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

## Domenico Di Gangi

Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126, Pisa, Italy

#### Giacomo Bormetti

Department of Mathematics, University of Bologna, Piazza di Porta San Donato 5, 40126 Bologna, Italy

#### Fabrizio Lillo

Department of Mathematics, University of Bologna, Piazza di Porta San Donato 5, 40126 Bologna, Italy

May 28, 2019

#### Abstract

Motivated by the evidence that real-world networks evolve in time and may exhibit non-stationary features, we propose an extension of the Exponential Random Graph Models (ERGMs) accommodating the time variation of network parameters. Within the ERGM framework, a network realization is sampled from a static probability distribution defined parametrically in terms of network statistics. Inspired by the fast growing literature on Dynamic Conditional Score-driven models, in our approach, each parameter evolves according to an updating rule driven by the score of the conditional distribution. We demonstrate the flexibility of the score-driven ERGMs, both as data generating processes and as filters, and we prove the advantages of the dynamic version with respect to the static one. Our method captures dynamical network dependencies, that emerge from the data, and allows for a test discriminating between static or time-varying parameters. Finally, we corroborate our findings with the application to networks from real financial and political systems exhibiting non stationary dynamics.

<sup>\*</sup>We are particularly thankful for comments and suggestions received by Fulvio Corsi and Giuseppe Buccheri. FL acknowledges partial support from the European Community H2020 Program under the scheme INFRAIA-1-2014-2015: Research Infrastructures, grant agreement #654024 SoBigData: Social Mining and Big Data Ecosystem (http://www.sobigdata.eu). Corresponding author domenico.digangi@sns.it

## 1 Introduction

A network, or graph<sup>1</sup>, is a useful abstraction for a system composed by a number of single elements that have some pairwise relation among them. The simplified description of social, economic, biological, transportation systems, often very complex in nature, in terms of nodes and links attracted and still attracts an enormous amount of attention, in a number of different streams of literature (Albert and Barabási, 2002; Bullmore and Sporns, 2009; Newman, 2010; Jackson, 2010; Easley et al., 2010; Allen and Babus, 2011). Formally, a graph 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 labelled and a link is identified by the pair of nodes it connects (i, j). To each G, we can assign one adjacency matrix  $\mathbf{Y}$  such that  $Y_{ij} = 1$  if link (i, j) is present in E and  $Y_{ij} = 0$  otherwise. In general, links may have an orientation and the corresponding adjacency matrix is not symmetric. In this case, network are dubbed directed networks. Moreover, 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, but will not consider the weighted variant.

Often systems that are fruitfully described as networks evolve in time. When pairwise interactions change over time, one usually speaks of temporal networks (Soramäki et al., 2010; Holme and Saramäki, 2012; Craig and Von Peter, 2014). In the time-varying setting, links can last for a finite interval of time, a quantity usually referred to as duration, or be instantaneous. In the latter case, one speaks of interaction or contact links, and different notations can used to describe them (Rossetti and Cazabet, 2018). In this paper, we will focus on the description of temporal networks as sets of links among nodes evolving over discrete times. Then, a dynamical network is a sequence of networks, each one 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  $\left\{Y_{ij}^{(t)}\right\}_{t=1}^{T}$ . In the following, we will present an approach to time-varying networks that is based on two main ingredients: i) a parametric probabilistic model, according to which one can sample a network realization, i.e. an adjacency matrix; ii) a simple mechanism to introduce time-variation on the model parameters and, consequently, to induce a dynamics on the network sequence. Concerning the former point, a natural choice is the class of statistical models for networks, known as Exponential Random Graph Models (ERGMs). As far as point ii) is concerned, a flexible candidate is suggested by the fast growing literature on the Dynamic Conditional Score-driven models (DCSs). The goal of this paper is to present a new class of models for temporal networks and to provide evidence that the novel approach is versatile and effective in capturing time-varying features. To the best of our knowledge, this is the first time the two frameworks are combined to provide a dynamic description of networks.

A statistical model for graphs can be specified providing the probability distribution over the set of possible graphs, i.e. all possible adjacency matrices (see Kolaczyk, 2009, for an introduction and a review of statistical models for networks). If the distribution belongs to the exponential family, than the model is named ERGM. To introduce it, let us mention the first and probably most famous example of this class: the Erdös-Rényi model of Erdős and Rényi (1959). In this model, fixed the

<sup>&</sup>lt;sup>1</sup>The two names are used interchangeably in this paper.

number of nodes N, each of the possible N(N-1)/2 links <sup>2</sup> is present with constant probability p, equal for all links. The probability to observe the adjacency matrix  $\mathbf{Y}$  is

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

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 from the others. Examples of more general ERGMs have been first proposed by Holland and Leinhardt (1981), under the name of log-linear, or  $p^*$ , models. Specifically, they named p1 the model defined by

$$\log P(\mathbf{Y}) = \sum_{ij} \left[ Y_{ij} Y_{ji} \rho_{ij} + Y_{ji} \phi_{ij} \right] - \log(\mathcal{K}(\boldsymbol{\rho}, \boldsymbol{\phi})), \tag{2}$$

where  $\rho$  and  $\phi$  are two matrices of parameters, and  $\mathcal{K}(\rho,\phi)$  is a normalization factor<sup>3</sup>, that ensures that the probabilities defined over all the possible adjacency matrices sum to one. 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 (Fienberg and Wasserman, 1981) or predetermined (exogenous or endogenous) covariates  $X_{ij}$  (Wasserman and Pattison, 1996). The requirement of independence among dyads has been relaxed since Frank and Strauss (1986) in order 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. In fact, in its most general form an ERGM is a model where the log-likelihood takes the form

$$\log P(\mathbf{Y}) = \sum_{s} \theta_{s} h_{s}(\mathbf{Y}) - \log(\mathcal{K}(\theta)), \tag{3}$$

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 the normalization  $\mathcal{K}$  is defined by

$$\mathcal{K}\left(\theta\right) = \sum_{\{\mathbf{Y}\}} e^{\theta_s h_s(\mathbf{Y})}.$$

The literature on ERGMs is extremely vast and still growing (see Schweinberger et al., 2018, for a recent literature review). Without being exhaustive, in section 2.1, we will focus only on aspects that are relevant for our extended approach, give some examples of network statistics to be used in ERGMs, and discuss their inference. It is worth to mention that the ERGM framework is

<sup>&</sup>lt;sup>2</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>&</sup>lt;sup>3</sup>Also known as partition function in the statistical physics literature.

intrinsically linked to the very well known principle of maximum entropy (Shannon, 2001) and its applications to statistical physics (Jaynes, 1957). 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  $h(\theta)$  (Park and Newman, 2004; Garlaschelli and Loffredo, 2008).

The second main ingredient of this work is the class of Dynamic Conditional Score-driven models introduced in Creal et al. (2013) and Harvey (2013), also known as Generalized Autoregressive Score (GAS) models <sup>4</sup>. In the language of Cox et al. (1981), DCSs belong to the class of observation-driven models. Specifically, one considers a conditional density  $P\left(y^{(t)}|f,\mathcal{F}_{t-1}\right)$  for the observation y at time t, depending on a vector of static parameters f and conditional on the information set  $\mathcal{F}_{t-1}$ . When a sequence of realizations  $\{y^{(t)}\}_{t=1}^T$  is observed, DCSs introduce an updating rule to promote the vector of parameters f to a time-varying sequence  $f^{(t)}$ . The update, described in details in Sec. 2.2, has an autoregressive component and an innovation term which depends on the scaled score of the conditional density. It turns out that many well known models in econometrics can be expressed as score-driven models. Famous examples are the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model of Bollerslev (1986), the Exponential GARCH model of Nelson (1991), the Autoregressive Conditional Duration (ACD) model of Engle and Russell (1998), and the Multiplicative Error Model (MEM) of Engle (2002). The introduction of this framework in its full generality opened the way to applications in various contexts. Our main contribution is an extension of the ERGM framework that allows model parameters to change over time in a scoredriven fashion. The result of our efforts is a class of models for time-varying networks where all information encoded in  $\mathcal{F}_{t-1}$  is exploited to filter the time-varying parameters  $\theta^{(t)}$  at time t. We refer to this class as Score-Driven Exponential Random Graph Models (SD-ERGMs). At this point, it is worth to comment that a generic SD-ERGM can be also used to generate synthetic sequences of graphs, i.e. it can be considered as a data generating process (DGP). However, we are more inclined to interpret it as an effective filter of latent time-varying parameters, regardless of what the true DGP might be.

