University of Kent

Kent Academic Repository

Ren, Sa, Liu, Peng, Zhang, Hai, Wang, Xue and Zhang, Jian (2026) *Dynamic Exponential Random Graph Models with the Fused Lasso Penalty*. Applied Network Science . ISSN 2364-8228.

Downloaded from
https://kar.kent.ac.uk/117279/ The University of Kent's Academic Repository KAR

The version of record is available from
https://doi.org/10.1007/s41109-026-00846-8

This document version
Publisher pdf

DOI for this version

Licence for this version
CC BY (Attribution)

Additional information

Versions of research works

Versions of Record
If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts
If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in ***Title of Journal*** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries
If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

RESEARCH | Article in Press

Dynamic exponential random graph models with the fused lasso penalty

Sa Ren · Peng Liu · Hai Zhang · Xue Wang · Jian Zhang

Received: 16 October 2025 / Accepted: 26 October 2026
© The Author(s) 2026

Abstract

This paper investigates how the parameters governing network structure vary over time across a sequence of observed dynamic networks. We focus on ensembles of networks that exhibit sudden structural changes at certain time points but remain stable or undergo minimal changes in between. Many real-world networks, such as international trade networks and political co-voting networks, display this behaviour. Given a sequence of observed networks, our objective is to estimate the corresponding time-varying parameters of an exponential random graph model (ERGM) and identify periods of parameter stability as well as abrupt changes in the underlying network-generating mechanisms. To this end, we introduce a novel fused lasso exponential random graph model (FL-ERGM), which encourages similar parameter values across consecutive time points while allowing abrupt changes when supported by the data. To address the intractability of the ERGM likelihood, we adopt a pseudo-likelihood approximation and optimise the resulting penalised objective using the alternating direction method of multipliers (ADMM). We prove that the solution obtained by ADMM is equivalent to the optimal solution of the original objective function. Monte Carlo simulations demonstrate that FL-ERGM outperforms existing methods in estimating time-varying ERGM parameters and detecting structural changes. Applications to international trade and political co-voting networks reveal interpretable and contextually relevant patterns of temporal parameter variation.

Keywords Exponential random graph models · Temporal variations · Fused Lasso · Alternating direction method of multipliers

Abbreviations

ERGM	Exponential random graph model
SABM	Stochastic actor-based model
TERGM	Temporal exponential random graph model
VCERGM	Varying coefficient exponential random graph model
FL-ERGM	Fused lasso exponential random graph model
ADMM	Alternating direction method of multipliers
LU	Lower-upper
gwdsp	Geometrically weighted dyad-wise shared partners

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



https://doi.org/10.1007/s41109-026-00846-8

Published online: 06 November 2026

Springer

Introduction

Networks, consisting of nodes and edges, are widely used to describe the interactions among different entities. Dynamic networks refer to an ensemble of networks obtained on the same or similar sets of nodes over different time points. With the recent development of technology, such data have become commonplace in many fields. Examples include trade networks built over different years of product exchange, brain networks constructed at different phases of disease, and friendship networks collected over different school years.

The analysis of network data is an emerging area in statistics. Over the last few decades, a significant number of statistical models have been proposed to infer the underlying structures of network data, including latent space models (Hoff et al. 2002; Sewell and Chen 2015), stochastic block models (Holland et al. 1983; Lee and Wilkinson 2019), and exponential random graph models (ERGMs) (Frank and Strauss 1986; Schweinberger et al. 2020). ERGMs are a class of versatile network models that are often used in single networks to describe the observed network structure through various local connectivity patterns and global network features. The ability to incorporate complex connectivity patterns through summary statistics makes ERGM an attractive model. There is a large and growing body of literature on the developments and inference of ERGMs, as reviewed by Schweinberger et al. (2020).

Stochastic actor-based models (SABMs) (Snijders et al. 2010) are one of the most well-known models for dynamic networks. They are specifically designed for modelling the network evolution using a stochastic process driven by nodes. SABMs explicitly simulate network dynamics by taking into account the actions of individual nodes and are particularly useful in understanding the network evolution at the node level. Recently, there are also several statistical models that utilise ERGMs in the dynamic network setting. Hanneke et al. (2010) proposed a temporal ERGM (TERGM) to capture the differences in network features between two consecutive networks through exponential family distributions. Krivitsky and Handcock (2014) extended the TERGM to a separable TERGM, which considers both the formation and dissolution processes of networks by separately parametrising the prevalence and duration of fluctuations. More recently, Lee et al. (2020) developed a model-based clustering framework based on the TERGMs so that the block structure of networks can be detected simultaneously. Through the TERGM framework, the changes between consecutive networks are modelled directly through an exponential family distribution, but the characteristics of each network are not captured in the model. Other methods, such as the varying coefficient ERGM (VCERGM) (Lee et al. 2020), are capable of modelling both the characteristics of individual networks and their evolution over time. The VCERGM is designed for networks with smoothly changing patterns.

Inferring the temporal variations of dynamic ensembles is challenging because it is difficult to simultaneously estimate both the structure within each network and the changes in the structure over time. Each network can display a wide range of dependency structures, and dynamic ensembles may manifest diverse types of changes. The range of possibilities includes a sudden shift of the entire network structure, a single node rewiring all of its connections, or even just one or two edges changing in the whole network. Therefore, any method should be flexible enough to discover many types of evolutionary patterns, while also be powerful enough to learn this temporal structure over a long time period.

In practice, a dynamic ensemble often exhibits a diverse changing pattern. For example, when fitting ERGMs to international trade networks, we observed that the ERGM parameters have a different changing trend before and after the global financial crisis in 2008. In the political co-voting networks, the affiliation parameters also remained stable from Congress 106 (1999–2001) to 109 (2005–2007), when Bush was the president, but then there was a sudden change at Congress 110 (2007–2009) due to the global financial crisis. While the VCERGM offer a flexible framework for modelling temporal networks, they typically assume that parameter trajectories evolve smoothly over time. This smoothness assumption limits their ability to detect important structural phenomena such as abrupt changes or periods of structural stability where the network remains unchanged. Thus, VCERGMs (Lee et al. 2020) can not adequately capture the temporal variations in the above examples.

The parameters of an ERGM quantify the effects of network features and covariates on the probability of observing a particular network structure. To account for diverse temporal patterns in network structure, we adopt the fused lasso approach by imposing an L_1 penalty (Tibshirani et al. 2005) on the differences between consecutive model parameter vectors. Specifically, we propose a fused lasso exponential random graph model (FL-ERGM) to describe the dynamic network evolution process. Unlike methods such as the graphical lasso (Friedman et al. 2008), which aim to recover the underlying network structure from observed nodal or dyadic data, our objective is to estimate the ERGM parameters from a sequence of observed networks. The fused lasso penalty encourages similar parameter estimates across consecutive time points while allowing abrupt changes between different parameter regimes. This regularisation provides flexibility in representing diverse temporal patterns in the ERGM parameters, allowing the FL-ERGM to capture a variety of changing patterns over time. Thus, the proposed FL-ERGM characterises within-network dependence through ERGMs and captures temporal evolution through changes in the corresponding time-varying ERGM parameters.

Though FL-ERGMs have many advantages over the existing methods as we described above, the estimation of the parameters in FL-ERGMs is challenging. First, the likelihood of ERGM is intractable. Second, the use of the fused lasso penalty in the dynamic networks also increases the difficulty of estimation as the fused lasso penalty is not separable and smooth, and we have to adapt the fused lasso to ERGMs to encourage sparsity on the differences of parameters. Third, when the observed dynamic ensemble has a large number of networks and nodes, strong computational power is needed due to the large sample size.

To address these challenges, we employed a pseudo-likelihood approximation to the likelihood function for its computational tractability (Schmid and Desmarais 2017). We optimised the penalised pseudo-likelihood objective using an iteratively reweighted least squares (IRLS) scheme, where each quadratic subproblem is efficiently solved via the alternating direction method of multipliers (ADMM). The ADMM framework decomposed the optimisation problem into several smaller subproblems (Boyd et al. 2010). Applying the ADMM algorithm to our problem enables us to leverage the known properties and solution methods to derive closed-form ADMM updates for each subproblem. However, obtaining these closed-form solutions involves solving a ridge regression in one of the subproblems, which significantly slows down computation due to the required matrix inversion. This issue is further exacerbated by the dynamic nature of the networks under consideration. To address this, we incorporate a lower-upper (LU) matrix decomposition technique into the ridge regression solution, which substantially accelerates our algorithm compared to a naive ADMM implementation.

The rest of this paper is organised as follows. We formulate the FL-ERGM in section "Model formulation". In section "Parameter estimation", we use the ADMM algorithm to solve the resulting optimisation problem. In section "Empirical results", we numerically evaluate the performance of our proposed method to the counterpart. We conclude this paper in section "Discussion".

Model formulation

In this section, we first briefly review the standard ERGM, which provides the foundation for our proposed FL-ERGM framework. We then introduce the proposed FL-ERGM, which extends the standard ERGM to a sequence of observed networks by allowing the ERGM parameters to vary over time.

Exponential random graph models

ERGMs provide a flexible framework for modelling diverse structural characteristics of networks (Robins et al. 2007; Schweinberger et al. 2020). The network is typically represented by a random adjacency matrix $Y \in \{0,1\}^{n\times n}$, with n for the number of nodes. We define $Y_{ij} = 1$ if there is an edge from node i to node j, and $Y_{ij} = 0$ otherwise. The realisation of Y is denoted by y and the set of all possible outcomes of Y is denoted by $\mathcal{Y}$. Let $X \in \mathcal{X}$ denote the observed network attributes, including nodal and dyadic attributes that may be associated with the formation

of network ties. The ERGM specifies the probability of observing a particular network y, conditional on the observed covariates X, as

$$P(Y = y|X;\theta) = \frac{\exp\{\theta^\top S(y,X)\}}{k(\theta)}, \tag{1}$$

where $\theta \in \mathbb{R}^d$ is the vector of ERGM parameters and $S(y,X)$ is a d-dimensional vector of network summary statistics (Morris et al. 2008). These statistics capture structural features of the network and associations with observed covariates, such as the number of edges, mutual ties, triangles, stars, or ties between nodes with shared attributes. Each parameter in θ determines the contribution of its associated statistic to the probability of observing a particular network, conditional on the other statistics. Thus, a positive parameter indicates that networks with larger values of the corresponding statistic are more likely to be observed, whereas a negative parameter indicates the opposite. Our objective is to estimate these ERGM parameters, which describe the strength and direction of the associations between the specified network statistics and the observed network structure. The normalising constant $k(\theta) = \sum_{y^* \in y} \exp\{\theta^\top S(y^*,X)\}$ is the sum over all potential graphs in the sample space, which is numerically intractable, except for very small graphs.

Fused lasso exponential random graph models

To study the changing patterns in network structures over time, we propose a FL-ERGM by adding a fused lasso constraint to the ERGM parameters. Specifically, we represent the marginal distribution of network Y_t at time t using an ERGM with parameter θ_t, and then we include the term of the sum of the absolute values of differences between parameters, $\theta_t - \theta_{t-1}$, to encourage similar estimates from neighbouring networks.

Thus, for a sequence of networks $\{Y_1, Y_2, \ldots, Y_T\}$ that are collected at T discrete time points with covariate $\{X_1, X_2, \ldots, X_T\}$, the FL-ERGM is formulated as follows,

$$\begin{aligned} P(Y_t = y_t|X_t;\theta_t) &= \frac{\exp\{\theta_t^\top S(y_t,X_t)\}}{k(\theta_t)} \quad (t = 1,\ldots,T) \\ \text{subject to} \quad &\sum_{t=2}^{T} \|\theta_t - \theta_{t-1}\|_1 \le m, \end{aligned} \tag{2}$$

where $\theta_t \in \mathbb{R}^d$ denotes the parameter vector at time t, $S(y_t,X_t)$ denotes the summary statistics of network y_t with covariates X_t, $k(\theta_t)$ denotes the normalising constant, $\|\cdot\|_1$ denotes the L_1 norm, and m controls the degree of similarity between θ_t and θ_{t-1} for each $2 \le t \le T$. In other words, it determines the amount of information to borrow from neighbouring networks. A larger value of m indicates less borrowing from adjacent networks, with each parameter θ_t relying mostly on the current individual network y_t. A smaller m results in more borrowing from adjacent networks, leading to greater similarity in the parameters across different time points.

In the proposed FL-ERGM framework, temporal structure is incorporated by penalising differences between successive parameter vectors, while the network structure at each time point is modelled through its own time-specific parameters. The L_1 penalty encourages sparsity in temporal differences, shrinking some differences exactly to zero and allowing certain ERGM parameters to remain unchanged over time while others vary. This formulation is particularly useful for characterising temporal evolution in which changes may occur abruptly and affect only specific aspects of network structure. FL-ERGM jointly estimates the temporal trajectories of the ERGM parameters and allows different parameters to have different temporal block structures.

The proposed FL-ERGM differs from existing approaches in several ways. TERGMs (Hanneke et al. 2010) provide a framework for modelling network evolution through temporal processes such as tie formation and dissolution. Rather than explicitly modelling these temporal processes, FL-ERGM focuses on how the ERGM parameters governing network structure evolve across a sequence of observed networks. The Bayesian fused lasso regression model by Betancourt et al. (2017) was designed for directed networks based on a model that

assumes conditional independence of pairs (Holland and Leinhardt 1981), and therefore cannot capture structures involving more than two nodes. In contrast, FL-ERGM can model complex network structures, such as triangles, stars, and clusters, in both directed and undirected networks. Moreover, the parameters in Betancourt et al. (2017) are edge-specific, whereas FL-ERGM focuses on network-level parameters. VCERGMs (Lee et al. 2020) impose smoothness on parameter trajectories, encouraging gradual changes over time, whereas FL-ERGM allows temporal changes to be sparse and potentially abrupt. Finally, the block-structured dynamic ERGMs of Jiang et al. (2023) adopt a piecewise polynomial model for structure discovery and trend estimation through a two-step procedure, where parameters are first estimated separately for each network and trends are then estimated using univariate trend filtering. In contrast, FL-ERGM jointly estimates the temporal trajectories of the ERGM parameters and allows different parameters to have different temporal block structures, whereas the block-structured dynamic ERGM imposes a common block structure across parameters.

Parameter estimation

Estimating FL-ERGMs is challenging due to the intractability of ERGMs, the specialty of the fused lasso penalty and the large scale of network data being collected. In this section, we propose to estimate FL-ERGMs using ADMM algorithm based on a pseudo-likelihood approximation. We begin by approximating the FL-ERGM likelihood using a pseudo-likelihood function in section "Pseudo-likelihood approximation". Then we discuss the selection of tuning parameters in section "Choice of tuning parameters". Next, we apply the ADMM algorithm to optimise the objective function by decomposing it into several smaller subproblems in section "ADMM algorithm". To address the computation cost related to matrix inversion in one subproblem, we employ the technique of LU decomposition, significantly accelerating the computation compared to a naive ADMM implementation. We also establish the convergence of the algorithm in section "Convergence of the ADMM algorithm".

Pseudo-likelihood approximation

Next, we consider inference for the proposed FL-ERGMs. Model parameters can be estimated by maximizing the penalised log-likelihood,

$$\arg\max_{\theta_t} \sum_{t=1}^{T} \{\theta_t^\top S(y_t, X_t) - \log\{k(\theta_t)\}\} - \lambda \sum_{t=2}^{T} \|\theta_t - \theta_{t-1}\|_1, \tag{3}$$

with λ being the tuning parameter. The intractability of the normalising constant $k(\theta_t)$ makes the exact likelihood-based inference computationally infeasible even for networks of moderate size.

To circumvent this issue, several approximation methods have been developed in the context of a single network (Geyer and Thompson 1992; Snijders 2002; Hunter and Handcock 2006)). One widely used approach is the Monte Carlo Maximum Likelihood Estimation (MCMLE) (Hunter and Handcock 2006)). The core idea of MCMLE is to approximate the ratio of normalising constants using importance sampling, based on a large sample of networks drawn from the ERGM distribution at an initial parameter value θ_0. While MCMLE provides an estimation to the exact likelihood, it is computationally intensive and can be sensitive to the choice of the initial parameter values and the convergence of the Markov chain. Extending the MCMLE approach to the FL-ERGM framework would further exacerbate these computational challenges as it would require approximating T normalising constant ratios at each iteration. Moreover, the included fused lasso penalty is not differentiable, making the inference of the exact-likelihood even more challenging.

Alternatively, pseudo-likelihood (Strauss and Ikeda 1990; Schmid and Desmarais 2017; Schmid and Hunter 2023) is often employed as a computationally efficient approximation to the exact likelihood. This approach approximates the likelihood using a product of the conditional probabilities for each edge. It enables efficient

parameter estimation without the need to calculate the normalising constant, making it particularly attractive for large-scale networks or when repeated estimation is required. Hereafter, we consider the pseudo-likelihood analogue corresponding to model (3).

In the pseudo-likelihood function, the conditional probability of each edge is described using the so-called change statistics. The change statistic for edge $y_{ij,t}$ is defined as $\Delta S_{ij,t} = S(y_{ij,t} = 1, y_{-ij,t}, X_t) - S(y_{ij,t} = 0, y_{-ij,t}, X_t)$. This quantity measures the change in global network structures when the status of one edge is altered from disconnected to connected. Here, $y_{ij,t}$ represents the edge between nodes and j at time t, and $y_{-ij,t} = \{y_{kl,t}, (k,l) \neq (i,j)\}$ denotes the remaining network excluding edge $y_{ij,t}$.

