Separating the wheat from the chaff

Karimova, Diana; Leenders, Roger; Meijerink-Bosman, Marlyne; Mulder, Joris

Published in:
Social Networks

DOI:
10.1016/j.socnet.2023.02.006

Publication date:
2023

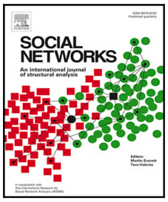
Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Karimova, D., Leenders, R., Meijerink-Bosman, M., & Mulder, J. (2023). Separating the wheat from the chaff: Bayesian regularization in dynamic social networks. *Social Networks*, *74*, 139-155. https://doi.org/10.1016/j.socnet.2023.02.006

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet



Separating the wheat from the chaff: Bayesian regularization in dynamic social networks

Diana Karimova [a,*], Roger Th.A.J. Leenders [b,c], Marlyne Meijerink-Bosman [a], Joris Mulder [a,b]

[a] *Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, Netherlands*
[b] *Jheronimus Academy of Data Science, Netherlands*
[c] *Department of Organization Studies, Tilburg School of Social and Behavioral Sciences, Tilburg University, Netherlands*

ARTICLE INFO

Keywords:
Bayesian regularization
Shrinkage priors
Bayesian lasso
Horseshoe prior
Relational event data

ABSTRACT

In recent years there has been an increasing interest in the use of relational event models for dynamic social network analysis. The basis of these models is the concept of an "event", defined as a triplet of time, sender, and receiver of some social interaction. The key question that relational event models aim to answer is what drives the pattern of social interactions among actors. Researchers often consider a very large number of predictors in their studies (including exogenous effects, endogenous network effects, and interaction effects). However, employing an excessive number of effects may lead to overfitting and inflated Type-I error rates. Moreover, the fitted model can easily become overly complex and the implied social interaction behavior difficult to interpret. A potential solution to this problem is to apply Bayesian regularization using shrinkage priors to recognize which effects are truly nonzero (the "wheat") and which effects can be considered as (largely) irrelevant (the "chaff"). In this paper, we propose Bayesian regularization methods for relational event models using four different priors for both an actor and a dyad relational event model: a flat prior model with no shrinkage, a ridge estimator with a normal prior, a Bayesian lasso with a Laplace prior, and a horseshoe prior. We apply these regularization methods in three different empirical applications. The results reveal that Bayesian regularization can be used to separate the wheat from the chaff in models with a large number of effects by yielding considerably fewer significant effects, resulting in a more parsimonious description of the social interaction behavior between actors in dynamic social networks, without sacrificing predictive performance.

1. Introduction

Relational event history data are becoming increasingly available, partly due to increasing use of technology-supported interaction, such as email, phone calls, online social networks (Twitter, Facebook, et cetera), or sociometric badges. Relational event data encode who does what with respect to whom at what point in time. Typically, relational event data contain information of the exact timing (or order) of interactions, who was the sender, who was the receiver, and, possibly, what was the mode of communication (e.g., face-to-face or digital), the sentiment (e.g., positive or negative), the content, et cetera. Additionally, information about attributes of the involved actors is also often available, for example gender, group memberships, or the hierarchical position of an actor in an organization. Finally, information about possible external influences to the event history, such as deadlines, the start of new projects, or time-of day/day-of-week may be available. Due to their high-resolution, relational event data can potentially greatly support our understanding of complex interaction processes, of temporal effects of interventions in networks, or of how past interactions may affect what will happen in the (nearby) future.

Relational event models (initially proposed by Butts (2008), and later extended by DuBois et al. (2013), Quintane et al. (2014), Leenders et al. (2016), Pilny et al. (2016), Vu et al. (2017), Stadtfeld et al. (2017), Mulder and Leenders (2019) and Lerner and Lomi (2020), among others) have become widely used over the past decade for analyzing relational event data. In the relational event model, the outcome variable is the rate of interaction between potential senders and receivers in the network at a given point in time. This rate of interaction is commonly assumed to be a log linear function of a set of predictor variables (at that given point in time). Following Leenders et al. (2016), predictor variables can be categorized as being endogenous (ie., summarizing the past activity of the actors in the network) or exogenous (ie., capturing actor attributes or external influences). By modeling the interaction rate between actors, the relational event

[*] Correspondence to: Warandelaan 2, Tilburg, 5037 AB, Netherlands.
E-mail address: d.karimova@uvt.nl (D. Karimova).

https://doi.org/10.1016/j.socnet.2023.02.006

Available online 28 March 2023

model can predict the next event: given the past interactions among the actors and the characteristics of the actors, at a given point in time, estimated rates predict who will be involved in the next interaction (and when will it occur). In relational event models, this is achieved by finding a loglinear combination of endogenous predictors, exogenous predictors, and possible interactions that best (or, sufficiently) model the rates of interaction of every event that is possible at a given time. In additional to directly modeling the interaction rate between dyads, actor relational event models are also available where we separately model who is going to be the sender of the next event and who is likely to become the receiver of the next event given this sender.

Building a model for social interaction is an inherently highly complex and complicated endeavor. Although researchers may build their model based on some theory or on previous findings, it is often extremely complicated to put together the exact effects that are able to not only capture who interacts with whom, but also correctly captures the order and timing at which each event occurs vis-à-vis the others. Just like with other statistical network models, such as (S)(T)ERGM's or SIENA-models, a well-fitting model usually requires active model selection and variable selection by the researcher. Often, researchers start with a set of predictors (inspired by theory or previous findings), remove those that are not statistically significant and add new ones that might improve accuracy further, until the researcher is satisfied with the model fit. Considering that relational event models (especially as the number of actors and the number of events grow) can take quite some time to run, this is a potentially time-consuming approach (and makes the interpretation of t-statistics questionable). Alternatively, researchers may specify a set of potential models and compare their performance based on measures such as AIC and BIC. Again, this can be a time and resource-intensive approach. Moreover the final model may depend on individual choices made by the researcher which may be difficult to reproduce.

Variable selection algorithms have not yet been thoroughly developed for relational event models, but researchers have developed some ways to help the variable selection process. For example, Butts (2008) proposed a model for explaining radio communication messages between emergency transponders during the 9/11 World Trade Center disaster. To make decisions about which variables to include, Butts (2008) utilized the BIC, a model selection criterion that balances model fit and model complexity (via the number of predictor variables). In most cases, however, it is not computationally feasible to compute the BIC for all possible models since the number of possible models to consider increases exponentially with the number of predictors K via 2^K. Therefore, this is generally not computationally feasible for relational event models with many predictor variables. Hence, in practice researchers only compare a few competing models. This choice of which models to compare is inherently somewhat arbitrary and may be driven by computational burden (the longer it takes for a model to run, the fewer models can feasibly be compared).

Another potential solution for variable selection problems exists in a form of penalized or regularized regression. In penalized regression, the optimization problem of finding the best fitting estimates for the coefficients (say, by minimizing the sum of squared errors) is replaced by a constrained optimization problem using a penalty with respect to the total magnitude of all estimated coefficients. For example, in the *least absolute shrinkage and selection operator* (lasso; Tibshirani, 1996), the penalty function is equal to the sum of the absolute values of all of the coefficients in the model, i.e., $\sum_p |\beta_p|$, and the following restricted maximization problem needs to be solved:

$$\text{maximize}_{\beta}\left\{\mathcal{L}(data|\boldsymbol{\beta})\right\} \text{ subject to } \sum_{p=1}^{P}|\beta_p| \leq t, \quad (1)$$

where $\mathcal{L}$ is the likelihood of the data given the unknown parameters, and t can be viewed as the "budget" the model needs to stay within (Hastie et al. (2016), 2016, Ch. 3). For example, when considering two parameters, (β_1, β_2) with unconstrained maximum likelihood estimates of $(.3, 2)$ and a budget of $t = 2$, the estimates $(0, 2)$ would be acceptable and $(.3, 2)$ would not be. In this small example, we obtain a simpler solution with only one nonzero estimate instead of two by applying penalized regression rather than standard regression. The resulting, more parsimonious model thus highlights which effects really matter, without the "clutter" of many effects that do not contribute much statistically. So even though the solution $(.3, 1.7)$ would also fit within the "budget" of the penalty above, such solutions generally result in a smaller likelihood of the data (i.e., a worse fit) than the more parsimonious solution $(0, 2)$ where the large effect of β_2 is left unaffected. Of course, the lasso is only one kind of penalized regression and many other penalty terms have been proposed in the literature. In this paper, we will explore three ways of penalizing the parameters in the relational model and show how they differ in which estimates are shrunk towards zero and how strongly. Moreover, both frequentist and Bayesian regularization approaches have been shown to effectively guard against overfitting and to result in good predictive performance (Tibshirani, 1996; Park and Casella, 2008; Kyung et al., 2010; Van Erp et al., 2019). In this paper, we therefore contend that a useful and statistically sound alternative to iterative model and variable selection approaches is to specify a single large model (i.e. include all potential predictors of interest, regardless of multicollinearity) and then use regularization to separate the wheat from the chaff and end up with a parsimonious model that is easier to interpret than the initial (much) larger model.

In this paper, we introduce Bayesian regularization approaches for relational event models. In Bayesian regularization, the prior distribution of the coefficients, $p(\boldsymbol{\beta})$, serves a similar purpose as the penalty function in classical penalization methods. The prior reflects which values of the unknown parameters are likely or unlikely before observing the data. Thus, if the prior is concentrated near zero, *on average* we expect coefficients to be close to 0 a priori. This is a realistic assumption when considering a relational event model with many predictor variables of which many can potentially have a negligible effect (but we may not know which ones, beforehand). Intuitively, such priors will therefore shrink small, negligible effects towards zero yielding a more parsimonious result. Moreover, if a prior is used with relatively thick tails, large estimated effects will remain mostly unaffected by the shrinkage behavior towards zero. As a result, small effects will vanish (i.e. they are shrunk towards zero), while large effects are retained. This is what we call to separate the wheat from the chaff (statistically).

In Bayesian regularization, statistical inferences are based on the posterior distribution which is proportional to the product of the likelihood function and the prior:

$$p(\boldsymbol{\beta}|data) \propto \mathcal{L}(data|\boldsymbol{\beta}) \times p(\boldsymbol{\beta}),$$

where $p(\boldsymbol{\beta}|data)$ is the posterior that reflects which values for $\boldsymbol{\beta}$ are likely after observing the data. The posterior mode is then obtained via

$$\text{maximize}_{\beta}\left\{\mathcal{L}(data|\boldsymbol{\beta}) \times p(\boldsymbol{\beta})\right\}. \quad (2)$$

As shown by Park and Casella (2008), a Bayesian counterpart to the lasso in (1) is obtained by using a prior with a Laplace distribution for each parameter, i.e., $Laplace(\beta_p|\lambda) = \lambda/2\exp(-\lambda|\beta_p|)$, where λ denotes a penalty or shrinkage parameter, which has a similar role as the budget t in the regular lasso. Fig. 1 (dotted line) shows the Laplace distribution with a clear peak at 0 and relatively thick tails. When using the Laplace prior, it can be shown that the posterior mode is identical to the solution of the lasso in standard penalized regression for a specific choice of λ and t. The Laplace prior as well as the other priors in this figure will be discussed in more detail later in this paper.

The Bayesian approach to regularization has several attractive properties. First, it performs competitively and sometimes better than its classical counter parts (in terms of *predictive mean squared error*) and results in more accurate uncertainty bounds using the full posterior

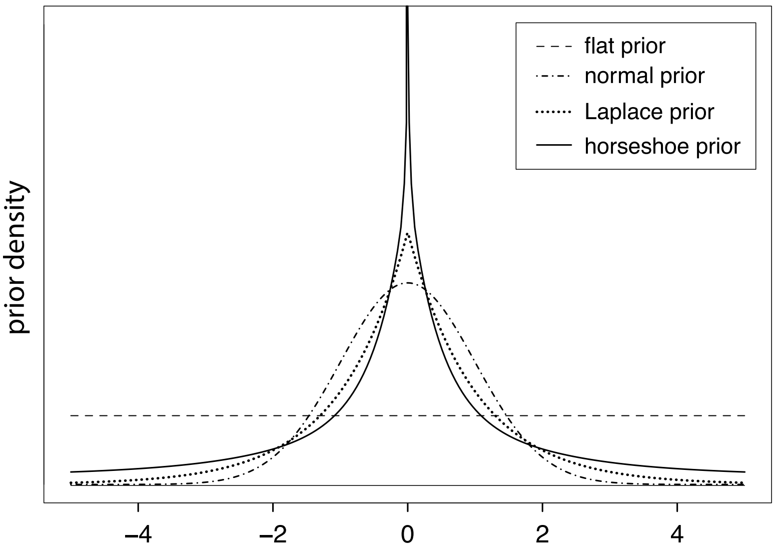


Fig. 1. Flat prior (dashed line), normal prior (dash-dotted line), Laplace prior (dotted line), and horseshoe prior (solid line).

instead of using bootstrapped standard errors in the regular lasso (Park and Casella, 2008). Second, Bayesian approaches are quite flexible regarding the choice of the prior. Standard penalized regression, on the other hand, can be challenging to optimize when the target function (i.e., the combination of the likelihood function and the penalty function) is not convex (Hastie et al., 2015, p. 2). Third, in the Bayesian approach we can learn all model parameters (including the penalty parameter λ) in one step while regular penalized regression requires two-step procedures where the penalty parameter needs to be learned in a second step using cross-validation where the data are split into arbitrarily-sized subsets. On the other hand, Bayesian regularization algorithms are generally computationally more expensive than their non-Bayesian counterpart. In this paper, we minimize the computational burden using so-called conjugate priors for which the conditional posteriors have known probability distributions that we can sample from directly.