#### 1.1 Literature Review

To the best of our knowledge, an ERGM with score-driven time-varying parameters has never been considered before. Nevertheless, we are by no means the first to discuss models for time-varying networks, for a review of latent space dynamical network models, see for example (Kim et al., 2018). Moreover, extensions of the ERGM framework for the description of dynamical networks exist in the literature. Two are the main streams. The first one was pioneered by Robins and Pattison (2001) and subsequently discussed in detail in Hanneke et al. (2010) and Cranmer and Desmarais (2011) and is known as TERGM. This approach builds on the ERGM, but allows the network statistics defining 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 TERGM

<sup>&</sup>lt;sup>4</sup>see http://www.gasmodel.com/index.htm for the updated collection of papers dealing with GAS models.

describes a stationary dynamics via the following dependencies among the links

$$\log P\left(\mathbf{Y}^{(t)}|\theta,\mathbf{Y}^{(t-K)},\ldots\mathbf{Y}^{(t)}\right) = \sum_{s} \theta_{s} h_{s}\left(\mathbf{Y}^{(t-K)},\ldots\mathbf{Y}^{(t)}\right) - \log(\mathcal{K}\left(\theta\right)).$$

A second approach, more related with our work, allows for the parameters of the ERGM to be time-varying. A notable example of this approach is the Varying-Coefficient-ERGM (VCERGM) proposed in Lee et al. (2017). There, the authors combine the varying-coefficient models' formalism (see Fan and Zhang, 2008, for a review) with ERGM to take into account the possibility of the ERGM parameters to be smoothly time-varying. The approach of VCERGM is different from ours in several respects. First, to infer parameter time-variation at time t, it uses all the available observations, including those from future times t' > t. Thus, it is a smoother and not a filter, in time series jargon. Second, as a consequence, it cannot generate sequences of time-evolving networks. At variance, our approach can be used as a filter, as a DGP for time-varying networks, when Monte Carlo scenarios are required, and, following Buccheri et al. (2018), can be extended as a smoother.

In a related, but different, approach Mazzarisi et al. (2017) consider the possibility of a random evolution of node specific parameters. Notably, the latter work accommodates for sender and receiver effect as well as link persistency. The model can also be used to filter the time-varying parameters in a very specific ERGM. Our contribution differs from the latter in flexibility, for the methods used, and for the general scope. In fact, we discuss and test our approach for a generic ERGM. More importantly, the authors of Mazzarisi et al. (2017) consider the random evolution of the parameters to be driven by an exogenous source of randomness. In this respect, following the language of Cox et al. (1981), they consider a class of model known as parameter driven. As stated above, we consider the dynamics of the parameters to be observation driven, i.e. the innovations are generated from observations, as it will be more clear in Section 3.

Finally, it is important to mention that frameworks alternative to latent space models and temporal extension of ERGM for modeling temporal networks have also been considered in the social science literature. Notable examples are the *Stochastic Actor Oriented model* (SAOM) of Snijders (1996) and the *Relational Event Model* (REM) of Butts (2008). For an overview of contributions in the social science community, we refer to the literature therein.

The rest of the paper is organized as follows. In Section 2 we detail key background concepts from the literature on ERGM and observation-driven models. In Section 3 we introduce the new class of models and validate it with extensive numerical experiments for two specific instances of the SD-ERGM. Section 4 presents the results from an application to two real temporal networks: The eMID interbank network for liquidity supply and demand and the U.S. Congress co-voting political network. Section 5 draws the relevant conclusions.

## 2 Preliminary notions

In this section, we review some concepts and notations from random graph modeling necessary to introduce our contribution. In the first part, we provide two examples of ERGMs, we highlight

some key difficulties in ERGMs inference, and review the standard approaches to circumvent them. These two examples will be functional to the introduction of our SD-ERGMs framework, since in Section 3 we will define their score-driven counterparts. Secondly, we discuss more formally the score-driven modeling framework. These topics come from different streams of literature: from the social sciences, statistics, and physics those in the former part, and from recent developments in econometrics the latter. Given the different origins of these ideas, and their central role in the rest of the paper, we deem appropriate a brief overview of both of them.

#### 2.1 ERGM Examples and Inference

ERGMs can be seen as an application of the family of discrete exponential distributions (Barndorff-Nielsen, 2014) to the description of graphs. The sufficient statistics, known as network statistics, are functions of the adjacency matrix  $h_s(\mathbf{Y})$  and the probability mass function (PMF) is defined by

$$\log P(\mathbf{Y}) = \sum_{s} \theta_{s} h_{s}(\mathbf{Y}) - \log(\mathcal{K}(\theta)). \tag{4}$$

The normalizing factor  $\mathcal{K}(\theta)$  is often not available as a closed-form function of the parameters  $\theta$ . Each matrix element is a binary random variable, and its probability depends only on the value of the network statistics appearing in (4). As common for exponential families, and discussed in detail for example in Park and Newman (2004), this probability distribution can be obtained from a direct application of the principle of maximum entropy.

In the following, we introduce two specific examples of ERGMs. They describe distinct features of the network and require different approaches to the parameter inference. The first statistic we will consider is meant to capture the heterogeneity in the number of connections that each link can have. It allows for straightforward maximum likelihood estimation. The second one describes transitivity in the formation of links, i.e. the tendency of connected nodes to have common neighbors. For this case, inference is instead complicated by the fact that the normalizing factor in (4) as a function of the parameters is not available in closed-form. The choice of these examples is instrumental to the main focus of this work, i.e. the time-varying parameter extension of the general ERGM in (4). In fact, they allow us to discuss different estimation techniques that will be crucial for our methodology: Maximum Likelihood Estimation (MLE) for node specific parameters and approximate pseudo-likelihood inference.

#### 2.1.1 The Beta (or Configuration) Model

The first example we consider is quite simple but, at the same time, largely employed in different streams of literature (Zermelo, 1929; Bradley and Terry, 1952; Holland and Leinhardt, 1981; Caldarelli et al., 2002; Park and Newman, 2004; Garlaschelli and Loffredo, 2008; Chatterjee et al., 2011). The range of applications for this model is so broad that researchers were often not aware of previous works using exactly 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:  $\overrightarrow{\theta}_i$ , that captures the propensity of node i to form outgoing connections, and  $\overleftarrow{\theta}_i$  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

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

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

$$\log P\left(\mathbf{Y}\right) = \sum_{i=1}^{N} \left( \overleftarrow{\theta}_{i} \overleftarrow{D}_{i} + \overrightarrow{\theta}_{i} \overrightarrow{D}_{i} \right) - \sum_{ij} \log \left( 1 + e^{\overleftarrow{\theta}_{i} + \overrightarrow{\theta}_{j}} \right). \tag{5}$$

This formulation is often used when the heterogeneity in the degrees is expected to play a prominent role in explaining the presence or absence of links. It is worth to notice that the static version of the beta model, in the directed case, is not identified. If we add any constant to each  $\theta_i$ and subtract it from each  $\overrightarrow{\theta}_i$ , the PMF remains unchanged. To fix it, one needs to introduce an identification restriction. This is essential to compare the parameter values estimated for the same network at different times. Appendix A comments on the restriction more extensively. Relevant for the discussion in Section 3.2, it is important to notice that the MLE can be performed using a fixed point algorithm, described for example in Yan et al. (2016), that reaches the optimal solution in a fast way. Moreover, we point out the existence of interesting results on the asymptotic behavior of the maximum likelihood estimates for  $(\overleftarrow{\theta}, \overrightarrow{\theta})$  when the number of nodes increases. Indeed, consistency results have been proved in Chatterjee et al. (2011) for the undirected case and in Yan et al. (2016) for the directed case (see also Graham, 2017; Yan et al., 2018; Jochmans, 2018, for discussions of the statistical properties of the beta model). 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 1/N. Networks belonging to this density regime are named sparse. Notably, no consistency results are known for large N in the sparse regime.

#### 2.1.2 A Statistic for Transitivity and Pseudo-Likelihood Estimation

Our second example is motivated by the need to demonstrate the applicability of our methodology to the widest possible set of network statistics and ERGM. In fact, as we will see in Section 3, the score of the likelihood plays a defining role in our approach. It is henceforth important to show how to deal with cases when the normalization function  $\mathcal{K}(\theta)$ , and thus the score, is not available as a closed-form function of the parameters  $\theta$ . It is well known that, when network

<sup>&</sup>lt;sup>5</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 link is dN, and the density is of order 1/N.

statistics involve products of matrix elements<sup>6</sup>, this is often the case. This lack of analytical tractability has been arguably the main obstacle in estimation and understanding of the properties of ERGMs. 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, in order to avoid statistical issues (as discussed for example in Handcock, 2003a,b). The main issue, with consequences on estimation, simulation, and interpretability of ERGMs, is known as degeneracy. An ERGM is degenerate if it concentrates a large portion of its probability on a small set of configurations, typically the uninteresting graphs that are 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. Indeed, a great effort has been dedicated to investigating this problem, and characterizing degeneracy (see for example Schweinberger, 2011, and references therein). A family of network statistics that, while describing properties of the whole network, is not plagued by degeneracy has been proposed in Snijders et al. (2006); Robins et al. (2007) and discussed in Hunter and Handcock (2006). This family is referred to as curved exponential random graphs, and one example of curved statistic, that we will use in Section 3.3, is the Geometrically Weighted Edgewise Shared Partners (GWESP). This function has recently been applied extensively to describe transitivity in social networks (see Hunter and Handcock, 2006). 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, that are connected by an edge, and count the number of nodes to which they are both connected, i.e. the number of neighbors that they share. Let us indicate with  $ESP_k(\mathbf{Y})$  the number of edgewise shared partners, i.e. connected node pairs <sup>7</sup> that share exactly k neighbors in the network described by Y. Then GWESP is defined as

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

To conclude this partial overview of background concepts about ERGMs, we need to discuss parameter inference, when the likelihood is not available in closed-form. The two standard approaches to ERGM inference consist in maximizing alternative functions that are known to share the same optimum as the exact likelihood. The first possibility (described, for example, in Snijders, 2002) 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 (4) necessary relies on Markov Chain Monte Carlo (MCMC) approaches (see Hunter et al., 2008, for a description of a popular software that implements it). The computational burden of MCMC-based estimation can

<sup>&</sup>lt;sup>6</sup>Examples of such statistics are the count of 2 stars present in the network, or the number of triangles (Wasserman and Pattison, 1996).