Under the ERGM framework, the conditional probability distribution of edge $y_{ij,t}$ is

$$P\left(y_{ij,t} = 1 | y_{-ij,t}, X_t\right) = \frac{\exp\left(\theta_t^\top \Delta S_{ij,t}\right)}{1 + \exp(\theta_t^\top \Delta S_{ij,t})}.$$

The pseudo-likelihood of network y_t is the joint conditional probability of all edges in the network. For directed networks, the pseudo-likelihood is

$$\prod_{i \neq j} P\left(y_{ij,t} | y_{-ij,t}, X_t\right) = \prod_{i \neq j} P\left(y_{ij,t} = 1 | y_{-ij,t}, X_t\right)^{y_{ij,t}} \left[1 - P\left(y_{ij,t} = 1 | y_{-ij,t}, X_t\right)\right]^{1 - y_{ij,t}}.$$

For undirected networks, $y_{ij} = y_{ji}$, so we only need to consider the cases for $i < j$. To keep the notation simple, we focus on the directed networks in the rest of the paper. Using change statistics, the log pseudo-likelihood of y_t is written as

$$\begin{aligned}
&\sum_{i \neq j} y_{ij,t} \log\left[P\left(y_{ij,t} = 1 | y_{-ij,t}, X_t\right)\right] + (1 - y_{ij,t}) \log\left[1 - P\left(y_{ij,t} = 1 | y_{-ij,t}, X_t\right)\right] \\
&= \sum_{i \neq j} \left\{ y_{ij,t} \theta_t^\top \Delta S_{ij,t} - \log\left[1 + \exp\left(\theta_t^\top \Delta S_{ij,t}\right)\right] \right\}.
\end{aligned}$$

Thus, the pseudo-likelihood based estimator for FL-ERGMs can be obtained by

$$\arg\max_{\theta_t} \sum_{t=1}^{T} \sum_{i \neq j} \left\{ y_{ij,t} \theta_t^\top \Delta S_{ij,t} - \log\left[1 + \exp\left(\theta_t^\top \Delta S_{ij,t}\right)\right] \right\} - \lambda \sum_{t=2}^{T} \|\theta_t - \theta_{t-1}\|_1, \tag{4}$$

where θ_t is the model parameter, $\Delta S_{ij,t}$ is the change statistic, reflecting the change in a variety of local and global connectivity patterns as well as network covariates.

To facilitate model inference and use closed-form updates in the estimation procedure, we use a Taylor expansion to form a quadratic approximation to the first part of formula (4) at current estimates $\tilde{\theta}_t$ (Friedman et al. 2010),

$$-\frac{1}{2} \sum_{t=1}^{T} \sum_{i \neq j} w_{ij,t} (\theta_t^\top \Delta S_{ij,t} - z_{ij,t})^2 + C(\tilde{\theta}_t) \quad \text{(second order Taylor expansion)}, \tag{5}$$

where

$$
\begin{aligned}
\tilde{p}_{ij,t} &= P(y_{ij,t} = 1|\tilde{\theta}_t, y_{-ij,t}, X_t) = \frac{\exp(\tilde{\theta_t}^\top \Delta S_{ij,t})}{1 + \exp(\tilde{\theta_t}^\top \Delta S_{ij,t})}, \\
w_{ij,t} &= \tilde{p}_{ij,t}(1 - \tilde{p}_{ij,t}) = \frac{\exp(\tilde{\theta_t}^\top \Delta S_{ij,t})}{(1 + \exp(\tilde{\theta_t}^\top \Delta S_{ij,t}))^2}, \\
z_{ij,t} &= \tilde{\theta_t}^\top \Delta S_{ij,t} + \frac{y_{ij,t} - \tilde{p}_{ij,t}}{w_{ij,t}}, \\
C(\tilde{\theta_t}) &= -\frac{(y_{ij,t} - \tilde{p}_{ij,t})^2}{2w_{ij,t}} + f(\tilde{\theta_t}).
\end{aligned} \tag{6}
$$

The last term of formula (5), $C(\tilde{\theta_t})$, can be omitted as it does not contain θ_t. Details of the expansion can be found in Appendix A.

With the quadratic approximation, the pseudo-likelihood estimator of FL-ERGMs can be obtained by solving the penalised weighted least square problem iteratively,

$$
\arg\min_{\theta_t} \frac{1}{2} \sum_{t=1}^{T} \sum_{i \neq j} w_{ij,t}(\theta_t^\top \Delta S_{ij,t} - z_{ij,t})^2 + \lambda \sum_{t=2}^{T} \|\theta_t - \theta_{t-1}\|_1. \tag{7}
$$

Choice of tuning parameters

The tuning parameter λ controls the similarity of coefficients at different time points. When $\lambda = 0$, parameters for each network observation are estimated individually. In this case, using FL-ERGMs is the same as estimating each network separately with standard ERGMs. The bigger the λ, the more information we take from neighbouring networks. As $\lambda \to \infty$, all estimates at different time points tend to be the same. The added fused lasso penalty makes it possible to estimate single network parameters while considering information from other networks in the meantime.

Including the penalty term with optimal tuning parameters have two advantages. The first advantage is that each observed network is a single sample from the real unknown network. Even when the underlying networks at different time points come from the same distribution, the network observations at different time points appear to be different because of the sampling variability. In this case, borrowing information from other networks provides a more accurate estimation. The second advantage is that the evolution of networks is a time-dependent process. The networks that are observed at close time points tend to exhibit similar values of network structural statistics, which can be incorporated in the model through the use of a penalty term. An optimal λ can tell how much information to borrow from neighbouring networks.

Generally, cross-validation is one approach to find the optimal tuning parameters. However, the data partition of network data in the cross-validation is complicated because edge variables are dependent on each other (Li et al. 2020). Splitting network nodes into groups would destroy the network structure. Also, the cross-validation method for the proposed model can be very computationally expensive. Therefore, we adopt Bayesian information criterion (BIC) to select the tuning parameters. To be specific, we first choose a sequence of different values of λ. Then we perform model estimations at each λ, and obtain one BIC value for each λ. The λ with the lowest BIC value is the optimal tuning parameter. We use the number of distinct parameter groups as an estimator for the degrees of freedom (Tibshirani and Taylor 2011, 2012). We use the total number of dyads across the network sequence as a practical sample-size term in PBIC, while recognising the dependence among dyads. Thus, the pseudo-based BIC (PBIC) for FL-ERGMs is expressed as

$$PBIC(\lambda) = -2\sum_{t=1}^{T}\sum_{i\neq j}\left\{y_{ij,t}\hat{\theta}_t(\lambda)^\top \Delta S_{ij,t} - \log\left[1+\exp\left(\hat{\theta}_t(\lambda)^\top \Delta S_{ij,t}\right)\right]\right\} + K_u(\lambda)\cdot\log(N),$$

where $\hat{\theta}_t(\lambda)$ is the estimation with tuning parameter λ, $K_u(\lambda)$ is the number of distinct parameter groups and N is the total number of dyads across the T observed networks.

ADMM algorithm

The ADMM algorithm (Boyd et al. 2010) is highly effective for solving large-scale optimisation problems, as it decomposes the overall task into several smaller, simpler subproblems. We employ ADMM to estimate the proposed FL-ERGMs, leveraging its well-established properties and solution techniques to derive closed-form updates for each subproblem in this section.

To begin with, we express the objective function of the quadratic approximation in (7) in a matrix form as follows:

$$\min \frac{1}{2}\|A\Theta - b\|_2^2 + \lambda\|F\Theta\|_1. \tag{8}$$

$A \in \mathbb{R}^{N\times(d\times T)}$ is a block diagonal matrix with A_t, $t = 1, 2, \ldots, T$ as diagonal entries. $N = \sum_{t=1}^{T} N_t$ is the total number of dyads in all networks, with N_t being the number of dyads and n_t being the number of nodes for network y_t. The diagonal block $A_t \in \mathbb{R}^{N_t\times d}$ including information on the weight and the change statistics, has the following specification,

$$A_t = \begin{bmatrix} \sqrt{w_{12,t}}\Delta S^1_{12,t} & \sqrt{w_{12,t}}\Delta S^2_{12,t} & \cdots & \sqrt{w_{12,t}}\Delta S^d_{12,t} \\ \sqrt{w_{13,t}}\Delta S^1_{13,t} & \sqrt{w_{13,t}}\Delta S^2_{13,t} & \cdots & \sqrt{w_{13,t}}\Delta S^d_{13,t} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{w_{n_t(n_t-1),t}}\Delta S^1_{n_t(n_t-1),t} & \sqrt{w_{n_t(n_t-1),t}}\Delta S^2_{n_t(n_t-1),t} & \cdots & \sqrt{w_{n_t(n_t-1),t}}\Delta S^d_{n_t(n_t-1),t} \end{bmatrix}.$$

Θ is a $(d\times T)$-dimensional vector

$$\Theta = (\theta_t^\top)^\top = (\theta_{1,t}, \theta_{2,t}, \ldots, \theta_{d,t})^\top,\ t = 1, \ldots, T.$$

b is an N-dimensional vector

$$b = (\sqrt{w_{12,t}}z_{12,t}, \sqrt{w_{13,t}}z_{13,t}, \ldots, \sqrt{w_{n_t(n_t-1),t}}z_{n_t(n_t-1),t})^\top,\ t = 1, \ldots, T.$$

The pairwise difference matrix F indicates the position of the penalty. It has $d\times(T-1)$ rows and $d\times T$ columns, with the following form

$$F = \begin{bmatrix} -1 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & -1 & \cdots & 1 \end{bmatrix}.$$

The first d rows correspond to the difference between θ_1 and θ_2, the next d rows to the difference between θ_2 and θ_3, and so on. Within each block, the corresponding entries are -1 and 1 for the parameters at the two consecutive time points, respectively, with all other entries equal to zero.

We then reformulate the matrix-form quadratic approximation in (8) as a constrained optimisation problem by introducing an auxiliary variable Q:

$$\begin{aligned} \min \quad & \frac{1}{2}\|A\Theta - b\|_2^2 + \lambda\|Q\|_1 \\ \text{subject to} \quad & F\Theta - Q = 0. \end{aligned} \tag{9}$$

Here, vectors $\Theta \in \mathbb{R}^{d\times T}$ and $Q \in \mathbb{R}^{d\times(T-1)}$ are two primal variables. This reformulation separates the quadratic loss from the fused lasso penalty, making the problem suitable for ADMM.

To facilitate the application of ADMM, we incorporate the constraint in (9) into the augmented Lagrangian. The augmented Lagrangian is given by

$$L_\rho(\Theta, Q, y) = \frac{1}{2}\|A\Theta - b\|_2^2 + \lambda\|Q\|_1 + y^\top(F\Theta - Q) + \frac{\rho}{2}\|F\Theta - Q\|_2^2, \tag{10}$$

where $y \in \mathbb{R}^{d\times(T-1)}$ is a dual variable and $\rho > 0$ is a penalty parameter.

For convenience, we use the equivalent scaled form of the augmented Lagrangian:

$$\begin{aligned} L_\rho(\Theta, Q, U) &= \frac{1}{2}A\Theta - b_2^2 + \lambda Q_1 + \rho U^\top(F\Theta - Q) + \frac{\rho}{2}F\Theta - Q_2^2 \\ &= \frac{1}{2}A\Theta - b_2^2 + \lambda Q_1 + \frac{\rho}{2}F\Theta - Q + U_2^2 - \frac{\rho}{2}U_2^2, \end{aligned} \tag{11}$$

where $U = (1/\rho)y$ is the scaled dual variable. The scaled form is then used to derive the alternating updates for Θ, Q and U in the ADMM algorithm.

Thus, the ADMM algorithm can be performed by updating variables Θ, Q, U in turn until convergence:

$$\Theta^{k+1} = \arg\min_\Theta L_\rho(\Theta, Q^k, U^k), \tag{12}$$

$$Q^{k+1} = \arg\min_Q L_\rho(\Theta^k, Q, U^k), \tag{13}$$

$$U^{k+1} = U^k + F\Theta^{k+1} - Q^{k+1}. \tag{14}$$

Specifically, the ADMM algorithm consists of the following steps:

(1) Θ updates,

$$\begin{aligned} \Theta^{k+1} &= \arg\min_\Theta \frac{1}{2}\|A\Theta - b\|_2^2 + \frac{\rho}{2}\|F\Theta - Q^k + U^k\|_2^2 \\ &= (A^\top A + \rho F^\top F)^{-1}[A^\top b + \rho F^\top(Q^k - U^k)]. \end{aligned} \tag{15}$$

(2) Q updates,

$$\begin{aligned} Q^{k+1} &= \arg\min_Q \ \lambda\|Q\|_1 + \frac{\rho}{2}\|F\Theta^{k+1} - Q + U^k\|_2^2 \\ &= S_{\lambda/\rho}(F\Theta^{k+1} + U^k). \end{aligned} \tag{16}$$

(3) U updates,

$$U^{k+1} = U^k + F\Theta^{k+1} - Q^{k+1}. \tag{17}$$

Here, k is the iteration number and $S_\delta(z)$ denote the soft-thresholding operator with value

$$S_\delta(z) = \begin{cases} z - \delta & \text{if } z > 0 \text{ and } z > \delta \\ z + \delta & \text{if } z < 0 \text{ and } z < -\delta \\ 0 & |z| \le \delta. \end{cases}$$

The termination criterion is that the primal and dual residuals must be smaller than the thresholds ε^{pri} and $\varepsilon^{\text{dual}}$,

$$\|F\Theta^{k+1} - Q^{k+1}\|_2 \leq \varepsilon^{\text{pri}} \text{ and } \| - \rho F^\top (Q^{k+1} - Q^k)\|_2 \leq \varepsilon^{\text{dual}},$$

where

$$\varepsilon^{\text{pri}} = \sqrt{(d-1) \times T}\varepsilon^{\text{abs}} + \varepsilon^{\text{rel}} \max\{\|F\Theta^k\|_2, \| - Q^k\|_2\},$$
$$\varepsilon^{\text{dual}} = \sqrt{d \times T}\varepsilon^{\text{abs}} + \rho\varepsilon^{\text{rel}}\|F^\top U^k\|_2.$$

We note that in the first step of ADMM algorithm (15), the calculation of $(A^\top A + \rho F^\top F)^{-1}$ is challenging as T becomes large. We propose to calculate the matrix inverse of $X = A^\top A + \rho F^\top F$ by using the LU decomposition technique. We show that X is a block tridiagonal matrix and the inverse of X, a square matrix with dimension $dT \times dT$, can be solved by calculating the inverse of T smaller matrices of dimension d in Appendix B. The use of LU decomposition reduce the computational burden to a large extent.

In summary, we estimated the proposed FL-ERGM using an IRLS–ADMM algorithm. The estimation procedure consists of an outer IRLS loop that updates working responses and weights, combined with an inner ADMM routine that solves the resulting quadratic optimisation problem. The algorithm iterated until changes in the parameter estimates fell below a specified tolerance, with recommended initializations, maximum iterations, and typical runtimes provided for practical guidance. The corresponding pseudo-code is presented in Algorithm 1, ensuring reproducibility and clarity of the computational procedure.

Algorithm 1: IRLS–ADMM Estimation Algorithm

Input: Sequence of networks $\{Y_1, \ldots, Y_T\}$, regularisation parameters λ, ρ, tolerances $\varepsilon_{\text{IRLS}}$ (10^{-4}), $\varepsilon_{\text{pri}}(10^{-4})$, $\varepsilon_{\text{dual}}$ (10^{-4}), maximum iterations K_{IRLS} (10^3), K_{ADMM} (10^4)
Preprocessing: compute change statistics $\Delta S_{ij,t}$ and responses $y_{ij,t}$
Initialize: $\Theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_T^{(0)})$, $k \leftarrow 0$;
repeat
  for $t = 1$ **to** T **do**
    forall dyads (i, j) **do**
      Compute probabilities: $p_{ij,t}^{(k)} = \frac{1}{1+\exp(-\theta_t^{(k)\top}\Delta S_{ij,t})}$
      Compute weights: $w_{ij,t}^{(k)} = p_{ij,t}^{(k)}(1 - p_{ij,t}^{(k)})$
      Compute working responses: $z_{ij,t}^{(k)} = \theta_t^{(k)\top}\Delta S_{ij,t} + \frac{y_{ij,t} - p_{ij,t}^{(k)}}{w_{ij,t}^{(k)}}$
  Compute $A^{(k)}$ and $b^{(k)}$ using $w_{ij,t}^{(k)}$ and $z_{ij,t}^{(k)}$ to rewrite the objective function based on specifications in (8);
  Initialize ADMM variables, $m \leftarrow 0$;
  repeat
    Update primal variable:
    $\Theta^{(m+1)} = (A^{(k)\top}A^{(k)} + \rho F^\top F)^{-1}[(A^{(k)\top}b^{(k)} + \rho F^\top (Q^{(m)} - U^{(m)})]$
    Update auxiliary variable: $Q^{(m+1)} = S_{\lambda/\rho}(F\Theta^{(m+1)} + U^{(m)})$
    Update dual variable: $U^{(m+1)} = U^{(m)} + F\Theta^{(m+1)} - Q^{(m+1)}$
    Compute primal residual: $r^{(m+1)} = \|F\Theta^{(m+1)} - Q^{(m+1)}\|_2$
    Compute dual residual: $s^{(m+1)} = \rho\|F^\top (Q^{(m+1)} - Q^{(m)})\|_2$
    $m \leftarrow m + 1$
  until ADMM convergence $(r^{(m)} \leq \varepsilon_{\text{pri}}$ and $s^{(m)} \leq \varepsilon_{\text{dual}})$ or $m = K_{\text{ADMM}}$;
  Update $\Theta^{(k)} \leftarrow$ ADMM solution;
  $k \leftarrow k + 1$;
until IRLS convergence $(\frac{\|\Theta^{(k)} - \Theta^{(k-1)}\|_2}{\max(1, \|\Theta^{(k-1)}\|_2)} \leq \varepsilon_{\text{IRLS}})$ or $k = K_{\text{IRLS}}$;
return $\hat{\Theta} = \Theta^{(k)}$

Convergence of the ADMM algorithm

Next, we demonstrate that the solution found by the ADMM algorithm converges to the optimal solution of (4), the FL-ERGM optimisation problem. We first show that the optimal solution of the quadratic approximation in (7) converges to that of the objective function in (4). Define

$$l(\theta_1,\dots,\theta_T)=\sum_{t=1}^{T}\sum_{i\neq j}-y_{ij,t}\theta_t^\top\Delta S_{ij,t}+\log\left\{1+\exp\left(\theta_t^\top\Delta S_{ij,t}\right)\right\},$$

$$\tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1,\dots,\theta_T)=\frac{1}{2}\sum_{t=1}^{T}\sum_{i\neq j}w_{ij,t}(\theta_t^\top\Delta S_{ij,t}-z_{ij,t})^2-C(\tilde{\theta}_t).$$

Lemma 1 *$l(\theta_1,\dots,\theta_T)$ and $\tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1,\dots,\theta_T)$ are strictly convex.*

We prove Lemma 1 by showing that the Hessian matrices of $l(\theta_1,\dots,\theta_T)$ and $\tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1,\dots,\theta_T)$ are positive definite. The proof details can be found in Appendix C The strict convexity established here ensures that any local minimum discovered by our numerical solver is globally optimal and unique. Consequently, it guarantees that the tracking framework is structurally stable and free from deceptive local traps or saddle points during optimisation.