In this paper, we discuss the implementation of Bayesian regularization algorithms and illustrate their use for both a dyadic and an actor relational event model. We use probit relational event models (which assume a normal latent scale) throughout the paper due to their computational benefit (they allow the shrinkage priors to be written as scaled mixture of normal distributions). We present three empirical relational event data sequences to illustrate Bayesian regularization: the Enron email data, the voice loops between actors during the infamous Apollo 13 mission to the moon, and a sequence of social interactions between team members based on proximity data. In Section 2 of this paper we explain two partial likelihood approaches to relational event modeling: one for a dyadic relational events and one for actor events. In Section 3 we discuss the three alternative priors for Bayesian regularization, and in Section 4 we illustrate the methodology with three empirical examples. We conclude and provide final discussions in Section 5.

In sum, the aim of this paper is to show how relational event models with a large number of effects can be fitted and be regularized to a more parsimonious size, where statistically unimportant effects (i.e. the "chaff") are shrunk to zero, while retaining the effects that are statistically important to the modem (i.e. the "wheat"). The researcher can use these methods even when the effects are highly collinear. The empirical illustrations show that this model reduction process need not reduce the predictive accuracy of the resulting, parsimonious model compared to the larger initial model and can sometimes even improve it by reducing noise from the model.

2. Relational event modeling using partial likelihoods

The relational event model was popularized by Butts (2008), who modeled a relational event history as a Poisson process where the event rate for each specific dyad depends on a set of endogenous and exogenous effects through a loglinear function. Instead of working with the full likelihood for a joint model for all event times, senders, and receivers, in this paper we adopt the idea of partial likelihoods (Cox, 1972; Perry and Wolfe, 2013) that only considers specific parts of a conditional likelihood. By working with partial likelihoods, we simplify the specification of the model by focusing on the outcome variables that are of most interest for a given application. Below, we first present a partial likelihood for the actor model, based on Perry and Wolfe (2013) and Stadtfeld et al. (2017). This approach may be preferred when a researcher is interested in modeling the choice of the receiver of an event conditional on the sender (Vu et al., 2017; Stadtfeld and Block, 2017; Hoffman et al., 2020; Hedström and Bearman, 2009). Second, we provide a partial likelihood for a dyadic relational event model. The dyadic REM directly models the dyad (i.e., the combination of sender and receiver) (Leenders et al., 2016; Brandes et al., 2009; Malang et al., 2019; Liang, 2014; Lerner and Lomi, 2018). For partial likelihoods we consider a probit regression model due to its Gaussian latent scale which, in combination with (conditional) Gaussian shrinkage priors (discussed in Section 3), results in a computationally efficient model using Markov chain Monte Carlo (MCMC) algorithms.

2.1. A partial likelihood for an actor model

Using the notation of events $e_m = (t_m, s_m, r_m)$, $m \in \{1, \ldots, M\}$ as a triplet of time t, sender s, and receiver r, we can write the likelihood of the sequence of events as a product of conditional likelihoods:

$$\begin{aligned} L(e_1, \ldots, e_M) =& L(e_1)L(e_2, \ldots, e_M | e_e) \\ =& L(e_1)L(e_2|e_1) \cdot \ldots \cdot L(e_M|e_1, \ldots, e_{M-1}) \\ =& L(t_1, s_1)L(r_1|t_1, s_1)L(t_2, s_2|t_1, s_1, r_1)L(r_2|t_2, s_2, e_1) \cdot \ldots \cdot \\ & \cdot L(t_M, s_M|e_{M-1}, \ldots, e_1)L(r_M|t_M, s_M, e_{M-1}, \ldots, e_1) \end{aligned} \tag{3}$$

In an actor approach, the focus is on the choice of the receiver for a given sender at a given point in time. In most research projects, understanding who will be the receiver of an event for a given sender is more informative than modeling who will be the sender. In this paper we therefore consider the following partial likelihood of the receivers of the events conditional on the senders and event times (following directly from Eq. (3)):

$$PL(\mathbf{r}|\mathbf{s},\mathbf{t}) = L(r_1|t_1,s_1)\cdot L(r_2|t_2,s_2,e_1)\cdot\ldots\cdot L(r_M|t_M,s_M,e_{M-1}) \tag{4}$$

The partial likelihood in (4) can be seen as a statistical choice model where the sender "chooses" the most suitable receiver from the set of possible receivers.

In this paper, we consider a Bayesian probit model using a Gaussian latent variable approach by extending the work of Imai and Van Dyk (2005) to relational event data. For each event e_i in a sequence $\{e_1,\ldots,e_M\}$, where M is a total number of events, we define a categorical outcome variable Y_i that represents a receiver of the event e_i. This receiver of event i can be any actor in the *risk set* $\mathcal{R}_{actor}$: the set of actors who are possible receivers of a given event. As is common in social network analysis, we assume that all actors, except for the sender, are at risk. In a latent variable approach, this means that the sender assigns a latent score to all potential receivers in the risk set. We denote the latent score that sender i assigns to potential receiver r by Z_{ir}. The receiver r with the largest score Z_{ir} will be the predicted receiver of the event:

$$Y_i(Z_i) = r, \text{ if } \max(Z_i) = Z_{ir}, \tag{5}$$

where $Z_i = (Z_{i1},\ldots,Z_{iN})$ is a multivariate latent variable for N actors. In the framework of the multivariate probit model, we can write

$$Z_i = X_i\boldsymbol{\beta} + \epsilon_i, \tag{6}$$

where X_i is a $N \times P$ matrix of observed predictor variables at time i, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_P)^T$ is a vector of network parameters, and ϵ_i is a Gaussian error term, centered at zero, having an identity covariance matrix (to ensure identifiability of the model). The matrix $X_i, i = 1,\ldots,M$ of predictor variables can include endogenous as well as exogenous predictors, defined for each actor in the risk set.

2.2. A partial likelihood for a dyadic model

Often, the interest is in jointly modeling the combination of sender and receiver (Leenders et al., 2016; Brandes et al., 2009; Malang et al., 2019; Liang, 2014; Lerner and Lomi, 2018). Thus, for a dyad model we build a statistical model for all possible dyads that can be observed at a given point in time. Starting from the full likelihood, but redefining the conditional likelihood in such a way that we condition on the time points of events, we get the following representation of the likelihood:

$$\begin{aligned} L(e_1,\ldots,e_M) =& L(e_1)L(e_2,\ldots,e_M)\\ =& L(e_1)L(e_2|e_1)\cdot\ldots\cdot L(e_M|e_1,\ldots,e_{M-1}) = \cdots\\ =& L(t_1)L(s_1,r_1|t_1)L(t_2|t_1,s_1,r_1)L(s_2,r_2|t_2,e_1)\cdot\ldots\cdot\\ &\cdot L(t_M|e_{M-1},\ldots,e_1)L(s_M,r_M|t_M,e_{M-1},\ldots,e_1)\end{aligned} \tag{7}$$

A partial likelihood for a dyad REM can then be written as follows:

$$PL(\mathbf{s},\mathbf{r}|\mathbf{t}) = L(s_1,r_1|t_1)\cdot L(s_2,r_2|t_2,e_2)\cdot\cdots\cdot L(s_M,r_M|t_M,e_{M-1}) \tag{8}$$

This can be viewed as a dyadic partial likelihood for the REM where the sender and receiver for event e_i are jointly modeled.

In contrast to the actor model, in the dyadic model the outcome variable is the rate of the occurrence of a dyad. In particular, Y_i is defined as an index of the dyad (s,r) from the risk set $\mathcal{R}_{dyad}$ of all the $N(N-1)$ possible ordered dyads. Following the idea of a multivariate probit model, we assume that all dyads that are at risk lie on a latent scale where the dyad with the largest latent score becomes the dyad that is predicted to occur next:

$$Y_i(W_i) = l(s_i,r_i), \text{if } \max(W_i) = l(s_i,r_i),$$

where $l(s_i,r_i) \in \{1,\ldots,N(N-1)\}$ is the index of dyad (s_i,r_i) in the ordered risk set $\mathcal{R}_{dyad}$. Therefore, the latent vectors W_i have length $N(N-1)$, under the condition that an actor cannot send an event to oneself. Note that the dyadic relational event model can be computationally costly for large networks as there will be many unknown latent variables to be estimated.

We write the regression for the dyad REM as follows:

$$W_i = X_i\boldsymbol{\beta} + \epsilon_i,\ i = 1,\ldots,M \tag{9}$$

where matrices X_i are of dimension $N(N-1)\times P$ (containing the dyadic predictor variables) and $\boldsymbol{\beta}$ is the corresponding vector that quantifies the relative importance of the predictors. As in the actor model, the set of potentially important dyadic predictor variables is allowed to be huge. To find the true nonzero effects, a Bayesian regularization algorithm will be proposed, as discussed in the next section.

3. Bayesian regularization via shrinkage priors

A complete Bayesian model combines the statistical model for the data with a prior distribution for the network parameters $\boldsymbol{\beta}$[1]:

$$\text{model}: \begin{cases} Y_i(Z_i) &= j, \text{ if } \max(Z_i) = Z_{ij},\\ Z_i &= X_i\boldsymbol{\beta} + \epsilon_i, \text{ with } \epsilon_i \sim N(0,I)\end{cases}$$
$$\text{prior}: p(\boldsymbol{\beta})$$

Even though one might think that priors are only used for including external information about the unknown parameters in the analysis (e.g. based on expert knowledge or previous empirical findings), specific choices of the prior can also result in desirable shrinkage behavior where negligible effects are shrunk towards zero while leaving large effects mostly unaffected. Such priors are often called "shrinkage priors". Using such priors is useful for Bayesian regularization as they result in more parsimonious models than when using noninformative priors that do not induce any shrinkage behavior. To allow the same type of shrinkage behavior for negative and positive effects, the prior should have a symmetric form with a peak at zero (to shrink small effects) while having sufficient probability mass allocated in the tails (to leave large effects largely unaffected).

Different types of priors can be used for this purpose. For a recent overview, see Van Erp et al. (2019). In the current paper, we consider three of the most popular shrinkage priors from the Bayesian regularization literature: a Gaussian (normal) prior (for Bayesian ridge regression; Fig. 1, dashed–dotted line), a Laplace prior (for Bayesian lasso regression; Fig. 1, peaked dotted line), and a horseshoe prior (Fig. 1, sharply peaked solid line). We add a flat (horizontal) prior (Fig. 1, dashed line) that does not perform any shrinkage (as it assumes that all values of the parameters are equally likely a priori). The results of the flat prior are comparable to the results from maximum likelihood estimation, making it a suitable prior for a reference analysis. The figure shows some intuition about the different shrinkage behavior of the three different priors. For example, we see that the horseshoe prior has the sharpest peak (in fact it has a pole at zero), followed by the Laplace prior, and finally the normal prior has the smallest peak. This suggests that the horseshoe prior results in heaviest shrinkage near zero, while the normal prior results in the least shrinkage. Furthermore, because the horseshoe prior has the thickest tails and the normal prior has the thinnest tails, the horseshoe prior will leave large effects mostly unaffected, followed by the Laplace prior, while the normal prior is expected to also shrink large effects a bit towards zero.

In addition to the shape of the shrinkage priors, the variance (or scale) of the prior also has a direct effect on the amount of shrinkage in the model: a large (small) prior variance induces little (considerable) shrinkage. This prior variance is controlled via the *shrinkage parameter* and is denoted as λ^2. Ideally, when there are many large effects, the shrinkage parameter should be large (so the large effects are left intact), and when there are hardly any large effects, the shrinkage parameter

[1] In this section we present the priors in the context of an actor model. The priors for the dyadic model are mathematically equivalent.

should be small (so the small effects are nudged towards zero). The optimal value of the shrinkage parameter for a given data set can be found using two-step approaches such as cross-validation or empirical Bayes methods (e.g., Park and Casella, 2008). In a Bayesian framework, however, it is a more natural choice to estimate the shrinkage parameter jointly with the other parameters in one single step. This does require one to specify a separate prior density for the shrinkage parameter λ^2, we will discuss this later in this paper.

To fit the Bayesian regularized relational event models, we use Markov Chain Monte Carlo (MCMC) methods to sample the parameters from the joint posterior. The approach is to sample the model parameters sequentially from their conditional posterior distributions. Gibbs sampling is an efficient MCMC algorithm where the conditional posterior distributions of the parameters belong to known distributional families from which we can directly sample. This is the case when the priors have the same distributional form as the likelihood (this is known as "conjugacy"). Because of the Gaussian distribution under the probit model in the partial likelihoods from Section 2, Gaussian priors for $\boldsymbol{\beta}$ result in conditional posteriors that also have Gaussian distributions. Interestingly, the Laplace prior and the horseshoe prior both can be written as scaled mixtures of Gaussian priors making posterior sampling in a Gibbs sampler straightforward and efficient. We use F priors for the shrinkage parameters as they are relatively vague while allowing easy posterior sampling using Gamma and Inverse distributions (Mulder et al., 2018). This is equivalent to choosing a half-Cauchy prior for λ, which is a common choice (Carvalho et al., 2009). In the next section, we discuss each Bayesian shrinkage model for relational event analysis in detail.

