Statistica Sinica Preprint No: SS-2022-0040	
Title	Grouped Network Poisson Autoregressive Model
Manuscript ID	SS-2022-0040
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0040
Complete List of Authors	Yuxin Tao, Dong Li and Xiaoyue Niu
Corresponding Authors	Xiaoyue Niu
E-mails	xiaoyue@psu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

Grouped Network Poisson Autoregressive Model

Yuxin Tao[1], Dong Li[1], and Xiaoyue Niu[2]

[1] *Tsinghua University,* [2] *The Pennsylvania State University*

Abstract: Although multivariate Poisson autoregressive models are popular for fitting count time series data, statistical inferences are quite challenging. The network Poisson autoregressive (NPAR) model reduces the inference complexity by incorporating network information into the dependence structure, where the response of each individual can be explained by its lagged values and the average effect of its neighbors. However, the NPAR model makes the strong assumption that all individuals are homogeneous and share a common autoregressive coefficient. Here, we propose a grouped network Poisson autoregressive (GNPAR) model, in which individuals are classified into groups, using group-specific parameters to describe heterogeneous nodal behaviors. We present the stationarity and ergodicity of the GNPAR model and study the asymptotic properties of the maximum likelihood estimation. We develop an expectation-maximization algorithm to estimate the unknown group labels, and investigate the finite-sample performance of our estimation procedure using simulations. We analyze Chicago Police Investigatory Stop Report data, and find distinct dependence patterns in different neighborhoods of Chicago, which may help with future crime prevention.

Key words and phrases: EM algorithm, Individual heterogeneity, Maximum likelihood estimation, Multivariate Poisson autoregression, Network data.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

1. Introduction

Count time series data are often observed in practice. The monograph of Weiß (2018) summarizes the development of count time series analysis, and Davis et al. (2021) give a comprehensive methodological review. Count time series have unique features, including being integer-valued, over-dispersed, zero-inflated, and time-dependent, and having a nonnegative autocorrelation. Most existing works on count time series modeling focus on univariate cases. For example, Du and Li (1991) propose an integer-valued AR model, and Ferland, Latour, and Oraichi (2006) propose an integer-valued GARCH model, also called the Poisson autoregression (PAR) mdoel. Others have studied variants of the PAR, along with their statistical inference and applications (see, e.g., Fokianos, Rahbek, and Tjøstheim (2009), Fokianos and Tjøstheim (2011, 2012), Neumann (2011), Wang et al. (2014), Ahmad and Francq (2016), and Davis and Liu (2016)). On the other hand, only a few theoretical results are available for multivariate count time series (see Latour (1997), Liu (2012), Pedeli and Karlis (2013), Andreassen (2013), Lee, Lee, and Tjøstheim (2018)), despite their important applications in many fields, such as environmental science, sociology, finance, marketing, and medicine, among others (Mahamunulu (1967), Aitchison and Ho (1989), Karlis and Meligkotsidou (2005, 2007), Weiß (2018), Fokianos et al. (2020), Davis et al. (2021)).

For a multivariate PAR model, a maximum likelihood-based statistical inference

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

is quite challenging, because the probability mass function of a multivariate Poisson random vector usually has a complicated functional form. To circumvent such difficulties, Fokianos et al. (2020) present a copula method, and develop a novel conceptual framework to handle multivariate count time series. To reduce the complex structure of a multivariate count time series model, following Zhu et al. (2017), Armillotta and Fokianos (2021) propose a network PAR (NPAR) model that incorporates the network structure into a multiple or high-dimensional PAR. This technique is widely used to reduce model complexity; see, for example, Zhu et al. (2019a,b, 2020), Huang et al. (2020), Zhou et al. (2020), and Zhu, Cai, and Ma (2021).

The NPAR model assumes that all individuals share a common dependence structure, which is often too stringent in practice. For example, if we consider district-level crime cases in Chicago, we find that crimes occur frequently in some districts, whereas others are relatively safer. Therefore, it is unreasonable to assume that the data-generating mechanism for all districts is the same. Furthermore, the intensity process in the NPAR model regresses only on the lagged observations.