Lemma 2 *For any $t\in\{1,\dots,T\}$, as $\tilde{\theta}_t\to\theta_t$, the surrogate objective function $\tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1,\dots,\theta_T)$ converges to the original loss function $l(\theta_1,\dots,\theta_T)$:*

$$\lim_{\tilde{\theta}_t\to\theta_t}\tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1,\dots,\theta_T)=l(\theta_1,\dots,\theta_T).$$

The pointwise convergence in Lemma 2 supports the use of the computationally efficient quadratic surrogate as a local proxy for the non-linear logistic loss. It ensures that as the reference parameter $\tilde{\theta}_t$ iteratively updates, the approximation error diminishes, so the surrogate function increasingly approximates the behaviour of the original loss.

Next, define

$$(\hat{\theta}_1,\dots,\hat{\theta}_T)=\arg\min l(\theta_1,\dots,\theta_T), \tag{18}$$

$$(\hat{\tilde{\theta}}_1,\dots,\hat{\tilde{\theta}}_T)=\arg\min \tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1,\dots,\theta_T). \tag{19}$$

Then we have the following proposition:

Proposition 1 *$(\hat{\theta}_1,\dots,\hat{\theta}_T)$ and $(\hat{\tilde{\theta}}_1,\dots,\hat{\tilde{\theta}}_T)$ defined in Eqs. (18) and (19) exist and are unique. Furthermore, as $\tilde{\theta}_t\to\theta_t$, we have $(\hat{\tilde{\theta}}_1,\dots,\hat{\tilde{\theta}}_T)\to(\hat{\theta}_1,\dots,\hat{\theta}_T)$.*

The strict convexity in Lemma 1 ensures uniqueness of the minimisers, while Lemma 2 establishes convergence of the surrogate to the original loss. Together, these results support the convergence of the corresponding minimisers stated in Proposition 1.

In the next proposition, we show that the ADMM solution converges to the optimal solution of quadratic approximation (8) under the following assumptions.

Assumption 1 The (extended-value-real) functions $1/2\|A\Theta-b\|_2^2:\mathbb{R}^{d\times T}\to\mathbb{R}\cup\{+\infty\}$ and $\lambda\|Q\|_1:\mathbb{R}^{d\times(T-1)}\to\mathbb{R}\cup\{+\infty\}$ are closed, proper and convex.

Assumption 2 The unaugmented Lagrangian $1/2\|A\Theta-b\|_2^2+\lambda\|Q\|_1+\rho U^\top(F\Theta-Q)$ has a saddle point, which means there exists a point (Θ^*,Q^*,U^*) (not necessarily unique), for which

$$\begin{aligned}&1/2\|A\Theta^*-b\|_2^2+\lambda\|Q^*\|_1+\rho U^\top(F\Theta^*-Q^*)\\&\le 1/2\|A\Theta^*-b\|_2^2+\lambda\|Q^*\|_1+\rho U^{*\top}(F\Theta^*-Q^*)\\&\le 1/2\|A\Theta-b\|_2^2+\lambda\|Q\|_1+\rho U^{*\top}(F\Theta-Q)\end{aligned}$$

holds for all (Θ,Q,U).

Assumption 3 The matrix F has a full row rank.

Assumption 1 holds by construction, since the quadratic loss and ℓ_1-norm are continuous, proper, and convex. Assumption 2 specifies the existence of a saddle point required for the standard ADMM convergence result, while Assumption 3 provides the additional condition for convergence of the primal sequence.

Proposition 2 *(Convergence of the ADMM algorithm) Suppose that Assumptions 1 and 2 hold. Then as* $k \to \infty$, $1/2\|A\Theta^k - b\|_2^2 + \lambda\|Q^k\|_1$ *converges to the optimal value of the ADMM reformulation (9). If assumption 3 also holds,* Θ^k *converges to a solution of the quadratic approximation (8).*

Proposition 2 provides a theoretical guarantee for the proposed computational procedure. Under Assumptions 1 and 2, the objective values generated by ADMM converge to the optimal value of the reformulation. When Assumption 3 additionally holds, the primal sequence Θ^k converges to a solution of (8). The proof details of Proposition 2 can be found in Appendix D.

Empirical results

In this section, we show the performance of the proposed FL-ERGM using both simulation studies and real-world case studies. We first apply FL-ERGM to simulated networks in section "Simulated networks" so that we can compare the model results with the true parameter values under various specifications. Then we demonstrate its ability in identifying real-world change patterns using international trade networks in section "International trade networks". At the same time, the comparison of FL-ERGM with other competing methods are also presented in this section. Finally, an application of FL-ERGM to political co-voting networks can be found in Appendix F.

Simulated networks

We assess the performance of FL-ERGM in identifying various patterns of parameter changes using simulated networks. To examine its ability to estimate both structural network effects and covariate effects, we consider categorical and continuous nodal covariates. Specifically, X^1 denotes a categorical nodal covariate representing group membership, while X^2 denotes a continuous nodal covariate. These covariates allow us to evaluate the estimation of parameter changes associated with categorical homophily and continuous covariate effects.

First, we specify three models with different combinations of network summary statistics. The commonly used ERGM summary statistics are defined as follows (Morris et al. 2008; Schweinberger et al. 2020):

- $S^1(y_i) = \sum_{r \neq s} y_{rs,i}$, the number of edges in network y_i.
- $S^2(y_i) = \sum_{r \neq s} y_{rs,i} y_{sr,i}$, the number of mutual connections in network y_i.
- $S^3(y_i) = e^{\phi} \sum_{k=1}^{n-2} \{1 - (1 - e^{-\phi})^k\} DP_k(y_i)$, $\phi = 0.25$, geometrically weighted dyad-wise shared partners (gwdsp), a representation for local clustering, where $DP_k(y_i)$ is the number of dyads that have k common neighbors in network y_i.
- $S^4(y_i) = \sum_{r \neq s} y_{rs,i} \mathbf{1}(X_r^1 = X_s^1)$, the number of connections between individuals with the same categorical covariate X^1.
- $S^5(y_i) = \sum_{r \neq s} y_{rs,i} (X_r^2 + X_s^2)$, a measurement associated with the continuous covariate X^2.

For model 1, three summary statistics are chosen, including the number of edges $S^1(y_i)$, the number of mutual edges $S^2(y_i)$ and the categorical covariate term $S^4(y_i)$, to represent the network density, the reciprocity feature and the covariate effect. For model 2, we replace the number of mutual edges $S^2(y_i)$ with gwdsp $S^3(y_i)$ to reflect the local clustering feature. For model 3, we include the number of edges $S^1(y_i)$, the categorical covariate term $S^4(y_i)$ and the continuous covariate term $S^5(y_i)$ to focus on the covariate effects.

Then, for each model, we generate dynamic network sequences under five different scenarios. These scenarios are designed to represent a range of temporal patterns in the underlying model parameters and to assess the ability of FL-ERGM to identify different types of parameter changes. In the first scenario, we generate $T = 10$ networks with constant underlying model parameters. In the second scenario, we generate $T = 20$ networks with spike-like change patterns in model parameters. In the third scenario, we generate $T = 30$ networks with piecewise constant parameters. In the fourth scenario, we generate $T = 40$ networks with smoothly varying parameters. In the last scenario, we generate $T = 50$ networks with sudden changes in parameters. The binary covariate X^1 is sampled from a Bernoulli distribution with probability $p = 0.5$ to provide balanced representation of the two categorical groups and the continuous covariate X^2 is sampled from a standard normal distribution to provide a continuous covariate centred at zero with unit variance. The number of nodes in each network is set as 30. Network samples are simulated using R package ergm (Hunter et al. 2008).

After the network ensembles are generated, we apply the proposed FL-ERGM to the simulated networks, with the tuning parameter selected using PBIC, and compare the resulting estimates and change points with the true parameters and change points under various settings. For comparison purpose, we also fit the competing method VCERGM (Lee et al. 2020) to the same simulated networks, as well as a separate ERGM to each individual network. Model estimates and true parameter values are shown in Figs. 1,2,3,4 and 5 for each scenario separately. The R code used for the simulation study is available on GitHub (https://github.com/SRenStats/FL-ERGMs).

To assess method performance, we compute the integrated absolute error (IAE) to measure the total deviation between estimated and true coefficients across all ERGM parameters and time points, defined as

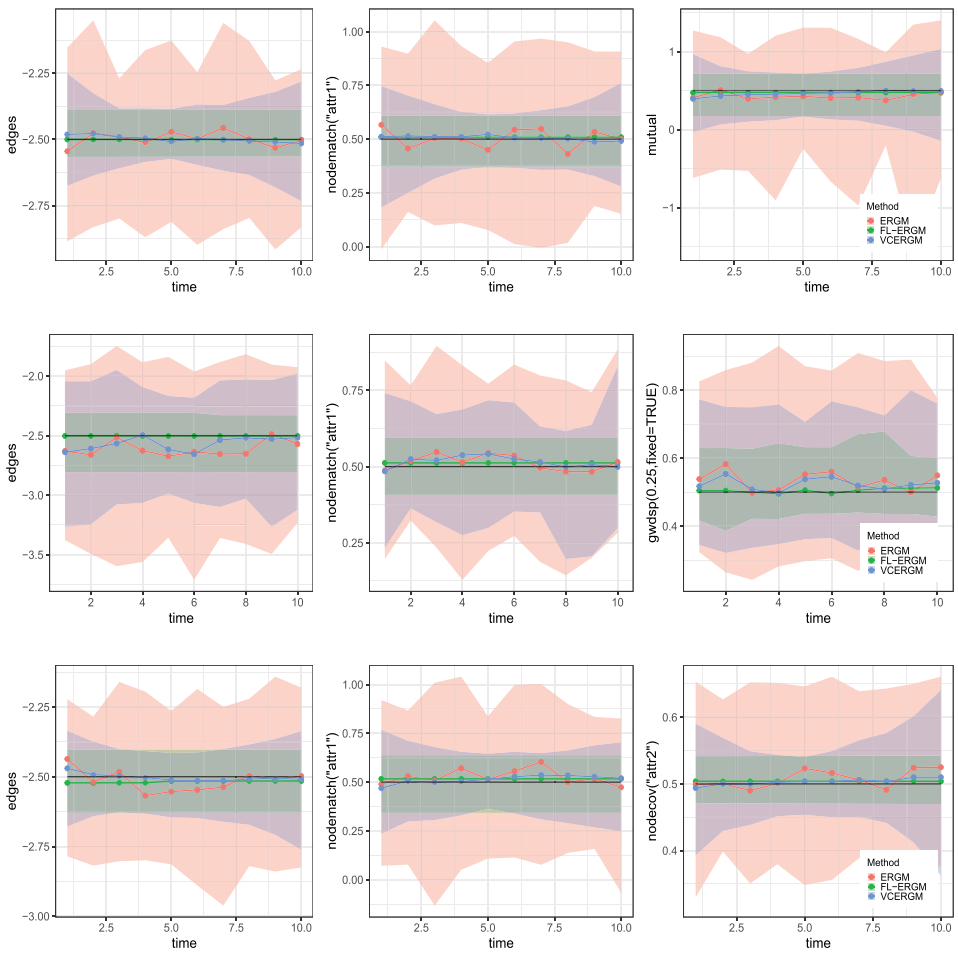


Fig. 1 Estimates for scenario 1 with constant underlying parameters. Black lines indicate the true parameters. Green lines show FL-ERGM estimates, blue lines stand for VCERGM estimates and red lines represent ERGM estimates. for each method, solid lines indicate the average of 50 estimated curves and the shaded bands illustrate the 0.025 and 0.975 quantiles

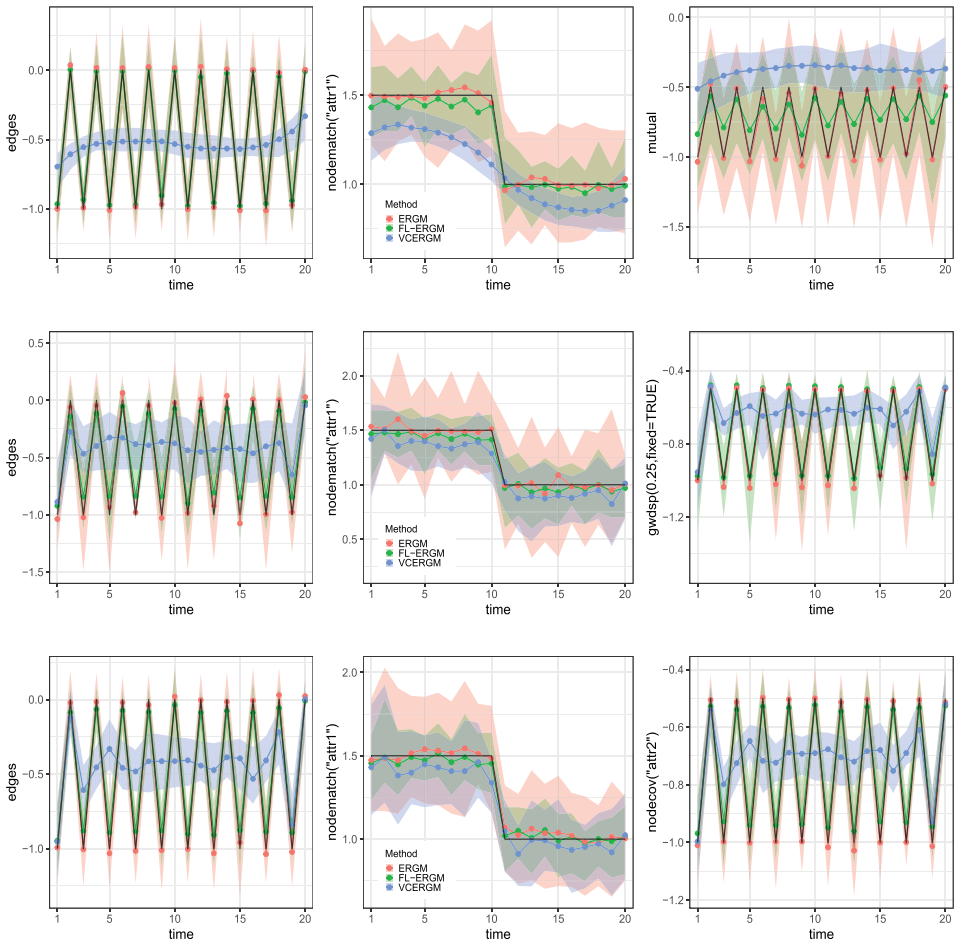


Fig. 2 Estimates for scenario 2 with spike-like change patterns in model parameters. Black lines indicate the true parameters. Green lines show FL-ERGM estimates, blue lines stand for VCERGM estimates and red lines represent ERGM estimates. for each method, solid lines indicate the average of 50 estimated curves and the shaded bands illustrate the 0.025 and 0.975 quantiles

$$\text{IAE} = \sum_{t=1}^{T} \|\hat{\theta}_t - \theta_t\|_1,$$

where $\|\cdot\|_1$ denotes the L_1 norm.

A time point t is defined as an abrupt change point if the absolute first-order difference exceeds a fixed threshold, $|\Delta\theta_t| > \delta_{abs}$, and the change at time t is substantially larger than the preceding change, $|\Delta\theta_t| > \delta_{rel} \cdot |\Delta\theta_{t-1}|$, where $\Delta\theta_t = \theta_t - \theta_{t-1}$. Change point detection performance is evaluated using sensitivity, specificity, and localisation error, defined as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$
$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$
$$\text{Localisation error} = \frac{1}{D_{\text{TP}}} \sum_{i=1}^{D_{\text{TP}}} |t_{\text{detected},i} - t_{\text{true},i}|,$$

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively, and D_{TP} is the number of correctly detected change points.

