0.009
95% interval	(-0.005, 0.004)	(-0.02, -0.01)	(-0.014, -0.004)
η_1 (slab dispersion)	-2.480	-0.129	-0.048
95% interval	(-2.657, -2.301)	(-0.24, -0.01)	(-0.198, 0.105)
η_2 (slab probability)	-1.798	0.313	0.734
95% interval	(-2.138, -1.416)	(0.092, 0.541)	(0.455, 1.046)

positive effect on the probability of a non-zero partial correlation, but its effect on their mean and variance is negligible. On the other hand, the economic network was positively associated with probability of a non-zero partial correlation, its mean and variance.

This rather surprising result is substantively important. Since the pioneering work of Hassan et al. (2019), economists have used word-count based approaches to measure and evaluate firm-level exposure to political and policy risks. But exposure to policy risks is in part a function of fundamental economic risks: for example, firms exposed to air travel via their business model will also be exposed to regulation of the Federal Aviation Administration. The finding that policy risk overlap is not associated with positive excess return co-movement after controlling for economic risk overlap suggests that much of its effect operates through firm fundamentals. This can also be seen in the last row of Table 4: when both networks are included in the GLASSO, the estimated impact of the economic risk network on the regularlization is stronger than that of the policy risk network.

For completeness, Table A.5 summarises the selected graphical model under a 0.5 and 0.95 posterior probability threshold for declaring an edge.

7 Discussion

In this paper, we explored how exogenous network information could be incorporated to improve graphical model inference. We considered two approaches for so doing. Firstly, under a penalised likelihood framework where the penalty term for the partial correlation between two variables is allowed to depend on the strength of their network connection(s). Secondly, a Bayesian hierarchical model

where the networks can characterise the probability that there is no connection and also model the behaviour of non-zero connections. The former approach was operationalised using the GOLAZO algorithm (Lauritzen and Zwiernik, 2020) and network hyper-parameters were learned using the BIC, while the Bayesian model was implemented in the `NumPyro` probabilistic programming language (Bingham et al., 2019; Phan et al., 2019) and empirical Bayes was used to estimate the network hyper-parameters. Our simulations showed that network information could be particularly helpful in situations where the network dimension p was large relative to the sample size n, as is often the case in real-world scenarios. Further, we observed that the ability to learn hyperparameters ameliorated the consequences in a worst-case scenario where one introduces an uninformative network.

We applied our methodology to model the evolution of COVID-19 cases across the $p = 99$ most populous counties in the US and to model the log-excess return of $p = 200$ US companies. In the former, we found that both a geographical distance network and a Facebook connectivity network were informative about the dynamics of COVID-19 cases, allowing for the estimation of a sparser graphical model that predicted better out-of-sample, and that of the two, the Facebook network appeared to have the greater effect. In the latter, we found that networks parameterising the similarity of two firm's risk exposures grouped into economic and policy risks were both informative of the graphical model and that while the policy network appeared to have a stronger relationship to whether two firms were connected, the economic network was more predictive of the behaviour of connected observations.

Further work could consider richer relationships for how the graphical models can depend on the networks, e.g. by considering non-parametric models. Other extensions are replacing our spike-and-slab prior with so-called '*non-local*' slabs using optimisation algorithms as in (Avalos-Pacheco et al., 2022), or developing tailored variational methods. Said methods could help scale Bayesian implementation to even higher dimensions, i.e. the setting where adding network information is more useful.

Finally, we remark that the general notion of adding external information to help improve inference extends well beyond our graphical model and network data applications. In this sense, we hope that our constructions can be helpful to researchers in (hopefully numerous) applied statistical modelling settings beyond the one considered here.

Acknowledgements

JJ, DR and PZ were partially funded by Government of Spain's Plan Nacional PGC2018-101643-B-I00 and the Ayudas Fundación BBVA Proyectos de Investigación Científica en Matemáticas 2021, grants. JJ was also partially funded by Juan de la Cierva Formación de la Agencia Estatal de Investigación FJC2020-046348-I, and DR by Europa Excelencia EUR2020-112096, the AEI/10.13039/501100011033 and European Union "NextGenerationEU"/PRT. LL acknowledges the financial support from the China Scholarship Council (CSC) (Grant No.202006240148). SH gratefully acknowledges funding from European Research Council Consolidator Grant 864863, which supported his time and the work of LB.

We thank Dennis Kristensen for helpful comments and Du Phan for advising us on our models' implementation in `NumPyro`.

References

Alejandra Avalos-Pacheco, David Rossell, and Richard S Savage. Heterogeneous large datasets integration using bayesian factor regression. *Bayesian Analysis*, 17(1):33–66, 2022.

Jonathan J Azose and Adrian E Raftery. Estimating large correlation matrices for international migration. *The annals of applied statistics*, 12(2):940, 2018.

Scott R. Baker, Nicholas Bloom, Steven J. Davis, and Kyle J. Kost. Policy News and Stock Market Volatility. *National Bureau of Economic Research Working Paper Series*, (w25720), March 2019.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

M.M. Barbieri and J.O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3): 870–897, 2004.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300, 1995.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20:28:1–28:6, 2019.

U.S. Census Bureau. Us 2019 population data, 2020. Available from github: https://www2.census.gov/programs-surveys/popest/tables/2010-2019/counties/totals/.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1):1–32, 2017.

Jack Storror Carter, David Rossell, and Jim Q Smith. Partial correlation graphical lasso. *arXiv preprint arXiv:2104.10099*, 2021.

Alberto Cassese, Michele Guindani, and Marina Vannucci. A Bayesian integrative model for genetical genomics with spatially informed variable selection. *Cancer informatics*, 13:S13784, 2014.

J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Ting-Huei Chen, Nilanjan Chatterjee, Maria Teresa Landi, and Jianxin Shi. *Journal of the American Statistical Association*, 116(533):133–143, 2021.

Sharon Chiang, Michele Guindani, Hsiang J Yeh, Zulfi Haneef, John M Stern, and Marina Vannucci. Bayesian vector autoregressive model for multi-subject effective connectivity inference using multimodal neuroimaging data. *Human brain mapping*, 38(3):1311–1332, 2017.

CSSE. Covid19 infection rates, 2020a. Available from github: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv.

CSSE. Government coronavirus policies, 2020b. Available from github: https://github.com/CSSEGISandData/COVID-19_Unified-Dataset.

CSSE. Daily average near-surface temperature, 2020c. Available from github: https://github.com/CSSEGISandData/COVID-19_Unified-Dataset/tree/master/Hydromet.

CSSE. U.s. vaccination data, 2020d. Available from github: https://github.com/govex/COVID-19/tree/master/data_tables/vaccine_data/us_data/time_series.

Steven J. Davis, Stephen Hansen, and Cristhian Seminario-Amez. Firm-Level Risk Exposures and Stock Returns in the Wake of COVID-19. Working Paper 27867, National Bureau of Economic Research, September 2020.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

Edwin J. Elton and Martin J. Gruber. Estimating the Dependence Structure of Share Prices–Implications for Portfolio Selection. *The Journal of Finance*, 28(5):1203–1232, 1973. ISSN 0022-1082. doi: 10.2307/2978758.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993. ISSN 0304-405X. doi: https://doi.org/10.1016/0304-405X(93)90023-5. URL https://www.sciencedirect.com/science/article/pii/0304405X93900235.

Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive LASSO and SCAD penalties. *Annals of Applied Statistics*, 3(2):521–541, 2009.

Yingying Fan and Cheng Yong Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society B*, 75(3):531–552, 2013.

Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. *Advances in Neural Information Processing Systems*, 23:604–612, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

L. Gan, N.N. Narisetty, and F. Liang. Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, just-accepted:1–14, 2018.

Domenico Giannone, Michele Lenza, and Giorgio E. Primiceri. Economic Predictions With Big Data: The Illusion of Sparsity. *Econometrica*, 89(5):2409–2437, 2021. ISSN 0012-9682. doi: 10.3982/ECTA17842.

Shingo Goto and Yan Xu. Improving Mean Variance Optimization through Sparse Hedging Restrictions. *The Journal of Financial and Quantitative Analysis*, 50(6):1415–1441, 2015. ISSN 0022-1090.

Sharmistha Guha and Abel Rodriguez. Bayesian regression with undirected network predictors with an application to brain connectome data. *Journal of the American Statistical Association*, (in press): 1–13, 2020.

Kathleen Weiss Hanley and Gerard Hoberg. Dynamic Interpretation of Emerging Risks in the Financial Sector. *The Review of Financial Studies*, 32(12):4543–4603, December 2019. ISSN 0893-9454. doi: 10.1093/rfs/hhz023.

Tarek A Hassan, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. Firm-Level Political

Risk: Measurement and Effects. *The Quarterly Journal of Economics*, 134(4):2135–2202, November 2019. ISSN 0033-5533. doi: 10.1093/qje/qjz021.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Zakaria S. Khondker, Hongtu Zhu, Haitao Chu, Weili Lin, and Joseph G. Ibrahim. The Bayesian covariance LASSO. *Statistics and its Interface*, 6:243–259, 2013.

Theresa Kuchler, Dominic Russel, and Johannes Stroebel. The geographic spread of covid-19 correlates with the structure of social networks as measured by facebook. *Journal of Urban Economics*, page 103314, 2021.

Markku Kuismin and Mikko J Sillanpää. Mcpese: Monte carlo penalty selection for graphical lasso. *Bioinformatics*, 37(5):726–727, 2021.

Steffen Lauritzen and Piotr Zwiernik. Locally associated graphical models and mixed convex exponential families. *arXiv*, 2008.04688:1–34, 2020. to appear in Annals of Statistics.

Harry Markowitz. Portfolio Selection. *The Journal of Finance*, 7(1):77–91, 1952. ISSN 0022-1082. doi: 10.2307/2975974.

P. Müller, G. Parmigiani, C. Robert, and J. Rousseau. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99(468): 990–1001, 2004.

Radford Neal. *MCMC using Hamiltonian dynamics*, pages 113–162. Chapman and Hall/CRC, 2011.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv*, 1912.11554:1–10, 2019.

MA Quintana and DV Conti. Integrative variable selection via Bayesian model uncertainty. *Statistics in medicine*, 32(28):4938–4953, 2013.