In this paper, we propose a grouped NPAR (GNPAR) model in which individuals are classified into groups, using group-specific parameters to describe heterogeneous nodal behaviors. Such an extension is of both theoretical and methodological importance, because it reduces the computational complexity of general multivariate PAR models, while providing a more realistic and flexible setup and interpretation than

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

those of NPAR models. Moreover, our model involves lags of the intensity process, in addition to the lagged observations and the average effect of its neighbors, which allows greater flexibility.

The remainder of the paper is organized as follows. Section 2 proposes a GNPAR model and gives its stationarity and ergodicity conditions. Section 3 studies the maximum likelihood estimation (MLE) of the GNPAR model, with its asymptotics, when prior information about the group label is known, and develops an expectation-maximization (EM) algorithm to estimate the group ratio and labels when they are unknown. Section 4 reports on our numerical studies conducted to assess the finite-sample performance of our estimation procedure. We study Chicago district-level crime data in Section 5.

Throughout the paper, we denote $\|\mathbf{x}\|_d = \left(\sum_{i=1}^p |x_i|^d\right)^{1/d}$ as the ℓ^d-norm of a p-dimensional vector $\mathbf{x}$. For a $q \times p$ matrix $\mathbf{A} = (a_{ij})$, the generalized matrix norm is defined by $|||\mathbf{A}|||_d = \max_{\|\mathbf{x}\|_d=1} \|\mathbf{A}\mathbf{x}\|_d$. In particular, $|||\mathbf{A}|||_1 = \max_{1\le j\le p} \sum_{i=1}^q |a_{ij}|$ and $|||\mathbf{A}|||_2 = \sqrt{\rho\left(\mathbf{A}'\mathbf{A}\right)}$, where $\rho(\cdot)$ denotes the spectral radius, and $'$ denotes the transpose of a matrix or vector. $|||\mathbf{A}|||_2$ is in fact the operator norm of $\mathbf{A}$. The Frobenius norm of $\mathbf{A}$ is denoted by $|||\mathbf{A}|||_F = \sqrt{\sum_{i,j} |a_{ij}|^2}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

2. GNPAR Model

2.1 Previous Models

Following Lee, Lee, and Tjøstheim (2018) and Fokianos et al. (2020), we assume that $\{\mathbf{Y}_t = (Y_{1t}, Y_{2t}, ..., Y_{Nt})', t \geq 1\}$ is an N-dimensional count time series, and $\{\boldsymbol{\lambda}_t = (\lambda_{1t}, \lambda_{2t}, ..., \lambda_{Nt})', t \geq 1\}$ denotes the corresponding N-dimensional intensity process. Here, N is fixed and finite, and $\mathscr{F}_t^{\mathbf{Y},\boldsymbol{\lambda}}$ is the σ-field generated by $\{\mathbf{Y}_t, \cdots, \mathbf{Y}_0, \boldsymbol{\lambda}_0\}$, with $\boldsymbol{\lambda}_0$ being an initial value of $\{\boldsymbol{\lambda}_t\}$. The multivariate PAR model is defined as follows: for each $i = 1, 2, \ldots, N$ and $t \geq 1$,

$$Y_{i,t} \mid \mathscr{F}_{t-1}^{\mathbf{Y},\boldsymbol{\lambda}} \sim \text{Poisson}\left(\lambda_{i,t}\right), \quad \boldsymbol{\lambda}_t = \mathbf{d} + \mathbb{A}\boldsymbol{\lambda}_{t-1} + \mathbb{B}\mathbf{Y}_{t-1}, \tag{2.1}$$

where $\mathbf{d}$ is an N-dimensional constant vector and $\mathbb{A}, \mathbb{B}$ are $N \times N$ matrices. The elements of $\mathbf{d}, \mathbb{A}$, and $\mathbb{B}$ are assumed to be positive to ensure $\lambda_{i,t} > 0$, for all i and t.