The mean and standard deviation of IAE across 50 replicates are presented in Table 1 for FL-ERGM, VCERGM and ERGM separately, with smaller values indicating higher accuracy. The results show that FL-ERGM achieves

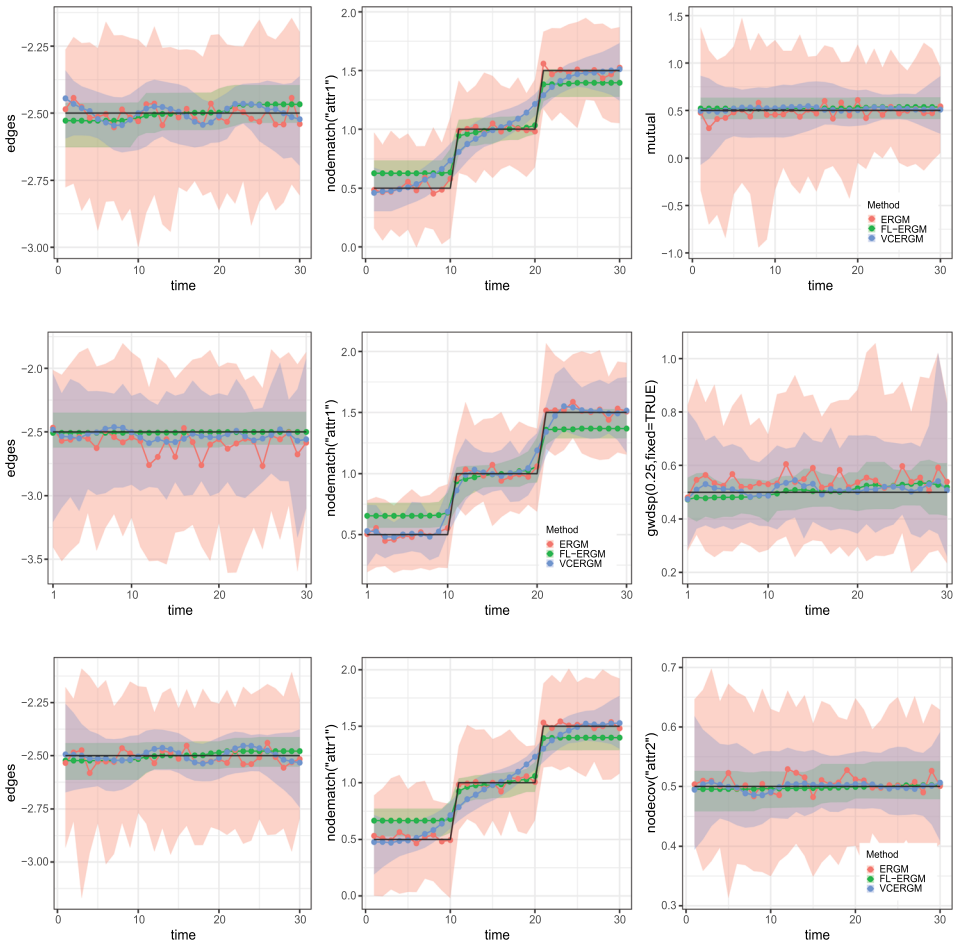


Fig. 3 Estimates for scenario 3 with piecewise constant underlying parameters. Black lines indicate the true parameters. Green lines show FL-ERGM estimates, blue lines stand for VCERGM estimates and red lines represent ERGM estimates. for each method, solid lines indicate the average of 50 estimated curves and the shaded bands illustrate the 0.025 and 0.975 quantiles

the lowest IAE in the constant, spike-like, piecewise-constant and sudden-change settings, whereas VCERGM performs better in the smoothly varying setting. Figures 1, 2, 3, 4 and 5 provide a visual comparison of the estimated parameter trajectories and help explain these differences in estimation accuracy.

When $T = 10$, the underlying parameters remain constant over time. FL-ERGM achieves the lowest IAE in Table 1. Estimates from FL-ERGM also stay constant, as demonstrated by the green, flat-shaded bands representing the 0.025 and 0.975 quantiles of the FL-ERGM estimates in Fig. 1. This behaviour is expected because the fused lasso penalty encourages consecutive parameter estimates to be identical when there is no evidence of temporal change. In contrast, the shaded bands for VCERGM and ERGM are not flat, indicating greater variation in their estimates across time.

When $T = 20$, the underlying parameters fluctuate over time, exhibiting a spike-like pattern, as shown in Fig. 2. FL-ERGM has the lowest IAE in Table 1 and successfully captures the spike-like pattern, whereas VCERGM produces a smoother trajectory with substantially less variation. This reflects the ability of the fused lasso penalty to preserve local changes when supported by the data, while the smoothness imposed by VCERGM tends to attenuate short-lived changes.

When $T = 30$, the underlying parameters are piecewise constant. Again, FL-ERGM achieves the lowest IAE and closely captures both the piecewise constant pattern and the turning points in Fig. 3, although some deviations from the true parameters are observed at certain time points. This performance reflects the sparsity induced in the temporal parameter differences, which allows changes to be concentrated around a small number of time points while maintaining stable estimates within each regime.

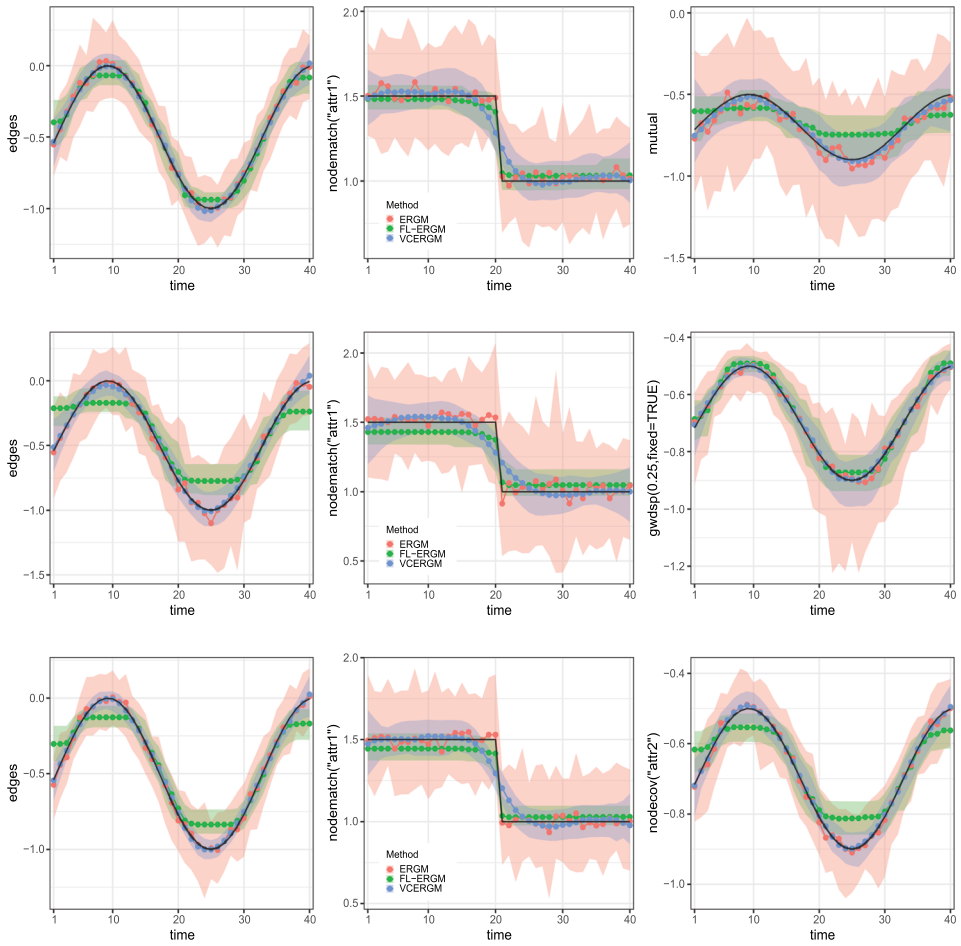


Fig. 4 Estimates for scenario 4 with smoothly varying parameters. Black lines indicate the true parameters. Green lines show FL-ERGM estimates, blue lines stand for VCERGM estimates and red lines represent ERGM estimates. For each method, solid lines indicate the average of 50 estimated curves and the shaded bands illustrate the 0.025 and 0.975 quantiles

When $T = 40$, the underlying parameters vary smoothly with time. In this setting, VCERGM achieves a lower IAE than FL-ERGM, consistent with the trajectories shown in Fig. 4. The smoother parameter estimates produced by VCERGM are better aligned with the gradual evolution of the underlying parameters. Although FL-ERGM does not reproduce the smooth turning points as accurately, it still captures the overall increasing and decreasing trends. This illustrates that the fused lasso approach is more suited to sparse temporal changes than to continuously varying parameter trajectories.

When $T = 50$, the underlying parameters exhibit sudden changes, as shown in Fig. 5. FL-ERGM again achieves the lowest IAE and successfully detects the changes in the underlying parameters, whereas VCERGM smooths over the abrupt shifts. This highlights the advantage of the fused lasso penalty in settings where parameter changes are concentrated around a small number of time points. Overall, the results demonstrate that FL-ERGM is particularly effective when the underlying parameter trajectories contain stable periods interspersed with local or abrupt changes, while VCERGM is better suited to smoothly evolving parameters.

The accuracy of change point detection are presented in Table 2 for FL-ERGM, VCERGM and ERGM separately. Results suggest that the proposed FL-ERGM has higher accuracy and smaller localisation error for detecting abrupt changes compared to VCERGM and separate ERGM fitting. Full results can be found in Appendix E. To compare the computational performance with and without LU decomposition, we record the total run time of each repetition for Model 3 at Table 3. Results suggest that the LU decomposition provides increasing efficiency gains as the time horizon grows.

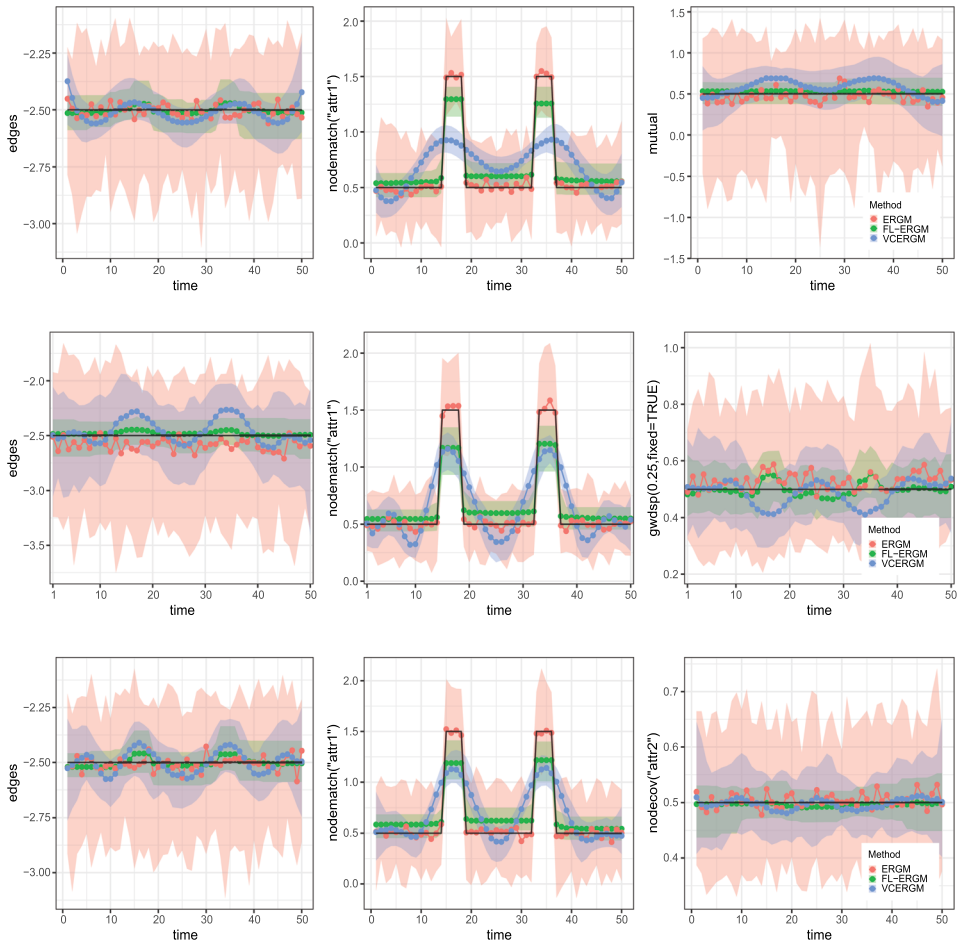


Fig. 5 Estimates for scenario 5 with sudden changes in parameters. Black lines indicate the true parameters. Green lines show FL-ERGM estimates, blue lines stand for VCERGM estimates and red lines represent ERGM estimates. for each method, solid lines indicate the average of 50 estimated curves and the shaded bands illustrate the 0.025 and 0.975 quantiles

Table 1 Simulation results with 50 replicates: mean (standard deviation) of IAE for each method

	T	FL-ERGM	VCERGM	ERGM
Model 1: $S^1(y_i) + S^2(y_i) + S^4(y_i)$	10	**2.14**(1.19)	3.18(1.16)	7.22(1.40)
	20	**7.36**(0.89)	20.60(1.40)	8.71(1.11)
	30	**5.39**(1.82)	7.62(1.46)	18.30(1.90)
	40	7.44(1.20)	**6.73**(1.01)	17.11(1.41)
	50	**10.53**(2.78)	23.00(2.07)	34.46(2.84)
Model 2: $S^1(y_i) + S^3(y_i) + S^4(y_i)$	10	**1.99**(1.20)	4.01(1.35)	6.03(1.06)
	20	**6.54**(1.10)	16.23(0.70)	9.65(3.39)
	30	**5.92**(2.00)	10.61(2.53)	18.24(1.97)
	40	9.27(1.91)	**6.10**(0.99)	15.59(1.40)
	50	**11.09**(2.40)	22.26(2.76)	29.25(2.56)
Model 3: $S^1(y_i) + S^4(y_i) + S^5(y_i)$	10	**1.20**(0.74)	1.73(0.69)	3.90(0.69)
	20	**4.95**(0.73)	14.85(0.51)	5.66(0.64)
	30	**4.71**(1.08)	5.58(0.79)	11.56(1.33)
	40	7.22(1.44)	**4.67**(0.81)	10.75(1.08)
	50	**9.15**(2.04)	15.22(1.05)	20.43(1.64)

Table 2 Simulation results for change point detection for $k = 2$ over 50 replicates for model 3: mean (standard deviation)

Method	Metric	T = 10	T = 20	T = 30	T = 40	T = 50
FL-ERGM	Sensitivity	**1.00 (0.00)**	0.92 (0.27)	0.81 (0.30)	0.72 (0.45)	0.97 (0.10)
	Specificity	**1.00 (0.00)**	**0.97 (0.04)**	**1.00 (0.01)**	**0.99 (0.01)**	**1.00 (0.00)**
	Localisation error	**0.00 (0.00)**	0.07 (0.33)	**0.04 (0.18)**	**0.03 (0.17)**	**0.00 (0.00)**
VCERGM	Sensitivity	**1.00 (0.00)**	0.52 (0.50)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Specificity	**1.00 (0.00)**	0.96 (0.04)	0.96 (0.00)	0.97 (0.00)	0.98 (0.00)
	Localisation error	**0.00 (0.00)**	**0.00 (0.00)**	—	—	—
ERGM	Sensitivity	0.02 (0.14)	**0.94 (0.24)**	**1.00 (0.00)**	**0.84 (0.37)**	**0.99 (0.05)**
	Specificity	0.65 (0.18)	0.86 (0.08)	0.69 (0.07)	0.84 (0.06)	0.85 (0.06)
	Localisation error	—	0.21 (0.51)	0.27 (0.41)	0.26 (0.54)	0.10 (0.17)

Table 3 Computational performance comparison: mean (standard deviation, in seconds) of runtime with and without LU decomposition and iteration count

T	without LU decomposition	with LU decomposition	iteration numbers
10	4.41 (0.24)	4.06 (0.25)	5
20	8.39 (0.80)	5.40 (0.24)	4
30	17.78 (3.34)	14.43 (0.98)	5
40	20.75 (3.63)	18.22 (3.75)	4
50	36.47 (3.02)	25.37 (3.27)	5

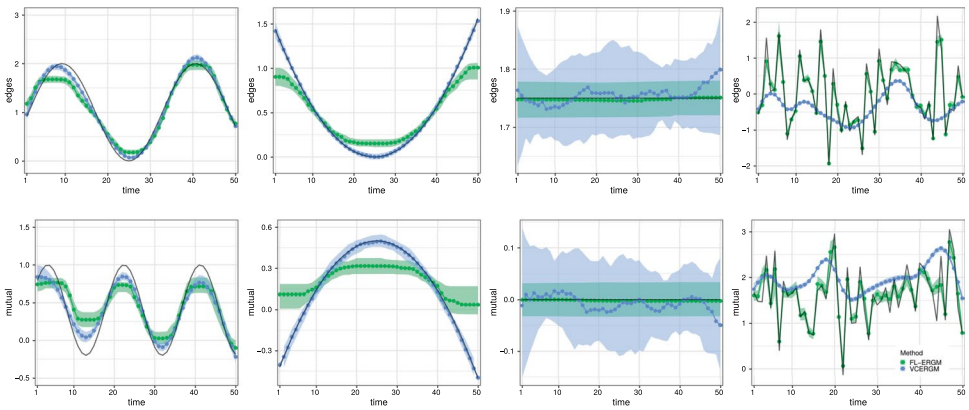


Fig. 6 Parameter estimates for simulated networks with same settings as Lee et al. (2020a). The four columns represent settings with (1) sinusoidal curves; (2) quadratic curves; (3) Erd"os-R'enyi graphs with constant parameters; (4) non-smooth curves respectively. Black lines indicate the true parameters. Green lines represent FL-ERGM estimates, and blue lines stand for VCERGM estimates. For each method, solid lines indicate the average of 100 estimated curves and the shaded bands illustrate the first and third quantiles

To further investigate the comparison of FL-ERGM and VCERGM, we also conduct simulation studies using the same settings as in Lee et al. (2020). In Lee et al. (2020), two network statistics, the number of edges, $S^1(y_i)$, and the number of mutual edges, $S^2(y_i)$, were included at four different scenarios. The four settings are (1) sinusoidal curves; (2) quadratic curves; (3) Erdős–Rényi graphs with constant parameters; (4) non-smooth curves with parameters sampled randomly from normal distributions $N(0, 1)$ and $N(1.5, 2.5)$ respectively. The true parameter values are presented in Fig. 6 along with the estimated coefficients from FL-ERGM and VCERGM. Mean and standard deviation of IAE across 50 replicates are presented in Table 4.