<sup>&</sup>lt;sup>7</sup> Edgewise precisely means that we count partners only if shared by nodes that are connected.

be prohibiting for large enough graphs. For this reason, a second approximate inference procedure, known as Maximum Pseudo-Likelihood Estimation (MPLE), (first proposed, in the context of ERGMs, in Strauss and Ikeda, 1990) is often used in empirical applications. Comparisons of the two inference procedures highlighted potential pitfalls of the pseudo-likelihood (Van Duijn et al., 2009), particularly in estimating appropriate confidence intervals. Nevertheless, the computational superiority of MPLE with respect to MCMC estimation is indisputable, and in many situations, as in our case, MPLE is the only viable solution (Desmarais and Cranmer, 2012; Schmid and Desmarais, 2017). MPLE is based on the optimization of the pseudo-likelihood function, that is in turn 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

$$PL(\mathbf{Y}) = \prod_{ij} \pi_{ij}^{Y_{ij}} (1 - \pi_{ij})^{(1 - Y_{ij})}$$
(6)

where  $\pi_{ij} = \left(1 + e^{-\sum_s \theta_s \delta_{ij}^s}\right)^{-1}$ . The maximum pseudo-likelihood estimates correspond to the parameter values  $\theta$  that maximize the pseudo-likelihood. This procedure is extremely faster than the exact MLE based on MCMC, even for networks of limited size.

#### 2.2 Score-Driven Models

In order to review the score-driven models as introduced by Creal et al. (2013) and Harvey (2013), 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  $P\left(y^{(t)}|f^{(t)}\right)$ , that depends on a vector of time-varying parameters  $f^{(t)} \in \mathbb{R}^K$ . Defining the score as  $\nabla^{(t)} = \frac{\partial \log P\left(y^{(t)}|f^{(t)}\right)}{\partial f^{(t)}}$ , a score-driven model assumes that the time evolution of  $f^{(t)}$  is ruled by the recursive relation

$$f^{(t+1)} = w + \beta f^{(t)} + \alpha S^{(t)} \nabla^{(t)}, \tag{7}$$

where 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, that is often chosen to be the inverse of the square root of the

Fisher information matrix associated with 
$$P\left(y^{(t)}|f^{(t)}\right)$$
, i.e.  $S^{(t)} = \mathbb{E}\left[\frac{\partial \log P\left(y^{(t)}|f^{(t)}\right)}{\partial f^{(t)'}}\frac{\partial \log P\left(y^{(t)}|f^{(t)}\right)'}{\partial f^{(t)'}}\right]^{-\frac{1}{2}}$ .

However, this is not the only possible specification and different choices for the scaling are discussed in Creal et al. (2013).

The most important feature of (7) is the role of the score as a driver of the dynamics of  $f^{(t)}$ . 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 a stochastic dynamics exactly thanks to the random sampling of  $y^{(t)}$ . A second look at eq. (7) reveals to the reader the similarity of the score-driven recursion with the iterative step from a Newton algorithm, whose objective function is precisely the log-likelihood function. Indeed, at each step the score pushes the parameter vector along the log-likelihood steepest direction. After scaling with the matrix S, the intensity of the push is modulated by the parameter  $\alpha$ , and its direction adjusted by the auto-regressive component.

Several are the reasons of the flexibility of a score-driven approach and of its success in timeseries modeling. In practical applications, the static parameters of (7) need to be estimated. As detailed in Harvey (2013), using the so-called prediction error decomposition, the likelihood of score-driven models can be readily expressed in closed form. In a univariate setting, Blasques et al. (2014) work 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 Blasques et al. (2015), the authors consider a true and unobserved DGP  $y^{(t)} \sim P\left(y^{(t)}|f^{(t)}\right)$ . They assume a given and in general mispecified conditional observation density  $\tilde{P}^{(t)} = \tilde{P}\left(\cdot |\tilde{f}^{(t)}\right)$ , and consider the Kullback-Leibler (K-L) divergence

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

where  $A \subseteq \mathbb{R}$ . Building on the minimum discrimination information principle (Kullback, 1997), 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 not known, an optimal update that minimizes  $\mathcal{D}_{\mathcal{KL}}$  cannot be defined in practice. For this reason, Blasques et al. (2015) 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  $\mathcal{D}_{\mathcal{KL}}$ 

$$\Delta_{t|t} \equiv \mathcal{D}_{\mathcal{KL}}\left(P^{(t)}, \tilde{P}^{(t+1)}\right) - \mathcal{D}_{\mathcal{KL}}\left(P^{(t)}, \tilde{P}^{(t)}\right) = \int_{A} P\left(y|f^{(t)}\right) \log \frac{\tilde{P}\left(y|\tilde{f}^{(t)}\right)}{\tilde{P}\left(y|\tilde{f}^{(t+1)}\right)} dy.$$

Based on this definition, a parameter update is realized K-L optimal when  $\Delta_{t|t} < 0$  for every  $\left(y^{(t)}, \tilde{f}^{(t)}, f^{(t)}\right)$ . The authors prove that, under reasonable assumptions, the updating rule (7) 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 Blasques et al. (2017). For the purposes of 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 clearly specific of 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 an infinite number of realized Kullback-Leibler optimal updates.

As a final aspect, which will be relevant in the application section of this paper, score-driven models allow for a test discriminating whether the observations are better described by a model with time-varying parameters or static ones. In fact, following Engle (1982), Calvori et al. (2017) discuss, and extensively evaluate, the performances of a test for parameter temporal variation tailored for score-driven models. For a detailed description of the test, we refer the reader to Calvori et al. (2017). Here, we shortly review the main idea. The method consists in a Lagrange Multiplier (LM) test for the parameter  $\alpha$  that multiplies the score in the one dimensional version of the recursion (7). The null hypothesis  $H_0$  is that the parameter  $f^{(t)}$  is actually static, i.e.  $\beta = \alpha = 0$ , and it corresponds to w. As explained in Davidson et al. (2004), 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 Lee (1991). As discussed in Calvori et al. (2017), the LM statistic can be written as the explained sum of squares from the regression

 $\mathbf{1} = c_w \nabla_w^{(t)} + c_\alpha S^{(t-1)} \nabla_w^{(t-1)} \nabla_w^{(t)\prime} + \text{residual},$ 

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 1 degree of freedom.

## 3 Score-Driven Exponential Random Graphs

In this section, we introduce the general SD-ERGM framework, discuss in detail the applicability of the score-driven approach to two different ERGMs, and validate their performances with extensive numerical simulations.

#### 3.1 Definition of SD-ERGM

We propose to apply the score-driven methodology to ERGMs, in order to allow any of the parameters  $\theta_s$  in (4) 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 the description of time-varying networks, more than in a single model, in very much the same way as ERGM is considered a modeling framework for static networks. We refer to such class of models as Score-Driven Exponential Random Graphs Models.

