

# Score-Driven Exponential Random Graphs: A New Class of Time-Varying Parameter Models for Temporal Networks

D. Di Gangi,<sup>1</sup> G. Bormetti,<sup>2</sup> and F. Lillo<sup>3</sup><sup>1</sup>*Domotz, via U. Forti 1, 56121 Pisa, Italy*<sup>2</sup>*Department of Economics and Management, University of Pavia, Via San Felice al Monastero 5, 27100, Pavia, Italy*<sup>3</sup>*Department of Mathematics, University of Bologna, Piazza di Porta San Donato 5, 40126, Bologna, Italy, and Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126, Pisa, Italy*

(\*Electronic mail: giacomo.bormetti@unipv.it)

(Dated: 17 October 2024)

Motivated by the increasing abundance of data describing real-world networks that exhibit dynamical features, we propose an extension of the Exponential Random Graph Models (ERGMs) that accommodates the time variation of its parameters. Inspired by the fast-growing literature on Dynamic Conditional Score models, each parameter evolves according to an updating rule driven by the score of the ERGM distribution. We demonstrate the flexibility of score-driven ERGMs (SD-ERGMs) as data-generating processes and filters and show the advantages of the dynamic version over the static one. We discuss two applications to temporal networks from financial and political systems. First, we consider the prediction of future links in the Italian interbank credit network. Second, we show that the SD-ERGM allows discriminating between static or time-varying parameters when used to model the U.S. Congress co-voting network dynamics.

**The paper introduces an innovative, dynamic model for temporal networks. The novel methodology is strongly interdisciplinary, and the estimation of model parameters is straightforward. We present two examples from financial and political systems, supporting the approach's flexibility and showcasing its potential widespread applicability.**

## I. INTRODUCTION

A network is a useful abstraction for a system composed of several single elements with some pairwise relations. The simplified description of social, economic, biological, and transportation systems, often very complex, in terms of nodes and links, attracted and still attracts an enormous amount of attention<sup>1–5</sup>. Formally, a network  $G$  is a pair  $(V, E)$  where  $V$  is a set of nodes and  $E$  is a set of node pairs named links. The nodes are labeled, and a link is identified by the pair of nodes it connects  $(i, j)$ . To each  $G$ , we can assign an adjacency matrix  $\mathbf{Y}$  such that  $Y_{ij} = 1$  if link  $(i, j)$  is present in  $E$  and  $Y_{ij} = 0$  otherwise. Links may have orientations. The corresponding network is dubbed directed. If the elements of the adjacency matrix are allowed to be different from 0 or 1, one speaks of weighted networks. In the following, we will focus on directed networks rather than consider the weighted variant.

Often, systems that are fruitfully described as networks evolve in time. When the number of nodes and/or pairwise interactions change over time, one usually speaks of temporal networks<sup>6–8</sup>. This paper will focus on temporal networks where links evolve in discrete time. A temporal network is a sequence of networks, each associated with an adjacency matrix and observed at  $T$  different points in time. The whole time series is given in terms of a sequence of matrices  $\left\{ Y_{ij}^{(t)} \right\}_{t=1}^T$ .

We introduce an innovative approach to temporal networks based on two main ingredients: (i) a parametric probabilistic model, according to which one can sample a network realization. A natural choice is the class of statistical models for networks known as Exponential Random Graph Models (ERGMs). (ii) A simple mechanism to induce dynamics on the network sequence by introducing time variation on the model parameters. The Dynamic Conditional Score (DCS) approach provides a flexible candidate. Our extension of the ERGM framework allows model parameters to change over time in a score-driven fashion. We develop a new class of temporal network models and show its versatility and effectiveness in capturing time-varying features. The information encoded in  $\mathcal{F}_{t-1}$  is exploited to filter the time-varying parameters (tvps)  $\theta^{(t)}$  at time  $t$ . We refer to this class as *Score-Driven Exponential Random Graph Models* (SD-ERGMs). A generic SD-ERGM can be used either as a data-generating process (DGP) to sample synthetic sequences of graphs or as an effective filter of latent tvps, regardless of the true DGP.

We are by no means the first to discuss models for temporal networks. For a review of latent space temporal network models, one can refer to Ref. 9 Extensions of the ERGM framework for the description of temporal networks exist in the literature. Two are the main streams. The first one, termed TERGM, was pioneered in Ref. 10 and further explored in Refs. 11–13. This approach builds on the ERGM but allows the network statistics to define the probability at time  $t$  to depend on current and previous networks up to time  $t - K$ . This  $K$ -step Markov assumption is a defining feature of the TERGMs. A second approach allows for the parameters of the ERGM to be time-varying. A notable example is the Varying-Coefficient-ERGM<sup>14</sup>, allowing for smooth parameter time variation. The approach differs from ours in several respects. Specifically, to infer parameter time-variation at time  $t$ , it uses all the available observations, including those

from future times  $t' > t^{15}$ . Consequently, it cannot be used to draw causal sequences of time-evolving networks. A related but different approach<sup>16</sup> considers the possibility of a random evolution of node-specific parameters. As a crucial difference with the SD-ERGM, the parameter evolution is driven by an exogenous source of randomness. Following the language of Ref. 17, the approach is *parameter driven*, while we consider an *observation driven* dynamics. Finally, it is important to mention that the social science literature has considered alternative frameworks for modeling temporal networks. Notable examples are the Stochastic Actor Oriented model<sup>18</sup> and the Relational Event Model<sup>19</sup>. For an overview of contributions, we refer to the literature therein.

The rest of the paper is organized as follows. In Section II, we review some key concepts on ERGMs and observation-driven models. In Section III, we introduce the new class of models and validate it with extensive numerical experiments for three specific instances of the SD-ERGM. Section IV presents the results from an application to two real temporal networks: The e-MID interbank network for liquidity supply and demand and the U.S. Congress co-voting political network. Section V draws the relevant findings.

## II. MAIN BUILDING BLOCKS

The two main ingredients in our approach are the ERGMs, a static class of network models well-known in physics, and the DCSs, a recent development in time-series econometrics.

### A. ERGM - Exponential Random Graph Models

A statistical network model can be specified by providing the probability distribution over the set of possible adjacency matrices<sup>20</sup>. If the distribution belongs to the exponential family<sup>21</sup>, then the model is named ERGM, and its log-likelihood takes the form

$$\log P(\mathbf{Y}) = \sum_s \theta_s h_s(\mathbf{Y}) - \log(\mathcal{K}(\theta)), \quad (1)$$

where  $h$  are network statistics,  $\theta$  is the vector of parameters whose component  $\theta_s$  is associated with the network statistic  $h_s(\mathbf{Y})$ , and  $\mathcal{K}(\theta) = \sum_{\{\mathbf{Y}\}} e^{\theta_s h_s(\mathbf{Y})}$ . The ERGMs literature is vast and still growing<sup>22</sup>. The ERGM framework is intrinsically linked to the very well-known *principle of maximum entropy*<sup>23</sup> and its applications to statistical physics<sup>24</sup>. Indeed, an ERGM with sufficient statistics  $h(\theta)$  naturally arises when looking for the probability distribution which maximizes the entropy under a linear equality constraint on the statistics  $h(\theta)$ <sup>25,26</sup>. The sufficient statistics  $h_s(\mathbf{Y})$ , known as network statistics, are functions of the adjacency matrix  $\mathbf{Y}$ , whose entries are binary random variables. The probability mass function (PMF) is defined by (1). The normalizing factor  $\mathcal{K}(\theta)$  is often unavailable as a closed-form function of the parameters  $\theta$ .

In the following, we will focus on two specific examples of ERGMs that describe distinct features of the network and

require different approaches to parameter inference. The first one is meant to capture the heterogeneity in the number of connections each node can have, and it allows for straightforward maximum likelihood estimation (MLE)<sup>27</sup>. It is known as *beta model*, *fitness model*, and *configuration model*<sup>25-30</sup>. The second one is an ERGM having as statistic the Geometrically Weighted Edgewise Shared Partners (GWESP). That is a network statistic describing transitivity in the formation of links, i.e., the tendency of connected nodes to have common neighbors and belongs to a family of network statistics referred to as curved exponential random graphs, proposed in Refs. 31 and 32 and discussed in Ref. 33. In the latter case, the inference is complicated because the normalizing factor in (1) is not available in closed form. In such cases, there are two standard approaches to ERGM inference, both consisting of maximizing alternative functions that are known to share the same optimum as the exact likelihood. In Appendix A, we provide more details on the beta model and GWESP ERGMs definitions as long as more details on the associated inference procedures.

### B. Score-Driven Models

The second main ingredient of this work is the class of DCS models<sup>34,35</sup>, also known as Generalized Autoregressive Score models<sup>36</sup>. In the language of Ref. 17, DCSs belong to observation-driven models. Let us consider a sequence of observations  $\{y^{(t)}\}_{t=1}^T$ , where each  $y^{(t)} \in \mathbb{R}^M$ , and a conditional probability density  $P(y^{(t)}|f^{(t)})$ , that depends on a vector of tvps  $f^{(t)} \in \mathbb{R}^K$ . Defining the score as

$$\nabla^{(t)} = \frac{\partial \log P(y^{(t)}|f^{(t)})}{\partial f^{(t)}}, \quad (2)$$

a Score-Driven model assumes that the recursive relation

$$f^{(t+1)} = w + \beta f^{(t)} + \alpha S^{(t)} \nabla^{(t)}, \quad (3)$$

rules the time evolution of  $f^{(t)}$ , with  $w$ ,  $\alpha$  and  $\beta$  are static parameters,  $w$  being a  $K$  dimensional vector and  $\alpha$  and  $\beta$   $K \times K$  matrices.  $S^{(t)}$  is a  $K \times K$  scaling matrix usually chosen as a power of the inverse of the Fisher information matrix associated with  $P(y^{(t)}|f^{(t)})$ .

The parameter updating rule can be intuitively motivated by general assumptions based on information theory principles. One can assume that the network behavior varies based on surprise: The more an observation of the network's state, i.e., the adjacency matrix, is "unexpected", the more the relations between its components will change. The most common measure of surprise is minus the logarithm of the likelihood of observing the current state conditional on the level of the model parameters. As a second principle, one assumes that the reaction to surprise is to adapt to it, making what had been unexpected at that moment less surprising in the future. This implies that the parameters change to minimize the surprise,

i.e., increase the log-likelihood of the last observation and thus move along the steepest direction the gradient provides. Then, the updated parameter value will be a linear combination of the current value and the log-likelihood score.

The structure of the conditional observation density determines the score, from which the dependence of  $f^{(t+1)}$  on the vector of observations  $y^{(t)}$  follows. When the model is viewed as a DGP, the update results in stochastic dynamics precisely thanks to the random occurrence of  $y^{(t)}$ . When the score-driven recursion is regarded as a filter, the update rule in (3) is used to obtain a sequence of filtered  $\left\{\hat{f}^{(t)}\right\}_{t=1}^T$ . In this setting, one estimates the static parameters by maximizing the log-likelihood of the whole sequence of observations.

A second look at eq. (3) reveals the similarity of the score-driven recursion with the iterative step from a Newton algorithm, whose objective function is precisely the log-likelihood function. As mentioned above, at each step, the score pushes the parameter vector along the log-likelihood steepest direction. Moreover, there are motivations, grounded on the variation of the Kullback-Leibler divergence, for the optimality of the score-driven updating rule<sup>37</sup>, as we review in Appendix B.

Many well-known econometrics models can be expressed as Score-Driven models. Famous examples are the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model<sup>38</sup>, the Exponential GARCH model<sup>39</sup>, the Autoregressive Conditional Duration model<sup>40</sup>, and the Multiplicative Error Model<sup>41</sup>. The introduction of this framework in its full generality opened the way to applications in various contexts.

Before moving to the most crucial section, let us mention two relevant technical aspects for the applications discussed in the paper: the computation of the confidence bands of the filtered parameters and testing for the parameter temporal variation. We postpone the technical discussion to Appendix B for brevity.