3.1. Flat prior (no shrinkage)

We first consider a benchmark model with no shrinkage effect. This model utilizes a flat improper prior that assumes that all values for the regression coefficients vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$ are equally likely a priori. Mathematically, this can be written as

$$p^{FLAT}(\boldsymbol{\beta}) \propto 1, \tag{10}$$

The prior density is constant over the complete real line (see the dashed line in Fig. 1). The prior does not shrink the regression coefficients: the estimates of $\boldsymbol{\beta}$ will be entirely driven by the data. The Bayesian flat prior model, therefore, behaves very similar to classical MLE estimation. Given an observed event history we can estimate the posterior distribution of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$. Because the latent variable has a multivariate normal distribution, the model can be estimated using a Gibbs sampler. We describe the Gibbs sampler algorithm in Appendix. This algorithm can be used to acquire a large sample from the posterior for all parameters, which can be used for statistical inference. For example, by taking 2.5% and 97.5% posterior quantiles, we obtain the bounds of the 95% Bayesian credibility interval. If zero is not contained in the interval this suggests that the parameter should be included in the model. To obtain a point estimate of each model parameter, we use the posterior mode (which reflects the most plausible value of a parameter after observing the data).

3.2. Bayesian ridge prior

Ridge regression was originally developed to improve estimates of the classic least squares model, especially in the case when there is high correlation among predictors. The model utilizes a modified variance matrix $X'X + \lambda^2 I$ that adds a quadratic penalty. In Bayesian ridge regression, a normal prior is used for the regression coefficients:

$$p^{RIDGE}(\boldsymbol{\beta}|\lambda^2) = \prod_{p=1}^{P} p(\beta_p|\lambda^2) = \prod_{p=1}^{P} \mathcal{N}(\beta_p|0, \lambda^2), \tag{11}$$

where $\mathcal{N}(\beta_p|0, \lambda^2)$ denotes a normal prior for β_p with mean 0 and variance λ^2.

We plot this prior in Fig. 1 using a dash-dotted line. The prior is centered around zero with relatively thin tails. Shrinkage is performed over the entire domain of parameters due to the structure of normal prior density: large values will be shrunk to the same degree as are small values.

To complete the Bayesian model, we need a prior for the shrinkage parameter λ^2. A common prior for this purpose is the gamma distribution (Park and Casella, 2008). However, the hyperparameters of the gamma prior may considerably affect its results (Kyung et al., 2010). For this reason, we use a half-Cauchy prior for λ instead, which is quite vague due to its thick tails. A half-Cauchy prior for λ is equivalent to an F prior for λ^2 (e.g., Mulder et al., 2018), which has density:

$$F(\lambda^2; \alpha_1, \alpha_2, b) = \frac{\Gamma(\frac{\alpha_1+\alpha_2}{2})}{\Gamma(\frac{\alpha_1}{2})\Gamma(\frac{\alpha_2}{2})} b^{-\alpha_2/2} \left(\frac{\lambda^2}{b}+1\right)^{-\frac{\alpha_1+\alpha_2}{2}} (\lambda^2)^{\alpha_2/2-1} \tag{12}$$

We set the hyperparameters to 1, which is the default minimally informative choice:

$$\lambda^2 \sim F(1,1,1)$$

Using a parameter expansion, the F distribution can be written as a scale mixture of inverse gamma distributions via

$$F(\lambda^2|\alpha_1, \alpha_2, b) = \int IG(\lambda^2|\tfrac{\alpha_2}{2}, \delta) G(\delta|\tfrac{\alpha_1}{2}, b) d\psi^2.$$

This makes the prior conditionally conjugate, making the MCMC algorithm quite efficient. The Gibbs Sampler algorithm can be found in Appendix.

3.3. Bayesian lasso prior

The classical lasso ("least absolute shrinkage and selection operator") regression model uses a L_1 norm as a penalty term, which is the sum of the absolute values of the regression coefficients. The Bayesian equivalent of the lasso penalty is obtained by using a Laplace prior for regression coefficients (Park and Casella, 2008):

$$p^{LASSO}(\boldsymbol{\beta}|\lambda^2) = \prod_{p=1}^{P} Laplace(\beta_p|\lambda^2). \tag{13}$$

To facilitate Bayesian computation, the Laplace prior can be written as a normal distribution where the scale has an exponential distribution; this results in a conditionally conjugate Bayesian model:

$$Laplace(\beta_p|\lambda^2) = \int \mathcal{N}(\beta_p|0, \tau_p^2\lambda^2) Exp(\tau_p^2|1) d\tau_p,$$

for $p = 1, \ldots, P$.

We plotted the Laplace prior as a dotted line in Fig. 1. As can be seen, the prior is more peaked around zero with thicker tails in comparison to the normal prior that is used for Bayesian ridge regression. As will be shown later, this results in stronger shrinkage of small estimated effects and, due to the Laplace prior having thicker tails than the normal prior, this results in less shrinkage of larger estimated effects.

Compared to the normal (ridge) prior, the lasso prior includes the parameter τ_p^2; this serves as a shrinkage parameter on a local level for effect β_p and varies across the β_p. The λ^2 parameter, on the other hand, controls global shrinkage and affects all β_p to the same degree. The idea of a separate global and a local shrinkage parameter was introduced by Carvalho et al. (2009) and allows a researcher to control the shrinkage behavior of the method precisely. To complete the model, we again set a vague F prior for the global shrinkage parameter:

$$\lambda^2 \sim F(1,1,1).$$

The Gibbs sampler can be found in Appendix.

3.4. Bayesian horseshoe prior

The horseshoe model was first introduced in Carvalho et al. (2010) and has an asymptote at zero combined with heavy tails. To construct this prior, the original model proposes to use a half-Cauchy distribution (i.e., a Student t distribution with 1 degree of freedom). This results in heavier shrinkage of small effects and less shrinkage of large effects compared to the Bayesian lasso. The prior can be written as follows:

$$p^{HORSESHOE}(\boldsymbol{\beta}|\lambda^2) = \prod_{p=1}^{P} Horseshoe(\beta_p|\lambda^2), \tag{14}$$

where λ^2 is a global shrinkage parameter. Again, to facilitate Bayesian computation the horseshoe prior is written as a scaled mixture of normals where λ^2 follows an F distribution:

$$Horseshoe(\beta_p|\lambda^2) = \int \mathcal{N}(\beta_p|0, \lambda^2\tau_p^2)F(\tau_p^2|1,1,1)d\tau_p^2,$$

for $p = 1, \ldots, P$.

A graphical representation of the horseshoe prior is given in Fig. 1 by the solid line. It is clear that the prior has a sharp peak at zero and has quite heavy tails. The name “horseshoe” comes from the observation that the shrinkage coefficient κ_p, defined as $\kappa_p = 1/(1+\tau_p^2)$, has a horseshoe-shaped $Beta(1/2, 1/2)$ distribution under the matrix-F prior for τ_p. This coefficient reflects the amount of weight that the posterior places around zero: $\kappa_p \approx 0$ (or τ_p^2 very large) corresponds to no shrinkage whereas $\kappa_p \approx 1$ (or $\tau_p^2 \approx 0$) corresponds to total shrinkage to zero.

Similar to the Bayesian lasso prior, τ_p^2 serves as a local shrinkage parameter for β_p, while λ^2 serves as a global shrinkage parameter. In contrast to the Bayesian lasso, the local shrinkage parameters now follows an F distribution instead of an exponential distribution. As the F distribution has thicker tails than the exponential in case of Bayesian lasso, the F prior will result in less shrinkage of large effects than the lasso model. The parameter λ^2 optimizes the overall level of sparsity, while the local shrinkage parameters τ_p^2 prevent large effects from being shrunk towards zero. Again, we finalize the Bayesian horseshoe model by setting a vague F prior on the global shrinkage parameter:

$$\lambda^2 \sim F(1,1,1).$$

The Gibbs sampler for fitting the model is described in Appendix.

3.5. A simple illustration of different shrinkage behaviors

In order to illustrate the shrinkage effect of the four models, we estimate the shrinkage models on relational event sequences of fixed length across increasing effect sizes. For this illustration we consider the actor model (but the shrinkage behavior is similar for the dyadic model). We create event sequences with different effect sizes by properly specifying the design matrix X. For example, consider a sequence of 30 relational events on a network of six actors, assuming a scalar network parameter β and a single predictor variable that is zero for all actors except for actor 2 (for whom it is equal to 0.2) for all events in the sequence ($i \in \{1, \ldots, 30\}$): $X_i' = (0, 0.2, 0, 0, 0, 0)$. Here we consider a situation where all events are sent by Actor 1. The number of actors who receive an event from actor 1 decreases across the sequences as shown in Table 1. In Sequence 1 all events are sent proportionally to all actors 2, 3, …, 6. In the last Sequence 29 all events but one (to avoid identification issues) are sent to actor 2.

By construction, in these relational event sequences the effect size of β is smallest for the Sequence 1, grows gradually, and reaches its maximum for the Sequence 29. This data structure allows a clear comparison of the different shrinkage priors. The objective of this example is to check the amount and the shape of the shrinkage that the ridge, lasso, and horseshoe priors impose on a network effect. For clarity of exposition, we fix the shrinkage parameter λ^2 to 1 for all models.

Table 1
Sequences of events with the allocation of the receivers. All senders are fixed to Actor 1. Receivers in Sequence 1 are distributed equally, in Sequence 29 all the receivers are Actor 2.

Event index	Sending actor	Receiver Sequence 1	Receiver Sequence 2	Receiver Sequence 3	⋯	Receiver Sequence 29
1	1	**2**	**2**	**2**		**2**
2	1	3	**2**	**2**		**2**
3	1	4	4	**2**		**2**
4	1	5	5	5		**2**
5	1	6	6	6		**2**
6	1	**2**	**2**	**2**		**2**
7	1	3	3	3		**2**
8	1	4	4	4		**2**
9	1	5	5	5		**2**
10	1	6	6	6		**2**
⋮	⋮	⋮	⋮	⋮		⋮
29	1	5	5	5		**2**
30	1	6	6	6		6

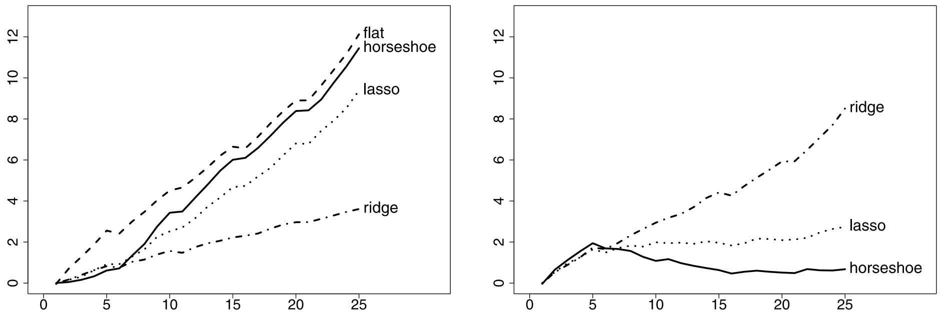
We estimate the network effect for each of the 29 relational event sequences with the actor model and show the resulting posterior means in Fig. 2. The left panel shows the estimate $\hat{\beta}$ across Sequences 1 to 29. The right panel shows the difference between the estimate under the unregularized model (using the flat prior) and those under each regularized model (using the different shrinkage priors). The estimate based on the flat prior model, which serves as a reference, increases over the sequence index as expected. The ridge model shows an approximately linear trend, heavily shrinking moderate to large effects. The lasso model shrinks small effects a bit more than the ridge prior and yields estimates that are approximately parallel to those of the flat model. The horseshoe shows most shrinkage for small effect sizes and the least shrinkage for larger effects where it gradually converges to the flat prior model. Note that if a larger value would be specified for the penalty parameter λ (which captures the scale of the prior), the induced shrinkage behavior would be rescaled to also shrink larger effects. Next, we explore how the shrinkage prior models perform in empirical settings where the penalty parameter is specified as an unknown parameter which is jointly estimated (or “optimized”) with the other parameters depending on the data at hand.

4. Empirical data applications

In this section we apply the Bayesian regularization algorithms to three empirical datasets. The goal of these studies is to explore whether the regularized relational event models result in more parsimonious models (i.e., fewer significant effects) while maintaining a good (or even better) predictive performance in comparison to an unregularized model. First we consider a relational event sequence based on the Enron corporate email data: the choice of the next receiver in this relational event sequence is analyzed using actor models with shrinkage priors. Second, we consider a sequence of communication from the infamous Apollo 13 mission to the moon. Third, we analyze undirected interactions between team members using the Wearable Sensors dataset. These latter two datasets are analyzed using dyadic relational event models. Each predictor variable in each relational event model is scaled to have a mean of 0 and a standard deviation of 1.

The estimation is performed using R code which is available at https://github.com/DianaKarim/bs. Based on the posterior distribution, we construct 95% credible intervals and evaluate the significance of the particular effect by checking whether that interval covers zero. Additionally, to obtain point estimates we look at the posterior mode of the effects based on the sample from the conditional posterior densities. Using these results, we can show that the shrinkage models result in many fewer significant coefficients compared to the flat prior model, which induces no shrinkage.