Conceptually, applying the score-driven approach is fairly straightforward. Given the observations  $\left\{Y_{ij}^{(t)}\right\}_{t=1}^{T}$ , we can apply the update rule in (7) to all or some elements of  $\theta$ , each of which is associated with a network statistic in (4). In order 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

$$\nabla_{s}^{(t)}(\theta) = \frac{\partial \log P\left(\mathbf{Y}^{(t)}|\theta\right)}{\partial \theta_{s}} = h_{s}\left(\mathbf{Y}^{(t)}\right) - \frac{\partial \log \mathcal{K}(\theta)}{\partial \theta_{s}}.$$

It follows that the vector of time-varying parameters evolves according to

$$\theta^{(t+1)} = w + \beta \theta^{(t)} + \alpha S^{(t)} \nabla^{(t)} \left( \theta^{(t)} \right). \tag{8}$$

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 a stochastic dynamics because, at each time t, the adjacency matrix is not known in advance. It is randomly sampled from  $P\left(\mathbf{Y}^{(t)}|\theta^{(t)}\right)$  and then used to compute the score that, as a consequence, becomes itself stochastic. When the sequence of networks  $\left\{\mathbf{Y}^{(t)}\right\}_{t=1}^{T}$  is observed, the static parameters  $(w, \boldsymbol{\beta}, \boldsymbol{\alpha})$ , that best fit the data, can be estimated maximizing the log-likelihood of the whole time series. Taking into account that each network  $\mathbf{Y}^{(t)}$  is independent from all the others conditionally on the value of  $\theta^{(t)}$ , the log-likelihood can be written as

$$\log P\left(\left\{\mathbf{Y}^{(t)}\right\}_{t=1}^{T} | w, \beta, \alpha\right) = \sum_{t=1}^{T} \log P\left(\mathbf{Y}^{(t)} | \theta^{(t)}\left(w, \beta, \alpha, \left\{\mathbf{Y}^{(t')}\right\}_{t'=1}^{t-1}\right)\right). \tag{9}$$

It is evident that the computation of the normalizing factor, and its derivative with respect to the parameters, is essential for the SD-ERGM. Not only it enters the definition of the update, but it is also required for the optimization of (9).

Our main motivation for the introduction of SD-ERGM is to describe the time evolution of a sequence of networks by means of the evolution of the parameters of an ERGM. We assume to know, from the context or from previous studies of static networks in terms of ERGM, which statistics are more appropriate in the description of a given network. Hence, we do not discuss the choice of statistics in the context of dynamical networks, but refer the reader to Goodreau (2007) and Hunter et al. (2008) for examples of feature selection and Goodness Of Fit (GOF) evaluation, as well as to Shore and Lubin (2015) for a recent proposal to quantify GOF specifically in network models.

In the rest of this section, we discuss in details the SD extension of ERGMs with a given statistics. The first example allows for the exact computation of the likelihood, but the number of parameters can become large for large network. In the second example, we discuss 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 time-varying parameters when the DGP is known and the score-driven model is employed as a misspecified filter. Moreover, we show a first application of the LM test in assessing the time-variation of ERGM parameters.

#### 3.2 Dynamical Node Specific Parameters: the SD Beta Model

Our first specific example is the score-driven version of the beta model, introduced in Section 2.1.1. We start with this model not only because of its wide applications and relevance in various streams of literature, but also because the likelihood of the ERGM can be computed exactly. As a consequence, the score can be computed straightforwardly. Moreover, the number of local statistics, the degrees, and parameters can become very large for large networks. Since we need to describe the dynamics of a large amount of parameters, this last feature poses a challenge to 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 an high dimensional dynamics.

Recall that, in this model, there are no impediments in writing the explicit dependence of the likelihood on the parameters  $\overrightarrow{\theta}$  and  $\overrightarrow{\theta}$ , that can be found in (5). Defining, for ease of notation,

$$p_{ij}^{(t)} = \frac{1}{1 + e^{-\overleftarrow{\theta}_{i}^{(t)} - \overrightarrow{\theta}_{j}^{(t)}}}$$

we can write the score as

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

Among the possible choices, we use as scaling the diagonal matrix  $S_{ij}^{(t)} = \delta_{ij} I_{ij}^{(t)^{-1/2}}$ , where  $\boldsymbol{I}^{(t)} = \mathbb{E}[\nabla^{(t)}\nabla^{(t)'}]$ , i.e. we scale each element of the score by the square root of its variance. To clarify the notation, we mention that for the rest of this Section<sup>8</sup>, when we write  $\theta$ , without any arrow, we mean the vector that includes both  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta}$ , that is  $\boldsymbol{\theta} = \left(\boldsymbol{\theta}', \boldsymbol{\theta}'\right)'$ .

#### 3.2.1 SD-ERGMs as filters: Numerical Simulations

As mentioned in the Introduction, SD-ERGMs (as other observation driven model, e.g. GARCH) can be seen either as DGP and estimated on real time series or as predictive filters (Nelson (1996)), since time-varying parameters are one-step-ahead predictable filters of the dynamics. In this Section we show the power of the ERGMs in this second setting. Specifically, we simulate generic non-stationary temporal evolution for the parameters  $\theta^{(t)}$  of temporal networks We then use the SD-ERGM to filter the paths of the parameters and evaluate its performances. It is important to note

<sup>&</sup>lt;sup>8</sup>This notation is indeed coherent with the one used in the rest of the paper, but we stress its meaning to avoid any confusion that might arise from the large number of parameters in the beta model.

that the simulated dynamics of the parameters is different from the score-driven one used in the estimation.

In practice, at each time t, we sample the adjacency matrix from the PMF of an ERGM with parameters  $\bar{\theta}^{(t)}$ , evolving according to known temporal patterns, that define different DGP. 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  $\bar{\theta}_i^{(t)}$  and  $\bar{\theta}_i^{(t)}$  evolving with predetermined patterns. We test four different DGPs. The first one is a naive case with constant parameters  $\bar{\theta}^{(t)} = \bar{\theta}_0$ . The elements of  $\bar{\theta}_0$  are chosen in order to ensure heterogeneity in the expected degrees of the nodes under the static beta model. Appendix B provides a description of how we fix their numerical values. For the remaining three DGPs, half of the parameters is static and half is time-varying 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 in order to allow fluctuations in the in and out degrees of the nodes. The details of the procedure used to fix them are given again in Appendix B. The dynamical DGPs are:

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

It is very common, in score-driven models with numerous time-varying parameters, to restrict the matrices  $\alpha$  and  $\beta$  of (7) 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  $\theta_s$ . It follows that each time varying parameter evolves according to

$$\theta_s^{(t+1)} = w_s + \beta_s \theta_s^{(t)} + \alpha_s \left( S^{(t)} \nabla^{(t)} \left( \theta^{(t)} \right) \right)_s.$$

<sup>&</sup>lt;sup>9</sup>In the following, the notation with a bar refers to the true parameters used in the DGP.

The resulting update rule for the beta model is

$$\overleftarrow{\theta}_{s}^{(t+1)} = w_{s}^{\text{in}} + \beta_{s}^{\text{in}} \overleftarrow{\theta}_{s}^{(t)} + \alpha_{s}^{\text{in}} \left( \frac{\sum_{i} Y_{is}^{(t)} - p_{is}^{(t)}}{\sqrt{\sum_{i} p_{is}^{(t)} \left( 1 - p_{is}^{(t)} \right)}} \right)$$

$$\overrightarrow{\theta}_{s}^{(t+1)} = w_{s}^{\text{out}} + \beta_{s}^{\text{out}} \overrightarrow{\theta}_{s}^{(t)} + \alpha_{s}^{\text{out}} \left( \frac{\sum_{i} Y_{si}^{(t)} - p_{si}^{(t)}}{\sqrt{\sum_{i} p_{si}^{(t)} \left( 1 - p_{si}^{(t)} \right)}} \right),$$

$$(10)$$

where the superscripts in and out indicate the first and second half of the parameter vectors, respectively. In order to simplify the inference procedure, we consider a two-step approach. First, we fix the node specific parameters  $w_i$  in order to target the unconditional means of  $\theta$  and  $\theta$  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.

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. Figure 1 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. It clearly emerges that the paths filtered with the SD beta model are on average much more accurate than those recovered from a standard beta model. In order to quantify the performance of the two approaches, we compute the Root Mean Square Error, that describes the distance between the known simulated path and the filtered one:

RMSE 
$$(\theta_s) = \frac{1}{T} \sqrt{\sum_t \left(\bar{\theta}_s^{(t)} - \hat{\theta}_s^{(t)}\right)^2}.$$

We then average the RMSE across all the time-varying parameters and 100 simulations, and report the results in Table 1. These results confirm that the SD beta model outperforms the standard beta model in recovering the true time-varying pattern. Notably, this holds true even when the DGP is inherently non stationary, as in the case of the DGP where each parameter has a step like evolution. Indeed the results of this Section and of Section 3.3 confirm that, while the SD update rule (7) defines a stationary DGP (see Creal et al., 2013), using SD models as filters, we can effectively recover non stationary parameters' dynamics.

#### 3.2.2 SD-Beta Model for Large N

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. This is not the case for many ERGMs, as for example the one that we will discuss in the following section. Consistently, when we use the score-driven

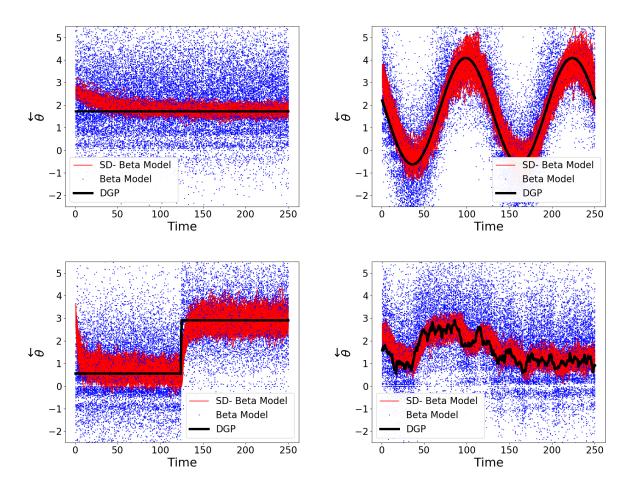


Figure 1: Temporal evolution of one of the parameters, randomly selected, for the considered DGPs. The black line is the true path of the parameter of the DGP, the red ones are those filtered using the SD-beta model, and the blue dots correspond to the cross sectional estimates of the beta model.