### III. SCORE-DRIVEN EXPONENTIAL RANDOM GRAPHS

This section describes the methodological innovation introduced by the manuscript. We present the general SD-ERGM framework, discuss the score-driven approach's applicability to three different ERGMs in detail, and validate their performances with extensive numerical simulations.

We apply the score-driven methodology to ERGMs to allow any of the parameters  $\theta_s$  in (1) to have a stochastic evolution driven by the score of the static ERGM model, computed at different points in time. This approach results in a framework for describing temporal networks, more than in a single model, in the same way ERGM is considered a modeling framework for static networks.

Conceptually, applying the score-driven approach is pretty straightforward. Given the observations  $\left\{Y_{ij}^{(t)}\right\}_{t=1}^T$ , we can apply the update rule in (3) to all or some elements of  $\theta$ , each of which is associated with a network statistic in (1). To do this, we need to compute the derivative of the log-likelihood at every time step, i.e., for each adjacency matrix  $\mathbf{Y}^{(t)}$ . For the

general ERGM, the elements of the score take the form

$$\nabla_s^{(t)}(\theta) = h_s\left(\mathbf{Y}^{(t)}\right) - \frac{\partial \log \mathcal{K}(\theta)}{\partial \theta_s},$$

and the vector of tvps evolves according to (3) with  $f^{(t)}$  replaced by  $\theta^{(t)}$ . Hence, conditionally on the value of the parameters  $\theta^{(t)}$  at time  $t$  and the observed adjacency matrix  $\mathbf{Y}^{(t)}$ , the parameters at time  $t+1$  are deterministic. When used as a DGP, the SD-ERGM describes stochastic dynamics because, at each time  $t$ , the adjacency matrix is not known in advance but must be randomly sampled from  $P\left(\mathbf{Y}^{(t)}|\theta^{(t)}\right)$  and used to

compute the score. When the sequence of networks  $\left\{\mathbf{Y}^{(t)}\right\}_{t=1}^T$  is observed, the static parameters  $(w, \beta, \alpha)$ , that best fit the data, can be computed via MLE. Taking into account that each network  $\mathbf{Y}^{(t)}$  is independent of all the others *conditionally* on the value of  $\theta^{(t)}$ , the log-likelihood can be written as

$$\log P\left(\left\{\mathbf{Y}^{(t)}\right\}_{t=1}^T | w, \beta, \alpha\right) = \sum_{t=1}^T \log P\left(\mathbf{Y}^{(t)} | \theta^{(t)}\left(w, \beta, \alpha, \left\{\mathbf{Y}^{(t')}\right\}_{t'=1}^{t-1}\right)\right). \quad (4)$$

The computation of the normalizing factor and its derivative with respect to the parameters is essential for the SD-ERGM. Not only does it enter the definition of the update, but it is also required to optimize (4).

Our primary motivation for introducing the SD-ERGM is to describe the time evolution of a sequence of networks using the evolution of the parameters of an ERGM. From the context or previous studies of static networks in terms of ERGM, we assume we know which statistics are more appropriate in describing a given network. Hence, we do not discuss the choice of statistics in the context of temporal networks but refer the reader to Refs. 42–44 for examples of feature selection and goodness-of-fit evaluation.

In this final paragraph, we anticipate the SD extensions of ERGMs with given statistics detailed in the following pages. The first example allows for the exact computation of the likelihood, but the number of parameters can become significant for a large network. The second example discusses how an SD-ERGM can be defined when the log-likelihood is not known in closed form. Using extensive numerical simulations, we show that SD-ERGMs are very efficient at recovering the paths of tvps when the DGP is known, and the score-driven model is employed as a misspecified filter. Moreover, we show the first application of the Lagrange Multiplier (LM) test<sup>45</sup> in assessing the time-variation of ERGM parameters.

#### A. Score-Driven Beta Model

Our first specific example is the Score-Driven version of the *beta model*, introduced in Sec. II A and further discussed in Appendix A. We start with this model because of its wide applications and relevance in various streams of literature and because the likelihood of the ERGM and its score can be computed exactly. Moreover, the number of local statistics, the

degrees, and parameters can become very large for large networks. Since we must describe the dynamics of many parameters, this last feature challenges any time-varying parameter version of the beta model. At the end of this Section, we will show how the SD framework allows for a parsimonious description of such high-dimensional dynamics.

As anticipated, the SD-beta model is defined by applying (3) to each of the  $\overrightarrow{\theta}$  and  $\overleftarrow{\theta}$  parameters. Among the possible choices, we use as scaling the diagonal matrix  $S_{ij}^{(t)} = \delta_{ij} I_{ij}^{(t)-1/2}$ , where  $I^{(t)} = \mathbb{E}[\nabla^{(t)} \nabla^{(t)'}]$ , i.e., we scale each element of the score by the square root of its variance. It is widespread, in score-driven models with numerous tvtps, to restrict the matrices  $\alpha$  and  $\beta$  of (3) to be diagonal. In this work, we consider a version of the score update having only three static parameters ( $w_s, \beta_s, \alpha_s$ ) for each dynamical parameter  $\theta_s$ . The resulting update rule for the beta model is

$$\begin{aligned} \overleftarrow{\theta}_s^{(t+1)} &= w_s^{\text{in}} + \beta_s^{\text{in}} \overleftarrow{\theta}_s^{(t)} + \alpha_s^{\text{in}} \frac{\sum_i (Y_{is}^{(t)} - p_{is}^{(t)})}{\sqrt{\sum_i p_{is}^{(t)} (1 - p_{is}^{(t)})}} \\ \overrightarrow{\theta}_s^{(t+1)} &= w_s^{\text{out}} + \beta_s^{\text{out}} \overrightarrow{\theta}_s^{(t)} + \alpha_s^{\text{out}} \frac{\sum_i (Y_{si}^{(t)} - p_{si}^{(t)})}{\sqrt{\sum_i p_{si}^{(t)} (1 - p_{si}^{(t)})}}, \end{aligned} \quad (5)$$

where the superscripts *in* and *out* indicate the first and second half of the parameter vectors, respectively. To simplify the inference procedure, we consider a two-step approach. First, we fix the node-specific parameters  $w_i$  to target the unconditional means of  $\overleftarrow{\theta}$  and  $\overrightarrow{\theta}$  resulting from an ERGM with static parameters. Conditionally on the target values, we estimate the remaining parameters  $\alpha^{\text{in}}$ ,  $\alpha^{\text{out}}$ ,  $\beta^{\text{in}}$ , and  $\beta^{\text{out}}$ . We verified that the bias introduced by the two-step procedure is negligible, and results remain similar when the joint estimation is performed.

### 1. SD-ERGMs as filters: Numerical Simulations

As mentioned in the Introduction, SD-ERGMs (as other observation-driven models, e.g., GARCH) can be seen as DGPs or predictive filters<sup>46</sup> since tvtps follow one-step-ahead predictable processes. In this Section, we show the ability of the ERGMs in the latter setting. Specifically, we simulate generic non-stationary evolution for temporal network parameters  $\theta^{(t)}$ . We then use the SD-ERGM to filter the paths of the parameters and evaluate its performances. It is important to note that the parameters' simulated dynamics differ from the score-driven ones specified for the estimation.

In practice, at each time  $t$ , we sample the adjacency matrix from the PMF of an ERGM with parameters<sup>47</sup>  $\overline{\theta}^{(t)}$ , evolving according to known temporal patterns that define different DGPs. We then use the realizations of the sampled adjacency matrices to filter the patterns. We consider a sequence of  $T = 250$  time steps for a network of 10 nodes, each with parameters  $\overleftarrow{\theta}_i^{(t)}$  and  $\overrightarrow{\theta}_i^{(t)}$  evolving with predetermined patterns. We test four different DGPs. The first one is a naive case with constant parameters  $\overline{\theta}^{(t)} = \overline{\theta}_0$ . The elements of  $\overline{\theta}_0$

TABLE I. RMSEs (on a percentage base) of the filtered paths averaged over all tvtps and all Monte Carlo replicas of the numerical experiment. Left column: results from the cross-sectional estimates of the beta model; right column: score-driven beta model results. Each row corresponds to one of the four DGPs.

DGPs	Average RMSE	
	beta model	SD-beta model
Const	1.75	0.20
Sin	2.76	0.34
Steps	2.46	0.28
AR(1)	1.82	0.24

are chosen to ensure heterogeneity in the expected degrees of the nodes under the static beta model. For the remaining three DGPs, half of the parameters are static, and half are time-varying, evolving with either a deterministic sinusoidal function, a deterministic step function, or a stochastic AR(1) dynamics. More details on the definition of such DGPs are given in the Appendix C.

In the following, we benchmark the performance of the SD-ERGMs with that of a sequence of cross-sectional estimates of static ERGMs, i.e., one ERGM estimated for each  $t$ . We quantify the performance of the two approaches computing the Root Mean Square Error  $\frac{1}{T} \sqrt{\sum_t (\overline{\theta}_s^{(t)} - \hat{\theta}_s^{(t)})^2}$ , that describes the distance between the known simulated path and the filtered. We then average the RMSE across all the tvtps and 100 simulations and report the results in Table I. These results confirm that the SD beta model outperforms the standard beta model in recovering the true time-varying pattern. Notably, this holds even when the DGP is inherently nonstationary, as in the case of the DGP, where each parameter has a step-like evolution. Indeed, the results of this Section and Section III B confirm that, while the SD update rule (3) defines a stationary DGP<sup>34</sup>, using SD models as filters, we can effectively recover nonstationary parameters' dynamics. Our last numerical simulations for the SD beta model explore its applicability and performance for networks of increasing size. We explore this setting for two reasons. The first one is that networks with a large number of nodes describe many real systems. The second reason is that we want to compare the performance of our approach with that of the standard beta model in regimes where the latter is known to perform better under suitable conditions. Indeed, as mentioned in Appendix A, asymptotic results on the single observation estimates<sup>27</sup> guarantee that, if the network density remains constant as  $N$  grows, the accuracy of the cross-sectional estimates increases. We want to check numerically that, within the regime of dense networks, the accuracy of the static and SD versions of the beta model reaches the same level. To check whether the SD approach provides any advantage for large networks, we perform numerical experiments similar to the previous ones but in a different and more realistic regime of sparse networks, i.e., keeping the average degree constant. Moreover, to ease the computational

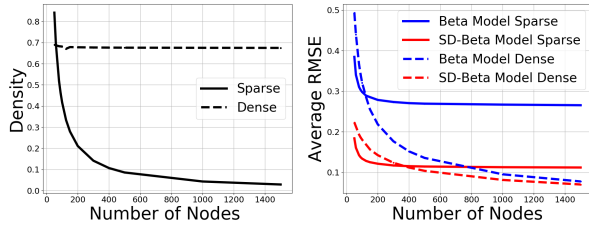


FIG. 1. Left panel: average density as a function of nodes  $N$  in the dense (dashed line) and sparse (solid line) regimes. Right panel: average RMSE of the filtered parameters with respect to the simulated DGP in both the dense (dashed lines) and sparse (solid lines) regimes. The average RMSE from the ERGM is plotted in blue, while the one from the SD-ERGM is in red.

burden for the estimates, we consider a restricted version of the SD-Beta model, as detailed in Appendix C 1, having only one set of parameters  $(\beta^{\text{in}}, \beta^{\text{out}}, \alpha^{\text{in}}, \alpha^{\text{out}})$  for the whole network, instead of one set per each node.

This analysis considers only one dynamical DGP and many different values of  $N$ . Among the DGPs used above, we focus on the one with smooth and periodic time variation. Most importantly, we set a maximum degree attainable for a node and let it depend on  $N$  in two distinct ways, each corresponding to a different density regime: one generating *sparse* networks and the other *dense* ones. It is worth noticing that the asymptotic results of Ref. 27 are expected to hold only in the dense case. The average densities for different values of  $N$  in the two regimes are shown in the left panel of Figure 1. Then, for both regimes and each value of  $N$ , we compute the average RMSE across all tvtps and all Monte Carlo replicas. In the right panel of Figure 1, the average RMSEs for different values of  $N$  indicate that, also for large networks, the SD version of the beta model attains better results compared with the cross-sectional estimates. As expected, both approaches reach the same accuracy in the dense network regimes as long as  $N$  becomes larger. However, in the more realistic sparse regime, the performance of the SD-ERGM remains much superior for both small and large network dimensions.