Results suggest that the proposed FL-ERGM performs the best in the constant Erdős-Rényi setting (3) and the non-smooth setting (4), as indicated by the smallest IAE values. This is consistent with the L_1 fused lasso penalty, which encourages temporal parameter differences to be zero while allowing abrupt changes when supported by the data. In contrast, the smoothness assumption of VCERGM is well suited to the sinusoidal and quadratic settings (1) and (2), where VCERGM provides more accurate estimates than FL-ERGM. Although FL-ERGM

Table 4 Simulation results with 50 replicates: mean (standard deviation) of IAE for each method under same settings as in Lee et al. (2020)

		FL-ERGM	VCERGM	ERGM
Edges	Sinusoidal	8.31(1.47)	**4.78**(1.01)	18.07(9.76
	Quadratic	7.18(2.12)	**2.31**(0.57)	6.34(0.71)
	Erdős-Rényi	**1.89**(1.38)	4.20(1.27)	14.58(2.51)
	Non-smooth	**9.54**(1.13)	33.63(0.47)	18.69(10.64)
Reciprocity	Sinusoidal	10.28(1.96)	**6.67**(1.15)	20.18(9.87)
	Quadratic	8.53(2.61)	**3.04**(0.82)	8.63(0.96)
	Erdős-Rényi	**2.07**(1.66)	4.73(1.40)	16.40(2.58)
	Non-smooth	**12.68**(1.42)	28.63(1.06)	22.24(10.75)

does not estimate the smooth parameter trajectories as accurately, it can still capture their overall trends. In the non-smooth settings, however, the smoothness imposed by VCERGM can smooth over abrupt changes and consequently provide misleading change patterns. These results indicate that FL-ERGM is not expected to outperform VCERGM in all settings; rather, its main advantage lies in scenarios with sparse temporal changes, locally invariant structures, or abrupt parameter shifts.

International trade networks

The international trade networks describe the trade relationships among 60 countries from 2001 to 2016. We apply the proposed FL-ERGM to international trade networks with the aim of analysing the dynamic evolution of trade relationships. The trade networks are built based on the annual import data between every two countries from the United Nations Commodity Trade Statistics Database[1]. A directed edge exists from node r to s, $y_{rs,i} = 1$, if the import amount from country r to s is more than 3 billion dollars at year . For each trade network, we choose the following statistics:

- $S^1(y_i) = \sum_{r \neq s} y_{rs,i}$, the total number of edges in trade network y_i, reflecting on the density of trade relationship.
- $S^2(y_i) = e^{\phi} \sum_{k=1}^{n-2} \{1 - (1 - e^{-\phi})^k\} DP_k(y_i)$, $\phi = 0.25$, geometrically weighted dyad-wise shared partners (GWDSP), a representation for local clustering phenomenon, measuring the extent to which the trade relationship localises. $DP_k(y_i)$ is the number of dyads that have k common neighbors.
- $S^3(y_i) = \sum_{r \neq s} y_{rs,i} X_{rs}$, measuring the effect of distance for the trade relationship. X_{rs} stands for the geographic distance between countries r and s, downloaded from CEPII database[2].

Next, we fit the FL-ERGM to the dynamic trade networks. The PBIC values of the FL-ERGM are shown in Fig. 7 for selecting the tuning parameter. The figure illustrates the PBIC values across candidate tuning parameters, with the value minimising the PBIC selected for the analysis. Parameter estimates of the FL-ERGM are presented in Fig. 8. In addition to the point estimates, we also presented the confidence intervals for each coefficient, which is estimated using the parametric bootstrap method (Schmid and Desmarais 2017). The parameter estimates in Fig. 8 reveal that the trade evolution has two phases, the increasing period before 2008 and the steady phase after 2008. This matches the fact of global financial crisis in 2008. Before 2008, the coefficient of edges keeps increasing, meaning that the trade relationship was expanding continuously. At the same time, the coefficient of gwdsp is decreasing, which implies that the local clustering phenomenon is slowly disappearing, and the trade relationships are becoming more universal. The distance has a negative and steady effect on trading during this period. After 2008, the coefficients of edges and distance effects are relatively stable over time. We also notice that the gwdsp coefficient has a slight increase from 2012, but we could still claim that the second phase is steady, as the value of gwdsp coefficient is relatively small. Moreover, comparing two phases, the second one has bigger edge coefficients but smaller distance effects, suggesting that the trade is more frequent and the reason might be the increased convenience of global transportation.

[1] https://comtrade.un.org/

[2] http://www.cepii.fr.

Fig. 7 PBIC of the FL-ERGM for dynamic trade networks

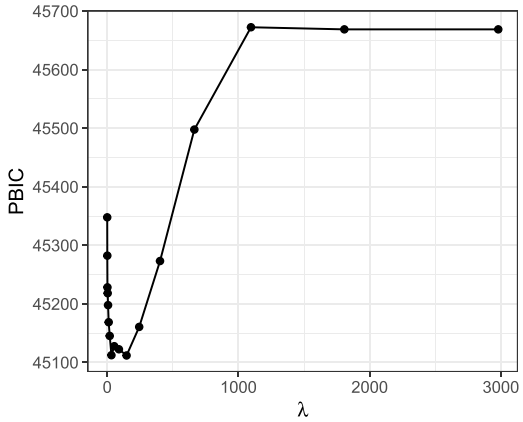


Fig. 8 Parameter estimates of the FL-ERGM for dynamic trade networks

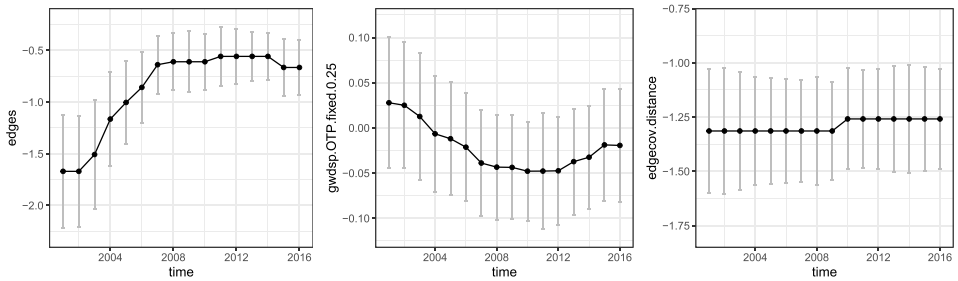


Fig. 9 Parameter estimates of different methods for dynamic trade networks. Red lines represent the results from the ERGM, blue lines stand for the VCERGM, and green lines show results from the proposed FL-ERGM

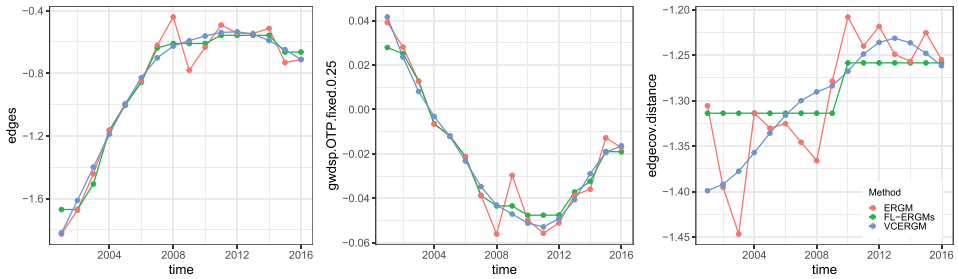


Table 5 Results of abrupt change point detection

	FL-ERGM	VCERGM	ERGM
Edges	2003, 2004, 2011, 2015	NULL	2009, 2010, 2012, 2015
GWDSP	NULL	NULL	NULL
Distance	2010	NULL	2004, 2009

We also apply VCERGM and ERGM to the dynamic trade networks. A comparison of the model results is shown in Fig. 9 and Table 5. The three models provide similar parameter estimates for the edges and gwdsp coefficients, apart from the large spikes observed for ERGM. VCERGM does not identify any abrupt changes according to our definition, whereas FL-ERGM and ERGM identify similar change points. For the distance effect, VCERGM suggests a continuously increasing trend, whereas FL-ERGM suggests two relatively stable phases. In the changepoint analysis, FL-ERGM also identifies additional changepoints around 2003 for the edges parameter, partly because the fused lasso regularisation shrinks the difference at 2002 towards zero.

The differences in the results are driven by the different model assumptions underlying the two approaches. In particular, VCERGMs assume that network parameters evolve smoothly over time, whereas FL-ERGMs encourage piecewise constant parameter trajectories through fused lasso regularisation. As illustrated in the simulation study, the suitability of these assumptions depends on the underlying temporal pattern of the parameters. In simulation scenario 3, where the underlying parameter structure is piecewise constant, FL-ERGMs capture this structure well, whereas VCERGM suggests a continuously increasing trend that differs from the known underlying pattern. For the real-data analysis, however, there is no ground truth and therefore no evidence to determine which method provides the correct temporal pattern. In the case of the 2008 global financial crisis, it is reasonable to consider the effect of geographic distance underwent an abrupt change at that time.

We also perform model assessments for FL-ERGMs and VCERGMs using simulated networks. Specifically, at each time point, we simulate 100 networks from the ERGM distribution with parameter estimate $\hat{\theta}_t$. We present the summary statistics of the simulated networks using box plots in Fig. 10. As is shown, the observed network statistics are close to the simulated network statistics, indicating both the FL-ERGM and VCERGM fit the dynamic trade networks well.

Discussion

In this paper, we propose to model a sequence of network samples observed at multiple time points using a fused lasso ERGM. This framework allows us to capture the contribution of each covariate or structural pattern to the creation of networks and network evolution through a piecewise constant function, enabling the identification of diverse and complex changing patterns. The model retains flexibility of ERGMs, meaning that the effects of network covariates and various connectivity patterns on the construction of the whole network are accounted for. Moreover, the use of ADMM algorithm with the theoretical guarantee ensures robust computational efficiency, making it suitable for estimating large-scale networks over extended observation periods. The comprehensive simulation studies and the applications to the real-world dataset demonstrate that the FL-ERGM is capable of providing interpretable changing patterns that are not discovered by the existing methods.

A well-known challenge in ERGMs is model degeneracy, whereby the model assigns a substantial proportion of its probability mass to a small number of unrealistic network configurations. In our framework, the fused lasso regularisation is introduced to control parameter complexity by shrinking differences between successive parameter vectors and limiting excessive temporal variation in the estimated coefficients. Although this regularisation does not eliminate or guarantee against model degeneracy, it can improve estimation stability and reduce the occurrence of extreme parameter estimates. We therefore recommend assessing the fitted model through simulation-based diagnostics, including generating networks from the estimated parameters and comparing their structural characteristics with those of the observed networks. While the current formulation assumes regularly spaced networks, the framework can be extended to irregularly spaced observations by incorporating temporal gaps into the weighting of the penalty.

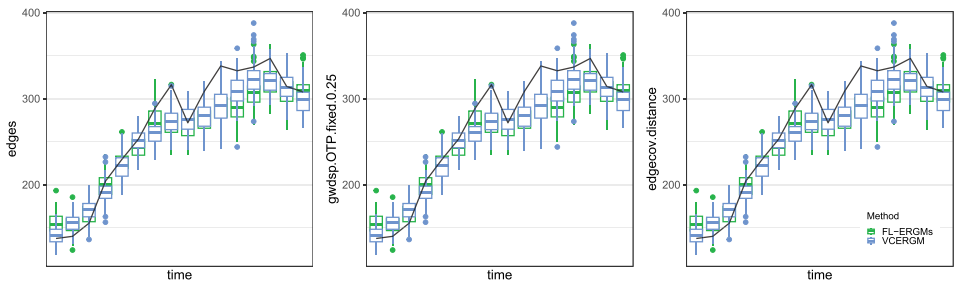


Fig. 10 Model assessments on dynamic trade networks. The solid line indicates the observed network statistics and the box plots represent the statistics of networks simulated from the estimated model

Several other potential directions for future work also remain. First, we have developed an efficient algorithm for the pseudo-likelihood-based estimator. However, developing an efficient algorithm for estimating FL-ERGMs based on the exact likelihood remains an important area for future research, whether within a frequentist or Bayesian framework. Further discussion on likelihood-based inference can be found in section "Pseudo-likelihood approximation". Obtaining a solution to the exact likelihood function in a dynamic setting is both a challenging and intriguing task. While progress has been made in approximating intractable likelihoods for a single network (Hunter and Handcock 2006; Caimo and Friel 2011; Bouranis et al. 2018; Tan and Friel 2020) or for a sequence of independent networks (Ren et al. 2023), solutions for dynamic network settings remain largely unknown. Second, we have included the fused lasso penalty in this paper for the purpose of trend estimation but there are also other types of penalty terms existing (Hallac et al. 2017). It would be interesting to extend the current framework to a more general setting to utilise the strength of different penalty terms. Given the of versatility of ADMM algorithm, the estimation of ERGMs with different penalty terms should be straightforward. Third, we have proved the convergence of the ADMM algorithm, but the theoretical property of the penalised pseudo-likelihood estimators remains unknown.

Supplementary materials

A Taylor expansion

The function to expand is

$$f(\theta_t) = y_{ij,t}\theta_t^\top \Delta S_{ij,t} - \log\left\{1 + \exp\left(\theta_t^\top \Delta S_{ij,t}\right)\right\},$$

with first and second derivatives as follows,

$$\frac{\partial f(\theta_t)}{\partial \theta_t} = y_{ij,t}\Delta S_{ij,t} - \frac{\exp(\theta_t^\top \Delta S_{ij,t})}{1 + \exp(\theta_t^\top \Delta S_{ij,t})}\Delta S_{ij,t},$$
$$\frac{\partial^2 f(\theta_t)}{\partial \theta_t^2} = -\frac{\exp(\theta_t^\top \Delta S_{ij,t})}{(1 + \exp(\theta_t^\top \Delta S_{ij,t}))^2}\Delta S_{ij,t}\Delta S_{ij,t}^\top.$$

The second Taylor expansion at $\tilde{\theta}_t$ is

$$
\begin{aligned}
& f(\tilde{\theta}_t) + (\theta_t - \tilde{\theta}_t)^\top f'(\tilde{\theta}_t) + \frac{1}{2}(\theta_t - \tilde{\theta}_t)^\top f''(\tilde{\theta}_t)(\theta_t - \tilde{\theta}_t) \\
=& f(\tilde{\theta}_t) + (\theta_t - \tilde{\theta}_t)^\top \Delta S_{ij,t}(y_{ij,t} - \frac{\exp(\tilde{\theta}_t{}^\top \Delta S_{ij,t})}{1 + \exp(\tilde{\theta}_t{}^\top \Delta S_{ij,t})}) \\
& - \frac{\exp(\tilde{\theta}_t{}^\top \Delta S_{ij,t})}{2(1 + \exp(\tilde{\theta}_t{}^\top \Delta S_{ij,t}))^2}(\theta_t - \tilde{\theta}_t)^\top \Delta S_{ij,t} \Delta S_{ij,t}^\top (\theta_t - \tilde{\theta}_t) \\
=& -\frac{1}{2} w_{ij,t}[(\theta_t - \tilde{\theta}_t)^\top \Delta S_{ij,t}]^2 + (\theta_t - \tilde{\theta}_t)^\top \Delta S_{ij,t}(y_{ij,t} - \tilde{p}_{ij,t}) + f(\tilde{\theta}_t) \\
=& -\frac{1}{2} w_{ij,t} \left\{ [(\theta_t - \tilde{\theta}_t)^\top \Delta S_{ij,t}]^2 - \frac{2(\theta_t - \tilde{\theta}_t)^\top \Delta S_{ij,t}(y_{ij,t} - \tilde{p}_{ij,t})}{w_{ij,t}} \right\} + f(\tilde{\theta}_t) \\
=& -\frac{1}{2} w_{ij,t} \left\{ [(\theta_t - \tilde{\theta}_t)^\top \Delta S_{ij,t}]^2 - \frac{2(\theta_t - \tilde{\theta}_t)^\top \Delta S_{ij,t}(y_{ij,t} - \tilde{p}_{ij,t})}{w_{ij,t}} + (\frac{y_{ij,t} - \tilde{p}_{ij,t}}{w_{ij,t}})^2 \right\} \\
& - \frac{(y_{ij,t} - \tilde{p}_{ij,t})^2}{2w_{ij,t}} + f(\tilde{\theta}_t) \\
=& -\frac{1}{2} w_{ij,t} \left\{ (\theta_t - \tilde{\theta}_t)^\top \Delta S_{ij,t} - \frac{y_{ij,t} - \tilde{p}_{ij,t}}{w_{ij,t}} \right\}^2 - \frac{(y_{ij,t} - \tilde{p}_{ij,t})^2}{2w_{ij,t}} + f(\tilde{\theta}_t) \\
=& -\frac{1}{2} w_{ij,t} \left\{ \theta_t^\top \Delta S_{ij,t} - \tilde{\theta}_t{}^\top \Delta S_{ij,t} - \frac{y_{ij,t} - \tilde{p}_{ij,t}}{w_{ij,t}} \right\}^2 - \frac{(y_{ij,t} - \tilde{p}_{ij,t})^2}{2w_{ij,t}} + f(\tilde{\theta}_t) \\
=& -\frac{1}{2} w_{ij,t} \left\{ \theta_t^\top \Delta S_{ij,t} - z_{ij,t} \right\}^2 + C(\tilde{\theta}_t),
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{p}_{ij,t} &= \frac{\exp(\tilde{\theta}_t{}^\top \Delta S_{ij,t})}{1 + \exp(\tilde{\theta}_t{}^\top \Delta S_{ij,t})}, \\
w_{ij,t} &= \tilde{p}_{ij,t}(1 - \tilde{p}_{ij,t}) = \frac{\exp(\tilde{\theta}_t{}^\top \Delta S_{ij,t})}{(1 + \exp(\tilde{\theta}_t{}^\top \Delta S_{ij,t}))^2}, \\
z_{ij,t} &= \tilde{\theta}_t{}^\top \Delta S_{ij,t} + \frac{y_{ij,t} - \tilde{p}_{ij,t}}{w_{ij,t}}, \\
C(\tilde{\theta}_t) &= -\frac{(y_{ij,t} - \tilde{p}_{ij,t})^2}{2w_{ij,t}} + f(\tilde{\theta}_t).
\end{aligned}
$$