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				

Table 1: RMSEs (on a percentage base) of the filtered paths averaged over all time-varying parameters 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 correspond to one of the four DGPs.

extension described so far, the length of the vectors w,  $\alpha$  and  $\beta$  increases too. Here we discuss the application of the SD beta model to large networks<sup>10</sup> for two reasons. The first one is that many real systems are described by networks with a large number of nodes. The second reason is that we want to compare the performances of our approach with those of the standard beta model in regimes where the latter is known to perform better, under suitable conditions. Indeed the asymptotic results, mentioned at the end of Section 2.1, on the single observation estimates guarantee that, if the network density remains constant as N grows larger, 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. In order to check whether the score-driven 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 constant the average degree. In this analysis, we consider only one dynamical DGP and many different values of N.

In the numerical experiment discussed in the previous section, 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. While nowadays estimating models with thousands of parameters is manageable, thanks for example to Automatic Differentiation (Baydin et al., 2018), we aim at defining a version of the SD beta model that can be estimated, in reasonable time, on a common laptop. For this reason we propose a further parameter restriction. Specifically, we assume that the parameters  $\alpha^{\text{out}}$  and  $\beta^{\text{out}}$  are common to all out-degree time-varying parameters  $\overrightarrow{\theta}^{(t)}$ . Similarly, all in-degree time-varying parameters  $\psi^{(t)}$  share the same  $\psi^{(t)}$  and  $\psi^{(t)}$ . The coefficients  $\psi^{(t)}$  and

<sup>&</sup>lt;sup>10</sup>We tested the applicability for networks having thousands of nodes. We believe that this limitation can be easily fixed and the network dimension further increased with a coding approach less naive than ours.

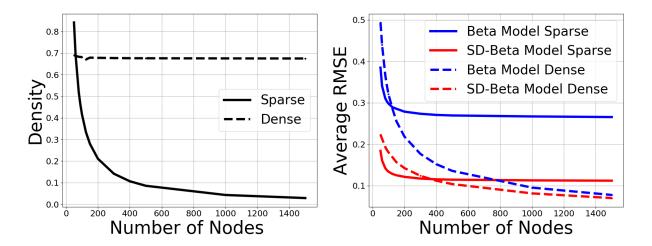


Figure 2: Left panel: average density as a function of the number of nodes N in the dense (dashed line) and sparse (solid line) regimes. Right panel: average RMSE of the filtered parameters with respect to the simulated DGP in both the dense (dashed lines) and sparse (solid lines) regimes. The average RMSE from the ERGM is plotted in blue, while the one from the SD-ERGM in red.

 $w_s^{\text{out}}$  remain node specific but are fixed, as before, by targeting. The resulting update rule is

$$\overleftarrow{\theta}_{s}^{(t+1)} = w_{s}^{\text{in}} + \beta^{\text{in}} \overleftarrow{\theta}_{s}^{(t)} + \alpha^{\text{in}} \left( \frac{\sum_{i} Y_{is}^{(t)} - p_{is}^{(t)}}{\sqrt{\sum_{i} p_{is}^{(t)} \left(1 - p_{is}^{(t)}\right)}} \right) .$$

$$\overrightarrow{\theta}_{s}^{(t+1)} = w_{s}^{\text{out}} + \beta^{\text{out}} \overrightarrow{\theta}_{s}^{(t)} + \alpha^{\text{out}} \left( \frac{\sum_{i} Y_{si}^{(t)} - p_{si}^{(t)}}{\sqrt{\sum_{i} p_{si}^{(t)} \left(1 - p_{si}^{(t)}\right)}} \right) .$$
(11)

Among the DGPs used in the previous section, we consider the one with smooth and periodic time variation. Recall that the numerical values of  $\theta_0$ ,  $\theta_1$  and  $\theta_2$  are chosen in order to fix the values of average degrees over time and the amplitude of their fluctuations, as described in Appendix B. For each value of N, we choose them in order 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 we let it depend on N in two distinct ways, each one corresponding to a different density regime: one generating *sparse* networks and the other *dense* ones. It is important to notice that the asymptotic results of (Chatterjee et al., 2011) 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 2. Then, for both regimes and each value of N, we compute the average RMSE across all time-varying parameters and all Monte Carlo replicas. In the right panel of Figure 2, the average RMSEs for different values of N clearly indicate that, also for large networks, the SD version of

the beta model attains better results compared with the cross sectional estimates. As expected, in the dense network regimes, both approaches reach the same accuracy 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.

#### 3.3 SD-ERGM With Unknown Normalization

As mentioned earlier, the dependence of the normalizing function on the  $\theta$  parameters is sometimes unknown. This fact prevents us from computing the score function and directly applying the update rule (7) to a large class of ERGMs. To circumvent this obstacle, we propose to use the score of the pseudo-likelihood, discussed in Sec. 2.1, instead of the unattainable score of the exact likelihood. This amounts to using the score of the pseudo-likelihood, or pseudo-score

$$\nabla^{(t)}(\theta) = \frac{\partial \log PL\left(\mathbf{Y}^{(t)}|\theta\right)}{\partial \theta_s^{(t)\prime}} = \sum_{ij} \delta_{ij}^s \left(Y_{ij}^{(t)} - \frac{1}{1 + e^{-\sum_l \theta_l \delta_{ij}^l}}\right),\tag{12}$$

in place of the exact score in the definition the SD-ERGM update (7). We remark that, as discussed in Section 2.2, from the information theoretic perspective of Blasques et al. (2015), the update based on the pseudo-score 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. Additionally, we use the pseudo-likelihood for each observation  $\mathbf{Y}^{(t)}$  in (9) for the inference of the static parameters.

Our approach, based on the score of the pseudo-likelihood, requires as input the change statistics for each function  $h_s(\mathbf{Y}^{(t)})^{-11}$ . In the following, we show that the update based on the score of the pseudo-likelihood is effective in filtering the path of time-varying parameters. Remarkably, this is true even when the probability distribution in the DGP is the exact one, i.e. when we sample from the exact likelihood and then use the SD-ERGM based on the pseudo-likelihood to filter.

In order to show the concrete applicability and performance of the approach based on the pseudo-score, We consider an ERGM with two global network statistics. The first one is the total number of links present in the network. The second statistics is the GWESP, discussed at the end of Section 2. The ERGM is thus defined by

$$\sum_{s} \theta_{s} h_{s} \left( \mathbf{Y}^{(t)} \right) = \theta_{1} \sum_{ij} Y_{ij}^{(t)} + \theta_{2} \text{GWESP} \left( \mathbf{Y}^{(t)} \right) . \tag{13}$$

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  $\bar{\theta}_s^{(t)}$ .

<sup>&</sup>lt;sup>11</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* Hunter et al. (2008).

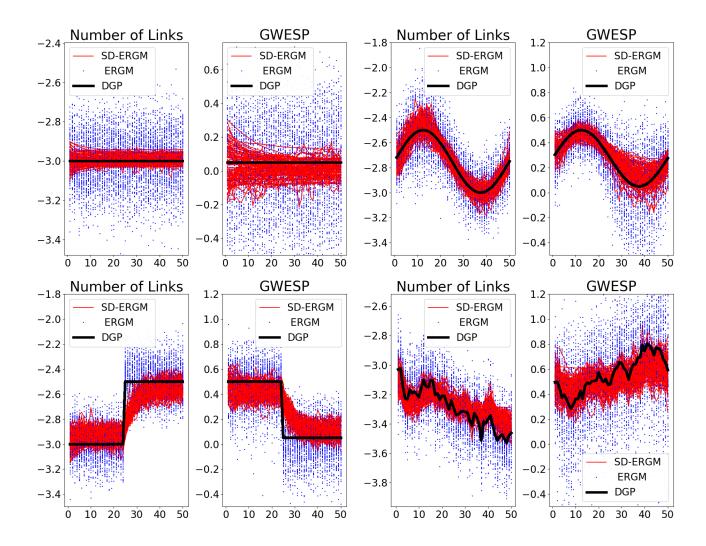


Figure 3: Filtered paths of the parameters of a ERGM with time-varying parameters. The path from the true DGP is in black. The blue dots are the cross sectional ERGM estimates, and the red lines the SD-ERGM filtered paths.