## B. Pseudo-Likelihood SD-ERGM

As mentioned earlier, the dependence of the normalizing function on the  $\theta$  parameters is often unknown. This fact prevents us from computing the score function and directly applying the update rule (3) to a large class of ERGMs. To circumvent this obstacle, instead of the unattainable score of the exact likelihood, we propose to use the score of the pseudo-likelihood, discussed in Sec. II A, that we refer to as pseudo-score

$$\frac{\partial \log PL(\mathbf{Y}^{(t)}|\theta)}{\partial \theta_s^{(t)'}} = \sum_{ij} \delta_{ij}^s \left( Y_{ij}^{(t)} - \frac{1}{1 + e^{-\sum_l \theta_l \delta_{ij}^l}} \right), \quad (6)$$

in place of the exact score in the definition of the SD-ERGM update (3). Additionally, we use the pseudo-likelihood for each observation  $\mathbf{Y}^{(t)}$  in (4) to infer the static parameters.

Our approach, based on the score of the pseudo-likelihood, requires as input the change statistics for each function  $h_s(\mathbf{Y}^{(t)})$ <sup>48</sup>. In the following, we show that the update based on the pseudo-likelihood score effectively filters the path of tvtps. Remarkably, this is true even when the probability distribution in the DGP is exact, i.e., when we sample from the exact likelihood and then use the SD-ERGM based on the pseudo-likelihood to filter.

### 1. SD-ERGM for Transitivity and Network Density

In this section, we discuss numerical simulations for an ERGM whose normalization is not known in closed form, which we also apply to real data in Section IV B. We show the concrete applicability of the SD-ERGM approach based on the pseudo-score and its performance as a filter compared with the cross-sectional MCMC estimates of the standard ERGM. The models we consider have two statistics. The first one is the total number of links present in the network. The second statistic is the GWESP, introduced in Section II A. The ERGM is thus defined by

$$\sum_s \theta_s h_s(\mathbf{Y}^{(t)}) = \theta_1 \sum_{ij} Y_{ij}^{(t)} + \theta_2 \text{GWESP}(\mathbf{Y}^{(t)}). \quad (7)$$

To test the efficiency of the SD-ERGM, we simulate a known temporal evolution for the parameters and, at each time step, we sample the exact PMF from the resulting ERGMs. Finally, we use the observed change statistics for each time step to estimate two alternative models: a sequence of cross-sectional ERGMs and the SD-ERGM. In what follows, we indicate the values from the DGP of parameter  $s$  at time  $t$  as  $\bar{\theta}_s^{(t)}$ .

We investigate four DGPs similar to those analyzed in Section III A. We sample and estimate the models 50 times for each DGP. Figure 2 compares the cross-sectional estimates and the score-driven filtered paths. Table II reports the RMSE of the GWESP tvtps, averaged over the different realizations for the whole sequence  $t = 1, 2, \dots, T$ . The SD-ERGM outperforms the cross-sectional ERGM estimates for all the investigated time-varying patterns. Moreover, when the constant DGP is considered, i.e.,  $\bar{\theta}_1^{(t)} = \bar{\theta}_1$  and  $\bar{\theta}_2^{(t)} = \bar{\theta}_2$ , the average RMSE of the SD-ERGM is larger but comparable, than the correctly specified ERGM that uses all the longitudinal observations to estimate the parameters. The latter result confirms that the SD-ERGM is a reliable and consistent choice even for the static case. It is worth noticing that, for sampling and cross-sectional inference, we employed the R package *ergm* that uses state-of-the-art MCMC techniques for both tasks<sup>49</sup> (for a description of the software). Hence, we compared the SD-ERGM based on the approximate pseudo-likelihood – both in the definition of the time-varying parameter update and inference of the static parameters – with a sequence of exact cross-sectional estimates that are in general known to be better performing than the pseudo-likelihood alternative, as mentioned in Section II A. Even if the cross-sectional estimates are based on the exact likelihood, while the SD approach is based on an approximation, the SD-ERGM remains

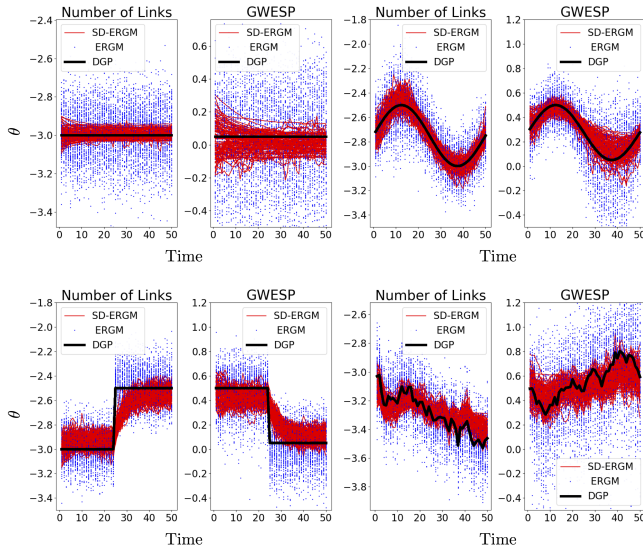


FIG. 2. Filtered paths of the parameters of the ERGM in (7) with tvps. The path from the true DGP is in black. The blue dots are the cross-sectional ERGM estimates, and the red lines are the SD-ERGM filtered paths.

TABLE II. First four columns: RMSEs for the filtered paths of the tvps, averaged over 50 repetitions, for the evolutions of Figure 2. The last three columns describe the accuracy of the test for dynamics in the parameters, considering the DGPs in Figure 2, as well as alternative DGPs where only one parameter is time-varying. We report the percentage of times that the LM test correctly identifies the parameter as time-varying (or static in the case of the first DGP). The chosen threshold for the p-values is 0.05.

DGP	Average RMSE				LM Test	
	ERGM		SD-ERGM		% Correct Results	
	$\theta_1^{(t)}$	$\theta_2^{(t)}$	$\theta_1^{(t)}$	$\theta_2^{(t)}$	$(\theta_1^{(t)}, \theta_2)$	$(\theta_1, \theta_2^{(t)})$
Const	0.02	0.1	0.0006	0.004		
Sin	0.02	0.04	0.003	0.005	94%	93%
Steps	0.02	0.03	0.01	0.001	92%	96%
AR(1)	0.02	0.2	0.007	0.01	93%	90%

the best-performing solution. This provides further evidence of the advantages of SD-ERGM as a filtering tool. Finally, the last column of Table II reports the percentage number of times the LM test of<sup>45</sup> applied to the SD-ERGM correctly classifies the parameters as time-varying (or static for the constant DGP). The test performs correctly in all the cases considered.

## 2. Comparison of Pseudo and Exact Likelihood SD-ERGM

To further investigate the proposed SD-ERGM and its version based on the pseudo-likelihood, in this section, we focus on the ERGM having the total number of links and the total

number of mutual links as network statistics:

$$\sum_s \theta_s h_s(\mathbf{Y}^{(t)}) = \theta_L \sum_{ij} Y_{ij}^{(t)} + \theta_M \sum_{ij} Y_{ij}^{(t)} Y_{ji}^{(t)}. \quad (8)$$

The static version of this model is known as *reciprocity  $p^*$  model*<sup>50</sup>. This model is relevant for our discussion because it allows us to compare the SD time-varying extension based on the pseudo-likelihood with the one based on the exact likelihood. Indeed, it is simple enough that the normalizing function is known in closed form, but it has enough structure that its pseudo-likelihood differs from its exact likelihood. The model results in dyads, i.e., pairs of mutual links  $(A_{ij}, A_{ji})$ , being independent, while the pseudo-likelihood amounts to assuming independent links. Moreover, since its partition function is available in closed form, such a model can be sampled efficiently without resorting to MCMC methods. This allows us to run extensive numerical simulations in reasonable time to investigate the properties of the confidence bands proposed by Ref. 51 in the context of SD-ERGM models.

In this section, we will refer to the pseudo-likelihood-based SD-ERGM as PML-SD-ERGM and to the exact likelihood case as ML-SD-ERGM. We compare the capacity of the two models, used as filters, to recover misspecified dynamics using the same approach as in the previous sections, i.e., we simulate a known DGP for  $\theta_L^{(t)}$  and  $\theta_M^{(t)}$ . We focus on a DGP where  $\theta_L$  and  $\theta_M$  follow two independent  $AR(1)$  processes, as the one discussed in III A. Each  $AR(1)$  has  $\Phi_1 = 0.98$  and  $\varepsilon \sim N(0, \sigma)$  with  $\sigma = 0.005$ . The  $\Phi_0$  parameters are chosen such that, on average, the network density equals 0.3, and the fraction of reciprocated pairs is 0.075. We select this value because it is between the maximum and minimum fraction of reciprocated links possible for a network of density 0.3, 0 and  $0.3(N^2 - N)/2$ , respectively. When comparing results for different network sizes, we keep the density fixed for all network sizes  $N$ , thus exploring a dense regime<sup>52</sup>. In our numerical experiment, we first sample sequences of synthetically generated observations repeatedly from different specifications of the DGP. We then estimate the PML and ML versions of the SD-ERGM on those observations and filter the tvps. Finally, we quantify their accuracy, with the average RMSE, across 50 samples with respect to the simulated DGP. In Table III, we report the RMSE for both PML-SD-ERGM and ML-SD-ERGM, divided by the RMSE of the cross-sectional standard ERGM, for various combinations of network size  $N$  and number of observations  $T$ . It emerges that both versions of SD-ERGM strongly outperform the cross-sectional ERGM. Moreover, the performances of PML-SD-ERGM are similar to the ones of the exact ML-SD-ERGM.

In the final part of this section, we investigate the possibility of using the method of Ref. 51, that we describe in Appendix B, to define confidence bands for the parameters filtered with SD-ERGM. The authors characterize the approximation error when the SD approach filters a set of latent parameters whose true DGP is an auto-regressive process. While we refer to the original manuscript for the details, we point out that their procedure rests upon the assumption that the SD filter approximates the true underlying DGP. The authors prove that this approximation becomes exact in the limit of

TABLE III. RMSEs of ML-SD-ERGM and PML-SD-ERGM, relative to that of the cross-sectional ERGM. The averages are obtained over 50 repetitions for the  $AR(1)$  DGP described in the text. For each value of  $T$  and  $N$ , we report the RMSE of the SD-ERGMs divided by the RMSE of the cross-sectional ERGM.

T \ N	PML-SD-ERGM			ML-SD-ERGM		
	50	100	500	50	100	500
100	0.016	0.011	0.006	0.015	0.011	0.006
300	0.015	0.011	0.006	0.014	0.011	0.007
600	0.014	0.012	0.007	0.014	0.011	0.006

TABLE IV. Coverages of the 95% confidence bands averaged over 50 repetitions, for the  $AR(1)$  DGP, described in the text, and  $N = 100$ .

T	ML-SD-ERGM	PML-SD-ERGM
300	99.1 %	99.9 %
3000	94.5%	95.7 %

small variance for the latent parameters. Hence, the confidence bands obtained with their method are theoretically guaranteed to be reliable only in this limit. In practice, assessing whether the application of the confidence bands is justified for a given value of the variance of the DGP is appropriate. Numerical experiments can do this to determine their coverage with a simulated DGP. For example, for the model and the DGP considered in this section, we check the coverage of the confidence bands obtained and report the results in Table IV, for  $N = 100$ . We find that the coverage of the confidence bands, for both ML-SD-ERGM and PML-SD-ERGM, approaches the nominal value in the limit of large  $T$ , while for short time series, their coverages are higher than the nominal value. Hence, in small samples, they should be interpreted as having a confidence of *at least* their nominal values.