The performance of the fitted models is assessed using the posterior predictive distribution (Gelman et al., 2013), using the posterior modes



(a) Estimated coefficient for the flat, ridge, lasso, and horseshoe models

(b) Estimated difference between the shrinkage models and the flat model

Fig. 2. Estimated $\hat{\beta}$ for each generated sequence for the flat model (dashed line), ridge (dash-dotted line), lasso (dotted line), and horseshoe (solid line).

as point estimates. Posterior predictive distributions make use of the complete posterior of the parameters by using the posterior draws as plausible realizations of the unknown parameters, and thus take the uncertainty of the model parameters into account. Using the posterior predictive distribution, we can evaluate the performance of the fitted model by comparing the events predicted by the model with the events that actually occurred. By looking at the predictive performance we can learn whether the Bayesian shrinkage models provide a more parsimonious description of the interaction behavior in the networks without sacrificing predictive performance.

To evaluate the predictive power of the fitted models, we use the latent variables $Z_i^{draw} = X_i\boldsymbol{\beta}^{draw}$ that are calculated on the sampled values $\boldsymbol{\beta}^{draw}$ generated by the Gibbs sampler. Each component of the vector Z_i^{draw} corresponds to an actor (or dyad) from the ordered risk set. Higher values imply a larger probability of a corresponding actor to be the next receiver (or the dyad to transpire next in the case of dyad model). Thus, we can compare the event that actually occurred with the set of "best predicted events", i.e. sorted values of the latent variable Z_i^{draw}, for each event in the sequence $e_i, i \in 1, \ldots, M$. Below, we consider how often events with the highest Z_i^{draw} component appears in the set of 5%, 10%, or 20% of the highest scoring events. Similarly, the prediction power is also calculated using the posterior mode estimates by using the vector $\boldsymbol{\beta}^{post.mode}$ instead of $\boldsymbol{\beta}^{draw}$. Both in-sample as well as out-of-sample predictive performance is explored using the above methods. For the out-of-sample predictions, the posterior is based solely on the training sample (i.e. the "in-sample") but the endogenous statistics are updated in the test sample (i.e. the "out-of-sample") based on the actually observed events in this period. For example, in the case of a training sample of 2000 events, and a test sample of the next 500 events, the posterior of the coefficients is solely based on the first 2000 events but the endogenous statistics are updated in the following 500 events, e.g. the endogenous statistics to predict the 2002nd event are based on the 1st until the 2001st event.

4.1. Bayesian regularization of an actor REM with directed relational events

4.1.1. Enron email data

To demonstrate the performance of the actor oriented shrinkage models, we use publicly available data of the Enron corpus from the repository of Carnegie Mellon University. These data contain the time-stamped emails of 156 users, mostly senior management of Enron Corporation, a former American energy, commodities, and services company. The data were made public in 2001 after Enron Corporation declared bankruptcy and the following public investigation. These data have been widely used in different fields from social network research to computer science (Keila and Skillicorn, 2005; Diesner et al., 2005; Wilson and Banzhaf, 2009; Peterson et al., 2011), and have already been analyzed in the context of relational events in Perry and Wolfe (2013) using maximum likelihood estimation with no shrinkage. We consider the dataset compiled by Zhou et al. (2007) which consists of 21,635 messages in total (data retrieved from https://github.com/patperry/interaction-proc). The original data span a long time period (1998.11 - 2002.6) and record the tumultuous dynamics of Enron, from glory to collapse, declaring bankruptcy in December 2001. It is very unlikely that the drivers of the social dynamics in the company remained unchanged over this 3.5 year period. Because our interest is in illustrating the effect of regularization methods, rather than showing how to build a relational event model that fits the changing dynamics inside Enron over this entire period, we consider a sample consisting of the first 2000 events in the year of 2001 (assuming that effects will be reasonably stable for this period), where we split the multicast messages into multiple dyadic observations (for more direct approaches to model multicast messages see Perry and Wolfe (2013), Lerner et al. (2021), or Mulder and Hoff (2021)). The subset that we use in this application can also be found on the github page (link suppressed for blind review). The time interval of this subset is 37 days. The events that happened before the start of the subset are considered in the computation of the endogenous statistics.

Table 2
Dichotomous variables indicating whether the actor works in the Legal department, trading department, is a junior, or is female.

Variable	Characteristics of actor i
L(i)	Member of the Legal department
T(i)	Member of the Trading department
J(i)	Seniority is Junior
F(i)	Gender is Female

The data also contain information about several actor traits, such as the actors' gender (male or female), department (Legal or Trading), and seniority (Junior or Senior). We use these actor traits to model homophily (when both sender and receiver belong to the same group) and cross-group effects (when sender and receiver come from different groups). We do this by including interaction variables $X(i) * Y(i)$, where X indicates that a sender belongs to group X, and Y indicates that a receiver belongs to group Y, while X and Y come from the set of dichotomous actor dependent attributes (L, T, J, F) – see Table 2 for the overview. Hence, the interaction variable $L(i) * L(i)$ is 1 if sender and receiver are both members of the Legal department and 0 otherwise. The interaction variable $L(i) * F(i)$ is 1 if the sender is a member of the legal department and the receiver is female. Such interactions can allow a researcher to estimate the effects of combinations of social categories on interaction tendencies inside Enron.

To summarize the past activity between actors in the network, we include six endogenous network effects in our analysis that are commonly used in relational event models (e.g. Leenders et al., 2016).

Table 3
Predictive performance of the four flat prior model, the Bayesian ridge, the Bayesian lasso, and the horseshoe (HS) prior model for the Enron email data. The results reflect the percentages of observed events that belong to the top 5%, the top 10%, and the top 20% of most likely events based on the estimated model using the full posterior or the posterior modes for making predictions. In each category, the best result is displayed in bold.

In-sample	5%				10%				20%			
	Flat	Ridge	Lasso	HS	Flat	Ridge	Lasso	HS	Flat	Ridge	Lasso	HS
Full post.	76.3	**76.6**	76.5	76.5	83.2	**83.4**	83.2	83.1	88.3	**88.6**	88.5	88.4
Post. modes	75.9	**76.7**	76.6	76.3	83.0	**83.7**	83.2	82.8	87.8	**88.7**	88.5	88.4
Out-of-sample	5%				10%				20%			
	Flat	Ridge	Lasso	HS	Flat	Ridge	Lasso	HS	Flat	Ridge	Lasso	HS
Full post.	79.1	79.2	**79.3**	**79.3**	84.9	85.2	**85.3**	**85.3**	90.8	**91.1**	91.0	**91.1**
Post. modes	79.0	**79.4**	79.2	79.0	84.2	**85.7**	85.0	85.0	90.3	**91.6**	91.3	**91.6**

First, we included inertia (I), which quantifies the tendency of an actor to keep sending emails to an actor as a function the volume of emails sent by this sender to this same receiver in the past: the more emails s has sent to r in the past, the higher is the expected future rate of s sending to r. Second, we include reciprocity (R), which quantifies the tendency of an actor to send messages to another actor as a function of the number of messages that this sender received from that actor in the past: the more emails r has sent to s in the past, the higher is the expected future rate of s sending to r. In addition, we included receiver's indegree and outdegree, also referred to as popularity (P) and activity (A), which quantify the tendency to receive messages as a function of the number of messages received from everybody in the network and the tendency to receive messages as a function of the number of sent messages in the past, respectively: this captures the effect that past popular receivers (regardless of who was the sender) are likely future receivers (for any sender) and that very active senders are likely future receivers. Lastly, higher level triadic effects were added to the model, such as outgoing two-paths (OTP), incoming two-paths (ITP), outgoing shared partners (OSP), and incoming shared partners (ISP). Furthermore, to account for a possible memory decay where recently observed emails may have a greater impact on what happens next than emails that were exchanged longer ago, we calculated all six the endogenous effects listed above based on the events that were observed within intervals of 1 day, 2 days, 1 week, 2 weeks, 1 month, and 3 months (for more elaborate memory decay models for relational event data see also Arena et al. and Perry and Wolfe (2013)). In total, this results in a relational event model consisting of 64 effects.

4.1.2. Results

We estimated the model (for each of the priors) using the Gibbs sampler algorithms described in Appendix with 10,000 iterations as a burn-in period followed by a total of 100,000 iterations, where only every tenth iteration is recorded (to eliminate the effect of autocorrelation in the posterior draws). We plot the estimated posterior distributions and 95% credible intervals in Fig. 3. As expected, the flat prior model with no shrinkage behavior returns the largest number of significant predictors: 27 in total. The shrinkage models result in 24 (Bayesian ridge), 22 (Bayesian lasso), and 18 (horseshoe) significant predictors, respectively. Thus, we see a considerable drop in the number of significant effects, especially for the horseshoe model.

It is informative to inspect the posterior intervals in Fig. 3. First, consider the cases where the unregularized model (with the flat prior) results in an interval where zero is just barely excluded: this represents either a potentially large effect with much posterior uncertainty (i.e., a wide interval) or a smaller effect with a narrow interval. In these cases, a regularized model generally shifts the interval towards zero, resulting in a non-significant effect.[2] For example, we can see this for the endogenous popularity effect (P) in the 2 weeks period and the 1 month period. For the exogenous effects, we see some similar shrinkage behavior when both the sender and receiver are from the trading department (TT) and when the sender is a junior and the receiver is from the trading department (JT).

Another informative observation is that the regularized models generally result in narrower intervals, implying higher posterior certainty. This is a preferred behavior when fitting a large model with many potential predictor variables that induces much uncertainty. From a Bayesian perspective this behavior is expected as the posterior combines the information in the prior with the information in the data, and thus, if an informative prior is used (as we do here), there will be less posterior uncertainty than when this prior information is excluded.

Table 3 reports the predictive performance of the four models. Using the full posterior and using the posterior modes, we calculated the estimated rates and then calculated whether the actually transpired event was in the top 5%, top 10%, or top 20% of the estimated/predicted rates. We did this both for in-sample predictions (i.e. we predicted each next event within the observation period)and out-of-sample predictions for the next 500 events in the sequence. Overall the predictive performance of the models is quite good. The prediction scores also support our hypothesis that shrinkage models attain a comparable or better predictive power while eliminating a considerable number of non-significant effects. Note that the performance for the out-of-sample predictions is a bit higher than the performance for the in-sample predictions. This is somewhat surprising as out-of-sample predictions are more challenging. This could be explained by temporal changes of interaction behavior which is more likely to occur over longer observational periods. In our situation, the data for training the models consisted of 2000 events, and thus the interaction behavior is more likely to change in this longer observational period than the out-of-sample sequence which consisted of only 500 events. The interaction behavior in the out-of-sample observational period may have been approximately constant, and possibly close to the average behavior in the longer observational period of the training sample.

4.2. Bayesian regularization of a dyadic REM with directed relational events

4.2.1. Voice loops during the Apollo mission

We analyze the recorded voice loops from NASA's infamous Apollo 13 mission. The data were retrieved from http://apollo13realtime.org/ and consist of recorded voice messages of the members of the Command and Service Module (CSM) Odyssey and the Lunar Module (LM) Aquarius. The data consist of the Flight directors' voice loop and the air-ground's voice loop: Flight directors (Houston's Mission Control Center) were located in Houston and the crew (astronauts) were connected to this control center via Capsule Communicator (CAPCOM). In the original data only the senders of the messages were recorded. We added the receivers manually based on the content of the text messages.

In total, the event sequence includes 5498 messages within an observational period of six hours. Around the beginning of this period

[2] Using a larger confidence level, e.g., 99%, would still result in more significant effects than when using shrinkage models. We come back to this in the Discussion.