we consider the four DGPs introduced and analyzed in Section 3.2. We sample and estimate the models 50 times for each DGP. Figure 3 compares the cross sectional estimates and the score-driven filtered paths. Table 2 reports the RMSE of the GWESP time-varying parameters, averaged over the different realizations for the whole sequence t = 1, 2, ..., T. It is evident that 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, even for the

DGP	Average RMSE			LM Test			
	ERGM		SD-ERGM		% Correct Results		
	$\theta_1^{(t)}$	$\theta_2^{(t)}$	$\theta_1^{(t)}$	$\theta_2^{(t)}$	$\left\  \left(  heta_1^{(t)} \;,  heta_2^{(t)}  ight)  ight.$	$\left(  heta_{1}^{(t)} \; ,  heta_{2}  ight)$	$\left(  heta_1 \; ,  heta_2^{(t)}  ight)$
Const	0.02	0.1	0.0006	0.004	86%		<u> </u>
Sin	0.02	0.04	0.003	0.005	100%	94%	93%
Steps	0.02	0.03	0.01	0.001	98%	92%	96%
AR1	0.02	0.2	0.007	0.01	20%	93%	90%

Table 2: First four columns: RMSEs for the filtered paths of the time-varying parameters, averaged over 50 repetitions, for the evolutions of Figure 3. The last three columns describe the accuracy of the test for dynamics in the parameters, considering the DGPs in Figure 3, 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.

static case, the SD-ERGM is a reliable and consistent choice.

It is worth noticing that, for sampling and cross sectional inference, we employed the R package ergm. It uses state of the art MCMC techniques for both tasks (see Hunter et al., 2008, for a description of the software). Hence, we compared the SD-ERGM based on the approximate pseudolikelihood – both in the definition of the time-varying parameter update and inference of the static parameters – with a sequence of exact cross sectional estimates. In general, the latter are known to be better performing than the pseudo-likelihood alternative, as discussed in Section 2.1. 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 the best performing solution. In our opinion, this provides further evidence of the advantages of SD-ERGM as a filtering tool. Finally, the last column of Table 2 reports the percentage number of times the LM test of Calvori et al. (2017) applied to the SD-ERGM correctly classifies the parameters as time-varying (or static for the constant DGP). The test performs correctly in almost all the cases considered. The only exception is the case of the AR(1) dynamics, when both parameters vary in time. The test is quite conservative and correctly identifies both parameters as time-varying only 20% of the times. The problem is that, even for degeneracy-free ERGMs, multi-collinearity of statistics often emerges. A rigorous discussion would require to take into account different signal-to-noise ratios in stochastic DGPs, as done in Calvori et al. (2017). That would necessitate a precise identification of areas of the parameter space that are free from inferential issues. The topic of multi-collinearity is recently receiving attention (see Duxbury, 2018), but a clear characterization for different values of the ERGM parameters is not yet available.

## 4 Applications to Real Data

After the analysis of synthetic data, this section presents two applications to real dynamical 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. Hence, in this section we show the applications of the framework to two real temporal networks, that have been the object of multiple studies in different streams of literature, and 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 that captures the tendency of U.S senators to cosponsor legislative bills.

#### 4.1 Link Prediction in Interbank Networks with SD beta model

Our first empirical application is to data from the electronic Market of Interbank Deposit (e-MID), a market where 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 (see, for example Iori et al., 2008; Finger et al., 2013; Mazzarisi et al., 2017; Barucca and Lillo, 2018, and references therein). Our dataset contains the list of all credit transactions in each day from June 6, 2009 to February 27, 2015. In our analysis, we investigate 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 description of the dataset, we refer the reader to Barucca and Lillo (2018).

In recent years, the amount of lending in e-MID has significantly declined. In particular, as discussed in Barucca and Lillo (2018), it abruptly declined at the beginning of 2012, as a consequence of important unconventional measures (Long Term Refinancing Operations) by the European Central Bank, that guaranteed an alternative source of liquidity to European banks. This structural change is evident by looking at the evolution of network density in Figure 4. The evident non-stationary nature of the evolution of the inter-bank network is of extreme interest for our purposes. In fact, as mentioned in Sections 3.2 and 3.3, one of the key strengths of SD-ERGM, used as a filter, is precisely the ability of recovering such non-stationary dynamics.

In the following, we use the SD beta model for links forecasting. Specifically, we consider the version with a restricted number of static parameters given in Eq. (11). We divide the data set in two samples. We consider rolling windows of 100 observations and estimate the parameters  $\alpha^{\text{out}}$ ,  $\beta^{\text{out}}$ ,  $\alpha^{\text{in}}$  and  $\beta^{\text{in}}$  of Eq. (11) on each one of those rolling windows. For each window, we then 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  $\overrightarrow{\theta}^{(t)}$  and  $\overrightarrow{\theta}^{(t)}$ , we plug the estimated static parameters and the matrix  $\mathbf{Y}^{(t)}$  in

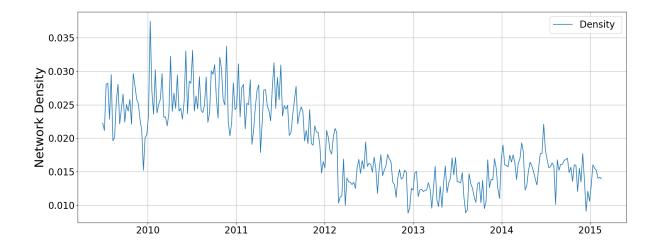


Figure 4: Density of the weekly aggregated interbank networks of overnight loans that occurred in e-MID.

the SD update (11) and compute the time-varying parameters  $\overleftarrow{\theta}^{(t+1)}$  and  $\overrightarrow{\theta}^{(t+1)}$ . From the latter, we readily obtain the forecast of the adjacency matrix

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

where t+1 is the first date of the test sample. The K-step-ahead forecast for the SD-ERGM model is obtained simulating the SD dynamics up to t+K 100 times<sup>12</sup>, thus obtaining  $\overrightarrow{\theta}_n^{(t+K)}$  and  $\overleftarrow{\theta}_n^{(t+K)}$  for  $n=1,\ldots,100$ , and then taking the average of the expected adjacency matrices  $\frac{1}{100}\sum_n\mathbb{E}\left[\mathbf{Y}^{(t+K)}|\overleftarrow{\theta}_n^{(t+K)},\overrightarrow{\theta}_n^{(t+K)}\right]$ . 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 to it the first element of the test sample. We repeat the forecasting exercise estimating the SD-ERGM parameters on the new train set and testing the performance on the new test sample. We name this procedure rolling estimate and iterate it until the test sample contains the last 8 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 Figure 5. 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 other three curves based on the static beta model. Specifically, the green curve results from a naive prediction, where the presence of 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

 $<sup>^{12}</sup>$ It is worth to stress that the results become stable after 20 simulations.

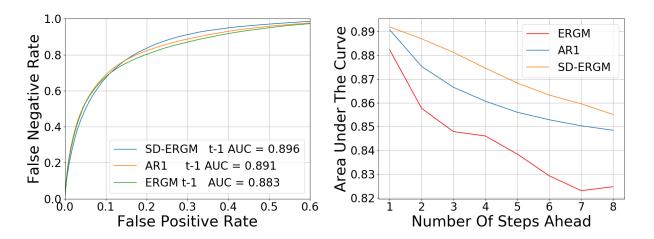


Figure 5: Left panel: ROC curves for one-step-ahead link forecasting. The blue and orange ROC curves describe the one-step-ahead forecasting with SD and cross sectional AR(1) beta model, respectively. The green curve corresponds to the forecast based on previous-time-step ERGM. Right panel: AUC for the multi-step-ahead forecast. The red curve corresponds to the cross sectional ERGM and the blue one to the SD-ERGM.

static ERGM is completed, we take the estimated values  $\widehat{\theta}^{(t)}$  and  $\widehat{\theta}^{(t)}$  as observed and model their evolution with an auto-regressive model of order one, AR(1). That amounts to assuming  $\widehat{\theta}^{(t+1)} = c_0 + c_1 \widehat{\theta}^{(t)} + \epsilon^{(t)}$ , where  $c_0$  and  $c_1$  are the static parameters of the AR(1), and  $\epsilon^{(t)}$  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, in the latter procedure we have to estimate the static parameters for each date in the train sample.

The left plot of Fig. 5 shows that the naive one-step-ahead forecast, in spite of its simplicity, provides a quite 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 it is slightly worst than the SD-ERGM. However, as commented before, it is more computationally intensive. More importantly, the right panel of Fig. 5 presents the result from a multi-step-head forecasting analysis. It emerges clearly that the performance of the naive forecast (red curve), tested up to K=8, rapidly deteriorates, while the SD-ERGM multi-step forecast remains the best performing.