#### IV. APPLICATIONS TO REAL DATA

After analyzing synthetic data, this section presents two applications to real temporal networks. Our goal is to show the value of SD-ERGM as a methodology to model temporal networks, irrespective of the specific system that a researcher wants to investigate. The two real networks that we consider have been the object of multiple studies in different streams of literature. They have been investigated in the context of ERGMs using different network statistics. We first consider a network of credit relations among Italian banks. The second real-world application focuses on a network of interest for the social and political science community, namely the network of U.S. senators cosponsoring legislative bills.

#### A. Link Prediction in Interbank Networks

Our first empirical application is to data from the electronic Market of Interbank Deposit (e-MID). In this market, banks can extend loans to one another for a specified term and/or collateral. Interbank markets are an important point of encounter for banks' supply and demand of extra liquidity. In particular, e-MID has been investigated in many papers<sup>16,53–55</sup>. Our dataset contains the list of all credit transactions on each day from June 6, 2009, to February 27, 2015. Our analysis investigates the interbank network of overnight loans aggregated weekly. We follow the literature and disregard the size of the exposures, i.e., the weights of the links. We thus consider a link from bank  $j$  to bank  $i$  present at week  $t$  if bank  $j$  lent money overnight to bank  $i$ , at least once during that week, irrespective of the amount lent. This results in a set of  $T = 298$  weekly aggregated networks. For a detailed dataset description, we refer the reader to Ref. 55.

In recent years, the amount of lending in e-MID has significantly declined. In particular, it abruptly decreased at the beginning of 2012 due to important unconventional measures (Long Term Refinancing Operations) by the European Central Bank that guaranteed an alternative source of liquidity to European banks. The evident non-stationary nature of the evolution of the interbank network is of extreme interest to us. As mentioned in Sections III A and III B, one of the key strengths of SD-ERGM, used as a filter, is precisely the ability to recover such non-stationary dynamics.

In the following, we use the SD beta model for link forecasting. Specifically, we consider the version with a restricted number of static parameters discussed at the end of Sec. III A 1. We divide the data set into two samples. We consider rolling windows of 100 observations and estimate the parameters  $\alpha^{\text{out}}$ ,  $\beta^{\text{out}}$ ,  $\alpha^{\text{in}}$  and  $\beta^{\text{in}}$  on each one of those rolling windows. For each window, we test the forecasting performances up to 8 steps ahead (i.e., roughly two months). The forecast works as follows. Assuming that at time  $t$ , the last date of the rolling window, we have filtered the value for the parameters  $\overleftarrow{\theta}^{(t)}$  and  $\overrightarrow{\theta}^{(t)}$ , we plug the estimated static parameters and the matrix  $\mathbf{Y}^{(t)}$  in the SD update and compute the tvps  $\overleftarrow{\theta}^{(t+1)}$  and  $\overrightarrow{\theta}^{(t+1)}$ . From the latter, we readily obtain the forecast of the adjacency matrix

$$\mathbb{E} \left[ \mathbf{Y}^{(t+1)} \mid \overleftarrow{\theta}^{(t+1)}, \overrightarrow{\theta}^{(t+1)} \right],$$

where  $t + 1$  is the first date of the test sample. The  $K$ -step-ahead forecast for the SD-ERGM model is obtained by simulating the SD dynamics up to  $t + K$  100 times<sup>56</sup>, thus obtaining  $\overrightarrow{\theta}_n^{(t+K)}$  and  $\overleftarrow{\theta}_n^{(t+K)}$  for  $n = 1, \dots, 100$ , and then taking the average of the expected adjacency matrices  $\frac{1}{100} \sum_n \mathbb{E} \left[ \mathbf{Y}^{(t+K)} \mid \overleftarrow{\theta}_n^{(t+K)}, \overrightarrow{\theta}_n^{(t+K)} \right]$ . Given the forecast values, we compute the rate of false positives and false negatives. Then, we drop the first element from the train set and add the first element of the test sample. We repeat the forecasting exercise, estimating the SD-ERGM parameters on the new train

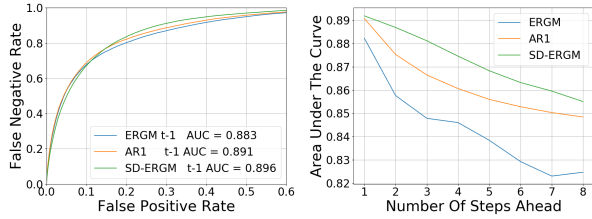


FIG. 3. Left panel: ROC curves for one-step-ahead link forecasting. The green and orange ROC curves describe the one-step-ahead forecasting with SD and cross-sectional AR(1) beta model, respectively. The blue curve corresponds to the forecast based on the ERGM at the previous time step. Right panel: AUC for the multi-step-ahead forecast.

set and testing the performance of the new test sample. We name this procedure a rolling estimate and iterate it until the test sample contains the last eight elements of the time series.

Given a forecast for the adjacency matrix, we evaluate the accuracy of the binary classifier by computing the Receiving Operating Characteristic (ROC) curve. All results are collected and presented in Fig. 3. The left panel reports the ROC curve for one-step-ahead link forecasting obtained according to the SD-ERGM rolling estimate. The panel also shows three other curves based on the static beta model. Specifically, the green curve results from a naive prediction, where a link tomorrow is forecasted, assuming that the  $t + 1$  ERGM parameter values are equal to those estimated at time  $t$ . Once the sequence of cross-sectional estimates of the static ERGM is completed, we take the estimated values  $\hat{\theta}^{(t)}$  and  $\hat{\theta}^{(t)}$  as observed and model their evolution with an autoregressive model of order one, AR(1). That amounts to assuming  $\hat{\theta}^{(t+1)} = c_0 + c_1 \hat{\theta}^{(t)} + \varepsilon^{(t)}$ , where  $c_0$  and  $c_1$  are the static parameters of the AR(1), and  $\varepsilon^{(t)}$  is a sequence of i.i.d. normal random variables with zero mean and variance  $\sigma^2$ . A similar equation holds for the out-degree parameters. Using the observations from the training sample, we estimate the parameters  $c_0$ ,  $c_1$ , and  $\sigma^2$  and use them for a standard AR(1) forecasting exercise on the test sample. The results correspond to the orange curve. It is important to stress that while the SD-ERGM forecast requires one static and one time-varying estimation on the train set, we must estimate the static parameters for each date in the train sample in the latter procedure.

The left plot of Fig. 3 shows that the naive one-step-ahead forecast, despite its simplicity, provides a reasonable result. The best performance corresponds, however, to the forecast based on the SD-ERGM. The AR(1) static ERGM improves on the naive forecast and is slightly worse than the SD-ERGM. However, as commented before, it is more computationally intensive. More importantly, the right panel of Fig. 3 presents a multi-step-ahead forecasting analysis result. It emerges that the naive forecast’s performance (blue curve), tested up to  $K = 8$ , rapidly deteriorates. In contrast, the SD-ERGM multi-step forecast remains the best performing<sup>57</sup>.

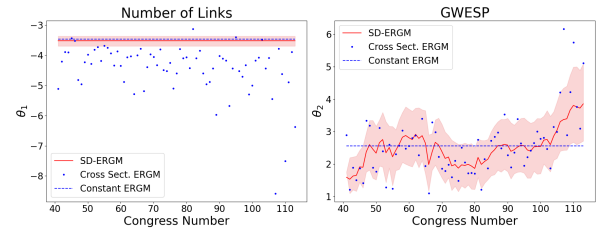


FIG. 4. Estimates for the tvtps associated with the number of links and the GWESP statistics. Blue dots correspond to the cross-sectional ERGM estimates. The red lines are the estimates from the SD-ERGM, with the corresponding 95% confidence intervals denoted by the red-shaded regions.

## B. Temporal Heterogeneity in U.S. Congress Co-Voting Political Network

Networks describing the U.S. Congress’ bills have been the object of multiple studies<sup>12,14,58–62</sup>. It is thus an appropriate real system for our second application of the SD-ERGM framework. In particular, we want to show that the update rule based on the pseudo-score defined in (6) can be concretely applied to a real network and draws a different picture when compared to the sequence of cross-sectional ERGM estimates. To build the network, we use the freely available data of voting records in the US Senate<sup>63</sup> covering the period from 1867 to 2015, for a total of 74 Congresses. We define the network of co-voting following Refs. 64 and 14, where a link between two senators indicates that they voted in agreement on over 75% of the votes among those held in a given senate when they were both present. This procedure results in 74 networks, one for each Congress, starting from the 40th. We consider the SD-ERGM with the two network statistics discussed in Section III B for this empirical application. As defined in (7), parameter  $\theta_1^{(t)}$  is associated with the number of edges, while  $\theta_2^{(t)}$  with the GWESP statistic. The fact that the number of nodes is not constant over time is not a problem for our application since we do not consider statistics associated with single nodes. That case – as, for instance, considering the degrees of the beta model – would require the number of tvtps to be different at each time step.

As we did for the numerical simulations and the previous empirical application, we compare our framework with a sequence of standard ERGMs. This empirical exercise does not aim to conclude the specific network at hand. We aim to show that the two approaches return a qualitatively different picture. The choice between the alternative models and combinations of statistics—possibly based on model selection techniques—is beyond the scope of our exercise.

Using the test for temporal heterogeneity based on SD-ERGM, only the parameter  $\theta_2$  turns out to be time-varying. Testing the null hypothesis that each parameter is static, we obtain a p-value of 0.1 for the link density and  $10^{-4}$  for GWESP. To check whether the sequence of cross-sectional estimates is consistent with the hypothesis that the parameters remain constant, we estimate the values  $\theta_1^c$ ,  $\theta_2^c$  from



an ERGM using all observations. This amounts to compute  $\theta^c = \arg \max_{\theta} \sum_{t=1}^{74} \log P(\mathbf{Y}^{(t)}, \theta)$ . Then, for each sequence of cross-sectional estimates  $\theta_1^{(t)}$  and  $\theta_2^{(t)}$ , we test the hypothesis of them being normally distributed around the constant values with unknown variance. The p-values resulting from the  $t$ -tests are  $1.4 \times 10^{-6}$  and 0.03 for parameters  $\theta_1$  and  $\theta_2$ , respectively. This simple test confirms that the two approaches imply quantitatively different parameter behaviors. This emerges from Fig. 4 that reports the estimates from the SD-ERGM (thick red lines), with their respective 95% confidence intervals (shaded red bands), as well as the cross-sectional ERGM estimates – one per date (blue dots) or using the entire sample (dashed blue line).

To compute the confidence bands as in Ref. 51, we numerically checked whether the data is compatible with a DGP with a small variance. In practice, we first estimate the SD-ERGM. Then we quantify the variance of the latent parameters by estimating an AR(1) on the filtered time series<sup>65</sup>. Finally, we repeatedly simulate such an AR(1) DGP, similarly to what is done at the end of section III B 2, and check the coverage of the confidence bands. We find that, for the current application, the coverage of the confidence bands is 99.9%, hence larger than the nominal value. These simulation-based results support the reliability of the approximate SD filter and provide a conservative estimate of the confidence bands. This allows us, for example, to deduce that the data is not compatible with a model where one of the two parameters is zero.

## V. CONCLUSIONS

In this paper, we proposed a framework for describing temporal networks that extends the well-known Exponential Random Graph Models. In the new approach, the parameters of the ERGM have stochastic dynamics driven by the conditional likelihood score. If the latter is unavailable in closed form, we showed how to adapt the score-driven updating rule to a generic ERGM by resorting to the conditional pseudo-likelihood. In this way, our approach can describe the dynamic dependence of the PMF from virtually all the network statistics usually considered in ERGM applications. We investigated two specific ERGM instances using an extensive Monte Carlo analysis of the SD-ERGM reliability as a filter for tvps. The chosen examples allowed us to highlight the applicability of our method to models with a large number of parameters and to models for which the normalization of the PMF is not available in closed form. The numerical simulations proved the clear superior performance of the SD-ERGM over a sequence of standard cross-sectional ERGM estimates. This is true not only in the sparse network regime but also in the dense case when the number of nodes is far from the asymptotic limit. Finally, we run two empirical exercises on real network data. The first application to the e-MID inter-bank network showed that the SD-ERGM provides a quantifiable advantage in a link forecasting exercise over different time horizons. The second example of the U.S. Congress co-voting political network enlightened that the ERGM and the

SD-ERGM could provide a significantly different picture describing the parameter dynamics.