In fact, for general $\mathbb{A}$ and $\mathbb{B}$, a statistical inference of model (2.1) is quite challenging when N is large. To reduce the complexity of model (2.1), following Zhu et al. (2017) and Zhou et al. (2020), we introduce a network structure on the observed counts into model (2.1). Assume a known adjacency matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{N\times N}$ is defined as $a_{ij} = 1$ if there is a directed edge from object i to object j, and $a_{ij} = 0$ otherwise. Let $a_{ii} = 0$, for $1 \leq i \leq N$. An NPAR model assumes that object i is affected only by the objects that it follows. It has the following form: for each

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

$i = 1, 2, \ldots, N$ and $t \geq 1$,

$$Y_{i,t} \mid \mathscr{F}_{t-1}^{\mathbf{Y},\boldsymbol{\lambda}} \sim \text{Poisson}\left(\lambda_{i,t}\right), \tag{2.2}$$
$$\lambda_{i,t} = \omega_0 + \alpha_0 Y_{i,t-1} + \rho_0 d_i^{-1} \sum_{j \neq i} a_{ij} Y_{j,t-1} + \beta_0 \lambda_{i,t-1},$$

where $\sum_{j \neq i}$ means $\sum_{j=1, j \neq i}^{N}$, and $d_i = \sum_{j=1}^{N} a_{ij}$ is the out-degree of i, which is the total number of objects to which i points. If there is no edge starting from object i, that is, $d_i = 0$, we define that $d_i^{-1} \sum_{j \neq i} a_{ij} Y_{j,t-1} = 0$. Note that α_0 measures the dependence on the previous count, ρ_0 measures the dependence on the network structure, that is, the average effect that the neighbors have on each object, and β_0 measures the dependence on the previous intensity. The network structure reduces the inference complexity and makes the model more interpretable. Model (2.2) differs from the NPAR model proposed by Armillotta and Fokianos (2021), because it includes the lags of the intensity process $\boldsymbol{\lambda}_t$.

In model (2.2), however, all individuals are treated homogeneously, because they share the same regression coefficients. This assumption is unrealistic in practice. For instance, the coefficient ρ_0 implies that all individuals are affected by their neighbors to the same extent, whereas in social networks, celebrities are less likely to be influenced by others than normal people are.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

2.2 GNPAR Model

To relax the homogeneous assumption, following Zhu and Pan (2020), we assume that all individuals can be classified into K groups, and each group is characterized by a specific set of parameters $\boldsymbol{\theta}_k = (\omega_k, \alpha_k, \rho_k, \beta_k)' \in \mathbb{R}^4$, for $1 \leq k \leq K$, with each parameter being positive. Define a latent variable $z_{ik} \in \{0, 1\}$ for each object i, where $z_{ik} = 1$ if object i is from the kth group, and $z_{ik} = 0$ otherwise. Assume $\{(z_{i1}, \ldots, z_{iK})', 1 \leq i \leq N\}$ is a sequence of independent and identically distributed (i.i.d.) multinomial random vectors, with number of events $n = 1$ and probability $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)'$. Here, γ_k represents the group proportion satisfying $\gamma_k \geq 0$ and $\sum_{k=1}^{K} \gamma_k = 1$. A GNPAR model can be constructed as

$$\begin{aligned} Y_{i,t} \mid \mathscr{F}_{t-1}^{\mathbf{Y},\boldsymbol{\lambda}} &\sim \text{Poisson}\left(\lambda_{i,t}\right), \\ \lambda_{i,t} &= \sum_{k=1}^{K} z_{ik}\Big(\omega_k + \alpha_k Y_{i,t-1} + \rho_k d_i^{-1} \sum_{j \neq i} a_{ij} Y_{j,t-1} + \beta_k \lambda_{i,t-1}\Big), \end{aligned} \tag{2.3}$$

for each $i = 1, \ldots, N$ and $t \geq 1$. Following the NPAR model, the parameters $\omega_k, \alpha_k, \rho_k$, and β_k represent the group-specific baseline effect, regression coefficient on past observations, network effect, and regression coefficient on past intensity process, respectively. Note that we assume the adjacency matrix $\mathbf{A}$ is asymmetric, which includes the special case of symmetric networks.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