<sup>&</sup>lt;sup>13</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 on the predictability of links in the data set. Importantly, the

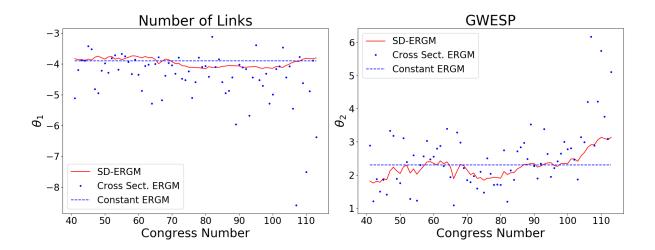


Figure 6: Estimates for the time-varying parameters associated with the number of links and the GWESP statistics. Blue dots correspond to the cross sectional ERGM estimates, while the red lines are the estimates from the SD-ERGM.

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

Networks describing the U.S. congress' bills have been the object of multiple studies (see, for example Fowler, 2006; Faust and Skvoretz, 2002; Zhang et al., 2008; Cranmer and Desmarais, 2011; Moody and Mucha, 2013; Wilson et al., 2016; Roy et al., 2017; Lee et al., 2017; Neal, 2018). 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 (12) can be concretely applied to a real network, and that it draws a different picture when compared to the sequence of cross sectional ERGM estimates. In order to build the network, we use the freely available data of voting records in the US Senate (see Lewis et al., 2019) covering the period from 1867 to 2015, for a total of 74 Congresses. We define the network of co-voting following Roy et al. (2017) and Lee et al. (2017), 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 a sequence of 74 networks, one for each different Congress starting from the 40th. For this empirical application, we consider the SD-ERGM with the two network statistics discussed in Section 3.3. As defined in (13), 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 to single nodes. That case – as for instance considering the degrees of the beta model – would require the number of time-varying parameters to be different at each time step.

ranking of the methods remain unaltered when we keep all links for performance evaluation.

As we did for the numerical simulations and the previous empirical application, we compare our framework with a sequence of standard ERGMs. The goal of this empirical exercise is not to draw conclusions about the specific network at hand. Our main aim is to show that the two approaches return a qualitatively different picture. The choice between the two alternatives – 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. In fact, 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. In order 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  $\theta^c = \arg\max_{\theta} \sum_{t=1}^{74} \log P\left(\mathbf{Y}^{(t)}, \theta\right)$ . 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 behaviors for the parameters. This clearly emerges from Figure 6 that reports the estimates from the SD-ERGM (bold red lines), as well as the cross sectional ERGM estimates – one per date (blue dots) or using the entire sample (dashed blue line).

## 5 Conclusions

In this paper, we proposed a framework for the description of temporal networks that extends the well known Exponential Random Graph Models. In the new approach, the parameters of the ERGM have a stochastic dynamics driven by the score of the conditional likelihood. If the latter is not available 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 by means of an extensive Monte Carlo analysis of the SD-ERGM reliability as a filter for time-varying parameters. 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 not only true 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 networks data. The first application to e-MID interbank network showed that the SD-ERGM provides a quantifiable advantage in a link forecasting exercise over different time horizons. The second example on the U.S. Congress co-voting political network enlightened that the ERGM and the SD-ERGM could provide a significantly different picture in describing the parameter dynamics.

Our work opens a number of 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 that are intrinsic to the ERGM context. Second, the SD-ERGM could be applied on multiple instances of real world dynamical networks. An interesting application would be the study of networks describing the dynamical correlation of neural activity in different parts of the brain (see, for example, Karahanoğlu and Van De Ville, 2017, for a review of the topic and list of references). In this context, the application of the static ERGM have already proven to be extremely successful (as, for example, in Simpson et al., 2011). The last future development that we plan to explore is the extension of the score-driven framework to the description of weighted dynamical network. Regretfully, this setting has not received enough attention in the literature (one isolate example is Giraitis et al., 2016), but it is of extreme relevance, particularly from the systemic risk perspective and its implication for financial stability.

## A Appendix A: Identification Restrictions

The beta model discussed in Section 2.1.1 suffers from an identification issue. Indeed, the probabilities remain unchanged after the application of the following transformation

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

The issue can be tackled by choosing one identification restriction that eliminates the possibility to shift all parameters by an arbitrary constant. 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 alternative choices are available, e.g.  $\sum_{i} \overleftarrow{\theta} = 0$  or  $\overrightarrow{\theta}_{i} = 0$ . However, and most importantly, the results presented in the paper do not change significantly when the identification condition changes.

## B Appendix B: Details about the Numerical Simulations

In this appendix, we describe how to build the vectors  $\overline{\theta}_{0_s}$ ,  $\overline{\theta}_{1_s}$  and  $\overline{\theta}_{2_s}$ , used in the DGPs considered in Section 3.2.

In the numerical experiments of Section 3.2, with N=10, the vector  $\overline{\theta}_0$  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 in- and out-degrees. In practice, we start with an empty matrix, select the largest out degree and set to one the matrix element between this node and the node with largest in degree. 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 to this adiacency matrix is guaranteed to be graphicable. The numerical values of  $\bar{\theta}_0$  follow from the estimation of the static beta model. Finally, in order 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

$$\overline{\theta}_{1_s} = \overline{\theta}_{0_s} + c_s \left( \overline{\theta}_{0_{s+1}} - \overline{\theta}_{0_s} \right)$$

$$\overline{\theta}_{2_s} = \overline{\theta}_{0_s} - c_s \left( \overline{\theta}_{0_{s+1}} - \overline{\theta}_{0_s} \right).$$

In Section 3.2.2, we consider networks of increasing size. We have to fix the vectors  $\overline{\theta}_{0_s}$ ,  $\overline{\theta}_{1_s}$ , and  $\overline{\theta}_{2_s}$  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. the maximum degree and the average degree both increase. This procedure produces the average densities plotted in the left panel of Figure 2.

## References

Albert, R. and A.-L. Barabási (2002, Jan). Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 47–97.

Allen, F. and A. Babus (2011). Networks in finance. In P. R. JKleindorfer, Y. J. R. Wind, and R. E. Gunther (Eds.), *The network challenge: strategy, profit, and risk in an interlinked world*, Chapter 21, pp. 367–379. Social Science Research Network.

Barndorff-Nielsen, O. (2014). Information and exponential families in statistical theory. John Wiley & Sons.

Barucca, P. and F. Lillo (2018). The organization of the interbank network and how ecb unconventional measures affected the e-mid overnight market. *Computational Management Science* 15(1), 33–53.

Baydin, A. G., B. A. Pearlmutter, A. A. Radul, and J. M. Siskind (2018). Automatic differentiation in machine learning: a survey. *Journal of Marchine Learning Research* 18, 1–43.

- Blasques, F., S. J. Koopman, and A. Lucas (2014). Maximum likelihood estimation for generalized autoregressive score models. Technical report, Tinbergen Institute Discussion Paper.
- Blasques, F., S. J. Koopman, and A. Lucas (2015). Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika* 102(2), 325–343.
- Blasques, F., A. Lucas, and A. van Vlodrop (2017). Finite sample optimality of score-driven volatility models. *Tinbergen Institute Discussion Paper*.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31(3), 307 327.
- Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika~39(3/4),~324-345.
- Buccheri, G., G. Bormetti, F. Corsi, and F. Lillo (2018). A general class of score-driven smoothers. Available at SSRN: https://ssrn.com/abstract=3139666.
- Bullmore, E. and O. Sporns (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience* 10(3), 186.
- Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology* 38(1), 155–200.
- Caldarelli, G., A. Capocci, P. De Los Rios, and M. A. Muñoz (2002, Dec). Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.* 89, 258702.
- Calvori, F., D. Creal, S. J. Koopman, and A. Lucas (2017). Testing for parameter instability across different modeling frameworks. *Journal of Financial Econometrics* 15(2), 223–246.
- Chatterjee, S., P. Diaconis, A. Sly, et al. (2011). Random graphs with a given degree sequence. The Annals of Applied Probability 21(4), 1400–1435.
- Cox, D. R., G. Gudmundsson, G. Lindgren, L. Bondesson, E. Harsaae, P. Laake, K. Juselius, and S. L. Lauritzen (1981). Statistical analysis of time series: Some recent developments. Scandinavian Journal of Statistics, 93–115.
- Craig, B. and G. Von Peter (2014). Interbank tiering and money center banks. *Journal of Financial Intermediation* 23(3), 322–347.
- Cranmer, S. J. and B. A. Desmarais (2011). Inferential network analysis with exponential random graph models. *Political Analysis* 19(1), 66–86.
- Creal, D., S. J. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics* 28(5), 777–795.