V. Rockova and E.I. George. EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014.

David Rossell and Piotr Zwiernik. Dependence in elliptical partial correlation graphs. *Electronic Journal of Statistics*, 15(2):4236–4263, 2021.

Lorenzo Schiavon, Antonio Canale, and David B Dunson. Generalized infinite factorization models. *arXiv*, 2103.10333:1–46, 2021.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

J.G. Scott and J.O Berger. Bayes and empirical Bayes multiplicity adjustment in the variable selection problem. *The Annals of Statistics*, 38(5):2587–2619, 2010.

Marc Senneret, Yannick Malevergne, Patrice Abry, Gerald Perrin, and Laurent Jaffrès. Covariance Versus Precision Matrix Estimation for Efficient Asset Allocation. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):982–993, September 2016. ISSN 1941-0484. doi: 10.1109/JSTSP.2016.2577546.

Roger W Sinnott. Virtues of the haversine. *Sky and telescope*, 68(2):158, 1984.

Francesco C Stingo, Yian A Chen, Marina Vannucci, Marianne Barrier, and Philip E Mirkes. A Bayesian graphical modeling approach to microRNA regulatory network inference. *The annals of applied statistics*, 4(4):2024–2048, 2010.

Francesco C Stingo, Yian A Chen, Mahlet G Tadesse, and Marina Vannucci. Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *The annals of applied statistics*, 5(3):1–24, 2011.

Hao Wang. Bayesian graphical LASSO models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886, 2012.

Lingxiao Wang, Xiang Ren, and Quanquan Gu. Precision matrix estimation in high dimensional Gaussian graphical models with faster rates. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, 51:177, 2016.

Tao Wang and Lixing Zhu. Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, 102(7):1141–1151, 2011.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Giacomo Zanella and Gareth Roberts. Multilevel linear models, gibbs samplers and multigrid decompositions (with discussion). *Bayesian Analysis*, 16(4):1309–1391, 2021.

Yiyun Zhang, Runze Li, and Chih-Ling Tsai. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323, 2010.

A Supplementary Material

Section A.1 provides further details for the implementation of our network GLASSO and network spike-and-slab models. Section A.2 contains further information related to our COVID-19 data application, including the data collection, preprocessing, linear model estimation and diagnostic checks, network specification and linearity check, as well as further results and figures. Section A.3 provides analogous information for the stock market data. Lastly, Section A.4 provides a performance comparison of the network GLASSO frequentist model with the network spike-and-slab Bayesian model using `Stan` and `NumPyro`. Code to implement all of our experiments and data pre-processing is available at https://github.com/llaurabat91/graphical-models-external-networks.

A.1 Implementation details for network GLASSO and network spike-and-slab

A.1.1 Bounding the region for optimal GOLAZO hyperparameters β.

We describe simple bounds to limit the grid search for the hyper-parameter $\hat{\beta}$ optimising the BIC/EBIC. The GOLAZO algorithm (Section 8.1 in Lauritzen and Zwiernik (2020)) is a block coordinate descent algorithm where the j-th row is optimized with other entries of Σ fixed by solving a quadratic program

$$\min_{d} \; d^T(\Sigma_{\setminus j})^{-1}d \qquad \text{subject to } |\Sigma_{ij} - S_{ij}| \leq \lambda_{ij} \text{ for all } i < j \text{ and } \Sigma_{ii} = S_{ii} \text{ for all } i, \tag{17}$$

where d contains the off-diagonal entries of the j-th row of Σ (the diagonal entry always satisfies $\Sigma_{jj} = S_{jj}$). The following lemma guarantees that for large enough λ_{ij} the solution is to set all parameter estimates to zero.

Lemma A.1. If $\lambda_{jk} \geq |S_{jk}|$ for all $k \neq j$ then $d = 0$ optimises (17).

Proof. Under the given condition $d = 0$ is always feasible. Since $d = 0$ is also the global minimum, the result follows. □

We can therefore assume that $\lambda_{jk} < |S_{jk}|$ for at least one pair (j, k). That is, we may restrict attention to β satisfying

$$\max_{j \neq k} \log(\lambda_{jk}) \;=\; \max_{j \neq k} \beta_0 + \sum_{q=1}^{Q} \beta_q \bar{a}_{jk}^{(q)} \;\leq\; \max_{j \neq k} \log |S_{jk}|.$$

Note that this expression bounds the range of possible optima for each β_q given the rest, and in particular for β_0 we obtain

$$\beta_0 \leq \max_{j \neq k} \{\log |S_{jk}| - \sum_{q=1}^{Q} \beta_q \bar{a}_{jk}^{(q)}\},$$

which is $\leq \max_{j \neq k} \log |S_{jk}|$ at the initialisation step where $\beta_1 = \ldots = \beta_Q = 0$.

The fact that Σ in (17) must be positive definite allows for the construction of further simple bounds. For every $i \neq j$ we necessarily have $\Sigma_{jk}^2 \leq \Sigma_{jj}\Sigma_{kk} = S_{jj}S_{kk}$, or equivalently, $\Sigma_{jk} \in [-\sqrt{S_{jj}S_{kk}}, \sqrt{S_{jj}S_{kk}}]$. It follows that, without loss of generality, we can restrict attention to that $\lambda_{jk} \leq \sqrt{S_{jj}S_{kk}} - |S_{jk}|$ giving

$$\beta_0 + \sum_{q=1}^{Q} \beta_q \bar{a}_{jk}^{(q)} \;\leq\; \log(\sqrt{S_{jj}S_{kk}} - |S_{jk}|) \qquad \text{for all } j \neq k.$$

A.1.2 Implementation of spike-and-slab

If the spike has a very small variance, or the slab has too bigger variance it can be difficult for an MCMC sampler to efficiently explore both spaces. We use a rescaling trick to facilitate efficient MCMC inference for the network spike-and-slab model. Rather than sample directly from $\pi(\rho)$ as defined by (9), for each ρ_{jk} we define latent variables $\tilde{\rho}_{jk}^{spike}$, $\tilde{\rho}_{jk}^{slab}$ and u_{jk}. We then sample

$$\tilde{\rho}_{jk}^{spike} \sim \mathrm{DE}\,(0,1)\,, \quad \tilde{\rho}_{jk}^{slab} \sim \mathrm{DE}\,(0,1) \;\text{ and } \quad u_{jk} \sim \mathrm{Unif}[0,1],$$

and set

$$\rho_{jk} = \mathrm{I}(u_{jk} > w_{jk}) \left(s_0 \times \tilde{\rho}_{jk}^{spike}\right) + \mathrm{I}(u_{jk} \leq w_{jk}) \left(\eta_0^T a_{jk} + s_{jk} \times \tilde{\rho}_{jk}^{slab}\right).$$

It is straightforward to see that the marginal distribution of ρ_{jk} matches that defined in (9). Lastly, to make such an implementation suitable for MCMC samplers that require differentiability, we approximate the indicator $\mathrm{I}(u_{jk} > w_{jk})$ with a sigmoid function

$$\mathrm{I}(x \geq 0) \approx \sigma_k(x) = \frac{1}{1 + \exp(-kx)} \text{ for large } k,$$

taking $k = 100$.

A.1.3 Prior elicitation

We elicit spike-and-slab prior parameters (η_0, η_1, η_2) that encourage sparse solutions, avoid pathological values, and maintain their specified intuition whilst being minimally informative. We finish this section with a table of the values used in the simulations and in our applications. For interpretability, we treat the spike's scale parameter s_0 as a constant. Recall that the spike captures partial correlations ρ_{jk} that are considered to be 0 for all practical purposes, which here we consider to be $|\rho_{jk}| < 0.01$. We hence

set s_0 such that the spike has most of its density below this threshold, i.e. $\Pi(\rho_{ij} \in (-\tau, \tau); s_0) = 0.95$, where $\tau = 0.01$. This gave the value $s_0 = 0.003$

Consider first the hyperparameters $(\eta_{00}, \eta_{10}, \eta_{20})$ defining the intercept of the regression of the slab's mean, variance, and prior probability on the network data. We set the priors

$$\begin{aligned}\eta_{00} &\sim \mathcal{N}\left(0, g_0^2\right) \\ \eta_{10} &\sim \mathcal{N}\left(m_1, g_1^2\right) \\ \eta_{20} &\sim \mathcal{N}\left(m_2, g_2^2\right).\end{aligned}$$

For the hyperparameters that capture the effect of each network $A^{(q)}$, where $q = 1, \ldots, Q$, we set

$$\begin{aligned}\eta_{0q} &\sim \mathcal{N}\left(0, g_0^2\right) \\ \eta_{1q} &\sim \mathcal{N}\left(0, g_1^2\right) \\ \eta_{2q} &\sim \mathcal{N}\left(0, g_2^2\right).\end{aligned}$$

Centering the prior of η_{00} at 0 encodes the absence of information about whether partial correlations are positive or negative on average. Similarly, centering the priors of $(\eta_{0q}, \eta_{1q}, \eta_{2q})$ at zero reflects no prior knowledge on whether the network data are predictive of ρ and in which direction. To set the remaining hyperparameters we assume the networks have been standardised and conduct the prior elicitation for the average value of the networks (i.e. $\bar{a}_{jk}^{(q)} = 0$ for all networks q). As a result, our prior elicitation is invariant to the network(s) considered.

The prior on η_2 was set based on sparsity and minimal informativeness considerations. Specifically, we set the prior expected number of edges (non-zero partial correlations) to scale linearly with p, so that each node is expected to have a constant degree as p grows. When all networks are at their average value the slab prior probability is $w = 1/(1 + e^{-\eta_{20}})$. A standard non-informative prior on slab prior probabilities is a Beta$(m_w v_w, m_w(1 - v_w))$ distribution (Scott and Berger, 2010), where m_w is the prior mean and v_w is often interpreted as the prior 'sample size'. We take the minimally informative choice $v_w = 1$. Regarding m_w, we set it such that the prior expected number of edges is p. Since the prior expected number of edges is

$$\mathbb{E}\left[\sum_{j=1}^{p}\sum_{k<j}\mathbb{I}(\rho_{jk} \in \text{slab})\right] = \frac{p(p-1)}{2}w,$$

for $m_w = \frac{2}{(p-1)}$ the expected number of edges is p. Based on these considerations, we set the (m_2, g_2^2) featuring in the prior of η_{20} and η_{2q} so that the implied prior on w has the same mean and variance as the Beta prior described above.