2.3 Stationarity and Ergodicity

In this subsection, we give a stationarity and ergodicity solution to model (2.3). Here, the dimension N is fixed throughout. Let $\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{N,t})'$, $\boldsymbol{\lambda}_t = (\lambda_{1,t}, \ldots, \lambda_{N,t})'$, $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_N)$, and $\mathbf{Z}_k = \operatorname{diag}(z_{ik} : 1 \leq i \leq N)$. Furthermore, define $\mathcal{B}_0 = \sum_{k=1}^{K} \omega_k \mathbf{Z}_k \mathbf{1}_N$, $\mathcal{B}_1 = \sum_{k=1}^{K} \alpha_k \mathbf{Z}_k$, $\mathcal{B}_2 = \sum_{k=1}^{K} \rho_k \mathbf{Z}_k$, and $\mathcal{B}_3 = \sum_{k=1}^{K} \beta_k \mathbf{Z}_k$, where $\mathbf{1}_N = (1, \ldots, 1)'$. Following Fokianos et al. (2020), model (2.3) can be rewritten as

$$\mathbf{Y}_t = \mathbf{N}_t(\boldsymbol{\lambda}_t), \quad \boldsymbol{\lambda}_t = \mathcal{B}_0 + (\mathcal{B}_1 + \mathcal{B}_2 \mathbf{D}^{-1} \mathbf{A}) \mathbf{Y}_{t-1} + \mathcal{B}_3 \boldsymbol{\lambda}_{t-1}, \tag{2.4}$$

where $\{\mathbf{N}_t(\cdot)\}$ is a sequence of independent N-dimensional copula-Poisson processes. See Subsection 4.1 for more details on generating $\mathbf{N}_t(\cdot)$.

Because $\mathbf{Y}_t$ is integer-valued, the ergodicity for model (2.4) is not sufficient to obtain the asymptotics of the parameter estimation, as discussed in Fokianos, Rahbek, and Tjøstheim (2009), Fokianos and Tjøstheim (2011), and Tjøstheim (2012). Thus, ergodicity should be strengthened to geometric ergodicity. However, it is very difficult to establish geometric ergodicity directly, particularly the ϕ-irreducibility of (2.4). To obtain the ϕ-irreducibility, a perturbation method is helpful, that is, adding a continuous component into the innovation; see Chapter 6 of Meyn and Tweedie (1993). Thus, following Fokianos, Rahbek, and Tjøstheim (2009), we define the perturbed model as

$$\mathbf{Y}_t^m = \mathbf{N}_t\left(\boldsymbol{\lambda}_t^m\right), \quad \boldsymbol{\lambda}_t^m = \mathcal{B}_0 + (\mathcal{B}_1 + \mathcal{B}_2 \mathbf{D}^{-1} \mathbf{A}) \mathbf{Y}_{t-1}^m + \mathcal{B}_3 \boldsymbol{\lambda}_{t-1}^m + \boldsymbol{\epsilon}_t^m, \tag{2.5}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

with $\boldsymbol{\epsilon}_t^m = c_m \mathbf{V}_t$, where the sequence c_m is strictly positive and tends to zero as $m \to \infty$, and $\mathbf{V}_t$ is an N-dimensional vector consisting of independent positive random variables, each with a bounded support of the form $[0, M]$, for some $M > 0$. From Lemma 1 in the Supplementary Material, the difference between the unperturbed model (2.4) and the perturbed model (2.5) can be arbitrarily small, in some sense.

The following proposition gives a sufficient condition for the geometric ergodicity of model (2.5), together with a stationary and ergodic condition for model (2.4). The proof of Proposition 1 is postponed to the Supplementary Material.