Our work opens several possibilities for future research. First, the applicability of the test for parameter instability in the context of SD-ERGM with multiple network statistics could be investigated much further. This would require an in-depth analysis of the multi-collinearity issues intrinsic to the ERGM context. Second, the SD-ERGM could be applied to multiple instances of real-world temporal networks. An interesting application would be the study of networks describing the dynamical correlation of neural activity in different parts of the brain<sup>66</sup>. In this context, applying the static ERGM has already proven highly successful<sup>67</sup>. The last future development we plan to explore is extending the score-driven framework to the description of weighted temporal networks. Regrettably, this setting deserves more attention in the literature<sup>68</sup>. Still, it is of extreme relevance, particularly from the financial stability perspective and its implications for systemic risk.

## ACKNOWLEDGMENTS

We are particularly thankful for the comments and suggestions received by Fulvio Corsi and Giuseppe Buccheri. Fabrizio Lillo acknowledges partial support by the European Program scheme 'INFRAIA-01-2018- 2019: Research and Innovation action', grant agreement #871042 'SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics.' Giacomo Borgetti and Fabrizio Lillo acknowledge financial support from the Italian Ministry MUR under the PRIN project *Dynamic models for a fast changing world: An observation driven approach to time-varying parameters* (grant agreement no. 20205J2WZ4).

## AUTHOR DECLARATIONS

### Conflict of Interest

The authors have no conflicts to disclose.

## DATA AVAILABILITY STATEMENT

We are not allowed to share the data from the e-MID inter-bank market, as a confidentiality agreement between the Scuola Normale Superiore of Pisa and the data provider, LIST SpA, binds us. The interested reader can contact LIST SpA for inquiries concerning data access. The second application's data are publicly available at <https://voteview.com/data>.

## Appendix A: Exponential Random Graphs Models

Here, we review some standard notions regarding ERGMs for the reader's convenience. A PMF over the set of graphs

as in eq. (1) defines an ERGM. To begin with, let us mention the first and probably most famous example of this class: the Erdős-Rényi model<sup>69</sup>. In this model, for a given number of nodes  $N$ , each of the possible  $N(N-1)/2$  links<sup>70</sup> is present with constant probability  $p$ , equal for all links. The probability to observe the adjacency matrix  $\mathbf{Y}$  is

$$P(\mathbf{Y}) = \prod_{i < j} p^{Y_{ij}} (1-p)^{(1-Y_{ij})}.$$

In the context of exponential distributions, it is possible to consider more general structures for the probability of a link to be present and even depart from the assumption that each link is independent of the others. Examples of more general ERGMs have been first proposed in Ref. 29, under the name of log-linear, or  $p^*$ , models. For instance, the  $p1$  model is defined by the PMF

$$\log P(\mathbf{Y}) = \sum_{ij} [Y_{ij} Y_{ji} \rho_{ij} + Y_{ji} \phi_{ij}] - \log(\mathcal{K}(\boldsymbol{\rho}, \boldsymbol{\phi})),$$

where  $\boldsymbol{\rho}$  and  $\boldsymbol{\phi}$  are two matrices of parameters, and  $\mathcal{K}(\boldsymbol{\rho}, \boldsymbol{\phi})$  is a normalization factor, also known as partition function in the statistical physics literature. This model can be estimated in parsimonious specifications, e.g.,  $\phi_{ij} = \phi_i + \phi_j$ , known as *sender plus receiver effect*, and  $\rho_{ij} = \rho$  that describes the tendency to reciprocate links. Additionally,  $p1$  models can be enriched with dependencies on node attributes<sup>71</sup> or predetermined (exogenous or endogenous) covariates  $X_{ij}$ <sup>72</sup>. The requirement of independence among dyads has been relaxed since Ref. 73 to take into account neighborhood effects, such as the tendency to form 2 stars, quantified by the function  $h_{2\text{-stars}} = \sum_{ijk} Y_{ik} Y_{jk}$  or triangles  $h_{\text{triangles}} = \sum_{ijk} Y_{ik} Y_{kj} Y_{ji}$ . These functions are examples of network statistics, i.e., functions of the adjacency matrix, that play a central role in ERGMs.

## 1. The Beta Model

The first example we consider is quite simple but, at the same time, largely employed in different streams of literature<sup>25-30</sup>. The range of applications for this model is so broad that researchers were often not aware of previous works using the same model. For this reason, it can be found under at least three different names: *beta model*, *fitness model*, and *configuration model*. They all refer to a probability distribution that can be rewritten as an ERGM where each node  $i$  has two parameters:  $\overleftarrow{\theta}_i$ , that captures the propensity of node  $i$  to form outgoing connections and  $\overrightarrow{\theta}_i$  those incoming. It is standard to indicate the number of connections a node has as its *degree*. For the directed network case considered here, we have – for node  $i$  – *out-degree*  $\overrightarrow{D}_i$  and *in-degree*  $\overleftarrow{D}_i$  defined as

$$\overrightarrow{D}_i = \sum_j Y_{ij}, \quad \overleftarrow{D}_i = \sum_j Y_{ji}.$$

With these definitions, and since it is possible to compute the normalization factor  $\mathcal{K}(\overleftarrow{\boldsymbol{\theta}}, \overrightarrow{\boldsymbol{\theta}})$ , the PMF reads

$$\log P(\mathbf{Y}) = \sum_{i=1}^N (\overleftarrow{\theta}_i \overleftarrow{D}_i + \overrightarrow{\theta}_i \overrightarrow{D}_i) - \sum_{ij} \log(1 + e^{\overleftarrow{\theta}_i + \overrightarrow{\theta}_j}). \quad (\text{A1})$$

In the main text, we extensively use the beta model's score to define its score-driven extension. Hence, we show its explicit expression here for the reader's convenience. Defining, for ease of notation,

$$p_{ij} = \frac{1}{1 + e^{-\overleftarrow{\theta}_i - \overrightarrow{\theta}_j}}$$

we can write the score as

$$\nabla(\overleftarrow{\boldsymbol{\theta}}, \overrightarrow{\boldsymbol{\theta}}) = \begin{pmatrix} \frac{\partial \log P(\mathbf{Y} | \overleftarrow{\boldsymbol{\theta}}, \overrightarrow{\boldsymbol{\theta}})}{\partial \overleftarrow{\boldsymbol{\theta}}} \\ \frac{\partial \log P(\mathbf{Y} | \overleftarrow{\boldsymbol{\theta}}, \overrightarrow{\boldsymbol{\theta}})}{\partial \overrightarrow{\boldsymbol{\theta}}} \end{pmatrix} = \begin{pmatrix} \sum_i (Y_{i1} - p_{i1}) \\ \vdots \\ \sum_i (Y_{iN} - p_{iN}) \\ \sum_i (Y_{1i} - p_{1i}) \\ \vdots \\ \sum_i (Y_{Ni} - p_{Ni}) \end{pmatrix}$$

The beta model is often used when the degree heterogeneity is expected to play a prominent role in explaining the presence or absence of links. It is worth noticing that the static version of the beta model, in the directed case, is not identified. Indeed, the probabilities remain unchanged after the application of the following transformation

$$\begin{cases} \overleftarrow{\boldsymbol{\theta}} \rightarrow \overleftarrow{\boldsymbol{\theta}} + c \\ \overrightarrow{\boldsymbol{\theta}} \rightarrow \overrightarrow{\boldsymbol{\theta}} - c. \end{cases}$$

The issue can be tackled<sup>74</sup> by choosing one identification restriction that eliminates the possibility of shifting all parameters by an arbitrary constant. This is essential to compare the parameter values estimated for the same network at different times. In all our investigations, both the numerical simulations and the empirical applications, we enforce the following condition:

$$\sum_i \overleftarrow{\theta}_i = \sum_i \overrightarrow{\theta}_i.$$

It is worth noticing that different choices are available, e.g.,  $\sum_i \overleftarrow{\theta}_i = 0$  or  $\overrightarrow{\theta}_i = 0$ . However, and most importantly, the results presented in the paper do not change significantly when the identification condition changes. It is also important to notice that the MLE can be performed using a fixed point algorithm, described for example in<sup>74</sup>, that reaches the optimal solution quickly. Moreover, we point out interesting results on the asymptotic behavior of the maximum likelihood estimates for  $(\overleftarrow{\boldsymbol{\theta}}, \overrightarrow{\boldsymbol{\theta}})$  when the number of nodes increases. Indeed, consistency results have been proved in Ref. 27 for the undirected case and in Ref. 74 for the directed case<sup>75,76</sup>. A

necessary condition for these results to hold is that the network density remains constant as  $N$  increases. An alternative, and often more realistic, possibility is that the average degree remains constant when  $N$  increases, implying that the density decreases as<sup>77</sup>  $1/N$ . Networks belonging to this density regime are named *sparse*. Notably, to our knowledge, no consistency results are known for large  $N$  in the sparse regime.

## 2. GWESP Statistic and Pseudo-Likelihood Estimation

Here, we give some details on an example of a network statistic for which we cannot compute the partition function and the associated inference procedures. We focus on the GWESP.

It is well known that when network statistics involve products of matrix elements<sup>78</sup>, this is often the case. This lack of analytical tractability has arguably been the main obstacle in estimating ERGMs and understanding their properties. Moreover, it is nowadays well known that, when dealing with ERGMs, the use of network statistics involving products of matrix elements, such as the number of triangles, requires some care to avoid statistical issues<sup>79,80</sup>. The main problem, with consequences on estimation, simulation, and interpretability of ERGMs, is *degeneracy*. An ERGM is degenerate if it concentrates a significant portion of its probability on a small set of configurations, typically the uninteresting graphs completely connected or void of links. When this phenomenon occurs, estimating the model becomes very hard, and often, the estimated model does not provide a meaningful description of real networks. A great effort has been dedicated to investigating this problem and characterizing degeneracy<sup>81</sup>. A family of network statistics that, while describing properties of the whole network, is not plagued by degeneracy has been proposed in Refs. 31 and 32 and discussed in Ref. 33. This family is called curved exponential random graphs, and one example of curved statistics is the GWESP. This function has recently been applied extensively to describe transitivity in social networks<sup>33</sup>. It captures the tendency of nodes to form triangles without the degeneracy issues that emerge when the direct triangle count is used as a statistic in ERGM. To get an intuition of the formula defining GWESP, let us consider two nodes connected by an edge and count the number of nodes to which they are connected, i.e., the number of neighbors they share. Let us indicate with  $\text{ESP}_k(\mathbf{Y})$  the number of edgewise shared partners, i.e., connected node pairs<sup>82</sup> that share exactly  $k$  neighbors in the network described by  $\mathbf{Y}$ . Then GWESP is defined as

$$\text{GWESP}(\mathbf{Y}) = e^\lambda \sum_{k=1}^{n-2} \left[ 1 - \left( 1 - e^{-\lambda} \right)^k \right] \text{ESP}_k(\mathbf{Y}).$$

In the following, we will be stuck to the usual approach in the literature, treating the parameter  $\lambda$  as fixed and known, i.e.,  $\lambda = 0.5$ .