B LU decomposition

We note that F is a banded matrix with $d \times (T-1)$ rows and $d \times T$ columns,

$$
F = \begin{bmatrix}
-I_d & I_d & & & \\
 & -I_d & I_d & & \\
 & & \ddots & \ddots & \\
 & & & -I_d & I_d
\end{bmatrix},
$$

where I_d denotes the diagonal matrix of order d. Then $F^\top F$ is a block tridiagonal matrix,

$$
F^\top F = \begin{bmatrix}
I_d & -I_d & & & \\
-I_d & 2I_d & -I_d & & \\
 & \ddots & \ddots & & \\
 & -I_d & 2I_d & -I_d & \\
 & & & -I_d & I_d
\end{bmatrix}.
$$

$X = A^\top A + \rho F^\top F$ is also a block tridiagonal matrix, as X can be written as

$$X = \begin{bmatrix} A_1^\top(A_1) + \rho I_d & -\rho I_d & & & \\ -\rho I_d & A_2^\top(A_2) + 2\rho I_d & -\rho I_d & & \\ & \ddots & \ddots & & \\ & -\rho I_d & A_{T-1}^\top(A_{T-1}) + 2\rho I_d & -\rho I_d & \\ & & -\rho I_d & A_T^\top(A_T) + \rho I_d \end{bmatrix}.$$

Next, we use the LU factorization to decompose the block tridiagonal matrix X as follows,

$$X = (\Delta + L)\Delta^{-1}(\Delta + L^\top).$$

L is the lower part of X,

$$L = \begin{bmatrix} 0 & & & \\ -\rho I_d & 0 & & \\ & \ddots & \ddots & \\ & -\rho I_d & 0 & \\ & & -\rho I_d & 0 \end{bmatrix}.$$

Δ is a block diagonal matrix with Δ_i as diagonal blocks,

$$\begin{aligned} &\Delta_1 = X_{11} = A_1^\top(A_1) + \rho I_d, \\ &\Delta_i = X_{ii} - \rho^2(\Delta_{i-1})^{-1} = A_i^\top(A_i) + 2\rho I_d - \rho^2(\Delta_{i-1})^{-1} \ (i = 2, \dots, T-1), \\ &\Delta_T = A_T^\top(A_T) + \rho I_d - \rho^2(\Delta_{T-1})^{-1}. \end{aligned}$$

As $\Delta = \Delta^\top$, then $\Delta + L = (\Delta + L^\top)^\top$, and

$$X^{-1} = (\Delta + L^\top)^{-1}\Delta(\Delta + L)^{-1} = (\Delta + L^\top)^{-1}\Delta((\Delta + L^\top)^{-1})^\top.$$

Let $D = (\Delta + L^\top)^{-1}$, and we then have $X^{-1} = D\Delta D^\top$. If we solve D, we solve X^{-1}. Since $\Delta + L^\top$ is a upper triangular matrix, D can be expressed in the following,

$$\begin{aligned} D = (\Delta + L^\top)^{-1} &= \begin{bmatrix} \Delta_1 & -\rho I_d & & & \\ & \Delta_2 - \rho I_d & & & \\ & & \ddots & & \\ & & & \Delta_{T-1} & -\rho I_d \\ & & & & \Delta_T \end{bmatrix}^{-1} \\ &= \begin{bmatrix} D_{11} = \Delta_1^{-1} & D_{12} = \rho\Delta_1^{-1}D_{22} & D_{13} = \rho\Delta_1^{-1}D_{23} & \cdots & D_{1T} = \rho\Delta_1^{-1}D_{2T} \\ & D_{22} = \Delta_2^{-1} & D_{23} = \rho\Delta_2^{-1}D_{33} & \cdots D_{2T} = \rho\Delta_2^{-1}D_{3T} & \\ & D_{33} = \Delta_3^{-1} & \cdots & D_{3T} = \rho\Delta_3^{-1}D_{4T} & \\ & & \ddots & \vdots & \\ & & D_{TT} = \Delta_T^{-1} & & \end{bmatrix}. \end{aligned}$$

Then we have

$$D\Delta = \left(\Delta + L^T\right)^{-1}\Delta = \begin{bmatrix} {\Delta_1}^{-1}\Delta_1 & \rho{\Delta_1}^{-1}D_{22}\Delta_2 & \rho{\Delta_1}^{-1}D_{23}\Delta_3 & \cdots & \rho{\Delta_1}^{-1}D_{2T}\Delta_T \\ & {\Delta_2}^{-1}\Delta_2 & \rho{\Delta_2}^{-1}D_{33}\Delta_3 & \cdots & \rho{\Delta_2}^{-1}D_{3T}\Delta_T \\ & & {\Delta_3}^{-1}\Delta_3 & \cdots & \rho{\Delta_3}^{-1}D_{4T}\Delta_T \\ & & & \ddots & \vdots \\ & & & & {\Delta_T}^{-1}\Delta_T \end{bmatrix},$$

$$X^{-1} = D\Delta D^T = \begin{bmatrix} I_d & D_{12}\Delta_2 & D_{13}\Delta_3 & \cdots & D_{1T}\Delta_T \\ & I_d & D_{23}\Delta_3 & \cdots & D_{2T}\Delta_T \\ & & I_d & \cdots & D_{3T}\Delta_T \\ & & & \ddots & \vdots \\ & & & & I_d \end{bmatrix} \cdot \begin{bmatrix} D_{11} & & & & \\ D_{12} & D_{22} & & & \\ D_{13} & D_{23} & D_{33} & & \\ \vdots & & \ddots & & \\ D_{1T} & D_{2T} & D_{3T} & \cdots & D_{TT} \end{bmatrix}.$$

Therefore, the inverse of $X = A^\top A + \rho F^\top F$, a square matrix with dimension $dT \times dT$, can be solved by calculating the inverse of T small matrices Δ_i of dimension d.

C Proof of lemma 1 and 2

To prove Lemma 1, we only need to show that the Hessian matrix for $l(\theta_1, \dots, \theta_T)$ and $\tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1, \dots, \theta_T)$ are positive definite.

For any $t \in \{1, \dots, T\}$, we have

$$\frac{\partial^2 l(\theta_1, \dots, \theta_T)}{\partial \theta_t^2} = \sum_{i \neq j} \frac{\Delta S_{ij,t} \Delta S_{ij,t}^\top \exp(\theta_t^\top \Delta S_{ij,t})}{(1 + \exp(\theta_t^\top \Delta S_{ij,t}))^2}.$$
$$\frac{\partial^2 \tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1, \dots, \theta_T)}{\partial \theta_t^2} = \sum_{i \neq j} w_{ij,t} \Delta S_{ij,t} \Delta S_{ij,t}^\top.$$

In addition, for $t \neq s, t \in \{1, \dots, T\}$, we have

$$\frac{\partial^2 l(\theta_1, \dots, \theta_T)}{\partial \theta_t \partial \theta_s} = 0,$$
$$\frac{\partial^2 \tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1, \dots, \theta_T)}{\partial \theta_t \partial \theta_s} = 0.$$

Thus the Hessian matrix for $l(\theta_1, \dots, \theta_T)$ is

$$\begin{bmatrix} \frac{\partial^2 l}{\partial \theta_1^2} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_T} \\ \frac{\partial^2 l}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 l}{\partial \theta_2^2} & \cdots & \frac{\partial^2 l}{\partial \theta_2 \partial \theta_T} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \theta_T \partial \theta_1} & \frac{\partial^2 l}{\partial \theta_T \partial \theta_2} & \cdots & \frac{\partial^2 l}{\partial \theta_T^2} \end{bmatrix}$$
$$= \sum_{i \neq j} \begin{bmatrix} \frac{\Delta S_{ij1} \Delta S_{ij1}^\top \exp(\theta_1^\top \Delta S_{ij1})}{(1+\exp(\theta_1^\top \Delta S_{ij1}))^2} & 0 & \cdots & 0 \\ 0 & \frac{\Delta S_{ij2} \Delta S_{ij2}^\top \exp(\theta_2^\top \Delta S_{ij2})}{(1+\exp(\theta_2^\top \Delta S_{ij2}))^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\Delta S_{ij,t} \Delta S_{ij,t}^\top \exp(\theta_T^\top \Delta S_{ij,t})}{(1+\exp(\theta_T^\top \Delta S_{ij,t}))^2} \end{bmatrix},$$

and the Hessian matrix of $\tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1, \dots, \theta_T)$ is

$$\begin{bmatrix} \frac{\partial^2 \tilde{l}}{\partial \theta_1^2} & \frac{\partial^2 \tilde{l}}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \tilde{l}}{\partial \theta_1 \partial \theta_T} \\ \frac{\partial^2 \tilde{l}}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \tilde{l}}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \tilde{l}}{\partial \theta_2 \partial \theta_T} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \tilde{l}}{\partial \theta_T \partial \theta_1} & \frac{\partial^2 \tilde{l}}{\partial \theta_T \partial \theta_2} & \cdots & \frac{\partial^2 \tilde{l}}{\partial \theta_T^2} \end{bmatrix}$$
$$= \sum_{i \neq j} \begin{bmatrix} w_{ij1} \Delta S_{ij1} \Delta S_{ij1}^\top & 0 & \cdots & 0 \\ 0 & w_{ij2} \Delta S_{ij2} \Delta S_{ij2}^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{ij,t} \Delta S_{ij,t} \Delta S_{ij,t}^\top \end{bmatrix}.$$

It is clear to see that the Hessian matrices for $l(\theta_1, \dots, \theta_T)$ and $\tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1, \dots, \theta_T)$ are positive definite, so $l(\theta_1, \dots, \theta_T)$ and $\tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1, \dots, \theta_T)$ are convex.

To establish the limit in Lemma 2, we analyse the structural relationship between the original loss function $l(\cdot)$ and the surrogate function $\tilde{l}(\cdot)$ by focusing on the individual pair components. Let $x_{ij,t} = \theta_t^\top \Delta S_{ij,t}$ and $\tilde{x}_{ij,t} = \tilde{\theta}_t^\top \Delta S_{ij,t}$. The component scalar loss function for a given pair (i, j) at time index t in the original objective function is defined via the logistic relationship:

$$f(x_{ij,t}) = -y_{ij,t} x_{ij,t} + \log\left(1 + \exp(x_{ij,t})\right)$$

Thus, the aggregate original loss function across all time horizons is written as:

$$l(\theta_1, \ldots, \theta_T) = \sum_{t=1}^{T} \sum_{i \neq j} f(\theta_t^\top \Delta S_{ij,t})$$

The surrogate function $\tilde{l}(\cdot)$ is designed as a second-order localised Taylor expansion (quadratic approximation) of $l(\cdot)$ constructed around the parameter anchor point $\tilde{\theta}_t$. By expanding $f(x_{ij,t})$ around the localised evaluation point $\tilde{x}_{ij,t} = \tilde{\theta}_t^\top \Delta S_{ij,t}$, Taylor's theorem states:

$$f(x_{ij,t}) = f(\tilde{x}_{ij,t}) + f'(\tilde{x}_{ij,t})(x_{ij,t} - \tilde{x}_{ij,t}) + \frac{1}{2} f''(\tilde{x}_{ij,t})(x_{ij,t} - \tilde{x}_{ij,t})^2 + R_2(x_{ij,t}; \tilde{x}_{ij,t})$$

where $R_2(x_{ij,t}; \tilde{x}_{ij,t})$ denotes the Lagrange form of the higher-order residual error. By matching the coefficients to the defined variables within the surrogate function framework, the tracking weights and localised parameters are explicitly designated as:

$$w_{ij,t} = f''(\tilde{x}_{ij,t})$$

$$z_{ij,t} = \tilde{x}_{ij,t} - \frac{f'(\tilde{x}_{ij,t})}{f''(\tilde{x}_{ij,t})}$$

Completing the square with respect to these terms permits the localised objective function to be structurally represented without the remainder term as:

$$\tilde{l}_{\tilde{\theta}_1, \ldots, \tilde{\theta}_T}(\theta_1, \ldots, \theta_T) = \frac{1}{2} \sum_{t=1}^{T} \sum_{i \neq j} w_{ij,t} (\theta_t^\top \Delta S_{ij,t} - z_{ij,t})^2 - C(\tilde{\theta}_t)$$

where $C(\tilde{\theta}_t)$ accumulates the remaining parameter-independent constants. Consequently, the discrepancy between the exact objective landscape and its tracking approximation is exactly equal to the total summation of the Taylor remainder residuals:

$$l(\theta_1, \ldots, \theta_T) - \tilde{l}_{\tilde{\theta}_1, \ldots, \tilde{\theta}_T}(\theta_1, \ldots, \theta_T) = \sum_{t=1}^{T} \sum_{i \neq j} R_2(\theta_t^\top \Delta S_{ij,t}; \tilde{\theta}_t^\top \Delta S_{ij,t})$$

Applying real analysis limit theorems, Taylor's theorem guarantees that the residual vanishes as the evaluation point matches the anchor centre:

$$\lim_{\tilde{x}_{ij,t} \to x_{ij,t}} R_2(x_{ij,t}; \tilde{x}_{ij,t}) = 0$$

Given that the projection mapping $\tilde{x}_{ij,t} = \tilde{\theta}_t^\top \Delta S_{ij,t}$ depends continuously on $\tilde{\theta}_t$, the parameter limit condition $\tilde{\theta}_t \to \theta_t$ strictly implies $\tilde{x}_{ij,t} \to x_{ij,t}$. Distributed across the finite sums, this yields:

$$\lim_{\tilde{\theta}_t \to \theta_t} \left[l(\theta_1, \dots, \theta_T) - \tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1, \dots, \theta_T) \right] = \sum_{t=1}^{T} \sum_{i \neq j} \lim_{\tilde{\theta}_t \to \theta_t} R_2(\theta_t^\top \Delta S_{ij,t}; \tilde{\theta}_t^\top \Delta S_{ij,t}) = 0$$

Hence, the localised surrogate function converges pointwise to the original objective function:

$$\lim_{\tilde{\theta}_t \to \theta_t} \tilde{l}_{\tilde{\theta}_1,\dots,\tilde{\theta}_T}(\theta_1, \dots, \theta_T) = l(\theta_1, \dots, \theta_T)$$

D Proof of proposition 2

To rigorously establish algorithmic stability, we construct a Lyapunov function V^k, which acts as a mathematical metric of the total tracking error. Structurally, V^k measures the combined distance between the current iteration's states and the true global optimum (U^*, Q^*). Proving convergence requires demonstrating that this overall energy distance steadily shrinks to zero as iterations progress.

$$V^k = \rho \|U^k - U^*\|_2^2 + \rho \| - (Q^k - Q^*)\|_2^2.$$

The first inequality is

$$V^{k+1} \leq V^k - \rho \|F\Theta^{k+1} - Q^{k+1}\|_2^2 - \rho \| - (Q^{k+1} - Q^k)\|_2^2, \tag{20}$$

which proves that the total error energy V^k is monotonically non-increasing. In addition, the amount of decrease depends on the norm of the prime residual $F\Theta^{k+1} - Q^{k+1}$ (how well the structural constraints are satisfied) and on the change of dual variables Q over one iteration. Since $V^k \leq V^0$, thus U^k and $-Q^k$ are bounded.

This guarantees that the algorithm can never oscillate wildly or diverge; instead, every single iteration is guaranteed to push the parameters closer to the optimal solution.

From inequality A(20, we have

$$\rho \sum_{k=0}^{\infty} (\|F\Theta^{k+1} - Q^{k+1}\|_2^2 + \| - (Q^{k+1} - Q^k)\|_2^2) \leq V^0, \tag{21}$$

which implies that $F\Theta^{k+1} - Q^{k+1} \to 0$ and $-(Q^{k+1} - Q^k) \to 0$ as $k \to \infty$. Practically, this guarantees that as the algorithm runs longer ($k \to \infty$), the primal residual $F\Theta^{k+1} - Q^{k+1}$ and the dual residual contract to absolute zero.