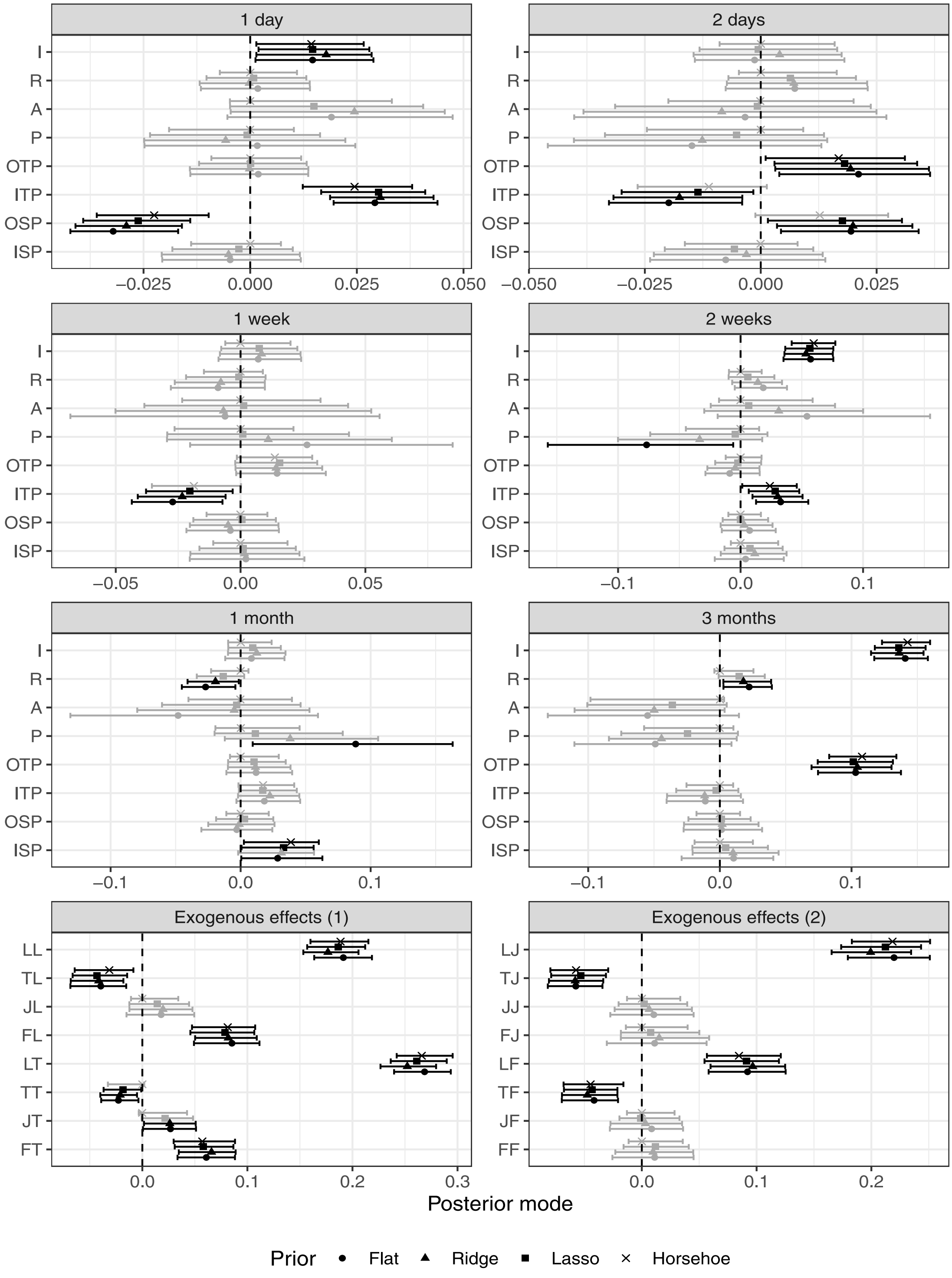


Fig. 3. Posterior modes and 95% credible intervals for all 64 effects estimated in the Enron email data with the flat, ridge, lasso, and horseshoe models. For clarity of exposition, the results are organized across separate panels, but the model was run with all effects together. Black and gray error-bars refer to significant and non-significant effects, respectively. The flat, ridge, lasso, and horseshoe prior model result in 27, 24, 22, and 18 significant effects, respectively. The abbreviations denote inertia (I), reciprocity (R), activity (A), popularity (P), law (L) and trading department (T), junior (J), female (F). The interactions are labeled as TT (sender and receiver both belong to the trading department), FT (sender is female, the receiver is in the trading department), et cetera.

an explosion occurred that damaged the oxygen tanks, followed by the rapid decrease of the oxygen levels and fluctuations in electrical power and control thrusters in the Lunar module (characterized by the famous quote “Houston, we’ve had a problem”). It is highly likely that the communication patterns before the incident (when the mission was still in “routine” mode) were drastically different from those after the incident (where the mission jumped into survival and problem-solving). To avoid the drastic changes in the parameters of the data in the middle of our observation period, we focus on the events after the incident for the analysis in this paper and removed the first 96 events from the dataset. This resulted in 5402 events, starting from the point when the problem was reported, until the successful splashdown. We

Table 4
Predictive performance of the four flat prior model, the Bayesian ridge, the Bayesian lasso, and the horseshoe (HS) prior model on the Apollo communication data. The results reflect the percentages of observed events that belong to the top 5%, the top 10%, and the top 20% of most likely events based on the estimated model using the full posterior or the posterior modes for making predictions. In each category, the best result is displayed in bold.

In-sample	5%				10%				20%			
	Flat	Ridge	Lasso	HS	Flat	Ridge	Lasso	HS	Flat	Ridge	Lasso	HS
Full post.	88.8	88.8	**88.9**	**88.9**	**96.9**	**96.9**	**96.9**	**96.9**	**99.4**	99.3	99.3	99.3
Post. modes	87.2	88.5	**89.0**	84.3	96.9	**97.0**	96.8	93.5	**99.4**	99.3	99.3	98.9
Out-of-sample	5%				10%				20%			
	Flat	Ridge	Lasso	HS	Flat	Ridge	Lasso	HS	Flat	Ridge	Lasso	HS
Full post.	92.7	92.7	92.8	**92.9**	97.7	97.7	**97.8**	**97.8**	**99.5**	99.3	99.4	99.4
Post. modes	90.0	91.6	**92.8**	84.4	97.6	97.8	**98.0**	93.6	**99.4**	99.2	99.2	**99.4**

leave the last 500 events out to evaluate the out-of-sample predictive performance of the models (resulting in a total of 4902 events that are used for model fitting).

In these relational event data, certain actors are substantively more active than other actors. This can be explained by the fact that strict protocols of communication had to be followed during the mission: only one actor (CAPCOM) was allowed to talk to the crew in the Lunar Module, while the main flight director (FLIGHT) was coordinating the team at the command and Service Module. As a result of these communication rules, around 40% of messages were sent by FLIGHT, 36% of the messages were received by FLIGHT, 13% of messages were sent by CAPCOM, 18% messages were received by CAPCOM. We adapted the risk set such that only those dyads were at risk that were allowed by the mission's protocols.

We specified a dyadic relational event model with twelve endogenous statistics: recency ranks of sender and receiver (rrankSnd, rrankRec), participation shifts (psABBA, psABBY, psABXA, psABAY), inertia (I), reciprocity (R), outdegree of sender (ODSnd), indegree of receiver (IDRec), and transitive closure effects of outgoing and incoming two-paths (otp and itp)–see Butts (2008) for their mathematical definitions. Further, we assigned each actor a general role: CAPCOM, FLIGHT, actors in the Lunar Module (Air), and the remaining actors on the ground (Ground). Because of the idiosyncratic role of each actor in this communication network, it is likely that communication behavior differs between different types of actors. We therefore interact each of the twelve endogenous variable with the eight role combinations (Air to air, Ground to CAPCOM, Air to CAPCOM, CAPCOM to ground, CAPCOM to air, ground to FLIGHT, FLIGHT to ground)–this allows to assess whether reciprocity is higher for actors in the lunar module (Air to Air) than for ground personnel (Ground–Ground) or communication from ground personnel to the CAPCOM, et cetera. Excluding the Ground–Ground fixed effect (for identifiability), this yields a relational event model with 103 network effects.

4.2.2. Results

We estimated the four models using the Gibbs sampler algorithms described in Appendix with 10,000 iterations in burn-in period and total 100,000 iterations, where only every tenth iteration is stored. We plot the estimated posterior modes and 95% credible intervals in Figs. 4 and 5. Of the 103 effects in total, the flat prior model, the Bayesian ridge model, the Bayesian lasso model, and the horseshoe prior model identified 62, 54, 53, and 45 significant effects, respectively, illustrating the impact of shrinkage algorithms to obtain more parsimonious model.

There are interesting patterns to be observed from these plots. For example, we see that otp is mostly significant based on the flat prior model but always never significant for the shrinkage models (with an acceptance of air-to-CAPCOM, which is the other way around). The point estimates are often practically equal to zero for the horseshoe model. Substantively, the results show a highly protocolized interaction pattern. In particular, it is clear that interaction between ground personnel (Ground–Ground) happens quite differently from the other interaction. The ground personnel was busy trying to come up with solutions that could be passed on to the crew in the air, which required a lot of going back-and-forth between them. As a result, for these interactions there are positive and significant effects for psABBA and psABBY (illustrating that the receiver of a message tends to immediately respond to it to the sender of the message or to immediately send a message on to another member of the ground team). Similarly, the development of discussions between the ground crew is shown by the significance of the recency effects rrankSnd and rrankRec (i.e. both the most recent senders and the most recent receivers tend to soon send a next message soon). Together, this is illustrative of a team that is frantically trying to make sense of the problem and find a solution to it. These statistics are rarely positive and significant for the interactions between other roles.

All models clearly show that interaction rates tend to be high among the ground crew and between the air crew and the CAPCOM. The flat model also flags CAPCOM-to-Ground and communication between Ground and FLIGHT as enhancing communication rates, but the horseshoe model shrinks all of these effects to zero, pointing to a role-based interaction dynamic that is governed by the problem-solving communication by the ground crew and the exchange of information about what is going on inside the capsule between the air crew and the CAPCOM.

Overall, the figures again show that the intervals tend to be narrower for the regularized models than for the unregularized model (with some exceptions).

Table 4 shows the results of the in-sample and out-of-sample predictive performance of the four models. Again we see a pattern where the models perform quite similarly. Overall, all models show good predictive performance. It is interesting that the predictive performance of the horseshoe model is somewhat lower when predicting events solely on the posterior modes, and thereby ignoring posterior uncertainty. This suggests that the use of the full posterior for making predictions would be advised for these data. Finally, note that the out-of-sample predictive performance is again a bit higher for all models than the in-sample predictive performance. This could be explained by the fact that the training sequence, which consists of 4902 events, is considerably larger than the testing sequence, which consists of 500 events, making the training sequence more difficult to predict in case of changes in the interaction behavior which are likely in longer observational periods.

4.3. Bayesian regularization of a dyadic REM with undirected relational events

4.3.1. Interactions between team members using proximity data

In the third empirical application, we analyze undirected interactions between members of a research team. The data were collected by Müller et al. (2018) with the aim to develop new tools and methods for doing research into the impact of gender diversity in Research & Development teams. Members of eight research teams wore sociometric badges that recorded close-range proximity between participants with Bluetooth sensors. Here, we analyze the proximity data of 'team 4', a team in the field of biomedical engineering in a public research center. The team consists of nine members: one team leader, two senior

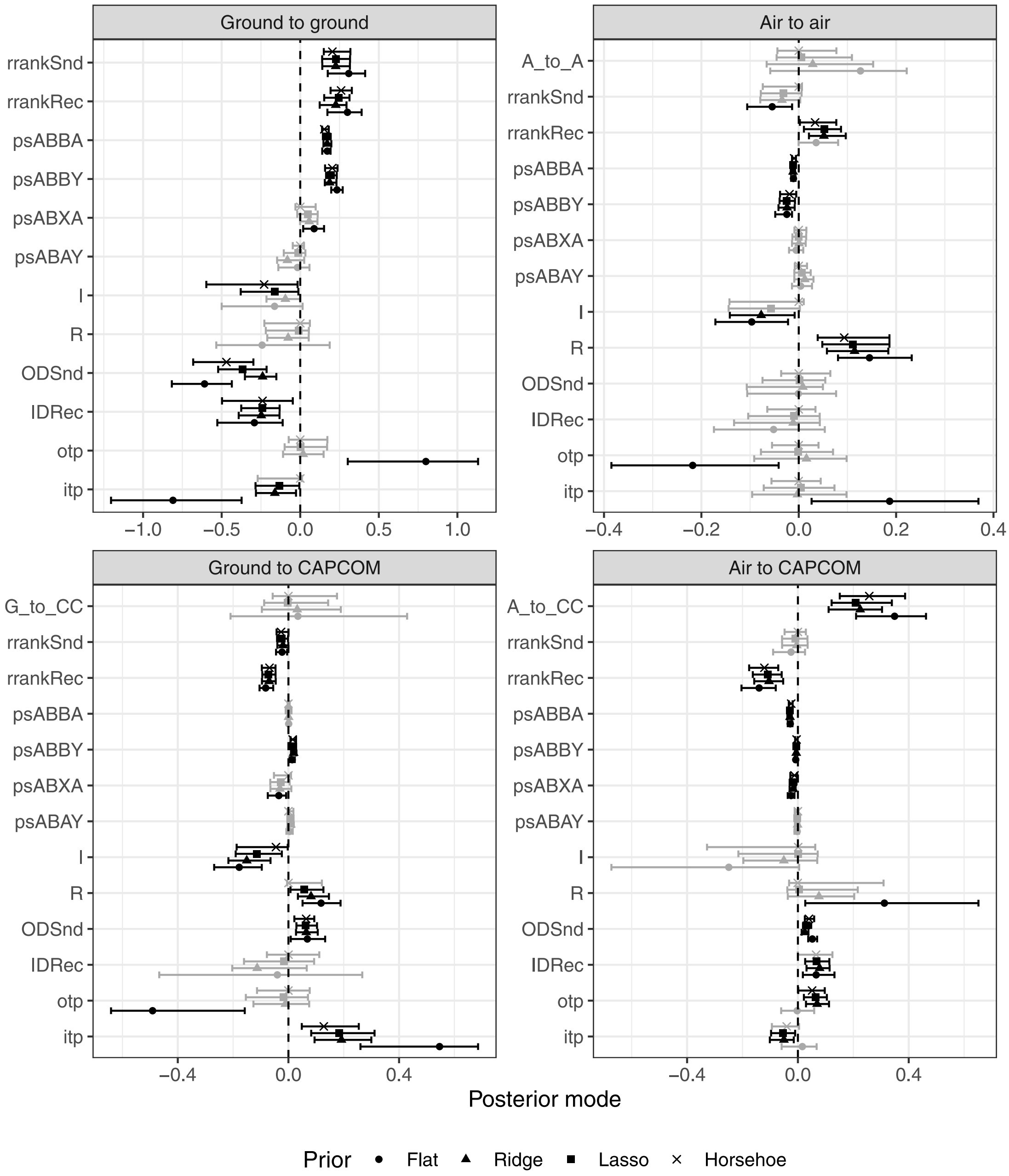


Fig. 4. Posterior modes and 95% credible interval for 51 effects estimated in the Apollo 13 mission data with the flat, ridge, lasso, and horseshoe prior model for Ground to ground, Air to air, Ground to CAPCOM, and Air to CAPCOM channels. Black and gray error-bars refer to significant and non-significant effects, respectively. The flat, ridge, lasso, and horseshoe prior model result in 33, 28, 28, and 24 significant effects, respectively.

researchers, four Ph.D.-students, one research assistant, and one master student.