As mentioned in the main text, there are two standard approaches to ERGM inference when the partition function is not available in closed form. The first possibility<sup>50</sup> is to

maximize an objective function obtained from a sufficiently large sample drawn from the PMF with an arbitrary (but close enough to the true one) parameter. As a consequence of the non-independence of the links in the general ERGM, sampling from (1) necessary relies on Markov Chain Monte Carlo (MCMC) approaches<sup>49</sup> (for a description of a popular software that implements it). The computational burden of MCMC-based estimation can be prohibitive for graphs that are large enough. For this reason, a second approximate inference procedure, known as Maximum Pseudo-Likelihood Estimation (MPLE), first proposed for ERGMs in the seminal work of Ref. 83, is often used in empirical applications. MPLE is based on optimizing the pseudo-likelihood function, defined from link-specific variables (one for each element of the adjacency matrix) named *change statistics*. Given an ERGM, the change statistic for the link between node  $j$  and  $i$ , associated with network statistic  $h_s$  is  $\delta_{ij}^s = h_s(\mathbf{Y}_{ij}^+) - h_s(\mathbf{Y}_{ij}^-)$ , where  $\mathbf{Y}_{ij}^+$  is a matrix such that  $Y_{ij}^+ = 1$  and it is equal to  $\mathbf{Y}$  in all other elements. Similarly,  $\mathbf{Y}_{ij}^-$  has  $Y_{ij}^- = 0$  and it is equal to  $\mathbf{Y}$  in all other entries. Given these definitions, the pseudo-likelihood reads

$$PL(\mathbf{Y}) = \prod_{ij} \pi_{ij}^{Y_{ij}} (1 - \pi_{ij})^{(1-Y_{ij})} \quad (\text{A2})$$

where  $\pi_{ij} = \left( 1 + e^{-\sum_s \theta_s \delta_{ij}^s} \right)^{-1}$ .

Pseudo-likelihood inference is of crucial importance when applying our methodology to any ERGM. Obtaining the pseudo-likelihood estimates is much faster than the MLE based on MCMC and easy to implement since the pseudo-likelihood boils down to the likelihood of a logistic regression. Then, it can be efficiently maximized with standard software for logistic regressions. However, the analogy with logistic regression is typically pushed too far. It has become widespread malpractice to associate with MPLEs the confidence intervals obtained from the maximum-likelihood theory for logistic regressions that are known to be theoretically unjustified, as already noted in Refs. 83 and 80, and thoroughly discussed in Ref. 84. It is nowadays common knowledge that such a naive approach to MPLE inference results in a systematic underestimation of confidence intervals' width<sup>85-87</sup>. More principled methods to estimate uncertainties of MPLEs, based on non-parametric and parametric bootstrap, have been proposed in Refs. 86 and 87, respectively. These contributions showed that the computational convenience of MPLE for ERGMs can be reconciled with a reliable estimation of statistical uncertainties.

## Appendix B: Score-Driven Models

Several are the reasons for the flexibility of a score-driven approach and its success in time-series modeling. Here, we review some key concepts mentioned in the main text that might prove useful to readers unfamiliar with the relative literature.

Score-Driven models have been introduced as Dynamic Conditional Score models by Ref. 35 and Generalized Au-

toregressive Score models by Ref. 34. Given a sequence of observations  $\{y^{(t)}\}_{t=1}^T$ , where each  $y^{(t)} \in \mathbb{R}^M$ , and a conditional probability density  $P(y^{(t)}|f^{(t)})$ , depending on a vector of tvps  $f^{(t)} \in \mathbb{R}^K$ , a Score-Driven model assumes that the time evolution of  $f^{(t)}$  is ruled by the recursive relation (3).

In practical applications, the static parameters of (3) must be estimated. As detailed in Ref. 35, the likelihood of Score-Driven models can be readily expressed in closed form using the so-called prediction error decomposition. In a univariate setting, Ref. 88 works out the required regularity conditions, ensuring the consistency and asymptotic normality for the maximum likelihood estimators of the parameter values.

There are motivations, originating in information theory, for the optimality of the score-driven updating rule. In Ref. 37, the authors consider a true and unobserved DGP  $y^{(t)} \sim P(y^{(t)}|f^{(t)})$ . They assume a given and, in general, misspecified conditional observation density  $\tilde{P}^{(t)} = \tilde{P}(\cdot | \tilde{f}^{(t)})$ , and consider the Kullback-Leibler (K-L) divergence

$$\mathcal{D}_{\mathcal{KL}}(P^{(t)}, \tilde{P}^{(t+1)}) = \int_A P(y|f^{(t)}) \log \frac{P(y|f^{(t)})}{\tilde{P}(y|\tilde{f}^{(t+1)})} dy,$$

where  $A \subseteq \mathbb{R}$ . Building on the minimum discrimination information principle<sup>89</sup>, they argue that when the new observation  $y_t$  becomes available,  $\tilde{f}^{(t+1)}$  should ideally be such that the updated density  $\tilde{P}^{(t+1)}$  is as close as possible to the true density  $P^{(t)}$ . Given that the real DGP is unknown, an optimal update that minimizes  $\mathcal{D}_{\mathcal{KL}}$  cannot be defined in practice. For this reason,<sup>37</sup> focus on the improvements of  $\mathcal{D}_{\mathcal{KL}}$  that an updating step produces irrespectively of the true DGP. One way of quantifying the improvement for a parameter update from  $\tilde{f}^{(t)}$  to  $\tilde{f}^{(t+1)}$  is to consider the realized variation of  $\mathcal{D}_{\mathcal{KL}}$

$$\begin{aligned} \Delta_{t|t} &\equiv \mathcal{D}_{\mathcal{KL}}(P^{(t)}, \tilde{P}^{(t+1)}) - \mathcal{D}_{\mathcal{KL}}(P^{(t)}, \tilde{P}^{(t)}) \\ &= \int_A P(y|f^{(t)}) \log \frac{\tilde{P}(y|\tilde{f}^{(t)})}{\tilde{P}(y|\tilde{f}^{(t+1)})} dy. \end{aligned}$$

Based on this definition, a parameter update is realized K-L optimal when  $\Delta_{t|t} < 0$  for every  $(y^{(t)}, \tilde{f}^{(t)}, f^{(t)})$ . The authors prove that, under reasonable assumptions, the updating rule (3) based on the score of  $\tilde{P}^{(t+1)}$  is locally realized K-L optimal. For more details and alternative definitions of optimality, we direct the reader to the original work and the more recent Ref. 90. For our definition of the SD-ERGM, we want to stress that realized optimality defines a class of updates; it does not represent a single update with a unique functional form. For instance,  $\Delta_{t|t}$  defined above is specific to the chosen  $\tilde{P}$ . A different choice of  $\tilde{P}$ , e.g., one inspired by the pseudo-likelihood specification, translates into an alternative optimal choice for the update. In general, there can be infinite realized Kullback-Leibler optimal updates. We remark that from the information-theoretic perspective of Ref. 37, an update based on the pseudo-score, as we propose in the main text, is not only admissible but also realized K-L optimal, i.e., at each

step, it diminishes the K-L distance of the pseudo-PMF, which assume independence of links, from the PMF of the true and unobserved DGP.

Any filtering tool should provide an estimation of the uncertainty and confidence bands for the estimates. Ref. 91 discussed methods to quantify the uncertainty associated with the Score-Driven filters when the DGP is a Score-Driven model. Specifically, they proposed a simulation-based method to define in-sample confidence bands around the filtered tvps. Their procedure starts from the maximum likelihood estimate of the static parameters, given observations  $\{y^{(t')}\}_{t'=1}^{t-1}$ . Given the MLE estimate, the method prescribes to repeatedly sample new parameters  $(w, \beta, \alpha)_i$  from a multivariate normal, centered around the MLE estimates, and variance-covariance matrix estimated with the Huber-White estimator<sup>92,93</sup>. Then one uses each sample to filter a different sequence of tvps, from the same time series of observations, thus obtaining a sample of filtered paths  $\hat{f}_i^{(t)} = \hat{f}_i^{(t)}(w_i, \beta_i, \alpha_i, \{y^{(t')}\}_{t'=1}^{t-1})$  for  $i = 1, \dots, K$ , where  $K$  is the number of samples. Finally, each time  $t$ , one uses the obtained distribution  $\hat{f}_i^{(t)}$  to calculate the appropriate percentiles defining the confidence bands. While this construction is intuitive and easy to implement in practice, it is meant to capture only the uncertainty due to the estimation of the static parameters, often referred to as *parameter uncertainty*. Hence, the confidence bands reliably quantify uncertainty only when the DGP is score-driven. In other words, these bands do not consider what is known as *filtering uncertainty*. This is the uncertainty because, in general, we do not know the true DGP and the score-driven filter may be regarded only as an approximate filter. Recently, Ref. 51 investigated the approximation error made by applying a score-driven filter to a time-varying parameter model following a different DGP. They found that, for a class of DGPs where the parameters follow an auto-regressive process, the approximation becomes exact in the limit of a small variance of the latent parameters. Moreover, they proposed a method to define confidence bands, inspired by Ref. 94, that accounts for filtering and parameter uncertainty in Score-Driven filters. While we refer to their paper for the derivation details, here we briefly describe the key steps of the procedure. The total conditional variance of the latent parameters is decomposed as the sum of two terms. One term captures the parameter uncertainty similarly to the approach of Ref. 91. The other term captures the filtering uncertainty and can be written in terms of the static parameters  $(w, \beta, \alpha)$  and the scaling matrix  $S^{(t)}$  from (3) as  $P^{(t)} = \beta^{-1} \alpha S^{(t)}$ . In practice, the procedure consists of sampling  $(w, \beta, \alpha)_i$  and obtaining a distribution of filtered paths, as in<sup>91</sup>. Then for each time step  $t$  the variance of the latent parameters is obtained as  $\frac{1}{K} \sum_i (\hat{f}_i^{(t)} - \hat{f}^{(t)})^2 + \frac{1}{K} \sum_i \beta_i^{-1} \alpha_i S_i^{(t)}$ , where  $\hat{f}^{(t)}$  is the path filtered using the maximum likelihood estimates.

Finally, we review the main idea behind the test for temporal parameter variation of Ref. 45. The method consists of a Lagrange Multiplier (LM) test for the parameter  $\alpha$  that multiplies the score in the one-dimensional version of the recursion

(3). The null hypothesis  $H_0$  is that the parameter  $f^{(t)}$  is static, i.e.,  $\beta = \alpha = 0$ , corresponding to  $w$ . As explained in Ref. 95, the LM statistic for the hypothesis  $H_0$ , versus the alternative  $\alpha = \beta \neq 0$ , can be conveniently obtained from an auxiliary regression. To allow for a coefficient  $\beta$  different from  $\alpha$ , one can use the same arguments as in Ref. 96. As discussed in Ref. 45, the LM statistic can be written as the explained sum of squares from the regression

$$\mathbf{1} = c_w \nabla_w^{(t)} + c_\alpha \mathcal{S}^{(t-1)} \nabla_w^{(t-1)} \nabla_w^{(t)'} + \text{residual},$$

where  $c_w$  and  $c_\alpha$  are regression coefficients that can be estimated with any statistical software. It is worth noticing that, under the null, the score of the conditional density with respect to  $f^{(t)}$  is equal to the score with respect to  $w$ . From standard asymptotic theory, it follows that the LM statistic is distributed as a  $\chi^2$  with one degree of freedom. For a detailed test description, we refer the reader to Ref. 45.

### Appendix C: Details of Numerical Simulations

The main text refers to a set of DGPs used for numerical simulations. Although the different numerical experiments that we presented differ in the meaning and number of parameters, in every experiment, each of the parameters can be constant or evolve according to one of the following dynamical DGPs:

- abrupt change of half the parameters at  $t = T/2$ , i.e., for odd  $s$  we have  $\bar{\theta}_s^{(t)} = \bar{\theta}_{1s}$  for  $t \leq T/2$  and  $\bar{\theta}_s^{(t)} = \bar{\theta}_{2s}$  for  $t > T/2$ , while for even  $s$  it is  $\bar{\theta}_s^{(t)} = \bar{\theta}_{0s}$  for  $t = 1 \dots T$ ;
- smooth periodic variation for half the parameters, i.e., for odd  $s$  we have  $\bar{\theta}_s^{(t)} = \bar{\theta}_{0s} + (\bar{\theta}_{2s} - \bar{\theta}_{1s}) \sin(4\pi t/T + \phi_s)$  for  $t = 1 \dots T$ , where the  $\phi_s$  are randomly chosen for each node, while for even  $s$  it is  $\bar{\theta}_s^{(t)} = \bar{\theta}_{0s}$  for  $t = 1 \dots T$ ;
- autoregressive of order 1 (AR(1)), i.e., for odd  $s$  we have  $\bar{\theta}_s^{(t)} = \Phi_0 + \Phi_1 \bar{\theta}_s^{(t-1)} + \varepsilon^{(t)}$  for  $t = 1 \dots T$ , where  $\Phi_1 = 0.99$ ,  $\Phi_0$  is chosen such that the unconditional mean is equal to  $\theta_{0s}$ ,  $\varepsilon \sim \mathcal{N}(0, \sigma)$  and  $\sigma = 0.1$ . As in the previous cases, for even  $s$  we keep  $\bar{\theta}_s^{(t)} = \bar{\theta}_{0s}$  for  $t = 1 \dots T$ .