Table A.1: Network spike-and-slab prior hyperparameters

	$p = 10$	$p = 50$	COVID-19 data ($p = 99$)	Stock data ($p = 200$)
s_0	0.003	0.003	0.003	0.003
g_0	0.145	0.152	0.145	0.140
m_1	-2.197	-2.197	-2.197	-2.197
g_1	0.661	0.661	0.661	0.661
m_2	-2.722	-6.737	-9.368	-9.158
g_2	3.278	3.395	4.184	3.658

Regarding the prior on η_1, we considered that for the slab to capture non-zero partial correlations its prior scale parameter at the average value of the networks $s_{jk} = s_0(1 + \exp\{-\eta_{10}\})$ should be significantly larger than that of the spike s_0. We hence set m_1 and g_1 such that the prior mode of s_1 is $10 \times s_0$, as well as $s_1 > 3 \times s_0$ with prior probability 0.99.

Finally, the prior on η_0 was set based on prior positive-definiteness considerations. Specifically, the positive-definiteness indicator $\mathrm{I}(\rho \succ 0)$ induces dependence in the spike-and-slab prior density, i.e. it can produce a joint prior that is vastly different from the product of independent priors on each ρ_{jk}. Such a discrepancy is undesirable for prior interpretation, particularly in our setting where the priors and their hyperparameters are objects of interest that describe how ρ_{jk} depends on network data. To address this issue we set prior parameters such that the prior probability of ρ being positive definite when independently sampling its elements is at least 0.95. Conditional on the priors specified for (η_1, η_2), g_0 was set to the largest value (i.e. least informative) that guarantees at least 0.95 probability that ρ is positive-definite under independent sampling from the unconstrained spike-and-slab prior components.

A.1.4 Elicited values

Table A.1 presents the elicited values used in our simulations and real data examples. Code to elicit priors following the specification above for further examples is available in the GitHub repository. As the dimension of the data increases, only the prior for η_2 changes greatly. This is a result of the assumption that the number of edges grows linearly with p, and therefore Θ is *a priori* assumed more sparse for larger p.

A.1.5 Reparametrisation of the network hyperparameters

An advantage of the Bayesian network spike-and-slab approach is that it allows us to do inference for the network hyperparameter as was done in Tables 3 and 5. Such inferences, however, require that the effective sample size (ESS) of the sampled hyperparameters is sufficiently high. We observed empirically that hyperparameters attain lower ESS. Although this phenomenon has not been studied in our graphical model settings, in hierarchical models it is well understood that parameters associated to higher levels have strictly slower MCMC mixing, and that said mixing can be improved by reparameterising the problem (Zanella and Roberts, 2021). We applied the following transformation of the hyperparameters to facilitate their sampling.

Rather than sample directly from the priors for the hyperparameters as outlined in Section A.1.3, we reparameterised and sampled

$$\tilde{\eta}_{iq} \sim \mathcal{N}\left(0, \frac{p(p-1)}{2n}\right), \quad i = 1, 2, 3, \quad q = 0, 1, \ldots, Q.$$

The original η hyperparameters can then be recovered as

$$\begin{aligned}
\eta_{i0} &= m_i + \tilde{\eta}_{i0} \times g_i / \sqrt{p(p-1)/2n}, \\
\eta_{iq} &= 0 + \tilde{\eta}_{iq} \times g_i / \sqrt{p(p-1)/2n}, \quad i = 1, 2, 3, \quad q = 1, \ldots, Q,
\end{aligned}$$

where $m_0 := 0$. The idea behind this is to first standardise the η's to all have mean 0 and variance 1, before adjusting the variance of the $\tilde{\eta}$'s by the square-root of the ratio of the number of ρ's ($p(p-1)/2$) from which the η's are learned, to the number of observations Y (n) from which the ρ's themselves are learned. Such a reparametrisation leaves the model completely unchanged, but we found this improved the ESS of the $\eta's$.

A.1.6 Simulation results using the EBIC

As a sensitivity check, we also consider using the EBIC (Chen and Chen, 2008) to select hyperparameters for the GLASSO and Network GLASSO models

$$\text{EBIC}(\lambda) = -2\ell_n(\hat{\Theta}(\lambda)) + |\mathbf{E}(\hat{\Theta}(\lambda))| \log n + 4|\mathbf{E}(\hat{\Theta}(\lambda))|\gamma_{\text{EBIC}} \log p \tag{18}$$

Compared with the BIC, (18) has an additional complexity penalty, controlled by hyperparameter γ. Foygel and Drton (2010) recommend $\gamma_{\text{EBIC}} \in [0, 0.5]$ where $\gamma_{\text{EBIC}} = 0$ recovers the BIC. Table A.2 presents the results of the experiments introduced in Section 5 when using the EBIC with $\gamma_{\text{EBIC}} = 0.5$

Table A.2: GLASSO and network GLASSO simulation results under non, mildly and strongly informative networks A_{ind}, $A_{0.75}$ and $A_{0.85}$ with EBIC rule ($\gamma_{\text{EBIC}} = 0.5$) for learning the β hyperparameters.

		$p = 10$			$p = 50$		
	n	MSE	FDR	FNR	MSE	FDR	FNR
GLASSO	100	0.474	0.243	0.176	6.628	0.163	0.566
Network GLASSO, $A_{ind.}$	100	0.556	0.163	0.253	7.008	0.128	0.632
Network GLASSO, $A_{0.75}$	100	0.383	0.138	0.162	5.691	0.112	0.504
Network GLASSO, $A_{0.85}$	100	0.195	0.103	0.153	4.566	0.098	0.414
GLASSO	200	0.254	0.283	0.060	2.726	0.224	0.241
Network GLASSO, $A_{ind.}$	200	0.265	0.223	0.082	2.678	0.227	0.248
Network GLASSO, $A_{0.75}$	200	0.200	0.176	0.058	2.155	0.206	0.216
Network GLASSO, $A_{0.85}$	200	0.108	0.118	0.120	1.837	0.188	0.207
GLASSO	500	0.101	0.281	0.004	0.958	0.327	0.138
Network GLASSO, $A_{ind.}$	500	0.099	0.235	0.011	1.002	0.286	0.142
Network GLASSO, $A_{0.75}$	500	0.074	0.185	0.000	0.781	0.272	0.153
Network GLASSO, $A_{0.85}$	500	0.051	0.116	0.096	0.698	0.214	0.158

to select hyperparameters. Comparing these results with Table 1 shows that using the EBIC reduced the FDR relative to the BIC, however, this generally results in much more conservative edge selection which damaged the MSE.

A.2 COVID-19 data analysis

This section provides additional details for the analysis of the COVID-19 infection rate data.

A.2.1 Data sources

To undertake our analysis, we collected and combined the following datasets.

1. U.S. population data

We selected the top 100 counties for analysis based on 2019 U.S. population data. Data were sourced from https://www2.census.gov/programs-surveys/popest/tables/2010-2019/counties/totals/. Due to the lack of statistics for the District of Columbia in the COVID-19 policy dataset, we proceed with the top 99 counties.

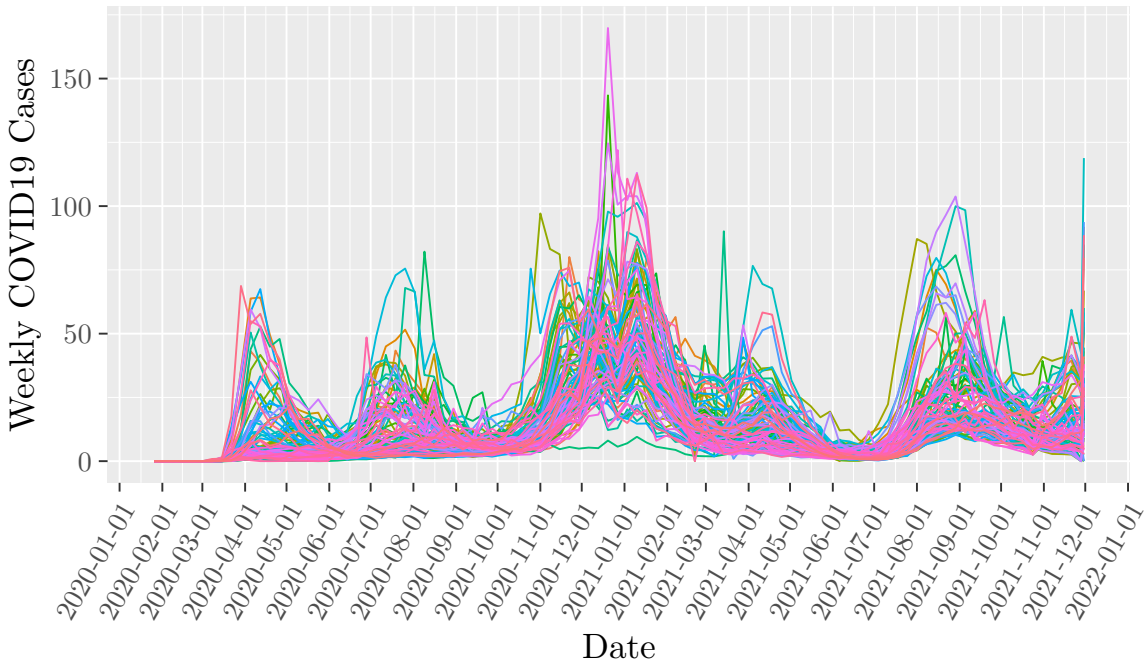


Figure A.1: Weekly COVID-19 Cases per county for the 99 biggest counties in the U.S.

2. FIPS code data

To allow for a better match between different datasets, we also extracted the "FIPS code" that uniquely identifies counties within the U.S. from the U.S. Bureau of Labor Statistics https://www.bls.gov/cew/classifications/areas/sic-area-titles.htm.