The proximity contacts between the nine team members were recorded for five days. Since differences in interaction dynamics can be expected across different days of the week (Amati et al., 2019), we chose to focus on the contacts of the second day of the observation period. In total, 2653 proximity contacts between the team members were detected on the second day of observation. The 2896 contacts on the first day of the observation period are included in the history of events used to train the endogenous statistics. In addition to the proximity data, team members filled out a questionnaire reporting on their social and advice-seeking relationships with their fellow team members. The team members rated each other in a round-robin design on the questions ‘I spend time socially with this person outside the lab/office’ (1 = never, 2 = some times a year, 3 = some times a month, 4 = some times a week, 5 = daily) and ‘I consult this person for work related advice’ (1 = never, 2 = rarely, 3 = sometimes, 4 = very often, 5 = always). Furthermore, we have information on the team members sex (five females, four males), educational level (three doctorate, four masters or postgraduate, one completed secondary education, and one bachelors), age (mean is 37 years, standard deviation is 9 years), and tenure (mean is 55 months, standard deviation is 35 months).

For these data, we estimate a dyadic relational event model. Since the proximity contacts are undirected, the risk set consists of 36 undirected pairs of team members. We include four endogenous network effects in our model: “degree” (D) (i.e., the total number of past interactions that any of the actors in the pair had with any of the other team members), “inertia” (I) (i.e., the number of past contacts between the actors in the pair), “recency” (R) (i.e., the time since the last contact between the actors in the pair), and “shared partners” (SP) (i.e., the number of past contacts with a third actor that both actors in the pair had been in contact with). In addition to the endogenous predictors,

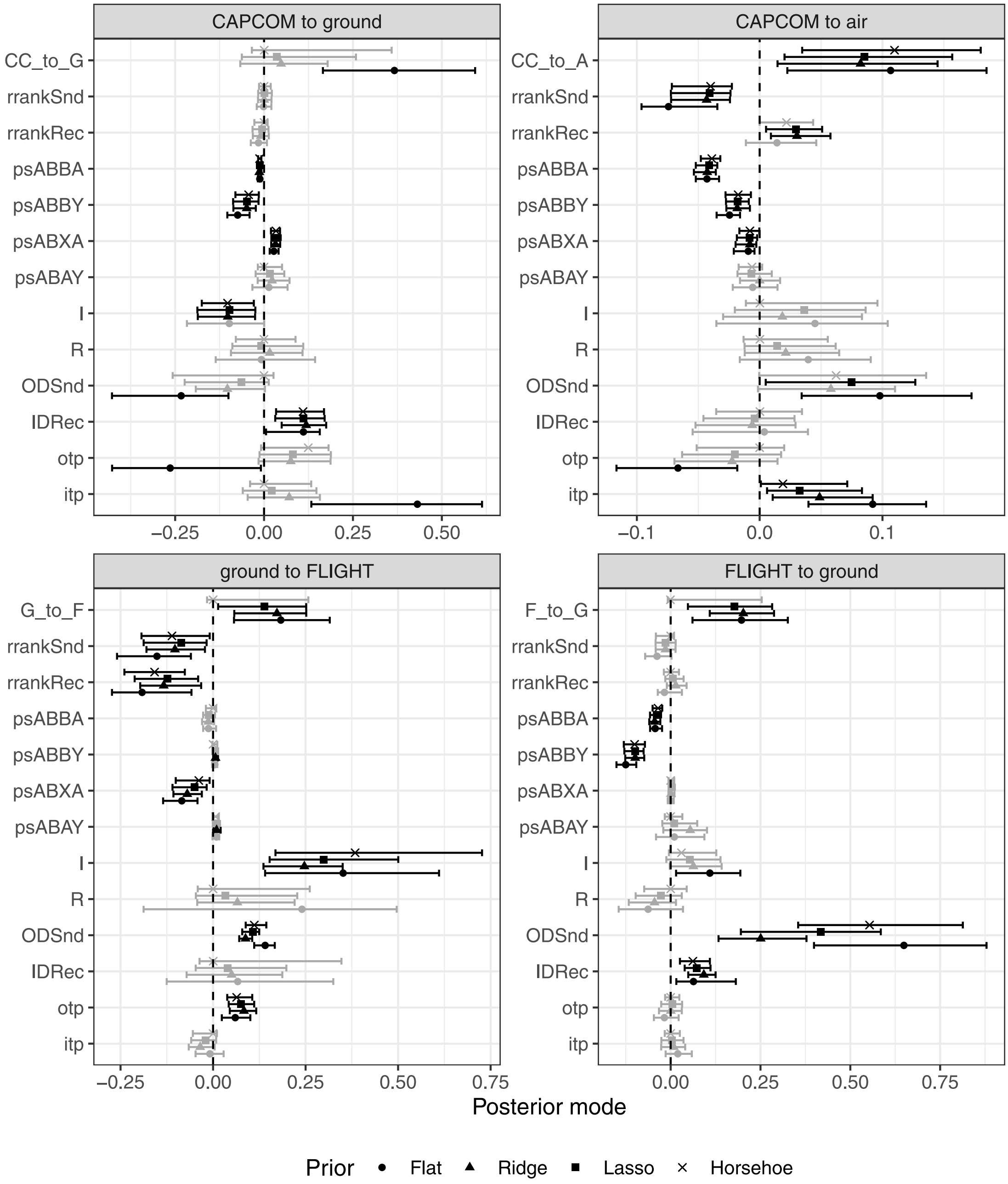


Fig. 5. Posterior modes and 95% credible interval for 52 effects estimated in the Apollo 13 mission data with the flat, ridge, lasso, and horseshoe prior model for CAPCOM to ground, CAPCOM to air, Ground to FLIGHT, and FLIGHT to ground channels. Black and gray error-bars refer to significant and non-significant effects, respectively. The flat, ridge, lasso and horseshoe prior model result in 29, 26, 25, and 21 significant effects, respectively.

we add eight exogenous predictor variables to our model. The "social average" (SA) variable describes the average of the rating the pair gave each other on the question about their social relationship, the "social difference" (SD) variable describes the absolute difference in the rating the pair gave each other on this question. Similarly, the "advice seeking average" (ASA) variable describes the average of the rating the pair gave each other on the question about their advice seeking relationship, the "advice seeking difference" (ASD) variable describes the absolute difference in the rating the pair gave each other on this question. The "age difference" (AD) variable describes the absolute difference in the age of the actors in the pair. The "tenure difference" (TD) variable describes the absolute difference in the tenure of the actors in the pair. The "same sex" (SS) variable is 1 if both actors in the pair have the same sex and 0 otherwise, and "same education" (SE) is 1 if both actors in the pair have the same educational level and 0 otherwise. Finally, we add all the interactions between the four endogenous predictors and eight endogenous predictors to our model. In total, we have 44 predictor variables in our model. All predictor variables in the model were standardized prior to the analysis.

4.3.2. Results

We estimated the four models on the 'team 4' data using the Gibbs samplers described in Section 2 with 10 000 iterations as a burn-in period followed by a total of 100 000 iterations, where only every tenth iteration is recorded. Fig. 6 shows the estimated posterior modes and 95% credible intervals. From Fig. 6, we can clearly observe the impact of the shrinkage models compared to the flat model: For most effects, the posterior mode is closer to zero for the models with shrinkage priors than for the model with a flat prior and, in general, the 95% credible interval of the posterior distribution is smaller for the shrinkage prior

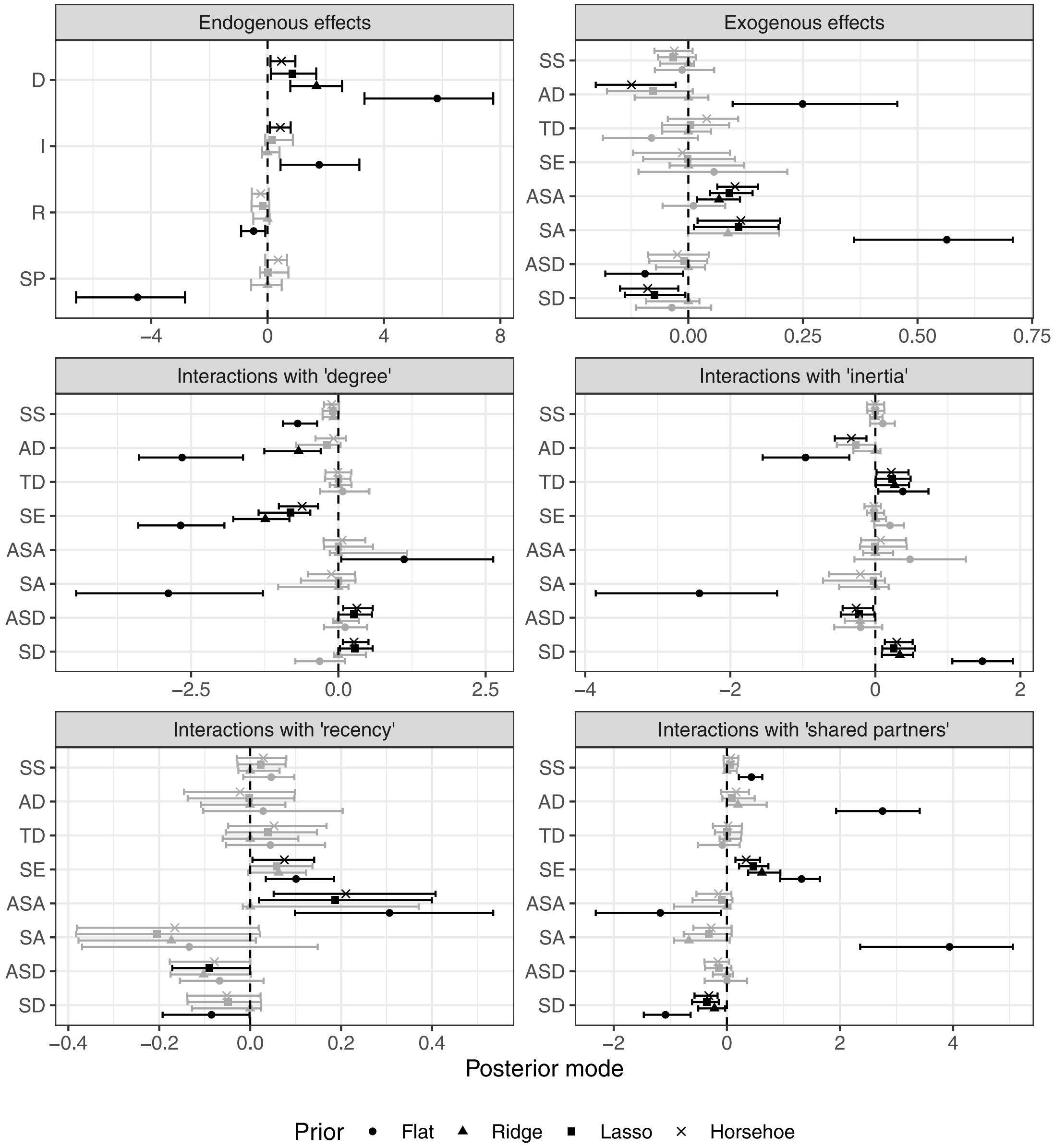


Fig. 6. Posterior modes and 95% credible interval for the 44 effects estimated in the "team4" data with the flat, ridge, lasso and horseshoe prior model. Black and gray error-bars refer to significant and non-significant effects, respectively. The flat, ridge, lasso and horseshoe prior model result in 25, 17, 14, and 8 significant effects, respectively.

models than for the flat prior model. Based on the 95% credible intervals, we can conclude that the number of significant predictors reduces from 25 in the flat prior model, to 17 in the ridge model, 14 in the lasso model and 8 in the horseshoe prior model.

The results (Fig. 6) show that the models with shrinkage priors yield more parsimonious models for the event rate than the model with the flat prior. As an example of how this affects the interpretation, we focus on the total degree effect in the upper left panel of Fig. 6 and the interactions between this degree effect and the exogenous effects in the middle left panel of Fig. 6. As shown in the upper left panel of Fig. 6, the model with the flat prior finds a strong, positive effect for the total degree of the dyad members on the event rate. Thus, pairs of team members who were (summed together) more active in the past, are more likely to interact again than pairs of team members who were jointly less active in the past. As shown in the middle left panel of Fig. 6, results from the model with flat prior indicate that this effect of degree on the event rate is stronger for pairs with team members that have a higher average advice seeking relationship. Furthermore, the model with flat prior finds that the effect of degree on the event rate is smaller for pairs with team members of the same sex, team members that differ more in age, team members that have the same level of education, and team members that have a higher average social relationship with each other. Results from the models with shrinkage priors suggest a much simpler model. Taking the results from the horseshoe prior as an example, the model also finds a positive effect for the total degree of the dyad members on the event rate, but much less strongly so. Also, the effect of degree on the event rate decreases for pairs who have the same level of education (SE) and for pairs who differ more in age (AD). Compared to the flat prior model, the horseshoe prior model yields no significant interaction effects between degree (D) and same sex (SS), average advice seeking relationship (ASD) and average rating of their social relationship (SA).