The second inequality is

$$\frac{1}{2}\|A^*\Theta^* - b^*\|_2^2 + \lambda \|Q^*\|_1 - \frac{1}{2}\|A^{k+1}\Theta^{k+1} - b^{k+1}\|_2^2 - \lambda \|Q^{k+1}\|_1 \leq \rho U^{*\top}(F\Theta^{k+1} - Q^{k+1}), \tag{22}$$

which mathematically formalises the exact net change in the total error tracking energy across a single iteration by decomposing it into the dynamic shifts of the dual variables, auxiliary variables, and core state parameters. This expansion serves as the structural foundation for the stability analysis, allowing us to explicitly track how the algorithm systematically reduces parameter discrepancy as it approaches the global equilibrium. The third inequality we are going to prove is

$$\begin{aligned} &\frac{1}{2}\|A^{k+1}\Theta^{k+1} - b^{k+1}\|_2^2 + \lambda \|Q^{k+1}\|_1 - \frac{1}{2}\|A^*\Theta^* - b^*\|_2^2 - \lambda \|Q^*\|_1 \\ &\leq -(\rho U^{k+1})^\top (F\Theta^{k+1} - Q^{k+1}) - \rho\{(Q^{k+1} - Q^k)\}^\top \{(F\Theta^{k+1} - Q^{k+1}) + (Q^{k+1} - Q^*)\}, \end{aligned} \tag{23}$$

this invokes the standard first-order optimality conditions for the true global minimisers, which serves as a fixed baseline to assess our algorithmic progress. By combining these saddle-point characterisations with the objective's convexity, we establish the necessary geometric inequality to bound our step-by-step updates against the ideal target state.

Using the results of inequality (21), we see that the righthand side of (22) also converges to 0 when $k \to \infty$, since the prime residual $F\Theta^{k+1} - Q^{k+1}$ converges to 0. The righthand side in (23) converges to 0 when $k \to \infty$, since $-(Q^{k+1} - Q^*)$ is bounded and $F\Theta^{k+1} - Q^{k+1}$ and $-(Q^{k+1} - Q^k)$ converge to 0. Thus we have $\lim_{k\to\infty}\{(1/2)\|A^{k+1}\Theta^{k+1} - b^{k+1}\|_2^2 + \lambda\|Q^{k+1}\|_1\} = (1/2)\|A^*\Theta^* - b^*\|_2^2 + \lambda\|Q^*\|_1$, which is the objective convergence. If assumption (3) holds, Θ^k converges to a solution of (8).

Next, we prove the three inequalities. We first prove inequality (22). Since (Θ^*, Q^*, U^*) is a saddle point for $(1/2)\|A\Theta - b\|_2^2 + \lambda\|Q\|_1 + \rho U^\top(F\Theta - Q)$, we have

$$\begin{aligned}&(1/2)\|A^*\Theta^* - b^*\|_2^2 + \lambda\|Q^*\|_1 + \rho U^{*\top}(F\Theta^* - Q^*)\\ &\le (1/2)\|A^{k+1}\Theta^{k+1} - b^{k+1}\|_2^2 + \lambda\|Q^{k+1}\|_1 + \rho U^{*\top}(F\Theta^{k+1} - Q^{k+1}).\end{aligned}$$

As $F\Theta^* - Q^* = 0$, the lefthand side of the above inequality equals to $(1/2)\|A^*\Theta^* - b^*\|_2^2 + \lambda\|Q^*\|_1$. Subtracting both parts with $(1/2)\|A^{k+1}\Theta^{k+1} - b^{k+1}\|_2^2 + \lambda\|Q^{k+1}\|_1$, then we have proved inequality (22).

Second, we prove inequality (23). Θ^{k+1} minimises the augmented Lagrangian $(1/2)\|A\Theta - b\|_2^2 + \lambda\|Q^k\|_1 + \rho U^{k\top}(F\Theta - Q^k) + (\rho/2)\|F\Theta - Q^k\|_2^2$. Because $(1/2)\|A\Theta - b\|_2^2$ is closed, proper, and convex, it is subdifferentiable, and so is the augmented Lagrangian $(1/2)\|A\Theta - b\|_2^2 + \lambda\|Q^k\|_1 + \rho U^{k\top}(F\Theta - Q^k) + (\rho/2)\|F\Theta - Q^k\|_2^2$. The (necessary and sufficient) optimality condition is

$$0 \in \partial\{(1/2)\|A^{k+1}\Theta^{k+1} - b^{k+1}\|_2^2\} + \rho F^\top U^k + \rho F^\top(F\Theta^{k+1} - Q^k). \tag{24}$$

Since $U^{k+1} = U^k + F\Theta^{k+1} - Q^{k+1}$, we plug in $U^k = U^{k+1} - (F\Theta^{k+1} - Q^{k+1})$ and rearrange (24):

$$0 \in \partial\{(1/2)\|A^{k+1}\Theta^{k+1} - b^{k+1}\|_2^2\} + F^\top\{\rho U^{k+1} + \rho(Q^{k+1} - Q^k)\}. \tag{25}$$

(25) shows that Θ^{k+1} minimises

$$\{(1/2)\|A\Theta - b\|_2^2\} + \{\rho U^{k+1} + \rho(Q^{k+1} - Q^k)\}^\top F\Theta.$$

Thus we have

$$\begin{aligned}&(1/2)\|A^{k+1}\Theta^{k+1} - b^{k+1}\|_2^2 + \{\rho U^{k+1} + \rho(Q^{k+1} - Q^k)\}^\top F\Theta^{k+1}\\ &\le (1/2)\|A^*\Theta^* - b^*\|_2^2 + \{\rho U^{k+1} + \rho(Q^{k+1} - Q^k)\}^\top F\Theta^*.\end{aligned} \tag{26}$$

A similar argument implies that Q^{k+1} minimizes $\lambda\|Q\|_1 - \rho(U^{k+1})^\top Q$, and we have

$$\lambda\|Q^{k+1}\|_1 - \rho(U^{k+1})^\top Q^{k+1} \le \lambda\|Q^*\|_1 - \rho(U^{k+1})^\top Q^*. \tag{27}$$

Adding inequalities (26) and (27), we obtain

$$\begin{aligned}&\frac{1}{2}\|A^{k+1}\Theta^{k+1} - b^{k+1}\|_2^2 - \frac{1}{2}\|A^*\Theta^* - b^*\|_2^2 + \lambda\|Q^{k+1}\|_1 - \lambda\|Q^*\|_1\\ &\le \{\rho U^{k+1} + \rho(Q^{k+1} - Q^k)\}^\top F\Theta^* - \{\rho U^{k+1} + \rho(Q^{k+1} - Q^k)\}^\top F\Theta^{k+1}\\ &- (\rho U^{k+1})^\top Q^* + (\rho U^{k+1})^\top Q^{k+1}.\end{aligned}$$

Then we plug in $F\Theta^* - Q^* = 0$ to the righthand side of the above inequality, and we obtain the righthand side of the (23)

$$\begin{aligned}&\{\rho U^{k+1} + \rho(Q^{k+1} - Q^k)\}^\top F(\Theta^* - \Theta^{k+1}) + (\rho U^{k+1})^\top(Q^{k+1} - Q^*)\\ =&\{\rho U^{k+1}\}^\top F(\Theta^* - \Theta^{k+1}) + \{\rho(Q^{k+1} - Q^k)\}^\top F(\Theta^* - \Theta^{k+1}) + (\rho U^{k+1})^\top(Q^{k+1} - Q^*)\\ =&\{\rho U^{k+1}\}^\top(Q^* - F\Theta^{k+1}) + (\rho U^{k+1})^\top(Q^{k+1} - Q^*) + \{\rho(Q^{k+1} - Q^k)\}^\top(Q^* - F\Theta^{k+1})\\ =&(\rho U^{k+1})^\top(Q^{k+1} - F\Theta^{k+1}) + \rho\{(Q^{k+1} - Q^k)\}^\top\{Q^* - F\Theta^{k+1})\}\\ =&-(\rho U^{k+1})^\top(F\Theta^{k+1} - Q^{k+1}) - \rho\{(Q^{k+1} - Q^k)\}^\top\{(F\Theta^{k+1} - Q^{k+1}) + (Q^{k+1} - Q^*).\}\end{aligned}$$

Thus (23) is proved.

The third inequality to prove is (20). To achieve this, we first add (22) and (23), and multiply the sum by 2,

$$\begin{aligned} &2(\rho U^{k+1} - \rho U^*)^\top (F\Theta^{k+1} - Q^{k+1}) + 2\rho(Q^{k+1} - Q^k)^\top (F\Theta^{k+1} - Q^{k+1}) \\ &+ 2\rho(Q^{k+1} - Q^k)^\top (Q^{k+1} - Q^*) \leq 0. \end{aligned} \tag{28}$$

Then inequality (20) can be obtained from (28) after some manipulation and rewriting.

Substituting $U^{k+1} = U^k + (F\Theta^{k+1} - Q^{k+1})$ in the first term of inequality (28) gives

$$2\rho(U^k - U^*)^\top (F\Theta^{k+1} - Q^{k+1}) + \rho\|F\Theta^{k+1} - Q^{k+1}\|_2^2 + \rho\|F\Theta^{k+1} - Q^{k+1}\|_2^2. \tag{29}$$

Replacing $F\Theta^{k+1} - Q^{k+1} = (U^{k+1} - U^k)$ in the first two terms of (29) gives

$$2\rho(U^k - U^*)^\top (U^{k+1} - U^k) + \rho\|U^{k+1} - U^k\|_2^2 + \rho\|F\Theta^{k+1} - Q^{k+1}\|_2^2. \tag{30}$$

Since $U^{k+1} - U^k = (U^{k+1} - U^*) - (U^k - U^*)$, (30) can be written as

$$\rho\left(\|U^{k+1} - U^*\|_2^2 - \|U^k - U^*\|_2^2\right) + \rho\|F\Theta^{k+1} - Q^{k+1}\|_2^2. \tag{31}$$

Substituting (31) in the first term of (28) we obtain:

$$\begin{aligned} &\rho\left(\|U^{k+1} - U^*\|_2^2 - \|U^k - U^*\|_2^2\right) + \rho\|F\Theta^{k+1} - Q^{k+1}\|_2^2 \\ &+ 2\rho(Q^{k+1} - Q^k)^\top (F\Theta^{k+1} - Q^{k+1}) + 2\rho(Q^{k+1} - Q^k)^\top (Q^{k+1} - Q^*) \leq 0. \end{aligned} \tag{32}$$

Physically, this step implies that the interaction between primal tracking errors and dual price adjustments can be geometrically decomposed. This decomposition lets us isolate the algorithm's step-by-step progress within a single iteration.

Replacing $Q^{k+1} - Q^* = (Q^{k+1} - Q^k) + (Q^k - Q^*)$ in the last term of (32), we obtain

$$\begin{aligned} &\rho\left(\|U^{k+1} - U^*\|_2^2 - \|U^k - U^*\|_2^2\right) \\ &+ \rho\|F\Theta^{k+1} - Q^{k+1} + (Q^{k+1} - Q^k)\|_2^2 + \rho\|Q^{k+1} - Q^k\|_2^2 + 2\rho(Q^{k+1} - Q^k)^\top (Q^k - Q^*) \leq 0. \end{aligned} \tag{33}$$

Substituting $Q^{k+1} - Q^k = (Q^{k+1} - Q^*) - (Q^k - Q^*)$ in the last two terms of (33), we obtain

$$\begin{aligned} &\rho\left(\|U^{k+1} - U^*\|_2^2 - \|U^k - U^*\|_2^2\right) + \rho\|F\Theta^{k+1} - Q^{k+1} + (Q^{k+1} - Q^k)\|_2^2 \\ &+ \rho\left(\|Q^{k+1} - Q^*\|_2^2 - \|Q^k - Q^*\|_2^2\right) \leq 0. \end{aligned} \tag{34}$$

Then we can rewrite (34) as

$$V^k - V^{k+1} \geq \rho\|F\Theta^{k+1} - Q^{k+1} + (Q^{k+1} - Q^k)\|_2^2. \tag{35}$$

(35) is equivalent to

$$V^{k+1} \leq V^k - \rho\|F\Theta^{k+1} - Q^{k+1}\|_2^2 - \rho\|(Q^{k+1} - Q^k)\|_2^2 - 2\rho(F\Theta^{k+1} - Q^{k+1})(Q^{k+1} - Q^k). \tag{36}$$

Next, we show that $2\rho(F\Theta^{k+1} - Q^{k+1})(Q^{k+1} - Q^k)$ is nonnegative. Note that Q^{k+1} minimises $\lambda\|Q\|_1 - (\rho U^{k+1})^\top Q$ and Q^k minimises $\lambda\|Q\|_1 - (\rho U^k)^\top Q$. Thus we have

$$\lambda\|Q^{k+1}\|_1 - (\rho U^{k+1})^\top Q^{k+1} \leq \lambda\|Q^k\|_1 - (\rho U^{k+1})^\top Q^k, \tag{37}$$

and

$$\lambda\|Q^k\|_1 - (\rho U^k)^\top Q^k \leq \lambda\|Q^{k+1}\|_1 - (\rho U^k)^\top Q^{k+1}, \tag{38}$$

From (37) and (38), we have

$$(\rho U^{k+1})^\top Q^k - (\rho U^{k+1})^\top Q^{k+1} \leq \lambda \|Q^k\|_1 - \lambda \|Q^{k+1}\|_1 \leq (\rho U^k)^\top Q^k - (\rho U^k)^\top Q^{k+1}. \tag{39}$$

From (39) we will have

$$\rho(U^{k+1} - U^k)^\top (Q^{k+1} - Q^k) \geq 0.$$

Since $U^{k+1} - U^k = F\Theta^{k+1} - Q^{k+1}$, thus we have proved that $2\rho(F\Theta^{k+1} - Q^{k+1})(Q^{k+1} - Q^k)$ is nonnegative. Thus, (36) leads to

$$V^{k+1} \leq V^k - \rho \|F\Theta^{k+1} - Q^{k+1}\|_2^2 - \rho \| - (Q^{k+1} - Q^k)\|_2^2. \tag{40}$$

With the non-negativity of the cross term verified, the negative sign in front of the final terms in (40) is preserved, thereby concluding the proof of the monotonic decay inequality (20). From a practitioner's standpoint, this complete derivation provides a rigorous guarantee that V^k strictly decreases at every iteration.

E Additional simulation results

Table 6 Simulation results for change point detection over 50 replicates for model 3: mean (standard deviation)

		FL-ERGM			VCERGM			ERGM		
T	**k**	Sens.	Spec.	Loc.	Sens.	Spec.	Loc.	Sens.	Spec.	Loc.
10	1	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	0.08 (0.27)	0.76 (0.17)	**0.00 (0.00)**
	2	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	0.02 (0.14)	0.65 (0.18)	—
	3	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	0.78 (0.42)	0.97 (0.06)	**0.00 (0.00)**
20	1	**0.00 (0.00)**	0.05 (0.00)	—	**0.00 (0.00)**	0.83 (0.03)	—	**0.00 (0.00)**	0.05 (0.00)	—
	2	0.92 (0.27)	0.97 (0.04)	0.07 (0.33)	0.52 (0.50)	0.96 (0.04)	**0.00 (0.00)**	0.94 (0.24)	0.86 (0.08)	0.21 (0.51)
	3	**0.00 (0.00)**	0.29 (0.16)	—	0.04 (0.20)	0.90 (0.04)	**0.00 (0.00)**	**0.00 (0.00)**	0.07 (0.03)	—
30	1	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	**0.00 (0.00)**	0.60 (0.10)	—
	2	0.81 (0.30)	1.00 (0.01)	0.04 (0.18)	0.00 (0.00)	0.96 (0.00)	—	**1.00 (0.00)**	0.69 (0.07)	0.27 (0.41)
	3	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	0.08 (0.27)	0.88 (0.06)	**0.00 (0.00)**
40	1	0.92 (0.27)	1.00 (0.01)	**0.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	0.04 (0.20)	0.93 (0.04)	**0.00 (0.00)**
	2	0.72 (0.45)	0.99 (0.01)	0.03 (0.17)	0.00 (0.00)	0.97 (0.00)	—	0.84 (0.37)	0.84 (0.06)	0.26 (0.54)
	3	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	0.84 (0.37)	1.00 (0.01)	**0.00 (0.00)**
50	1	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	**0.00 (0.00)**	0.86 (0.05)	—
	2	0.97 (0.10)	**1.00 (0.00)**	**0.00 (0.00)**	0.00 (0.00)	0.98 (0.00)	—	0.99 (0.05)	0.85 (0.06)	0.10 (0.17)
	3	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	0.84 (0.37)	1.00 (0.01)	**0.00 (0.00)**

F Application with political Co-voting networks

We also apply the FL-ERGM to the political co-voting networks to study the co-voting patterns of senators. The co-voting networks consist of a dynamic sequence of networks from Congress 40 (1867–1869) to Congress 113 (2013–2015). In each network, nodes represent senators, and edges indicate that the corresponding senators vote concurringly (both yay or both nay) on at least 80% of the bills. This dataset has been studied by Lee et al. (2020) using the VCERGM method. In this paper, we apply the FL-ERGM to the co-voting networks with the same statistics listed as follows:

- $S^1(y_i) = e^{\phi} \sum_{k=1}^{n-2} \{1 - (1 - e^{-\phi})^k\} EP_k(y_i)$, $\phi = 1$, geometrically weighted edgewise shared partners (gwesp), representing transitivity. It measures the extent to which the decisions of two senators are the same as their shared co-voters. $EP_k(y_i)$ is the number of connected pairs that have k common neighbours.