3. COVID-19 infection data

Time series data of confirmed COVID-19 infections in each U.S. county was obtained from https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_U.S..csv. Figure A.1 plots of the weekly aggregated confirmed COVID-19 infections

4. COVID-19 vaccination data

State-level vaccination data was obtained from https://github.com/govex/COVID-19/tree/master/data_tables/vaccine_data/us_data/time_series.

5. Policy data

The Oxford COVID-19 Government Response Tracker https://github.com/CSSEGISandData/COVID-19_Unified-Dataset tracks individual policy measures across 20 indicators. They also calculate several indices to give an overall impression of government activity. We used their Containment and Health indices to summarise the policy variables.

6. Temperature data

We extracted the daily average near-surface air temperature from the 'Hydromet' folder of the above repository https://github.com/CSSEGISandData/COVID-19_Unified-Dataset/tree/master/Hydromet.

7. U.S. area data

Population densities were obtained by dividing the county population by the area of the region. Area data were obtained from the U.S. Census Bureau https://tigerweb.geo.census.gov/tigerwebmain/TIGERweb_main.html.

8. Geodistance data

To measure the Geographical distance between two counties we use the Haversine distance (Sinnott, 1984) which assumes the earth is spherical. The latitude and longitude of each county were downloaded from the U.S. Census Bureau https://tigerweb.geo.census.gov/tigerwebmain/TIGERweb_main.html.

9. Facebook connectivity data

The Facebook Social Connectedness Index (SCI), obtained from https://data.humdata.org/dataset/social-connectedness-index, uses an anonymized snapshot of all active Facebook users and their friendship networks to measure the intensity of connectedness between locations. Specifically, it measures the relative probability that two individuals across two locations are friends with each other on Facebook.

A.2.2 Data processing

Once the data was collected, some minimal data preprocessing was required to prepare the data for our analysis. This consisted mainly of variable transformation and imputing of missing values.

Variables transformation Natural logarithms were taken of the variables '*confirmed case*', '*population density*' and '*number of vaccinations*'.

Missing values In addition, there were missing values in covariates Containment and Health Index data (CHI) as well as the Temperature (Temp) data and the Vaccination data. We imputed these missing values as follows

1. The CHI values were calculated as a function of different policy measures (https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/index_methodology.md). On several occasions either these policy measures or their flags were missing. We imputed these as follows

    - Missing flags were imputed as 0's, i.e. no flags
    - Missing values before the first recorded value were imputed as 0, i.e. assuming no measures were in place before the first recorded measure
    - Missing values in between two recorded values were imputed as an average of the before and after measures
    - Missing values after the last recorded value were imputed as the last seen measure, i.e. assuming a continuation

2. For the Temp data, the temperatures for San Francisco were not recorded at all. We imputed these using the nearest county geographically, San Mateo County.

3. The vaccination data was only recorded from the 14th of December 2020 and therefore all vaccination counts before this date were imputed as 0's.

A.2.3 Model description

Our final response variable is the log of the weekly COVID-19 infections per 10,000 members of the population (i.e. cases / population $\times$ 10,000). This results in data $y_1, \ldots, y_n$ where $y_i = (y_{i1}, ..., y_{ip})$ is the log of the standardised weekly COVID-19 infections at week i in the $p = 99$ counties with largest population. The sample interval is from 22 January 2020 to 30 November 2021 resulting in $n = 97$ weeks of data.

Our graphical model posits $y_i \sim \mathcal{N}_p(\mu_i, \Theta^{-1})$ where $\mu_i = (\mu_{i1}, ..., \mu_{ip})$. For convenience, we decouple the estimate of μ_i from Θ. We pose a regression model for μ_{ij} and then estimate Θ using

the residuals of this model assuming zero mean as in Section 3.1. Our generalised additive regression model for y_{ij} can be summarised as follows

$$log(confirmed)_{ij} = b_0 + b_1 \times log(Lag_{confirmed})_{ij} + b_2 \times log(popdensi)_j + b_3 \times Cum\_vaccinated_{ij} + b_4 \times CHI_{ij} + s(Temp)_{ij} + \gamma_2 \times Time_2 + ... + \gamma_T \times Time_T + \epsilon_{ij}$$

where

(1) $log(confirmed)_{ij}$ represents the natural logarithm of weekly per 10,000 people confirmed case in county j at time i.

(2) $log(Lag_{confirmed})_{ij}$ a first-order auto-regressive term measuring the infection rate at the previous time point $i-1$ for each county j

(3) $log(popdensi)_j$ is the population density for county j

(4) $Cum\_vaccinated_{ij}$ is the cumulative number of vaccinated individuals in county j by time i

(5) CHI_{ij} represented the Containment and Health Index summarising COVID-19 policies/measures put in place for county j and time i (wearing masks, closing schools, etc.)

(6) $s(Temp)_{ij}$ is a non-parametric smooth of the average temperature for county j at time i implemented in `mgcv` package in R

(7) $Time_i$ is a weekly fixed effect term estimating the mean infections across all counties at time i

(8) ϵ_{ij} are the residuals of county j at time i

With such a model we aim to remove the effect of the most relevant covariates that drive the mean number of infections, allowing Θ^{-1} to capture dependencies unexplained by these covariates.

A.2.4 Checking model goodness-of-fit

The main assumptions behind our assumed model require that the residuals ϵ_i are Gaussian distributed and independent across $i = 1, \ldots, n$ time points. We provide diagnostic plots to check these assumptions.

Figure A.2 plots the fitted values $\hat{y}_{ij}$ and each of the predictors against the residuals ϵ_{ij}. This demonstrates that the assumption that the covariates are linearly related to the response is satisfactory and that the residuals appear reasonably homoskedastic. Figure A.3 shows a histogram of the standardised residuals and Q-Q-normal plots for ϵ_{ij}. The Gaussian assumption is tenable here.

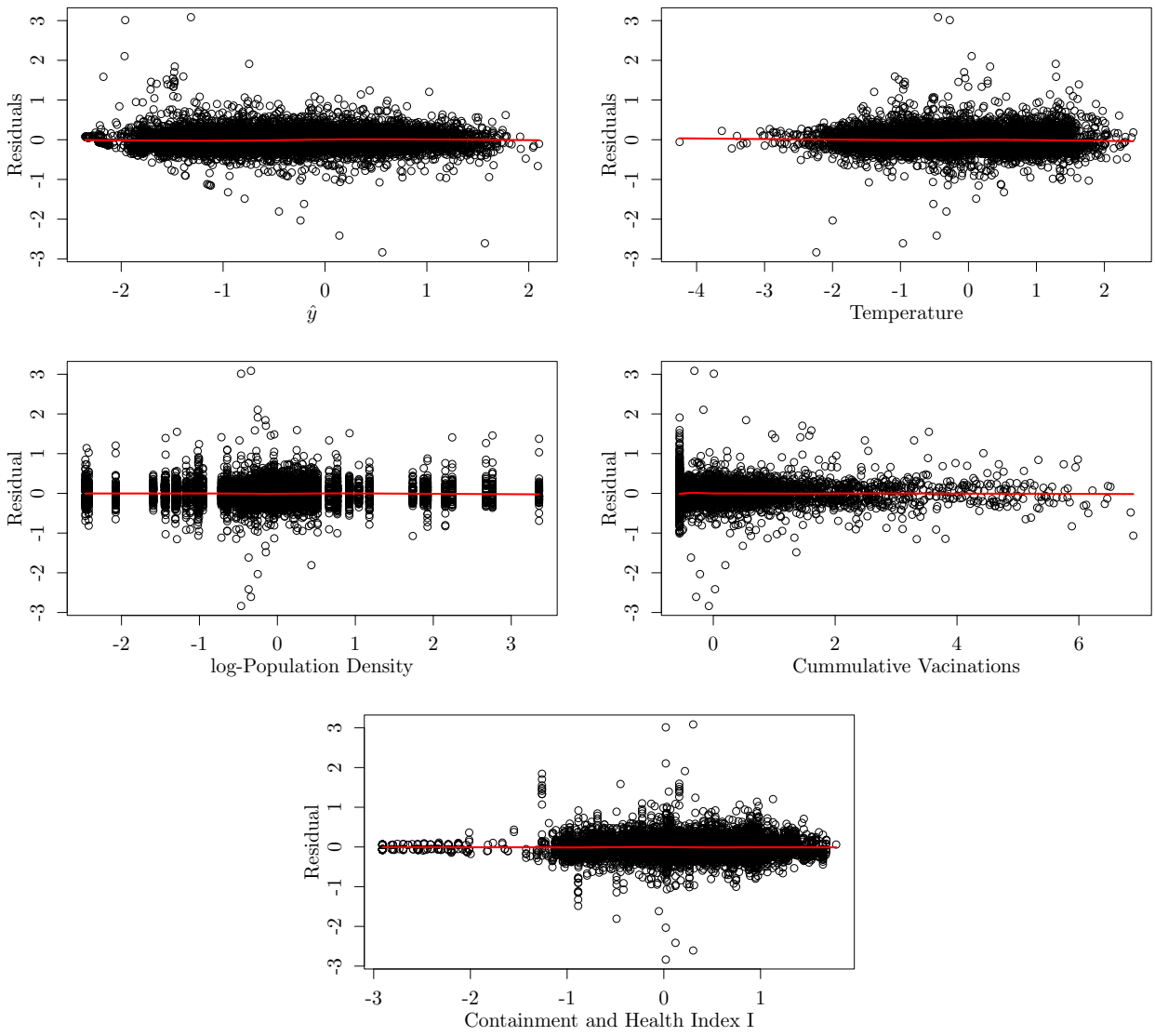


Figure A.2: Plots of the fitted values and each covariateagainst the residuals for the COVID-19 data. The **red** line corresponds to the LOWESS smooth.