The dynamics considered are such that element  $s$  of vector  $\theta$  remains bounded between  $\theta_{1s}$  and  $\theta_{2s}$ . The values of  $\theta_1$  and  $\theta_2$  are fixed to allow fluctuations in the in and out degrees of the nodes, as follows. The vector  $\bar{\theta}_0$  is obtained by first generating two-degree sequences (in and out) such that the degrees linearly interpolate between a minimum degree  $D_m = 3$  and a maximum of  $D_M = 8$ . Then, we need to ensure that the degree sequence is graphicable, i.e., such that it exists one matrix of zeros and ones from which it can be obtained. We iteratively match links that make up the out-degree sequence with those that make up the in-degree sequence, starting with the largest

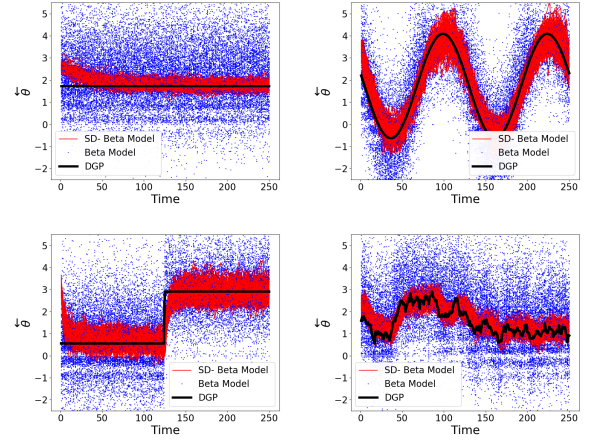


FIG. 5. Temporal evolution of one randomly selected parameter for the considered DGPs. The black line is the true path of the parameter of the DGP, the red ones are those filtered using the SD-beta model, and the blue dots correspond to the cross-sectional estimates of the beta model.

in- and out-degrees. In practice, we start with an empty matrix, select the largest out-degree, and set the matrix element between this node and the node with the largest in-degree to one. If, at some point, we cannot entirely allocate a given out-degree, we disregard the leftover links outgoing from that node and move to the next one. This procedure amounts to populating the adjacency matrix until no more links can be allocated. The degree sequence associated with this adjacency matrix is guaranteed to be graphicable. The numerical values of  $\bar{\theta}_0$  follow from the estimation of the static beta model. Finally, to gain additional heterogeneity in the amplitude of the fluctuations, we define  $N$  values evenly spaced between 0.4 and 1, i.e.,  $c_s$  for  $s = 1 \dots N$ . We use them to define

$$\bar{\theta}_{1s} = \bar{\theta}_{0s} + c_s (\bar{\theta}_{0_{s+1}} - \bar{\theta}_{0s})$$

$$\bar{\theta}_{2s} = \bar{\theta}_{0s} - c_s (\bar{\theta}_{0_{s+1}} - \bar{\theta}_{0s}).$$

Figure 5 shows the temporal evolution for one randomly chosen parameter of the beta model for all the DGPs, together with the paths filtered from the observations using the SD beta model and the sequence of cross-sectional static estimates. The score-driven filtering and cross-sectional estimation are repeated over 100 simulated network sequences. As discussed in the main text, the paths filtered with the SD beta model are, on average, much more accurate than those recovered from a standard beta model.

#### 1. SD-Beta Model for large $N$

In one of the numerical simulations we presented in the main text, we consider networks of increasing size. Here, we present some additional details on how the DGPs are defined for networks of increasing size. Practically, we have to fix the

vectors  $\bar{\theta}_0$ ,  $\bar{\theta}_{1_s}$ , and  $\bar{\theta}_{2_s}$  in a similar way, with the only difference being the numerical values for  $D_m$  and  $D_M$ . Specifically, in the sparse case we keep for each  $N$   $D_m = 10$  and  $D_M = 40$ . In the dense case, we set  $D_M = 0.8N$ , i.e., both the maximum degree and the average degree increase.

One peculiarity of the beta model is that the number of parameters, i.e., the length of the vectors  $\overleftarrow{\theta}$  and  $\overrightarrow{\theta}$ , increases with the number of nodes. Consistently, when we use the score-driven extension described so far, the length of the vectors  $w$ ,  $\alpha$ , and  $\beta$  increases too.

Recall that the numerical values of  $\theta_0$ ,  $\theta_1$ , and  $\theta_2$  are chosen to fix the values of average degrees over time and the amplitude of their fluctuations, as described in Appendix C. For each value of  $N$ , we choose them to guarantee heterogeneity in the degrees across nodes and significant fluctuation in time. Most importantly, we set a maximum degree attainable for a node and let it depend on  $N$  in two distinct ways, each corresponding to a different density regime: one generating *sparse* networks and the other *dense* ones. It is crucial to notice that the asymptotic results of Ref. 27 are expected to hold only in the dense case.

In the first numerical experiment testing the SD-beta model as a misspecified filter, we estimated a total of 60 parameters, 6 static parameters for each one of the ten nodes, 3 for the time-varying in-degree and 3 for the time-varying out-degree. Here, we present the further parameter restriction mentioned in the main text that proved useful when the number of nodes increases. Specifically, we assume that the parameters  $\alpha^{\text{out}}$  and  $\beta^{\text{out}}$  are common to all out-degree tvps  $\overrightarrow{\theta}^{(t)}$ . Similarly, all in-degree tvps  $\overleftarrow{\theta}^{(t)}$  share the same  $\alpha^{\text{in}}$  and  $\beta^{\text{in}}$ . The coefficients  $w_s^{\text{in}}$  and  $w_s^{\text{out}}$  remain node specific. The resulting update rule is

$$\begin{aligned}\overleftarrow{\theta}_s^{(t+1)} &= w_s^{\text{in}} + \beta^{\text{in}} \overleftarrow{\theta}_s^{(t)} + \alpha^{\text{in}} \frac{\sum_i (Y_{is}^{(t)} - p_{is}^{(t)})}{\sqrt{\sum_i p_{is}^{(t)} (1 - p_{is}^{(t)})}} \\ \overrightarrow{\theta}_s^{(t+1)} &= w_s^{\text{out}} + \beta^{\text{out}} \overrightarrow{\theta}_s^{(t)} + \alpha^{\text{out}} \frac{\sum_i (Y_{si}^{(t)} - p_{si}^{(t)})}{\sqrt{\sum_i p_{si}^{(t)} (1 - p_{si}^{(t)})}}.\end{aligned}$$

## REFERENCES

- <sup>1</sup>R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” *Rev. Mod. Phys.* **74**, 47–97 (2002).
- <sup>2</sup>E. Bullmore and O. Sporns, “Complex brain networks: graph theoretical analysis of structural and functional systems,” *Nature Reviews Neuroscience* **10**, 186 (2009).
- <sup>3</sup>M. Newman, *Networks: an introduction* (Oxford University Press, 2010).
- <sup>4</sup>D. Easley, J. Kleinberg, *et al.*, *Networks, crowds, and markets*, Vol. 8 (Cambridge University Press Cambridge, 2010).
- <sup>5</sup>F. Allen and A. Babus, “Networks in finance,” in *The network challenge: strategy, profit, and risk in an interlinked world*, edited by P. R. Kleindorfer and Y. J. Wind (Pearson Education, 2009) Chap. 21, pp. 367–379.
- <sup>6</sup>P. Holme and J. Saramäki, “Temporal networks,” *Physics reports* **519**, 97–125 (2012).
- <sup>7</sup>B. Craig and G. Von Peter, “Interbank tiering and money center banks,” *Journal of Financial Intermediation* **23**, 322–347 (2014).
- <sup>8</sup>G. Rossetti and R. Cazabet, “Community discovery in dynamic networks: a survey,” *ACM Computing Surveys (CSUR)* **51**, 35 (2018).
- <sup>9</sup>B. Kim, K. H. Lee, L. Xue, X. Niu, *et al.*, “A review of dynamic network models with latent variables,” *Statistics Surveys* **12**, 105–135 (2018).
- <sup>10</sup>G. Robins and P. Pattison, “Random graph models for temporal processes in social networks,” *Journal of Mathematical Sociology* **25**, 5–41 (2001).
- <sup>11</sup>S. Hanneke, W. Fu, E. P. Xing, *et al.*, “Discrete temporal models of social networks,” *Electronic Journal of Statistics* **4**, 585–605 (2010).
- <sup>12</sup>S. J. Cranmer and B. A. Desmarais, “Inferential network analysis with exponential random graph models,” *Political Analysis* **19**, 66–86 (2011).
- <sup>13</sup>P. N. Krivitsky and M. S. Handcock, “A separable model for dynamic networks,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 29–46 (2014).
- <sup>14</sup>J. Lee, G. Li, and J. D. Wilson, “Varying-coefficient models for dynamic networks,” *Computational Statistics & Data Analysis* **152**, 107052 (2020).
- <sup>15</sup>In time series jargon, it is a smoother and not a filter.
- <sup>16</sup>P. Mazzarisi, P. Barucca, F. Lillo, and D. Tantari, “A dynamic network model with persistent links and node-specific latent variables, with an application to the interbank market,” *European Journal of Operational Research* **281**, 50–65 (2020).
- <sup>17</sup>D. R. Cox, G. Gudmundsson, G. Lindgren, L. Bondesson, E. Harsaae, P. Laake, K. Juselius, and S. L. Lauritzen, “Statistical analysis of time series: Some recent developments,” *Scandinavian Journal of Statistics*, 93–115 (1981).
- <sup>18</sup>T. A. Snijders, “Stochastic actor-oriented models for network change,” *Journal of Mathematical Sociology* **21**, 149–172 (1996).
- <sup>19</sup>C. T. Butts, “A relational event framework for social action,” *Sociological Methodology* **38**, 155–200 (2008).
- <sup>20</sup>E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*, 1st ed. (Springer Publishing Company, Incorporated, 2009).
- <sup>21</sup>O. Barndorff-Nielsen, *Information and exponential families in statistical theory* (John Wiley & Sons, 2014).
- <sup>22</sup>M. Schweinberger, P. N. Krivitsky, C. T. Butts, and J. R. Stewart, “Exponential-family models of random graphs: Inference in finite, super and infinite-population scenarios,” *Statistical Science* **35**, 627–662 (2020).
- <sup>23</sup>C. E. Shannon, “A mathematical theory of communication,” *SIGMOBILE Mob. Comput. Commun. Rev.* **5**, 3–55 (2001).
- <sup>24</sup>E. Jaynes, “Information theory and statistical mechanics,” *Phys. Rev.* **106**, 620–630 (1957).
- <sup>25</sup>J. Park and M. E. J. Newman, “Statistical mechanics of networks,” *Phys. Rev. E* **70**, 066117 (2004).
- <sup>26</sup>D. Garlaschelli and M. I. Loffredo, “Maximum likelihood: Extracting unbiased information from complex networks,” *Phys. Rev. E* **78**, 015101 (2008).
- <sup>27</sup>S. Chatterjee, P. Diaconis, A. Sly, *et al.*, “Random graphs with a given degree sequence,” *The Annals of Applied Probability* **21**, 1400–1435 (2011).
- <sup>28</sup>E. Zermelo, “Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung,” *Mathematische Zeitschrift* **29**, 436–460 (1929).
- <sup>29</sup>P. W. Holland and S. Leinhardt, “An exponential family of probability distributions for directed graphs,” *Journal of the American Statistical Association* **76**, 33–50 (1981).
- <sup>30</sup>G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz, “Scale-free networks from varying vertex intrinsic fitness,” *Phys. Rev. Lett.* **89**, 258702 (2002).
- <sup>31</sup>T. A. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock, “New specifications for exponential random graph models,” *Sociological Methodology* **36**, 99–153 (2006).
- <sup>32</sup>G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison, “Recent developments in exponential random graph (p\*) models for social networks,” *Social Networks* **29**, 192–215 (2007).
- <sup>33</sup>D. R. Hunter and M. S. Handcock, “Inference in curved exponential family models for networks,” *Journal of Computational and Graphical Statistics* **15**, 565–583 (2006).
- <sup>34</sup>D. Creal, S. J. Koopman, and A. Lucas, “Generalized autoregressive score models with applications,” *Journal of Applied Econometrics* **28**, 777–795 (2013).
- <sup>35</sup>A. C. Harvey, *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*, Vol. 52 (Cambridge University Press, 2013).
- <sup>36</sup><http://www.gasmodel.com/index.htm> for the updated collection of papers dealing with GAS models.
- <sup>37</sup>F. Blasques, S. J. Koopman, and A. Lucas, “Information-theoretic opti-