- Davidson, R., J. G. MacKinnon, et al. (2004). *Econometric theory and methods*. Oxford University Press New York.
- Desmarais, B. A. and S. J. Cranmer (2012). Statistical mechanics of networks: Estimation and uncertainty. *Physica A: Statistical Mechanics and its Applications* 391(4), 1865–1876.
- Duxbury, S. W. (2018). Diagnosing multicollinearity in exponential random graph models. *Sociological Methods & Research*.
- Easley, D., J. Kleinberg, et al. (2010). *Networks, crowds, and markets*, Volume 8. Cambridge University Press Cambridge.
- Engle, R. (2002). New frontiers for arch models. Journal of Applied Econometrics 17(5), 425–446.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987–1007.
- Engle, R. F. and J. R. Russell (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 1127–1162.
- Erdős, P. and A. Rényi (1959). On random graphs i. Publ. Math. Debrecen 6, 290–297.
- Fan, J. and W. Zhang (2008). Statistical methods with varying coefficient models. *Statistics and its Interface* 1(1), 179.
- Faust, K. and J. Skvoretz (2002). Comparing networks across space and time, size and species. Sociological methodology 32(1), 267–299.
- Fienberg, S. E. and S. S. Wasserman (1981). Categorical data analysis of single sociometric relations. Sociological methodology 12, 156–192.
- Finger, K., D. Fricke, and T. Lux (2013). Network analysis of the e-mid overnight money market: the informational value of different aggregation levels for intrinsic dynamic processes. *Computational Management Science* 10(2-3), 187–211.
- Fowler, J. H. (2006). Connecting the congress: A study of cosponsorship networks. *Political Analysis* 14(4), 456–487.
- Frank, O. and D. Strauss (1986). Markov graphs. Journal of the american Statistical association 81(395), 832–842.
- Garlaschelli, D. and M. I. Loffredo (2008, Jul). Maximum likelihood: Extracting unbiased information from complex networks. *Phys. Rev. E* 78, 015101.

- Giraitis, L., G. Kapetanios, A. Wetherilt, and F. Žikeš (2016). Estimating the dynamics and persistence of financial networks, with an application to the sterling money market. *Journal of Applied Econometrics* 31(1), 58–84.
- Goodreau, S. M. (2007). Advances in exponential random graph (p\*) models applied to a large social network. *Social networks* 29(2), 231–248.
- Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica* 85(4), 1033–1063.
- Handcock, M. S. (2003a). Assessing degeneracy in statistical models of social networks. (Working Paper No. 39).
- Handcock, M. S. (2003b). Statistical models for social networks: Inference and degeneracy. In Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers, pp. 229. National Academies Press.
- Hanneke, S., W. Fu, E. P. Xing, et al. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics* 4, 585–605.
- Harvey, A. C. (2013). Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series. Econometric Society Monographs. Cambridge University Press.
- Holland, P. W. and S. Leinhardt (1981). An exponential family of probability distributions for directed graphs. J. Amer. Statistical Assoc. n 76 (373), 33–50.
- Holme, P. and J. Saramäki (2012). Temporal networks. *Physics reports* 519(3), 97–125.
- Hunter, D. R., S. M. Goodreau, and M. S. Handcock (2008). Goodness of fit of social network models. *Journal of the American Statistical Association* 103 (481), 248–258.
- Hunter, D. R. and M. S. Handcock (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics* 15(3), 565–583.
- Hunter, D. R., M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software* 24(3).
- Iori, G., G. De Masi, O. V. Precup, G. Gabbi, and G. Caldarelli (2008). A network analysis of the italian overnight money market. *J. Econ. Dyn. Control* 32(1), 259–278.
- Jackson, M. O. (2010). Social and economic networks. Princeton university press.
- Jaynes, E. (1957, May). Information theory and statistical mechanics. *Phys. Rev.* 106, 620–630.

- Jochmans, K. (2018). Semiparametric analysis of network formation. *Journal of Business & Economic Statistics* 36(4), 705–713.
- Karahanoğlu, F. I. and D. Van De Ville (2017). Dynamics of large-scale fmri networks: Deconstruct brain activity to build better models of brain function. *Current Opinion in Biomedical Engineering* 3, 28–36.
- Kim, B., K. H. Lee, L. Xue, X. Niu, et al. (2018). A review of dynamic network models with latent variables. *Statistics Surveys* 12, 105–135.
- Kolaczyk, E. D. (2009). Statistical Analysis of Network Data: Methods and Models (1st ed.). Springer Publishing Company, Incorporated.
- Kullback, S. (1997). Information theory and statistics. Courier Corporation.
- Lee, J., G. Li, and J. D. Wilson (2017). Varying-coefficient models for dynamic networks. arXiv preprint arXiv:1702.03632.
- Lee, J. H. (1991). A lagrange multiplier test for garch models. *Economics Letters* 37(3), 265–271.
- Lewis, J. B., P. Keith, R. Howard, B. Adam, R. Aaron, and S. Luke (2019). Voteview: Congressional roll-call votes database. https://voteview.com/.
- Mazzarisi, P., P. Barucca, F. Lillo, and D. Tantari (2017). A dynamic network model with persistent links and node-specific latent variables, with an application to the interbank market. arXiv preprint arXiv:1801.00185.
- Moody, J. and P. J. Mucha (2013). Portrait of political party polarization. *Network Science* 1(1), 119–121.
- Neal, Z. P. (2018). A sign of the times? weak and strong polarization in the us congress, 1973–2016. Social Networks.
- Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, 347–370.
- Nelson, D. B. (1996). Asymptotically optimal smoothing with arch models. *Econometrica* 64, 561573.
- Newman, M. (2010). Networks: an introduction. Oxford University Press.
- Park, J. and M. E. J. Newman (2004, Dec). Statistical mechanics of networks. *Phys. Rev. E* 70, 066117.
- Robins, G. and P. Pattison (2001). Random graph models for temporal processes in social networks. Journal of Mathematical Sociology 25(1), 5–41.

- Robins, G., T. Snijders, P. Wang, M. Handcock, and P. Pattison (2007). Recent developments in exponential random graph (p\*) models for social networks. *Social networks 29*(2), 192–215.
- Rossetti, G. and R. Cazabet (2018). Community discovery in dynamic networks: a survey. *ACM Computing Surveys (CSUR)* 51(2), 35.
- Roy, S., Y. Atchadé, and G. Michailidis (2017). Change point estimation in high dimensional markov random-field models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79(4), 1187–1206.
- Schmid, C. S. and B. A. Desmarais (2017). Exponential random graph models with big networks: Maximum pseudolikelihood estimation and the parametric bootstrap. In *Big Data (Big Data)*, 2017 IEEE International Conference on, pp. 116–121. IEEE.
- Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. Journal of the American Statistical Association 106 (496), 1361–1370.
- Schweinberger, M., P. N. Krivitsky, C. T. Butts, and J. Stewart (2018). Exponential-family models of random graphs: Inference in finite-, super-, and infinite-population scenarios.
- Shannon, C. E. (2001, January). A mathematical theory of communication. SIGMOBILE Mob. Comput. Commun. Rev. 5(1), 3–55.
- Shore, J. and B. Lubin (2015). Spectral goodness of fit for network models. *Social Networks* 43, 16 27.
- Simpson, S. L., S. Hayasaka, and P. J. Laurienti (2011). Exponential random graph modeling for complex brain networks. *PloS one* 6(5), e20039.
- Snijders, T. A. (1996). Stochastic actor-oriented models for network change. *Journal of mathematical sociology* 21(1-2), 149–172.
- Snijders, T. A. (2002). Markov chain monte carlo estimation of exponential random graph models. Journal of Social Structure 3(2), 1–40.
- Snijders, T. A., P. E. Pattison, G. L. Robins, and M. S. Handcock (2006). New specifications for exponential random graph models. *Sociological methodology* 36(1), 99–153.
- Soramäki, K., A. Wetherilt, and P. Zimmerman (2010). The sterling unsecured loan market during 2006–08: insights from network theory. *Unpublished working paper. Bank of England*.
- Strauss, D. and M. Ikeda (1990). Pseudolikelihood estimation for social networks. *Journal of the American statistical association* 85(409), 204–212.

- Van Duijn, M. A., K. J. Gile, and M. S. Handcock (2009). A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks* 31(1), 52 62.
- Wasserman, S. and P. Pattison (1996). Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p. *Psychometrika* 61(3), 401–425.
- Wilson, J. D., N. T. Stevens, and W. H. Woodall (2016). Modeling and estimating change in temporal networks via a dynamic degree corrected stochastic block model. arXiv preprint arXiv:1605.04049.
- Yan, T., B. Jiang, S. E. Fienberg, and C. Leng (2018). Statistical inference in a directed network model with covariates. *Journal of the American Statistical Association*, 1–12.
- Yan, T., C. Leng, J. Zhu, et al. (2016). Asymptotics in directed exponential random graph models with an increasing bi-degree sequence. *The Annals of Statistics* 44(1), 31–57.
- Zermelo, E. (1929). Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift 29*(1), 436–460.
- Zhang, Y., A. J. Friend, A. L. Traud, M. A. Porter, J. H. Fowler, and P. J. Mucha (2008). Community structure in congressional cosponsorship networks. *Physica A: Statistical Mechanics and its Applications* 387(7), 1705–1712.