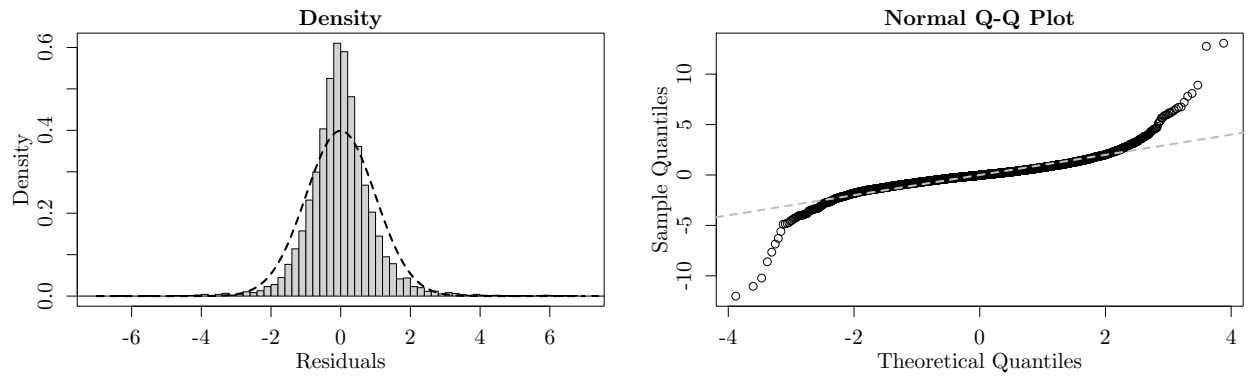


Figure A.3: COVID-19 data. **Left** Histogram of the standardised residuals compared with the standard Gaussian density. **Right** Q-Q Normal plot of the standardised residuals.

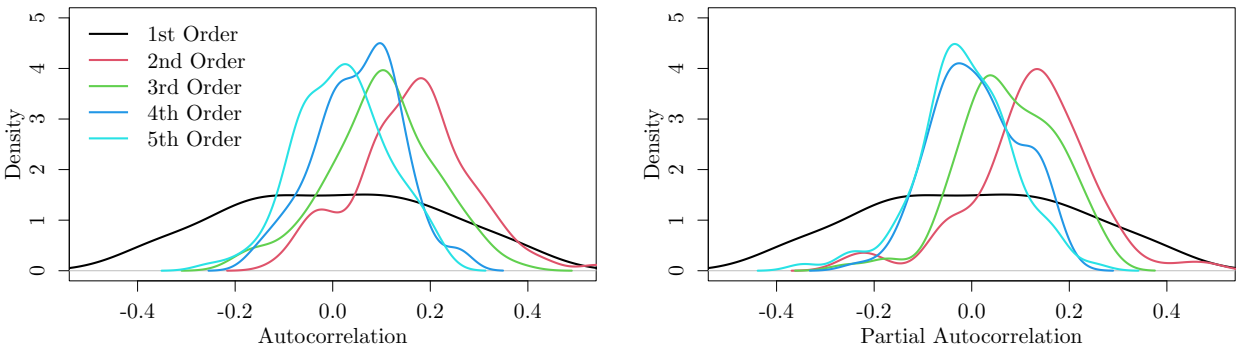


Figure A.4: Residual autocorrelation functions and partial autocorrelation functions after incorporating the AR1 term for the COVID-19 data.

The raw COVID-19 data exhibited strong serial correlation. To address this issue we added a first-order auto-regressive term. Figure A.4 plots the autocorrelation functions and partial autocorrelation functions for further lags after incorporating the AR1 term. These indicate that higher-order terms are unnecessary. After adding an AR1 term the interpretation of the errors (and their covariance) changes: they measure the infection rate relative to the covariates and to the infection rate of the previous week, i.e. they capture whether certain counties are growing faster/slower than expected (relative to the next week). So the model is investigating the growth rates, rather than absolute infection numbers.

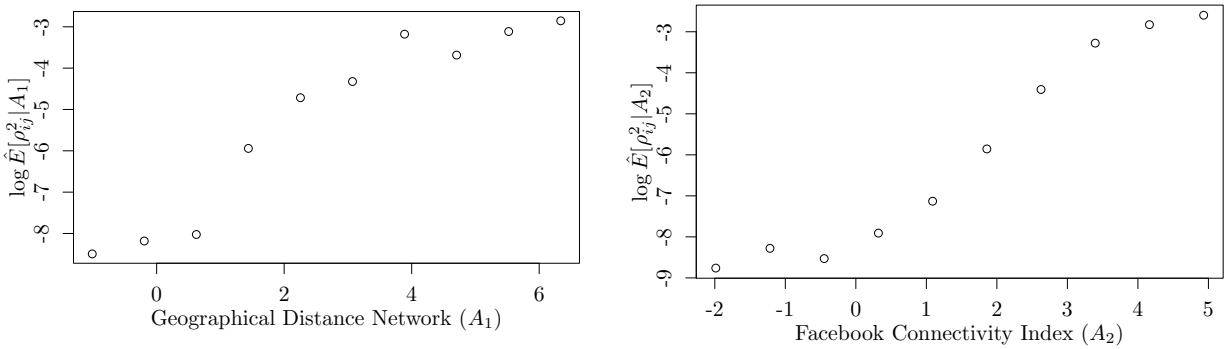


Figure A.5: Assessing the linear relation between $\log \mathbb{E}[\hat{\rho}_{jk}^2|A]$ and the network matrices, where $\hat{\rho}_{jk}$ is the GLASSO estimate. The points represent the log-mean values of $\hat{\rho}_{jk}^2$ within 10 equispaced bins defined for each network.

A.2.5 The network predictors

A further assumption of our proposed network GLASSO models, as discussed in Section 3.2, is that there is a linear relation between $\log \mathbb{E}[\rho_{jk}^2|A]$ and the network entries $a_{jk}^{(q)}$. To achieve linearity we defined our two network predictors as

$$A_1 := 1/\log(Geodist), \quad A_2 := \log(Facebook).$$

Figure A.5 illustrates that after such transformations, the assumption of linearity is reasonably satisfied.

A.2.6 Supplementary figures

Table A.3 summarises the estimated graphical model under the network spike-and-slab model using a posterior slab probability threshold of > 0.5 and > 0.95. The number of edges estimated under the 0.5 slab probability threshold is similar to the number of edges estimated under the network GLASSO models. Under the 0.95 slab probability threshold, the estimated number of edges is considerably more conservative.

Figure A.6 shows how the estimated network hyperparameters of Table 3 affect the location of the slab and the probability of being in the slab marginally for each network when fixing the other network to its mean. We see that while as both networks increase the location of the slab increases, the probability of being in the slab increases with the Facebook network and decreases with the geographical network.

Table A.3: COVID-19 data: Edge counts of the network spike-and-slab model when declaring an edge for posterior slab probability > 0.5 and > 0.95

	Edges (> 0.5)	Non-Edges (> 0.5)	Edges (> 0.95)	Non-Edges (> 0.95)
Network SS	280	4571	68	4783

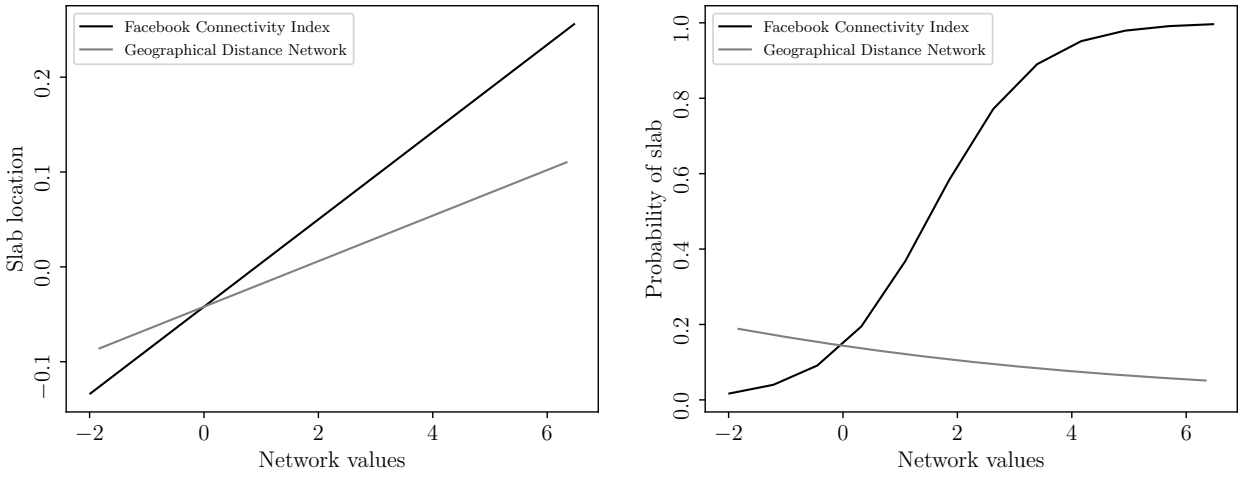


Figure A.6: COVID-19 data: Slab location (**left**) and slab probability (**right**) as a function of both networks estimated by empirical Bayes. We consider the slab here as the location of the partial correlations, the negation of the ρ model parameters, which is why negative hyperparameter for η_0 (Table 3) result in positive effects.

This demonstrates the benefits of the more complex spike-and-slab framework, while the network GLASSO estimated negative coefficients for both networks indicating positive correlation between the networks and the size of the partial correlations. The spike-and-slab disentangles this effect, showing that while the mean of those partial correlations that are non-zero increases with the geographical network, the probability of being non-zero does not.

A.2.7 U.S. map plots

Figure A.7 visualises the network given by non-zero elements of the GLASSO estimated Θ with no network information (top) and the network GLASSO estimate of Θ obtained when using both A_1 and A_2 (bottom) on top of a U.S. map. The network GLASSO estimates a much sparser network, but we

see there are still edges present between counties that are close in terms of geographical distance as well as those that are farther away.

A.2.8 Results using the EBIC

Similarly to the simulations, we also investigate the sensitivity of our COVID-19 data results by considering selecting hyperparameters using the EBIC with $\gamma_{\text{EBIC}} = 0.5$. Table A.4 presents these results. From the number of edges, we can see that using the EBIC estimates sparser networks than under the BIC, but the out-of-sample test set estimate suggests these estimates may be too sparse. Importantly, we see that the improvement of the network GLASSO methods over standard GLASSO is still apparent when using the EBIC selection criteria.