Table 5 shows the in-sample and out-of-sample prediction scores for the Team 4 data. For out-of-sample prediction, the first 500 events for the third (i.e., the next) day were predicted. Results in Table 5 show that the in-sample prediction scores for the four models are very similar based on the full posteriors. All four models correctly predict around 11% of the 2653 observed proximity in the top 5% predicted events,

Table 5
Predictive performance of the four flat prior model, the Bayesian ridge, the Bayesian lasso, and the horseshoe (HS) prior model on the proximity data of Team 4. The results reflect the percentages of observed events that belong to the top 5%, the top 10%, and the top 20% of most likely events based on the estimated model using the full posterior or the posterior modes for making predictions. In each category, the best result is displayed in bold.

In-sample	5%				10%				20%			
	Flat	Ridge	Lasso	HS	Flat	Ridge	Lasso	HS	Flat	Ridge	Lasso	HS
Full post.	**10.8**	**10.8**	**10.8**	**10.8**	**27.8**	27.2	27.4	27.2	**54.7**	54.1	54.1	54.2
Post. modes	8.2	9.8	**10.7**	10.3	25.8	**27.9**	26.5	26.6	52.4	**54.7**	52.9	52.3
Out-of-sample	5%				10%				20%			
	Flat	Ridge	Lasso	HS	Flat	Ridge	Lasso	HS	Flat	Ridge	Lasso	HS
Full post.	**2.7**	2.6	2.6	2.6	**5.4**	5.1	5.1	5.3	16.8	15.5	16.2	**18.4**
post. modes	2.2	**2.6**	**2.6**	**2.6**	2.6	6.6	4.0	**7.6**	16.6	12.8	**19.6**	18.0

around 27%–28% in the top 10% predicted events, and around 54%–55% in the top 20% predicted events. Even though the shrinkage priors resulted in fewer significant effects and in posterior modes closer to zero, the in-sample prediction performance of these models is similar to that of the flat prior model. The table also shows that all four models predict around 3% of the 500 events on the next day to be in the top 5% predicted events, and around 5% in the top 10% predicted events. Note that the lower out-of-sample performance suggests that the interaction dynamics in the network of team members was likely quite different on the next day. Because the model predicts the next events based on the fitted parameters of the interactions on day 2, it does so for day 3 based on values that are not in line with that day's social dynamics. In conclusion, the results of the predictive performance indicate that the full posterior, which takes posterior uncertainty into account, generally results in the best predictions. Furthermore, all four methods show comparable performance results on average. As before, regularization yields more parsimonious results, while generally retaining model fit and predictive accuracy.

5. Discussion

The temporal dynamics of relational event networks is governed by a large variety of (complex) effects such as endogenous network effects, exogenous effects, as well as interaction effects between them. There is no single and common formulation of the relational event model that can exactly explain who interacts with whom in what order and at what time. Similarly, there is no theory that suffices to formulate such a model. Due to the complexity of the phenomenon, it is only logical that relational event models with a decent fit can be quite complex themselves. And if modeling high-resolution temporal human interaction patterns were not complex enough, this complexity is even further exacerbated when exogenous attributes change (e.g., when the actors move to a different organizational department or when their mood changes), when there are cycles in interaction patterns (e.g., when time-of-day or time-of-week matters), when outside influences play a role (e.g., the start of a new project, holiday time, the arrival of a new manager, an explosion in the lunar capsule during the Apollo 13 mission) or the composition of the network changes (some members leave the organization–permanently or just for a while due to sickness–and others join)–all of these influences are quite common when studying real life events. With or without any of these increasingly complexity situations, modeling temporal social interaction behavior in a network, while achieving good and consistent model fit, can easily lead to relational event models with a large number of predictors. Further, it is important to acknowledge that the endogenous network effects can be highly collinear: for example, strong reciprocity gives rise to high inertia. Each single network effect may be (and typically is) a component of multiple effects and many event sequences of interest are related by a "parent-descendant" relation (e.g., reciprocity may solidify two-paths, two-paths are antecedents of three-cycles and other kind of triadic structures, et cetera). As a result, effects may correlate highly with one another and interpretation of model results becomes quite complicated when many effects are included in the model that have intricate mutual dependencies with each other. Therefore, we contend, there is a need for (statistical) algorithms that can separate the wheat from the chaff and can recognize the true nonzero effects (for the dataset and model at hand). This results in more parsimonious relational event models that are easier to interpret and may even have better predictive performance than models obtained using standard statistical methods such as maximum likelihood estimation.

As shown in this paper, Bayesian shrinkage models form an effective solution to this problem. The proposed methods result in considerably fewer significant effects and less posterior uncertainty (due to narrower interval estimates) without sacrificing predictive performance of the fitted models. This was shown for a variety of relational event models including an actor model for directed observations, a dyadic model for directed observations, and a dyadic model for undirected observations. By considering relational event data of different natures, including email messages between colleagues in a large organization, radio communication messages during the infamous Apollo mission, and physical proximity data between colleagues in a small team, we illustrated that the methodology can be quite generally applied to relational event data of different kinds.

We refrain from suggesting which specific regularization technique (i.e., which shrinkage prior) should be used in general as different empirical analyses can require different shrinkage behavior to optimally balance between model parsimony and predictive performance. For this reason, we plotted the priors explicitly in Fig. 1, so a researcher can guide this choice based on the type of regularization that is preferred in a specific research project. Alternatively, researchers may well choose to apply all three methods, i.e., the Bayesian ridge, the Bayesian lasso, and the horseshoe prior model, and use the results to determine which is most suitable for the research at hand by investigating predictive performance and the (number of) significant effects in the resulting models. If a researcher has a clear preference for a maximally parsimonious model while maintaining good predictive performance, we recommend using the horseshoe prior model and to use the full posterior to do predictions.

The work on (Bayesian) regularization of relational event models is still in its infant stages. Our intent has been to illustrate its promise and to provide researchers with the tools to apply the state-of-the-art models in their own research (the appropriate R code is available at github page; link suppressed for blind review). There are many ways in which this research can be extended. One direction is to not shrink all parameters in the model according to the same regime, but to distinguish between types of parameters by using different penalty parameters for different subsets of the effects. The group lasso (Kyung et al., 2010), for instance, can be used for this purpose. For instance, one could penalize the endogenous effects, the exogenous effects, and, possibly the interaction effects differently. Or, in a network of employees in an organization, one could apply different levels/regimes of penalization on within-department effects than between-department effects. A related extension would be to use different types of shrinkage priors for different types of parameters (e.g., dyadic effects, triadic effects, degree, etc.). One set of parameters could then be shrunk according to a lasso and another set of parameters would be shrunk according to a horseshoe prior.

Shrinkage priors may also prove useful when modeling network effects that may change over time. When the parameter is approximately constant over time, the difference is shrunk to zero such that the estimated parameter remains constant. If the parameter does change, e.g., due to a switch in interaction regime between the actors, no shrinkage should be applied and we would be able to identify a change of the interaction regime in the network (see also Shafiee Kamalabad and Grzegorczyk, 2020, for such decoupling techniques).

Another interesting topic for further exploration is to optimize the width of the credible intervals that are used to flag parameters as significant, based on the data at hand (e.g., Liu et al., 2020). Finally, the exploration of additional shrinkage priors (e.g., the Bayesian elastic net, or spike-and-slab priors; Van Erp et al. (2019)) as well as the application of shrinkage priors to other kinds of relational event models, such as the Cox model, will be important extensions of the current work.

In the meantime, the methods we presented in this paper enable a researcher to include all effects that might be relevant in explaining the relational event hierarchy, even when they are highly collinear (as is often the case in large network models), and then use these methods to recognize which effects really matter statistically and which are effectively zero. This makes the model more interpretable and increases the confidence a researcher can have in the results, since the effects that remain after regularization can be seen as indeed meaningful to the relational event model. This can help save a researcher from erroneously applying significance to noise and unimportant factors, but rather focus interpretation of the effects that matter. Hopefully, this can assist in developing deeper and more robust insight into the drivers of social dynamics in networks.

Acknowledgment

This work was supported by an ERC Starting Grant (758791).

Appendix. Gibbs samplers

Fitting a Bayesian relational event model using a noninformative, flat prior

The Gibbs sampling algorithm has the following steps

1. Set initial values for $\boldsymbol{\beta}^{(0)}$, and $Z^{(0)}$.
2. Given $Z^{(s-1)}$, draw $\boldsymbol{\beta}^{(s)}$ from its conditional posterior distribution that follows the multivariate normal distribution:

   $\boldsymbol{\beta}^{(s)}|Z^{(s-1)} \sim N(\mu_\beta, \Sigma_\beta)$, where

   $$\mu_\beta = \left(\sum_{i=1}^{M} X_i^T X_i\right)^{-1} \sum_{i=1}^{M} X_i^T Z_i^{(s-1)},\ \Sigma_\beta = \left(\sum_{i=1}^{N} X_i^T X_i\right)^{-1}$$

3. Given $\boldsymbol{\beta}^{(s)}$, draw $Z_{ir}^{(s)}$ from its conditional posterior distribution that follows the truncated normal distribution:

   $Z_i^{(s)}|\boldsymbol{\beta}^{(s)} \sim t\mathcal{N}(X_i\boldsymbol{\beta}^{(s)}, I_N)$.

   To reduce the degrees of freedom when choosing the latent variables, the first component of each Z_i is fixed to zero. The value of the element of Z_i that corresponds to the observed actor is generated from a truncated density on the interval $(\max_{r\neq r_i} Z_{ir}, \infty)$. Alternatively, the values of the elements that correspond to actors that are not observed are generated from the truncated density on the opposite interval $(-\infty, Z_{ir_i})$. This way we can guarantee that latent variables resemble the given relational event data.
4. Repeat steps 2 and 3 for $s = 1, \ldots, S$.

The initial set of draws is discarded as they are part of the burn-in period and would depend on the arbitrarily chosen initial values.

Fitting a Bayesian relational event model with a normal prior (Bayesian ridge regression)

The Gibbs sampling algorithm has the following steps

1. Set initial values for $\boldsymbol{\beta}^{(0)}$, and $Z^{(0)}$, as well as for parameters $\lambda^{2(0)}$, $\delta^{(0)}$.
2. Draw $\boldsymbol{\beta}^{(s)}$ from its conditional posterior distribution given $Z^{(s-1)}$,

   $\boldsymbol{\beta}^{(s)}|Z^{(s-1)} \sim \mathcal{N}(\mu^{ridge}, \Sigma^{ridge})$, where

   $$\mu^{ridge} = \left(\sum_{i=1}^{M} X_i^T X_i + \frac{1}{\lambda^2} I_P\right)^{-1} \sum_{i=1}^{M} X_i^T Z_i^{(s-1)}$$

   $$\Sigma^{ridge} = \left(\sum_{i=1}^{M} X_i^T X_i + \frac{1}{\lambda^{2(s-1)}} I_P\right)^{-1}$$

3. Draw $Z_{ir}^{(s)}$ from its conditional posterior given $\boldsymbol{\beta}^{(s)}$, which is a truncated normal distribution.

   $Z^{(s)}|\boldsymbol{\beta}^{(s)} \sim t\mathcal{N}(X_i\boldsymbol{\beta}^{(s)}, I_N)$

   To ensure that the sampled values of the latent variables correspond to the given data and satisfy Eq. (5), we sample latent variables in the following way: first element of each vector Z_i is set to zero to reduce degrees of freedom; the element that corresponds to the observed actor is sampled from a truncated normal density on the interval $(\max_{r\neq r_i} Z_{ir}, \infty)$; and the elements which correspond to non-observed actors are sampled from truncated density on the complement interval $(-\infty, Z_{ir_i})$.
4. Draw the shrinkage parameter λ^2 and the expanded parameter δ:

   $$\lambda^{2(s)}|\boldsymbol{\beta}^{(s)}, \delta^{(s-1)} \sim \mathrm{IG}(\alpha_1 + \frac{P}{2}, \delta^{(s-1)} + \frac{1}{2}\sum_{p=1}^{P}(\beta_p^{(s)})^2)$$

   $$\delta^{(s)}|\lambda^{2(s)} \sim \mathrm{G}(\alpha_1 + \alpha_2, \frac{1}{\lambda^{2(s)}} + \frac{1}{b})$$

5. Repeat steps 2 to 4 for $s = 1, \ldots, S$.

Fitting a Bayesian relational event model with a Laplace prior (the Bayesian lasso)

The steps in the Gibbs sampler are as follows:

1. Set initial values for $\boldsymbol{\beta}^{(0)}, Z^{(0)}, \lambda^{2(0)}, \tau_1^{2(0)}, \ldots, \tau_P^{2(0)}, \delta^{(0)}$
2. Draw $\boldsymbol{\beta}^{(s)}$ from its conditional posterior distribution given $Z^{(s-1)}$, $\tau_1^{2(s-1)}, \ldots, \tau_P^{2(s-1)}$, $\lambda^{2(s-1)}$

   $\boldsymbol{\beta}^{(s)}|Z^{(s-1)}, \tau_1^{2(s-1)}, \ldots, \tau_P^{2(s-1)}, \lambda^{2(s-1)} \sim \mathcal{N}(\mu^{lasso}, \Sigma^{lasso})$, where

   $$\mu^{lasso} = \left(\sum_{i=1}^{M} X_i^T X_i + D_\tau^{-1}\right)^{-1} \sum_{i=1}^{M} X_i^T Z_i^{(s-1)}$$

   $$\Sigma^{lasso} = \left(\sum_{i=1}^{M} X_i^T X_i + D_\tau^{-1}\right)^{-1},$$

   $$D_\tau = diag\{\lambda^{2(s-1)}\tau_1^{2(s-1)}, \ldots, \lambda^{2(s-1)}\tau_P^{2(s-1)}\}$$

3. Update latent variables by sampling $Z_{ir}^{(s)}$ from its conditional posterior given $\boldsymbol{\beta}^{(s)}$, which is a truncated normal distribution, such that

   $Z^{(s)}|\boldsymbol{\beta}^{(s)} \sim t\mathcal{N}(X_i\boldsymbol{\beta}^{(s)}, I_N)$.

   and for an element of Z_i that corresponds to the observed actor the truncated interval is $(\max_{r\neq r_i} Z_{ir}, \infty)$ while elements that conform the actors that are not observed are truncated in the interval $(-\infty, Z_{ir_i})$. These conditions will guarantee that sampled latent variables fit the observed categorical data and according to Eq. (5). In addition, the first element of each Z_i is fixed to zero, such that there are less degrees of freedom in the generating process.