- mality of observation-driven time series models for continuous responses,” *Biometrika* **102**, 325–343 (2015).
- <sup>38</sup>T. Bollerslev, “Generalized autoregressive conditional heteroskedasticity,” *Journal of Econometrics* **31**, 307–327 (1986).
- <sup>39</sup>D. B. Nelson, “Conditional heteroskedasticity in asset returns: A new approach,” *Econometrica: Journal of the Econometric Society*, 347–370 (1991).
- <sup>40</sup>R. F. Engle and J. R. Russell, “Autoregressive conditional duration: a new model for irregularly spaced transaction data,” *Econometrica*, 1127–1162 (1998).
- <sup>41</sup>R. Engle, “New frontiers for ARCH models,” *Journal of Applied Econometrics* **17**, 425–446 (2002).
- <sup>42</sup>S. M. Goodreau, “Advances in exponential random graph (p\*) models applied to a large social network,” *Social Networks* **29**, 231–248 (2007).
- <sup>43</sup>D. R. Hunter, S. M. Goodreau, and M. S. Handcock, “Goodness of fit of social network models,” *Journal of the American Statistical Association* **103**, 248–258 (2008).
- <sup>44</sup>J. Shore and B. Lubin, “Spectral goodness of fit for network models,” *Social Networks* **43**, 16–27 (2015).
- <sup>45</sup>F. Calvori, D. Creal, S. J. Koopman, and A. Lucas, “Testing for parameter instability across different modeling frameworks,” *Journal of Financial Econometrics* **15**, 223–246 (2017).
- <sup>46</sup>D. B. Nelson, “Asymptotically optimal smoothing with ARCH models,” *Econometrica* **64**, 561–573 (1996).
- <sup>47</sup>In the following, the notation with a bar refers to the true parameters used in the DGP.
- <sup>48</sup>For practical applications, it is very convenient that, for a large number of network functions, an efficient implementation to compute change statistics is made available in the R package *ergm*<sup>49</sup>.
- <sup>49</sup>D. R. Hunter, M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris, “*ergm*: A package to fit, simulate and diagnose exponential-family models for networks,” *Journal of Statistical Software* **24** (2008).
- <sup>50</sup>T. A. Snijders, “Markov Chain Monte Carlo estimation of exponential random graph models,” *Journal of Social Structure* **3**, 1–40 (2002).
- <sup>51</sup>G. Buccheri, G. Bormetti, F. Corsi, and F. Lillo, “Filtering and smoothing with score-driven models,” Available at SSRN: <https://ssrn.com/abstract=3139666> (2018).
- <sup>52</sup>We found the conclusions of this section to hold also in a sparse network density regime.
- <sup>53</sup>G. Iori, G. De Masi, O. V. Precup, G. Gabbi, and G. Caldarelli, “A network analysis of the italian overnight money market,” *J. Econ. Dyn. Control* **32**, 259–278 (2008).
- <sup>54</sup>K. Finger, D. Fricke, and T. Lux, “Network analysis of the e-mid overnight money market: the informational value of different aggregation levels for intrinsic dynamic processes,” *Computational Management Science* **10**, 187–211 (2013).
- <sup>55</sup>P. Barucca and F. Lillo, “The organization of the interbank network and how ECB unconventional measures affected the e-MID overnight market,” *Computational Management Science* **15**, 33–53 (2018).
- <sup>56</sup>It is worth stressing that the results become stable after 20 simulations.
- <sup>57</sup>In all the results on link forecasting – one- or multi-step-ahead – we excluded the links that are always zero, i.e., they never appear in the train and test samples. The reason is that those are extremely easy to predict, and keeping them would give an unrealistically optimistic picture of the predictability of links in the data set. Notably, the ranking of the methods remains unaltered when we keep all links for performance evaluation.
- <sup>58</sup>J. H. Fowler, “Connecting the congress: A study of cosponsorship networks,” *Political Analysis* **14**, 456–487 (2006).
- <sup>59</sup>K. Faust and J. Skvoretz, “Comparing networks across space and time, size and species,” *Sociological Methodology* **32**, 267–299 (2002).
- <sup>60</sup>Y. Zhang, A. J. Friend, A. L. Traud, M. A. Porter, J. H. Fowler, and P. J. Mucha, “Community structure in congressional cosponsorship networks,” *Physica A: Statistical Mechanics and its Applications* **387**, 1705–1712 (2008).
- <sup>61</sup>J. Moody and P. J. Mucha, “Portrait of political party polarization,” *Network Science* **1**, 119–121 (2013).
- <sup>62</sup>J. D. Wilson, N. T. Stevens, and W. H. Woodall, “Modeling and detecting change in temporal networks via the degree corrected stochastic block model,” *Quality and Reliability Engineering International* **35**, 1363–1378 (2019).
- <sup>63</sup>J. B. Lewis, P. Keith, R. Howard, B. Adam, R. Aaron, and S. Luke, “Voteview: Congressional roll-call votes database,” <https://voteview.com/> (2019).
- <sup>64</sup>S. Roy, Y. Atchadé, and G. Michailidis, “Change point estimation in high dimensional markov random-field models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 1187–1206 (2017).
- <sup>65</sup>We extensively tested via simulation that, for the model at hand and  $T$  and  $N$  taken from the data, estimating the variance of the latent parameters in such a way results, on average, in a small underestimation. In checking the coverage of the confidence bands, we considered a DGP with variance increased, with respect to the one estimated on the filtered time series, to compensate for this bias.
- <sup>66</sup>F. I. Karahanoğlu and D. Van De Ville, “Dynamics of large-scale fmri networks: Deconstruct brain activity to build better models of brain function,” *Current Opinion in Biomedical Engineering* **3**, 28–36 (2017).
- <sup>67</sup>S. L. Simpson, S. Hayasaka, and P. J. Laurienti, “Exponential random graph modeling for complex brain networks,” *PLoS one* **6**, e20039 (2011).
- <sup>68</sup>L. Giraitis, G. Kapetanios, A. Wetherilt, and F. Žikeš, “Estimating the dynamics and persistence of financial networks, with an application to the sterling money market,” *Journal of Applied Econometrics* **31**, 58–84 (2016).
- <sup>69</sup>P. Erdős and A. Rényi, “On random graphs i,” *Publ. Math. Debrecen* **6**, 290–297 (1959).
- <sup>70</sup>In the whole paper, we do not allow for links that start and end at the same node, so named *self-loops*. However, including them would be trivial.
- <sup>71</sup>S. E. Fienberg and S. S. Wasserman, “Categorical data analysis of single sociometric relations,” *Sociological Methodology* **12**, 156–192 (1981).
- <sup>72</sup>S. Wasserman and P. Pattison, “Logit models and logistic regressions for social networks: I. An introduction to markov graphs and p,” *Psychometrika* **61**, 401–425 (1996).
- <sup>73</sup>O. Frank and D. Strauss, “Markov graphs,” *Journal of the American Statistical Association* **81**, 832–842 (1986).
- <sup>74</sup>T. Yan, C. Leng, J. Zhu, *et al.*, “Asymptotics in directed exponential random graph models with an increasing bi-degree sequence,” *The Annals of Statistics* **44**, 31–57 (2016).
- <sup>75</sup>T. Yan, B. Jiang, S. E. Fienberg, and C. Leng, “Statistical inference in a directed network model with covariates,” *Journal of the American Statistical Association*, 1–12 (2018).
- <sup>76</sup>K. Jochmans, “Semiparametric analysis of network formation,” *Journal of Business & Economic Statistics* **36**, 705–713 (2018).
- <sup>77</sup>For a network with  $N$  nodes, the number of possible links is of order  $N^2$ . Instead, when all nodes have a fixed average degree  $d$ , the number of present links is  $dN$ , and the density is of order  $1/N$ .
- <sup>78</sup>Examples of such statistics are the count of 2 stars present in the network or the number of triangles<sup>72</sup>.
- <sup>79</sup>M. S. Handcock, G. Robins, T. Snijders, J. Moody, and J. Besag, “Assessing degeneracy in statistical models of social networks,” *Tech. Rep. (Working paper, 2003)*.
- <sup>80</sup>M. S. Handcock, “Statistical models for social networks: Inference and degeneracy,” in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers* (National Academies Press, 2003) p. 229.
- <sup>81</sup>M. Schweinberger, “Instability, sensitivity, and degeneracy of discrete exponential families,” *Journal of the American Statistical Association* **106**, 1361–1370 (2011).
- <sup>82</sup>*Edgewise* precisely means that we count partners only if shared by nodes that are connected.
- <sup>83</sup>D. Strauss and M. Ikeda, “Pseudolikelihood estimation for social networks,” *Journal of the American Statistical Association* **85**, 204–212 (1990).
- <sup>84</sup>C. Varin, N. Reid, and D. Firth, “An overview of composite likelihood methods,” *Statistica Sinica*, 5–42 (2011).
- <sup>85</sup>M. A. Van Duijn, K. J. Gile, and M. S. Handcock, “A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models,” *Social Networks* **31**, 52–62 (2009).
- <sup>86</sup>B. A. Desmarais and S. J. Cranmer, “Statistical mechanics of networks: Estimation and uncertainty,” *Physica A: Statistical Mechanics and its Applications* **391**, 1865–1876 (2012).
- <sup>87</sup>C. S. Schmid and B. A. Desmarais, “Exponential random graph models with big networks: Maximum pseudolikelihood estimation and the parametric bootstrap,” in *Big Data (Big Data), 2017 IEEE International Con-*

- ference on (IEEE, 2017) pp. 116–121.
- <sup>88</sup>F. Blasques, S. J. Koopman, and A. Lucas, “Maximum likelihood estimation for generalized autoregressive score models,” Tech. Rep. (Tinbergen Institute Discussion Paper, 2014).
- <sup>89</sup>S. Kullback, *Information theory and statistics* (Courier Corporation, 1997).
- <sup>90</sup>F. Blasques, A. Lucas, and A. C. van Vlodrop, “Finite sample optimality of score-driven volatility models: Some Monte Carlo evidence,” *Econometrics and Statistics* (2020).
- <sup>91</sup>F. Blasques, S. J. Koopman, K. Łasak, and A. Lucas, “In-sample confidence bands and out-of-sample forecast bands for tvtps in observation-driven models,” *International Journal of Forecasting* **32**, 875–887 (2016).
- <sup>92</sup>P. J. Huber *et al.*, “The behavior of maximum likelihood estimates under nonstandard conditions,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1 (University of California Press, 1967) pp. 221–233.
- <sup>93</sup>H. White, “A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity,” *Econometrica: Journal of the Econometric Society*, 817–838 (1980).
- <sup>94</sup>J. D. Hamilton, “A standard error for the estimated state vector of a state-space model,” *Journal of Econometrics* **33**, 387–397 (1986).
- <sup>95</sup>R. Davidson, J. G. MacKinnon, *et al.*, *Econometric theory and methods* (Oxford University Press New York, 2004).
- <sup>96</sup>J. H. Lee, “A Lagrange multiplier test for GARCH models,” *Economics Letters* **37**, 265–271 (1991).