Table A.4: Four models for the COVID-19 data when using the EBIC ($\gamma_{\text{EBIC}} = 0.5$) to learn the network hyperparameters. A_1 and A_2: networks defined by $1/\log(Geodistance)$ and $\log(Facebook)$. EBIC values account for the extra hyper-parameters in the network GLASSO models. 10-fold: 10-fold cross-validated log-likelihood

Method	EBIC	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	Edges	10-fold
GLASSO	9556.655	-0.789			175	59.67
Network GLASSO- A_1	7252.871	1.276	-1.132		132	246.07
Network GLASSO- A_2	**6795.846**	3.605		-1.947	91	262.9537
Network GLASSO- A_1 & A_2	6809.573	3.556	0.278	-2.278	97	**263.12**

A.3 Stock market data preparation

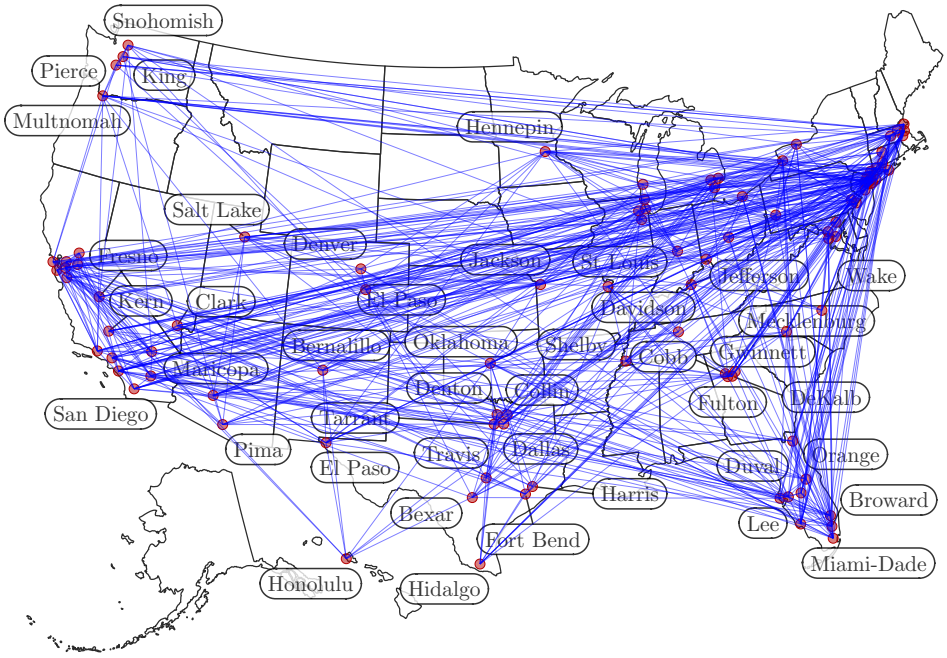
This section provides additional details for the analysis of the stock market excess returns data.

A.3.1 Data sources

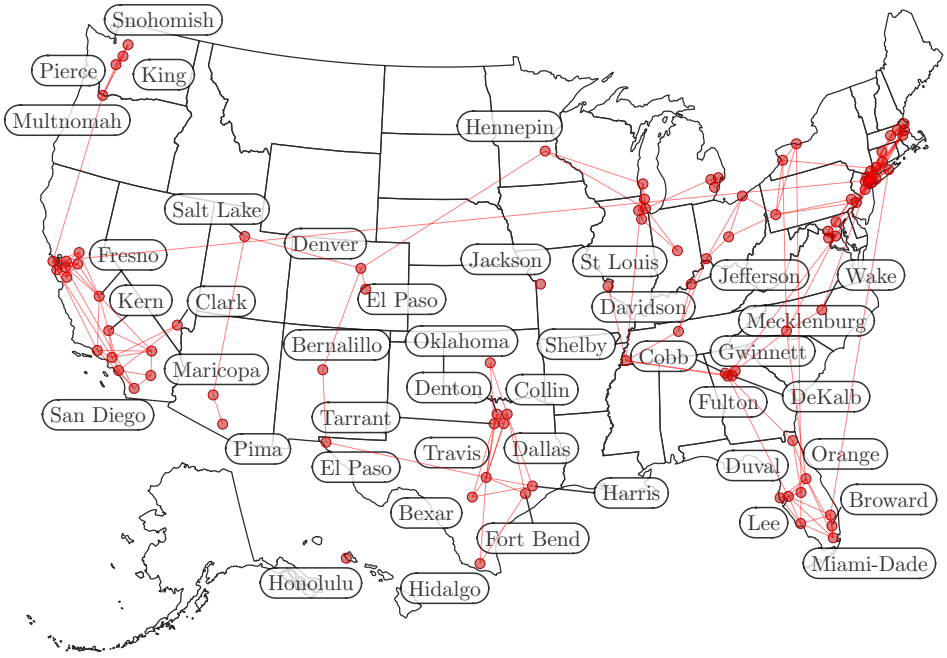
To undertake our analysis, we collected and combined the following datasets.

1. Stock price data

We extracted the daily closing stock price for a random selection of $p = 200$ firms satisfying the following criteria: we could retrieve their 10-K filings, they were sufficiently large; and had available financial records in the Compustat database between 2 January 2019 to 31 December 2019 (leaving $n = 252$ time points). The data was downloaded from the Center for Research in Security Prices



(a) Edges identified in GLASSO with no network



(b) Edges identified in network GLASSO with networks A_1 and A_2

Figure A.7: Edges identified by GLASSO and network GLASSO with the geographical distance and Facebook networks. The coordinates of the county Honolulu (Hawaii) have been adjusted from (-164.44361, 23.87280) to (-158.2019740, 21.4613654) for presentation.

(CRSP) database accessed via Wharton Research Data Services (WRDS).

2. CIK-TIC crosswalk data

The stock price data use the TIC identifier while the risk data from which we derive our networks uses the CIK identifier. We manually created a CIK-TIC crosswalk from the Compustat database accessed via WRDS.

3. Fama/French Three-Factor Model

We constructed excess returns using the Fama-French three-factor model (Fama and French, 1993). The three factors are the 1) overall market return, 2) a measure of firm size, and 3) a measure of book-to-market ratio. The daily Fama/French factors were downloaded from `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. For each stock, we regress 2019 daily returns on the three factors (plus a constant) and extract the residual as the excess return.

4. Risk measures

Our network data measures the similarity of two companies' risk exposures stratified into Economic and Policy risks. The 2019 10-K risk exposure data counts for each risk category the number of sentences within a company's 10-K filings that contained any member of a dictionary associated with that risk category. We manually construct these using the dictionary terms listed in Baker et al. (2019). From this data, we can construct a $p \times p$ network matrix for firms, where each entry X_{ij} represents the degree of "closeness" between firm X_i and firm X_j.

A.3.2 Data processing

1. The price data from CRSP is arranged by TIC, while the risk measures are arranged by CIK. So we need to use the TIC-CIK crosswalk table for matching.

2. There are some negative values in the stock data. The negative signs are to "indicate that it is a bid/ask average and not an actual closing price" when the "closing price is not available" `https://faq.library.princeton.edu/econ/faq/11159`. We, therefore, took the absolute values of the returns before the log return calculations.

A.3.3 Model description

Our final response variable is the log daily returns for $p = 200$ U.S. firms throughout 2019, resulting in $n = 252$ observations. We are, however, interested in the graphical model, $\mathcal{N}_p(0, \Theta^{-1})$, of the 'excess returns', defined as the residuals of a linear model regressing the log-returns on the Fama-French factors.

The 'excess returns' are therefore estimated using the following model

$$log(return)_{ij} = b_0 + b_1 \times SMB_i + b_2 \times HML_i + b_3 \times (Rm - Rf)_i + \epsilon_{ij} \tag{19}$$

where

(1) $log(return)_{ij}$ is the log daily return of stock j at time i

(2) SMB_i (Small Minus Big) is the average return on the three small portfolios minus the average return on the three big portfolios at time i https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_factors.html

(3) HML_i (High Minus Low) is the average return on the two value portfolios minus the average return on the two growth portfolios at time i https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_factors.html

(4) $(Rm - Rf)_i$, the excess return on the market at time i, value-weighted return of all CRSP firms incorporated in the U.S. and listed on the NYSE, AMEX, or NASDAQ that have a CRSP share code of 10 or 11 at the beginning of i's month, good shares and price data at the beginning of i's month, and good return data for i minus the one-month Treasury bill rate. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_factors.html

A.3.4 Checking model goodness-of-fit

Similarly to Section A.2.4, we produce diagnostic plots to confirm the validity of the linear-model and the Gaussianity and independence of its residuals.

Figure A.8 plots autocorrelation functions and partial autocorrelation functions, demonstrating that the observations can be considered independent and that there is no need to consider auto-regressive terms. Figure A.9 plots the fitted values $\hat{y}_{ij}$ and each of the predictors against the residuals ϵ_{ij}, demonstrating that the assumption that the covariates are linearly related to the response is satisfactory and that the residuals appear reasonably homoskedastic.

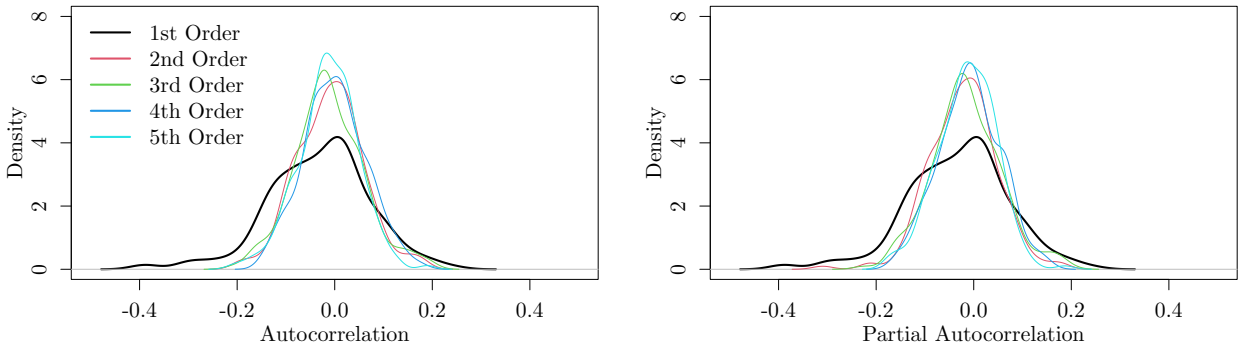


Figure A.8: Residual autocorrleation function and partial autocorrelation function for the stock market data.