4. Draw the value of parameter $\boldsymbol{\tau}^{2(s)} = (\tau_1^{2(s)}, \dots, \tau_P^{2(s)})$ from its conditional posterior distribution, which in this case is an inverse-Gaussian distribution:

$$\frac{1}{\tau_p^{2(s)}} | \beta_p^{(s-1)}, \lambda^{2(s-1)} \sim \text{Inv-Gauss}(\mu' = \sqrt{\frac{\lambda^{2(s-1)}}{\beta_p^{2(s-1)}}}, \lambda' = 1), p = 1, \dots, P$$

5. Update the values of $\lambda^{2(s)}$ (similarly to the step 4 for the ridge model):

$$\lambda^{2(s)} | \tau_1^{2(s-1)}, \dots, \tau_P^{2(s-1)}, \delta^{(s-1)} \sim IG(\alpha_1 + \frac{P}{2}, \delta + \frac{1}{2} \sum_{p=1}^{P} \frac{\beta_p^2}{\tau_p^2})$$

$$p(\delta^{(s)} | \lambda^{2(s-1)}) \propto G(\alpha_1 + \alpha_2, \frac{1}{\lambda^{2(s-1)}} + \frac{1}{b})$$

6. Repeat steps 2 to 5 for $s = 1, \dots, S$.

Fitting a Bayesian relational event model with the horseshoe prior

The Gibbs sampling algorithm has the following steps:

1. Set initial values for $\boldsymbol{\beta}^{(0)}, Z^{(0)}, \boldsymbol{\lambda}^{2(0)} = (\tau_1^{2(0)}, \dots, \tau_P^{2(0)}), \lambda^{2(0)}$, and mixing parameters of the F densities: $\boldsymbol{\psi}^{\mathbf{2(0)}} = (\psi_1^{2(0)}, \dots, \psi_P^{2(0)})$, $\gamma^{2(0)}$.
2. Draw $\boldsymbol{\beta}^{(s)}$ from its conditional posterior distribution given $Z^{(s-1)}$, $\boldsymbol{\tau}^{2(s-1)}, \lambda^{2(s-1)}$

$$\boldsymbol{\beta}^{(s)} | Z^{(s-1)}, \boldsymbol{\tau}^{2(s-1)}, \lambda^{2(s)} \sim \mathcal{N}(\mu^{horseshoe}, \Sigma^{horseshoe}), \text{ where}$$

$$\mu^{horseshoe} = \left(\sum_{i=1}^{M} X_i^T X_i + D_\tau^{-1} \right)^{-1} \sum_{i=1}^{M} X_i^T Z_i^{(s-1)}$$

$$\Sigma^{horseshoe} = \left(\sum_{i=1}^{M} X_i^T X_i + D_\tau^{-1} \right)^{-1},$$

$$D_\tau = diag\{\lambda^{2(s-1)} \tau_1^{2(s-1)}, \dots, \lambda^{2(s-1)} \tau_P^{2(s-1)}\}$$

3. Draw the values of latent variables by sampling $Z_{ir}^{(s)}$ from conditional posterior given $\boldsymbol{\beta}^{(s)}$, which is a truncated normal distribution

$$Z^{(s)} | \boldsymbol{\beta}^{(s)} \sim t\mathcal{N}(X_i \boldsymbol{\beta}^{(s)}, I_N).$$

under the conditions that for an observed element of Z_i the truncated interval is $(\max_{r \neq r_i} Z_{ir}, \infty)$ and for elements that are not observed the truncated interval is $(-\infty, Z_{ir_i})$, while the first element of Z_i is always set to zero. Such conditions form latent variables that correspond to the observed categorical data by the definition of Eq. (5).
4. Draw the value of parameter $\boldsymbol{\tau}^{2(s)}$ from its conditional posterior-inverse gamma distribution

$$\tau_p^{2(s)} | \psi_p^{2(s-1)}, \lambda^{2(s-1)}, \beta_p^{(s)} \sim \text{IG}\left(\alpha_3 + \frac{1}{2}, \psi_p^{2(s-1)} + \frac{\beta_p^{2(s)}}{2\lambda^{2(s-1)}} \right), p = 1, \dots, P$$

5. Draw the value of parameter $\lambda^{2(s)}$ from its conditional posterior-inverse gamma distribution

$$\lambda^{2(s)} | \gamma^{2(s-1)}, \boldsymbol{\tau}^{2(s)}, \boldsymbol{\beta}^{(s)} \sim \text{IG}\left(\alpha_1 + \frac{P}{2}, \gamma^{2(s-1)} + \frac{1}{2} \sum_{p=1}^{P} \frac{\beta_p^{2(s)}}{\tau_p^{2(s)}} \right)$$

6. Update the mixing parameters $\boldsymbol{\psi}^{\mathbf{2(s)}}$ and $\gamma^{2(s)}$:

$$\gamma^{2(s)} | \lambda^{2(s)} \sim \text{G}\left(\alpha_1 + \alpha_2, \frac{1}{b_1} + \frac{1}{\lambda^{2(s)}} \right)$$

$$\psi_p^{s(s)} | \tau_p^{2(s)} \sim \text{G}\left(\alpha_3 + \alpha_4, \frac{1}{b_2} + \frac{1}{\tau_p^{2(s)}} \right), p = 1, \dots, P$$

7. Repeat steps 2 to 6 for $s = 1, \dots, S$.

References

Amati, V., Lomi, A., Mascia, D., 2019. Some days are better than others: Examining time-specific variation in the structuring of interorganizational relations. Social Networks 57 (December 2018), 18–33.

Arena, G., Mulder, J., Leenders, R.T.A., A Bayesian Semi-Parametric Approach for Modeling Memory Decay in Dynamic Social Networks, Sociological Methods & Research.

Brandes, U., Lerner, J., Snijders, T.A., 2009. Networks evolving step by step: Statistical analysis of dyadic event data. In: 2009 International Conference on Advances in Social Network Analysis and Mining. IEEE, pp. 200–205.

Butts, C.T., 2008. A relational event framework for social action. Sociol. Methodol. 38 (1), 155–200.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2009. Handling sparsity via the horseshoe. In: Artificial Intelligence and Statistics. pp. 73–80.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2010. The horseshoe estimator for sparse signals. Biometrika 97 (2), 465–480.

Cox, D.R., 1972. Regression models and life-tables. J. R. Stat. Soc. Ser. B Stat. Methodol. 34 (2), 187–202.

Diesner, J., Frantz, T.L., Carley, K.M., 2005. Communication networks from the enron email corpus "it's always about the people. Enron is no different". Comput. Math. Organ. Theory 11 (3), 201–228.

DuBois, C., Butts, C.T., McFarland, D., Smyth, P., 2013. Hierarchical models for relational event sequences. J. Math. Psych. 57 (6), 297–309.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. CRC Press.

Hastie, T., Martin, R., Hastie, W., 2016. Tibshirani, wainwright. Statistical Learning with Sparsity The Lasso and Generalizations Statistical Learning with Sparsity.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical learning with sparsity. Monogr. Statist. Appl. Probab. 143, 143.

Hedström, P., Bearman, P., 2009. What is analytical sociology all about? An introductory essay. Oxf. Handb. Anal. Sociol. 3–24.

Hoffman, M., Block, P., Elmer, T., Stadtfeld, C., 2020. A model for the dynamics of face-to-face interactions in social groups. Netw. Sci. 8 (S1), S4–S25.

Imai, K., Van Dyk, D.A., 2005. A Bayesian analysis of the multinomial probit model using marginal data augmentation. J. Econometrics 124 (2), 311–334.

Keila, P.S., Skillicorn, D.B., 2005. Structure in the enron email dataset. Comput. Math. Organ. Theory 11 (3), 183–199.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al., 2010. Penalized regression, standard errors, and Bayesian lassos. Bayesian Anal. 5 (2), 369–411.

Leenders, R.T.A., Contractor, N.S., DeChurch, L.A., 2016. Once upon a time: Understanding team processes as relational event networks. Organ. Psychol. Rev. 6 (1), 92–115.

Lerner, J., Lomi, A., 2018. Let's talk about refugees: Network effects drive contributor attention to wikipedia articles about migration-related topics. In: International Conference on Complex Networks and their Applications. Springer, pp. 211–222.

Lerner, J., Lomi, A., 2020. Reliability of relational event model estimates under sampling: How to fit a relational event model to 360 million dyadic events. Netw. Sci. 8 (1), 97–135.

Lerner, J., Lomi, A., Mowbray, J., Rollings, N., Tranmer, M., 2021. Dynamic network analysis of contact diaries. Social Networks 66, 224–236.

Liang, H., 2014. The organizational principles of online political discussion: A relational event stream model for analysis of web forum deliberation. Hum. Commun. Res. 40 (4), 483–507.

Liu, H., Xu, X., Li, J.J., 2020. A bootstrap lasso+ partial ridge method to construct confidence intervals for parameters in high-dimensional sparse linear models. Statist. Sinica 30 (3), 1333–1355.

Malang, T., Brandenberger, L., Leifeld, P., 2019. Networks and social influence in European legislative politics. Br. J. Political Sci. 49 (4), 1475–1498.

Mulder, J., Hoff, P.D., 2021. A latent variable model for relational events with multiple receivers. arXiv preprint arXiv:2101.05135.

Mulder, J., Leenders, R.T.A., 2019. Modeling the evolution of interaction behavior in social networks: A dynamic relational event approach for real-time analysis. Chaos Solitons Fractals 119, 73–85.

Mulder, J., Pericchi, L.R., et al., 2018. The matrix-*F* prior for estimating and testing covariance matrices. Bayesian Anal. 13 (4), 1193–1214.

Müller, J., Günther, E., Humbert, A., 2018. GEDII wearable sensors dataset of 8 research teams. http://dx.doi.org/10.5281/zenodo.1434706.

Park, T., Casella, G., 2008. The Bayesian lasso. J. Amer. Statist. Assoc. 103 (482), 681–686.

Perry, P.O., Wolfe, P.J., 2013. Point process modelling for directed interaction networks. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (5), 821–849.

Peterson, K., Hohensee, M., Xia, F., 2011. Email formality in the workplace: A case study on the enron corpus. In: Proceedings of the Workshop on Language in Social Media (LSM 2011). pp. 86–95.

Pilny, A., Schecter, A., Poole, M.S., Contractor, N., 2016. An illustration of the relational event model to analyze group interaction processes.. Group Dyn.: Theory Res. Pract. 20 (3), 181.

Quintane, E., Conaldi, G., Tonellato, M., Lomi, A., 2014. Modeling relational events: A case study on an open source software project. Organ. Res. Methods 17 (1), 23–50.

Shafiee Kamalabad, M., Grzegorczyk, M., 2020. Non-homogeneous dynamic Bayesian networks with edge-wise sequentially coupled parameters. Bioinformatics 36 (4), 1198–1207.

Stadtfeld, C., Block, P., 2017. Interactions, actors, and time: Dynamic network actor models for relational events. Sociol. Sci. 4, 318–352.

Stadtfeld, C., Hollway, J., Block, P., 2017. Dynamic network actor models: Investigating coordination ties through time. Sociol. Methodol. 47 (1), 1–40.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

Van Erp, S., Oberski, D.L., Mulder, J., 2019. Shrinkage priors for Bayesian penalized regression. J. Math. Psych. 89, 31–50.

Vu, D., Lomi, A., Mascia, D., Pallotti, F., 2017. Relational event models for longitudinal network data with an application to interhospital patient transfers. Stat. Med. 36 (14), 2265–2287.

Wilson, G., Banzhaf, W., 2009. Discovery of email communication networks from the enron corpus with a genetic algorithm using social network analysis. In: 2009 IEEE Congress on Evolutionary Computation. pp. 3256–3263.

Zhou, Y., Goldberg, M., Magdon-Ismail, M., Wallace, A., 2007. Strategies for cleaning organizational emails with an application to enron email dataset. In: 5th Conf. of North American Association for Computational Social and Organizational Science, number 0621303.