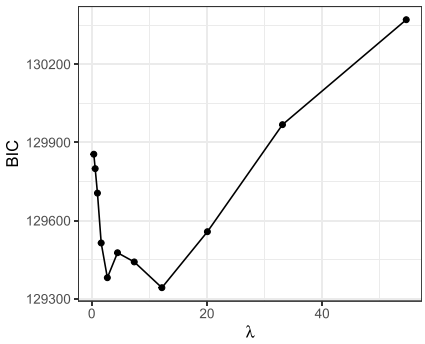


Fig. 11 PBIC of the FL-ERGM for dynamic co-voting networks

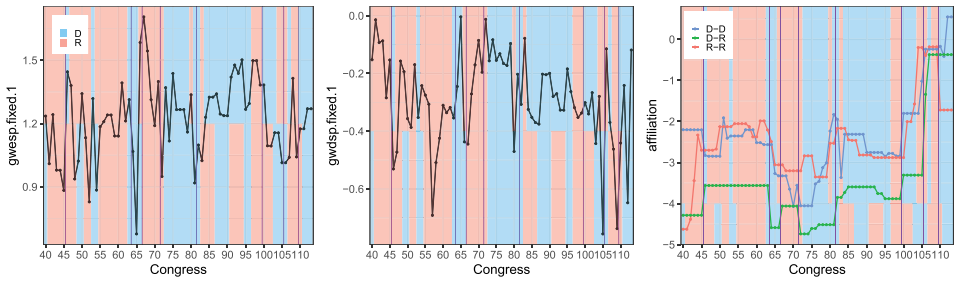


Fig. 12 Parameter estimates of the FL-ERGM for co-voting networks. in all three plots, the background colour of the top panel indicates the majority party of the Senate, and the background colour of the bottom panel represents the party of the US president. in the third plot, blue lines stand for the effects of Democrats-Democrats, green lines stand for the effects of Democrats-Republicans, and red lines show the effects of Republicans-Republicans

- $S^2(y_i) = e^{\phi} \sum_{k=1}^{n-2} \{1 - (1 - e^{-\phi})^k\} DP_k(y_i)$, $\phi = 1$, geometrically weighted dyad-wise shared partners (gwdsp), representing local clustering phenomenon. It measures the extent to which the senators agree on each other. $DP_k(y_i)$ is the number of dyads that have k common neighbours.
- $S^3(y_i) = \sum_{r \neq s} y_{rs,i} \mathbf{1}(X_r = X_s = D)$, the number of connections within Democratic senators.
- $S^4(y_i) = \sum_{r \neq s} y_{rs,i} \mathbf{1}(X_r \neq X_s)$, the number of connections between Democratic and Republican senators.
- $S^5(y_i) = \sum_{r \neq s} y_{rs,i} \mathbf{1}(X_r = X_s = R)$, the number of connections within Republican senators.

Next, we investigate the results of the FL-ERGM. PBIC is shown in Fig. 11 and parameter estimates of the FL-ERGM are presented in Fig. 12. The coefficient estimates from the FL-ERGM are functions over time. θ_1 and θ_2, representatives of the local clustering property, change irregularly with time. θ_3, θ_4, θ_5, the effects of affiliation, exhibit piecewise constant changing patterns. The big positive gwesp coefficient and small negative gwdsp coefficient suggest high transitivity of co-voting networks, meaning that there exist groups of senators that agree with each other on many bills. The negative coefficients of affiliation terms indicate the amount of decrease in the log odds of two senators being connected. In other words, senators being in different parties are most likely disconnected since θ_4 has the biggest absolute value of the three coefficients most time. The effects of Democrats and Republicans are close, and we also notice that the comparison of θ_3 and θ_5 is closely related to the majority party in the Senate. In order to show this relationship more clearly, we indicate the majority party of the Senate through the background colour of the top panel, and the party of the US president through the background colour of the bottom panel. Between Congress 47 and 72, Republican is the majority party in Congress for 22 times (out of 27 times) and the coefficient of Republicans, θ_5, is bigger than the coefficient of Democrats, θ_3, for 24 times. Between Congress 78 and 102, the coefficient of Democrats, θ_3, is bigger than the coefficient of Republicans, θ_5, for 23 times (out of 25 times), and Democrat is also the majority party in Congress for 20 times.

Our model found that the sequence of co-voting networks can be divided into 9 periods. We have been able to identify each of these periods with events in American history. Networks at each period have different

https://doi.org/10.1007/s41109-026-00846-8

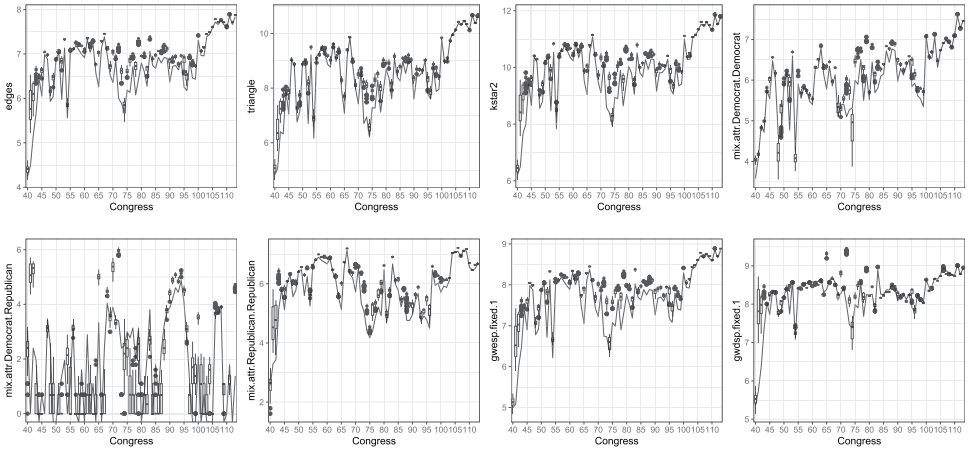


Fig. 13 Model assessments of the FL-ERGM on dynamic co-voting networks. The solid lines indicate the observed network statistics and the box plots represent the statistics of networks simulated from the estimated model

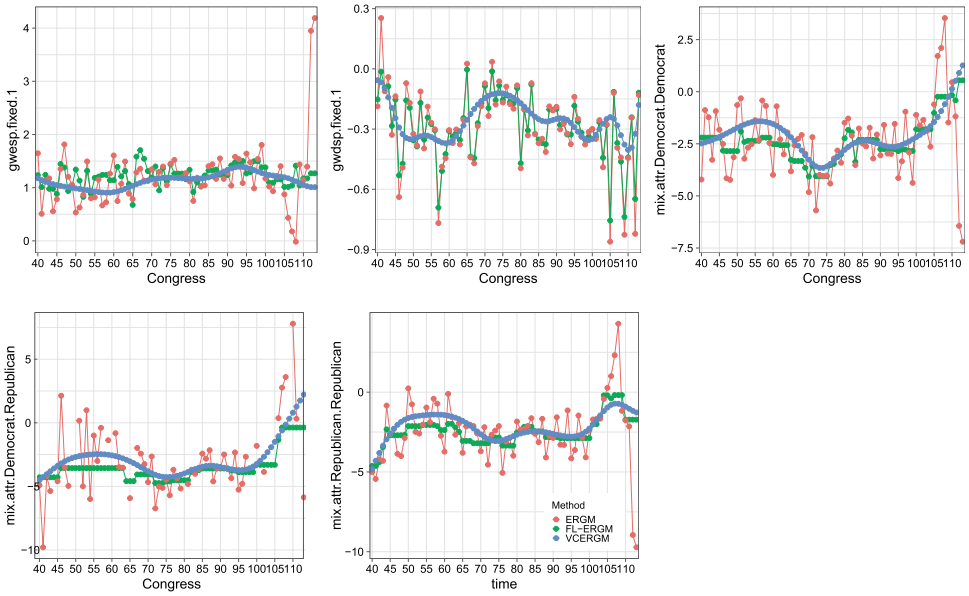


Fig. 14 Parameter estimates of different methods for co-voting networks. Red lines represent the estimates from the ERGM, blue lines stand for the estimates from the VCERGM, and green lines show results from the proposed FL-ERGM

characteristics represented by distinct coefficients of each period. The first period starts from Congress 40 (1867–1869) to 45 (1877–1879), when the US went through the Reconstruction period after the Civil War. The Congress is gradually dominated by Republican, indicated by the increasing coefficient of θ_5(R-R), while θ_3(D-D) and θ_4 (D-R) remain constant during this time. We also notice that there is an abrupt change from Congress 45 to 46, suggested by the sharp increase on θ_1(gwesp) and decrease on θ_2(gwdsp), meaning that new connections are built because of transitivity. The second phase, form Congress 46 (1879–1881) to 63 (1913–1915), started with the Gilded Age. During this period, US finished the Second Industrial Revolution and became a world economic and military power. The Congress was dominated by Republicans and more agreement is reached inside the Republican party than the Democratic party. There is also an abrupt change on θ_3 and θ_5, at around Congress 50 (1887–1889) and 51 (1889–1891), which matches the fact of the change in the affiliation of the president after a long time.

The third period, from Congress 64 (1915–1917) to 66 (1919–1921), corresponds to time of the World War I. $\theta_3, \theta_4, \theta_5$ all decrease compared with the last period, meaning that the effect of affiliation increases during wartime. Also, θ_1 reaches the minimum and θ_2 reaches its maximum at Congress 65. The 4th period, Congress 67 (1921–1923) to 71 (1929–1931), matches the Roaring Twenties when US had the great economic prosperity after World War I. θ_3 and θ_5 decrease but θ_4 increases compared with the last period. The 5th period is between Congress 72

(1931–1933) and 81 (1949–1951), when US experienced the Great Depression and World War II. We notice that the coefficient of Democrats increases dramatically and becomes bigger than the coefficient of Republicans. The position of Democrats is gained during the Great Depression and kept during the World War II.

The 6th period, from Congress 82 (1951–1953) to 99 (1985–1987), is related to the Cold War. The coefficient of Democrats is bigger than the coefficient of Republicans most of the time. We also notice that θ_3 has a dip at Congress 83 (1953–1955), which matches the time that the famous president Eisenhower was elected as the first Republican president since 1932. The 7th period, from Congress 100 (1987–1989) to 105 (1998–1999) is the post-Cold War era. $\theta_3, \theta_4, \theta_5$ all increase compared with the last period, indicating that affiliation plays a less important role in the decision of senators. θ_4 increases sharply at Congress 106, meaning that there is a lot of agreement between Democrats and Republicans. The 8th period, from Congress 106 (1999–2001) to 109 (2005–2007), lasts for 8 years. $\theta_3, \theta_4, \theta_5$ all remained stable during this time when Bush was the president. The change from phase 8 to 9 is marked by the financial crisis and at the same time, Democrats became the Senate majority again from Congress 110. During the 9th phase, from Congress 110 (2007–2009) to 115 (2013–2015), when Obama was the president, the coefficient of Democrats increased but the coefficient of Republicans decreased.

The model assessment is shown in Fig. 13, suggesting that the FL-ERGM fits the networks well. The results of the VCERGM and ERGM are also presented in Fig. 14. Similar to what we found in the simulation studies, ERGM estimates are spiky and VCERGM estimates are smooth. In comparing the FL-ERGM with competing methods, we see that the coefficient estimates from the FL-ERGM can account for the characteristics of each network as well as show the abrupt change points on the network structure.

Acknowledgements The authors would like to thank the editor and the reviewers for their constructive comments and suggestions that have helped to improve the manuscript. The authors are grateful to Dr. Jihui Lee for sharing their code and dataset on GitHub. The authors would also like to thank Dr. Xiao Guo from Northwest University for the helpful discussion on the paper.

Author contributions SR developed the model and drafted the manuscript. PL contributed to algorithm development and manuscript drafting. HZ conceived and designed the research. JZ and XW assisted with data analysis, interpretation of results, and provided methodological support. All authors contributed to editing and revising the manuscript and read and approved the final version.

Funding SR gratefully acknowledges the financial support provided by the University of Kent through a PhD scholarship during the period in which most of the research was undertaken.

Data availability The trade data are downloaded from the United Nations Commodity Trade Statistics Database (https://comtrade.un.org/). The R code and the data used in this study will be made available in a GitHub repository upon acceptance of the manuscript.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Betancourt B, Rodríguez A, Boyd N (2017) Bayesian fused lasso regression for dynamic binary networks. J Comput Graph Stat 26(4):840–850. https://doi.org/10.1080/10618600.2017.1341323

Bouranis L, Friel N, Maire F (2018) Bayesian model selection for exponential random graph models via adjusted pseudolikelihoods. J Comput Graph Stat 27(3):516–528. https://doi.org/10.1080/10618600.2018.1448832

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2010) Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn 3(1):1–122

Caimo A, Friel N (2011) Bayesian inference for exponential random graph models. Soc Networks 33(1):41–55. https://doi.org/10.1016/j.socnet.2010.09.004

Frank O, Strauss D (1986) Markov graphs. J Am Stat Assoc 81(395):832–842. https://doi.org/10.1080/01621459.1986.10478342

Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441. https://doi.org/10.1093/biostatistics/kxm045

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22. https://doi.org/10.18637/jss.v033.i01

Geyer C, Thompson E (1992) Constrained monte carlo maximum likelihood for depedent data. J R Stat Soc Ser B 54(3):657–683. https://doi.org/10.1111/j.2517-6161.1992.tb01443.x

Hallac D, Park Y, Boyd S, Leskovec J (2017) Network inference via the time-varying graphical lasso. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 205–213

Hanneke S, Fu W, Xing EP (2010) Discrete temporal models of social networks. Electron J Stat 4(none):585–605. https://doi.org/10.1214/09-EJS548

Hoff PD, Raftery AE, Handcock MS (2002) Latent space approaches to social network analysis. J Am Stat Assoc 97(460):1090–1098. https://doi.org/10.1198/016214502388618906

Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: first steps. Soc Networks 5(2):109–137. https://doi.org/10.1016/0378-8733(83)90021-7

Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs. J Am Stat Assoc 76(373):33–50. https://doi.org/10.1080/01621459.1981.10477598

Hunter DR, Handcock MS (2006) Inference in curved exponential family models for networks. J Comput Graph Stat 15(3):565–583. https://doi.org/10.1198/106186006X133069

Hunter DR, Handcock MS, Butts CT, Goodreau SM, Morris M (2008) Ergm: a package to fit, simulate and diagnose exponential-family models for networks. J Stat Softw 24(3):1–29. https://doi.org/10.18637/jss.v024.i03

Jiang X, Margarita Livas S, Yin F, Banerjee S, Butts CT, Shen W (2023) Structure recovery and trend estimation for dynamic network analysis. Stat 12(1):e593. https://doi.org/10.1002/sta4.593

Krivitsky PN, Handcock MS (2014) A separable model for dynamic networks. J R Stat Soc Ser B 76(1):29–46. https://doi.org/10.1111/rssb.12014

Lee C, Wilkinson DJ (2019) A review of stochastic block models and extensions for graph clustering. Appl Netw Sci 4(1):1–50. https://doi.org/10.1007/s41109-019-0232-2

Lee J, Li G, Wilson JD (2020) Varying-coefficient models for dynamic networks. Comput Stat Data Anal 152:107052. https://doi.org/10.1016/j.csda.2020.107052

Lee KH, Xue L, Hunter DR (2020) Model-based clustering of time-evolving networks through temporal exponential-family random graph models. J Multivar Anal 175:104540. https://doi.org/10.1016/j.jmva.2019.104540

Li T, Levina E, Zhu J (2020) Network cross-validation by edge sampling. Biometrika 107(2):257–276. https://doi.org/10.1093/biomet/asaa006

Morris M, Handcock MS, Hunter DR (2008) Specification of exponential-family random graph models: terms and computational aspects. J Stat Softw 24(4):1548. https://doi.org/10.18637/jss.v024.i04

Ren S, Wang X, Liu P, Zhang J (2023) Bayesian nonparametric mixtures of exponential random graph models for ensembles of networks. Soc Networks 74:156–165. https://doi.org/10.1016/j.socnet.2023.03.005

Robins G, Pattison P, Kalish Y, Lusher D (2007) An introduction to exponential random graph (p^*) models for social networks. Soc Networks 29(2):173–191

Schmid CS, Desmarais BA (2017) Exponential random graph models with big networks: maximum pseudolikelihood estimation and the parametric bootstrap. In 2017 IEEE International Conference on Big Data (Big Data), IEEE, pp. 116–121

Schmid CS, Hunter DR (2023) Computing pseudolikelihood estimators for exponential-family random graph models. J Data Sci 21(2):295–309. https://doi.org/10.6339/23-JDS1094

Schweinberger M, Krivitsky PN, Butts CT, Stewart JR (2020) Exponential-family models of random graphs: inference in finite, super and infinite population scenarios. Statis Sci 35(4):627–662. https://doi.org/10.1214/19-STS743

Sewell DK, Chen Y (2015) Latent space models for dynamic networks. J Am Stat Assoc 110(512):1646–1657. https://doi.org/10.1080/01621459.2014.988214

Snijders TA, de Bunt GGV, Steglich CE (2010) Introduction to stochastic actor-based models for network dynamics. Soc Networks 32(1):44–60
Snijders TAB (2002) Markov chain monte carlo estimation of exponential random graph models. J Soc Struct 3(2):1–40
Strauss D, Ikeda M (1990) Pseudolikelihood estimation for social networks. J Am Stat Assoc 85(409):204–212. https://doi.org/10.1080/01621459.1990.10475327
Tan LS, Friel N (2020) Bayesian variational inference for exponential random graph models. J Comput Graph Stat 29(4):910–928. https://doi.org/10.1080/10618600.2020.1740714
Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. J R Stat Soc Ser B 67(1):91–108. https://doi.org/10.1111/j.1467-9868.2005.00490.x
Tibshirani RJ, Taylor J (2011) The solution path of the generalized lasso. Ann Stat 39(3):1335–1371. https://doi.org/10.1214/11-AOS878
Tibshirani RJ, Taylor J (2012) Degrees of freedom in lasso problems. Ann Stat 40(2):1198–1232. https://doi.org/10.1214/12-AOS1003

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Sa Ren[1] · Peng Liu[2] · Hai Zhang[3] · Xue Wang[4] · Jian Zhang[5]

✉ Sa Ren
sarah.ren@sheffield.ac.uk

[1] School of Medicine and Population Health, University of Sheffield, Sheffield, UK

[2] School of Science, Loughborough University, Loughborough, UK

[3] School of Mathematics, Northwest University, Xi'an, Shaanxi, China

[4] Walsn Limited, Canterbury, UK

[5] School of Engineering, Mathematics and Physics, University of Kent, Canterbury, UK