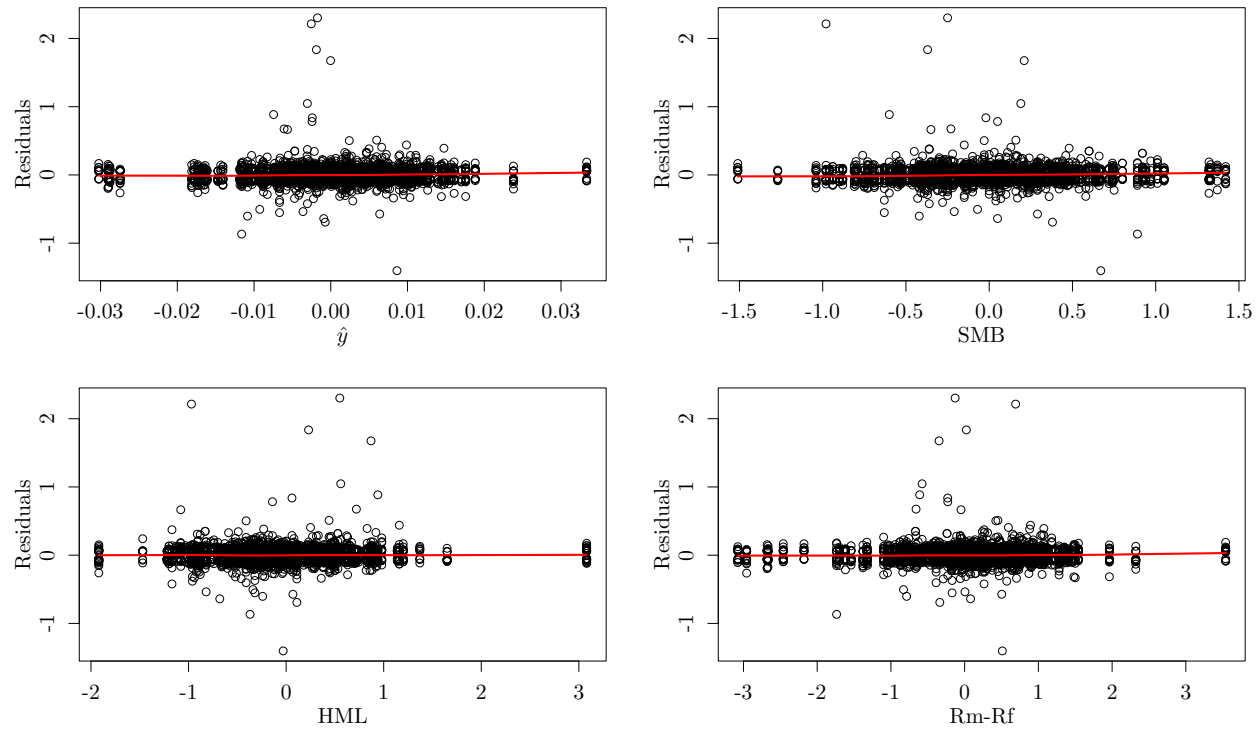


Figure A.9: Plots of the fitted values and each covariate against the residuals for the stock market data. The **red** line corresponds to the LOWESS smooth.

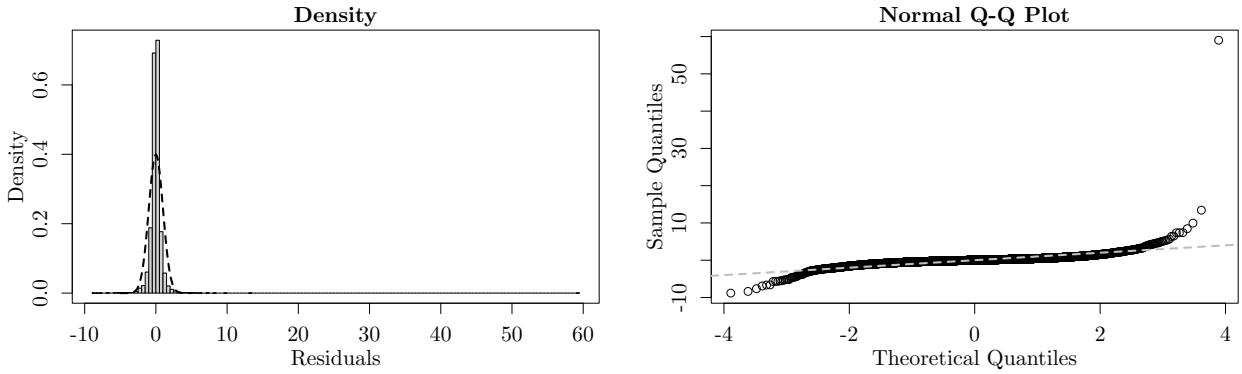


Figure A.10: Stock market data. **Left** Histogram of the standardised residuals compared with the standard Gaussian density. **Right** Q-Q Normal plot of the standardised residuals.

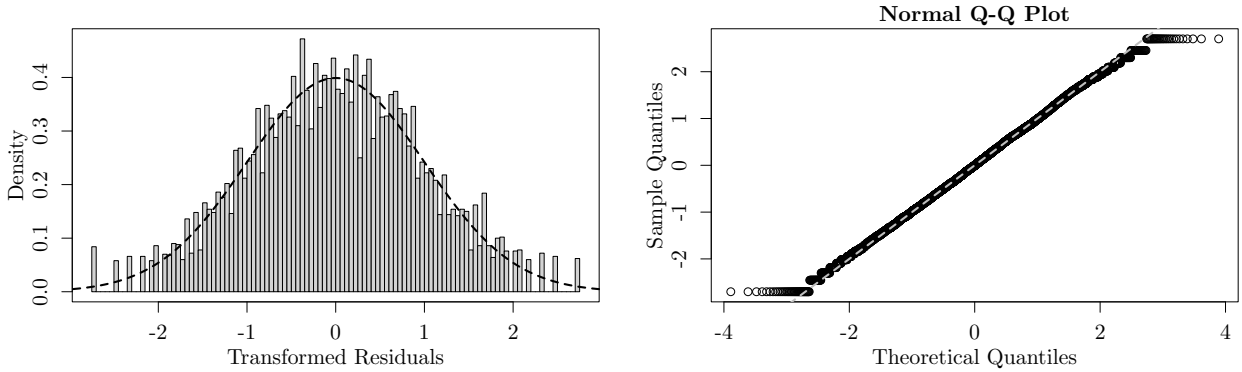


Figure A.11: Stock market data. **Left** Histogram of the transformed residuals compared with the standard Gaussian density. **Right** Q-Q Normal plot of the transformed residuals.

While the Gaussian assumption was tenable for the COVID-19 data, Figure A.10 shows that this is not the case for the stock market data. There is evidence of considerably heavier tails than Gaussianity. To address this issue we fit a non-paranormal model based on transforming the data into $f(\epsilon_i) := (f_1(\epsilon_{i1}), \ldots, f_p(\epsilon_{ip}))$, where $\hat{f}$ was estimated using the R package `huge`. Figure A.11 shows a histogram and Q-Q-normal plot for $f(\epsilon_i)$, where the Gaussian assumption is more tenable.

A.3.5 The network predictors

Based on the construction of Baker et al. (2019), we divided the 37 risk factors into two categories: the economic risks (containing 17 risks) and the policy risks (containing 20 risks). Then, for each risk type, we centered the $\log(1 + counts)$ and evaluated the Pearson's correlation between all pairs of

Table A.5: Stock market data: Edge counts of the network spike-and-slab model when declaring an edge for posterior slab probability > 0.5 and > 0.95

	Edges (> 0.5)	Non-Edges (> 0.5)	Edges (> 0.95)	Non-Edges (> 0.95)
Network SS	746	19154	80	19820

companies to obtain two network matrices E_{pears} and P_{pears}.

Figure A.12 demonstrates that under both networks there appears to be an increased chance of having positive partial correlation if the two firms have highly correlated risk factors. Figure A.13 demonstrates that no further transformation of the networks is required to satisfy the network GLASSO assumption of linearity.

The fitted spike-and-slab priors illustrated at the bottom of Figure A.12 show how the relationship between the networks and the graphical model is captured. The estimates parameterising these spike-and-slab distributions are available in Table 5. We see that as the economics network increases, the slab location increases, the slab scale increases, and also the probability of being in the slab increases. Alternatively, when accounting for the economic network, an increase in the policy network corresponds to an increased probability of being in the slab, but not an increase in the slab location or scale.

A.3.6 Supplementary figures

Table A.5 summarises the estimated graphical model under the network spike-and-slab model using a posterior slab probability threshold of > 0.5 and > 0.95. The number of edges estimated under the 0.5 slab probability threshold is slightly greater than the number of edges estimated under the network GLASSO models. Under the 0.95 slab probability threshold, the estimated number of edges is considerably more conservative.

Figure A.14 plots the marginal effect of each network on the slab location and the probability of being in the slab for each network, fixing the other at its mean, given the estimates of Table 5. This shows that while both networks cause both the slab location and the probability of being in the slab to increase, the Economic network increases the location of non-zero partial correlations to a greater extent than the Policy network, the Policy network increases the probability of having a non-zero partial correlation to a greater extent.

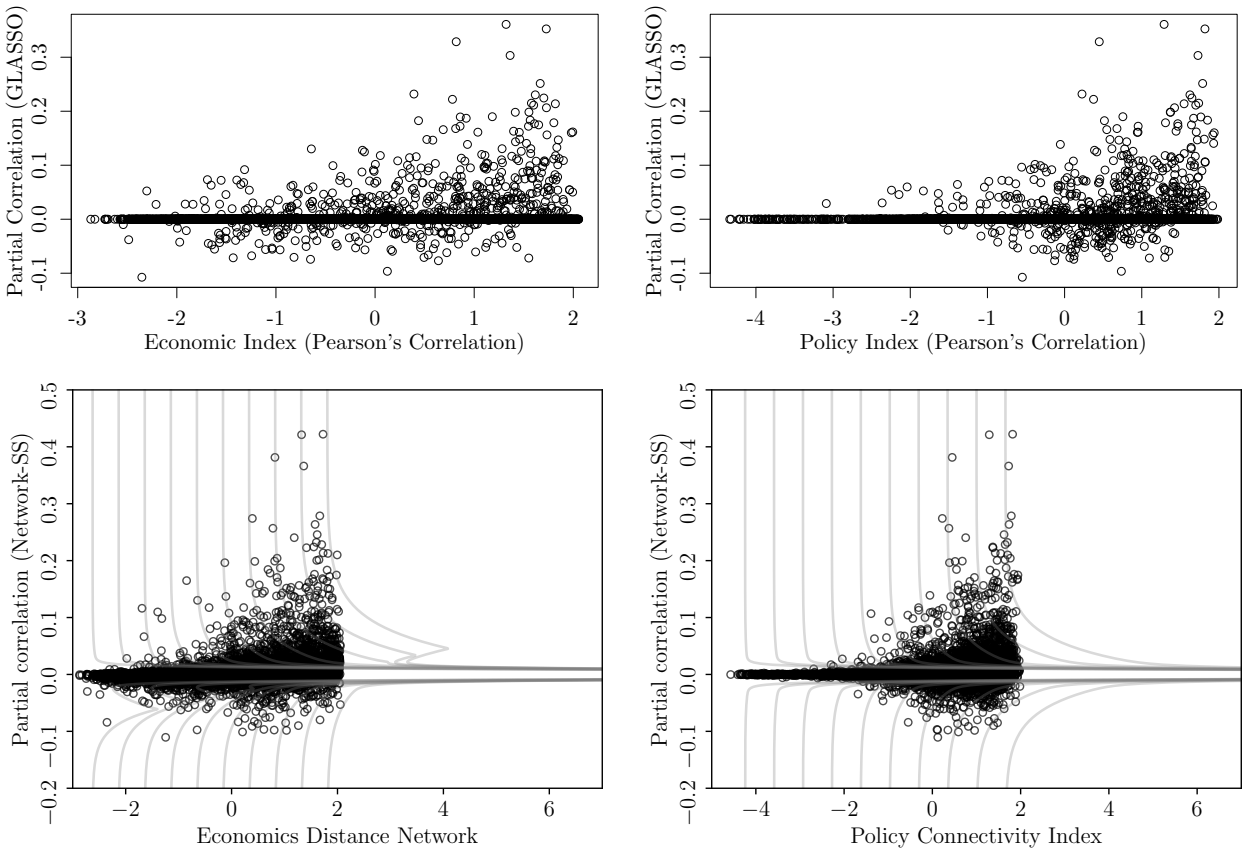


Figure A.12: Residual partial correlations of the stock market excess returns across firms vs Economy risk (left) and Policy risk (right). Top panel: Partial correlations were estimated with GLASSO, with penalization parameter set via BIC. Bottom panel: fitted spike-and-slab distributions and fitted partial correlations estimated with network spike-and-slab model.

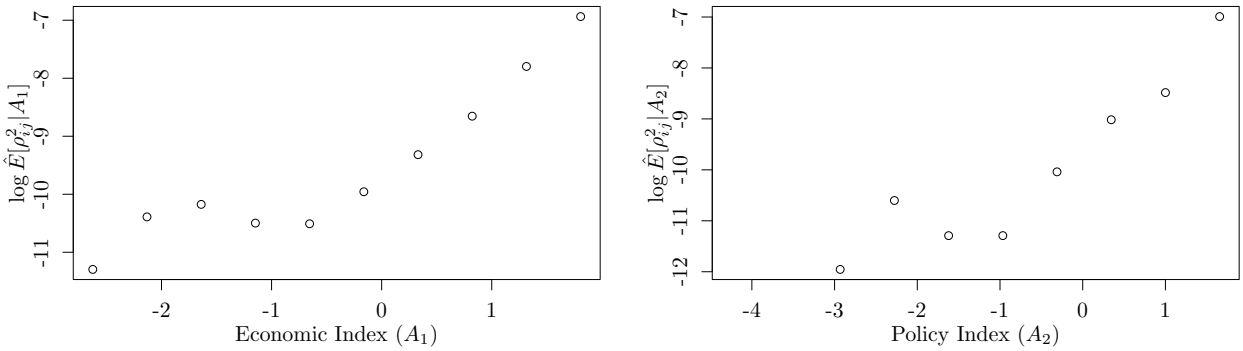


Figure A.13: Assessing the linear relation between $\log \mathbb{E}[\hat{\rho}_{jk}^2|A]$ and the network matrices, where $\hat{\rho}_{jk}$ is the GLASSO estimate. The points represent the log-mean values of $\hat{\rho}_{jk}^2$ within 10 equispaced bins defined for each network.

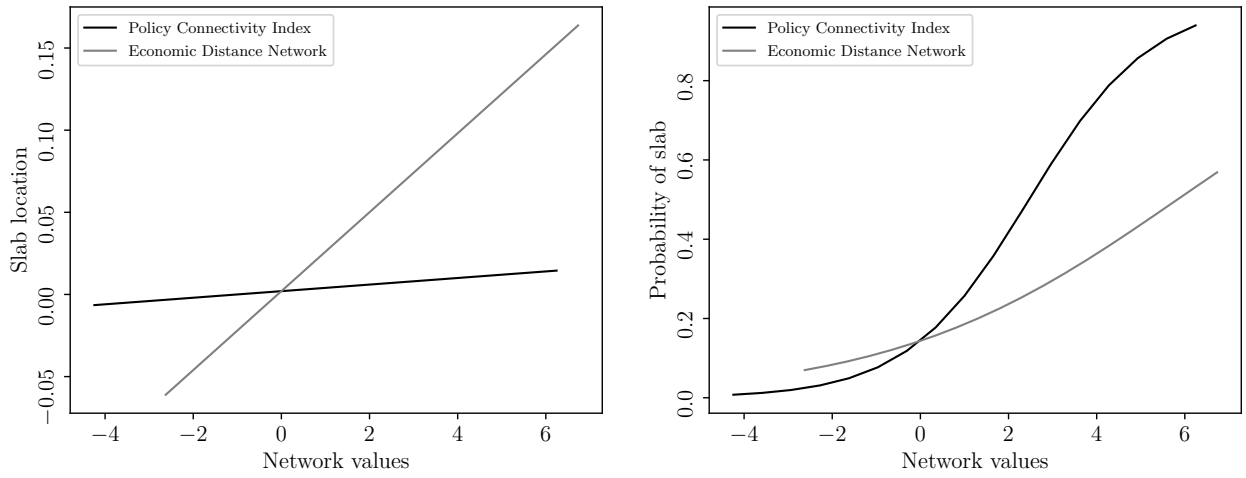


Figure A.14: Stock-market data: Slab location (**left**) and slab probability (**right**) as a function of both networks estimated by empirical Bayes. We consider the slab here as the location of the partial correlations, the negation of the ρ model parameters, which is why negative hyperparameter for η_0 (Table 5) result in positive effects.

A.3.7 Results using the EBIC

Table A.4 presents results investigating the stability of our stock market data analysis to selecting hyperparameters using the EBIC with $\gamma_{\text{EBIC}} = 0.5$ rather than the BIC. The EBIC continues to estimate sparser networks than the BIC, but to the detriment of the out-of-sample test set score. Importantly, we see that the improvement of the network GLASSO methods over standard GLASSO is still apparent when using the EBIC selection criteria.

Table A.6: Four models for the stock market data when using the EBIC ($\gamma_{\text{EBIC}} = 0.5$) to learn the network hyperparameters. A_1 is the Economic network, A_2 the Policy network. EBIC values account for the extra hyper-parameters in the network GLASSO models. 10-fold is the 10-fold cross-validation log-likelihood.

Method	EBIC	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	Edges	10-fold
GLASSO	50008.018	-0.868			93	-7020.84
Network GLASSO- A_1	49632.192	-0.237	-0.605		109	**-6993.37**
Network GLASSO- A_2	49723.101	1.632		-1.868	100	-7022.70
Network GLASSO- A_1 & A_2	**49587.239**	0.889	-0.361	-1.028	78	-7025.53

A.4 `Stan` vs `NumPyro`

We estimated our network spike-and-slab models using the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014), an extension of Hamiltonian Monte Carlo (HMC, Duane et al. 1987) that automates the setting of the step-size in the Hamiltonian discretisation. Two probabilistic programming implementations of NUTS are `Stan` (Carpenter et al., 2017) and `NumPyro` (Bingham et al., 2019; Phan et al., 2019). We provide implementations of our algorithm in both languages, but for our experiments, we found `NumPyro`'s ability to take advantage of parallel computing for automatic differentiation provided a considerable speed up.

We illustrate this using one of our simulated examples from Section 5. We consider $p = 10$, $n = 100$ and network matrix $A_{0.85}$. We ran both `Stan` and `NumPyro` for 2000 warmup iterations and 2000 sampling iterations. Table A.7 compares the time taken to sample and the effective sample size (ESS) of the resulting sample averaged across 10 repeat datasets. We present the ESS as averaged across the ρ model parameters and the network hyperparameter η. We see that both methods produce similar ESS but that `NumPyro` does so over ten times faster.

We also take this opportunity to demonstrate how efficient the network GLASSO is when implemented as a special case of the GOLAZO algorithm (Lauritzen and Zwiernik, 2020). For the same datasets considered above, we implement the network GLASSO using 50×50 grid search to estimate the network hyperparameters. We see that the GOLAZO algorithm takes a fraction of the time to run as the Bayesian implementation even when using a rudimentary grid-search optimisation scheme.

Table A.7: Comparison of time taken for network GLASSO implemented using the GOLAZO algorithm and the network spike-and-slab sampling algorithms in `Stan` and `NumPyro`.

	Time (s)	ESS ρ's	ESS η's
GOLAZO	8.58	-	-
`Stan`	537.05	1089	555
`NumPyro`	47.15	1056	601

Lastly, above we limited `NumPyro`'s access to only 6 cores on one machine. Using more cores, for example on a GPU, provides the potential for `NumPyro` to achieve even greater speed-ups for higher dimensional problems beyond the simple one considered here.