Proposition 1. (i). *The process* $\{\boldsymbol{\lambda}_t^m, t > 0\}$ *is a geometrically ergodic Markov chain with finite* r*th moments, for any* $r > 0$, *if* $|||\max_{1\le k\le K}(\alpha_k+\beta_k)\mathbf{I}_N+\max_{1\le k\le K}\rho_k\mathbf{D}^{-1}\mathbf{A}|||_2 < 1$. *Moreover, the process* $\{(\mathbf{Y}_t^m, \boldsymbol{\lambda}_t^m, \boldsymbol{\epsilon}_t), t > 0\}$ *is a* $V_{\mathbf{Y},\boldsymbol{\lambda},\boldsymbol{\epsilon}}$*-geometrically ergodic Markov chain with* $V_{\mathbf{Y},\boldsymbol{\lambda},\boldsymbol{\epsilon}} = 1 + ||\mathbf{Y}||_2^r + ||\boldsymbol{\lambda}||_2^r + ||\boldsymbol{\epsilon}||_2^r$, *for* $r > 0$.

(ii). *If* $|||\big(\max_{1\le k\le K}\alpha_k\big)\mathbf{I}_N + \big(\max_{1\le k\le K}\rho_k\big)\mathbf{D}^{-1}\mathbf{A}|||_1 + \max_{1\le k\le K}\beta_k < 1$, *then there exists a unique stationary and ergodic solution* $\{(\mathbf{Y}_t, \boldsymbol{\lambda}_t)\}$ *to model (2.4) that is nonanticipative and satisfies* $\mathrm{E}||\mathbf{Y}_t||_r^r < \infty$ *and* $\mathrm{E}||\boldsymbol{\lambda}_t||_r^r < \infty$, *for any* $r > 0$.

Remark 1. In Proposition 1, (i) is developed using the perturbation technique, and (ii) is based on the notion of weak dependence. The latter does not require a perturbed model, but the obtained sufficient conditions are much stronger. In what follows, we prefer the sufficient stationarity and ergodicity condition (i) for the perturbed process, and use the closeness between the perturbed model and the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

unperturbed one to obtain the asymptotic normality of the MLE of model (2.4). Note that the geometric ergodicity of the perturbed process makes it possible to employ classical statistical inference theory, similarly to GARCH models.

Remark 2. Proposition 1 is constructed using a fixed N, which does not necessarily hold if N is diverging, because no stationarity and ergodicity conditions are available when $\min\{N, T\} \to \infty$. In fact, how to define the stationarity of a time series with a diverging dimension in general remains an open problem. Moreover, the ergodicity conditions in Proposition 1 differ from those of the NPAR model proposed by Armillotta and Fokianos (2021), because the latter does not contain lags of the intensity process $\boldsymbol{\lambda}_t$.

3. Parameter Estimation

This section studies the MLE of the GNPAR model and establishes its asymptotics. Because there exists a latent variable z_{ik} as a group label, the parameter estimation and group classification need to be conducted simultaneously. We first study the MLE of the model parameter when the group labels are known, and then develop an EM algorithm for estimating the group labels when they are unknown. The former is useful if we have prior information for group classification, and the latter is more practical when there is little prior information available.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

3.1 MLE when group labels are known

Suppose z_{ik} is known, and define $\mathcal{G}_k = \{i \leq N : z_{ik} = 1\}$ and $N_k = |\mathcal{G}_k|$, for $1 \leq k \leq K$, denoting the group member and group size, respectively. Assume that the observations $\{\mathbf{Y}_t, t = 1, \ldots, T\}$ are from model (2.4), with true parameter $\boldsymbol{\theta}_0 = (\omega_{k0}, \alpha_{k0}, \rho_{k0}, \beta_{k0} : 1 \leq k \leq K)' \in \mathbb{R}_+^{4K}$, where $\mathbb{R}_+ = (0, \infty)$. Let $\mathbf{Y}_t^{(k)} = (Y_{i,t} : i \in \mathcal{G}_k)' \in \mathbb{R}^{N_k}$, for $t = 1, \ldots, T$, be in the kth group. Define $\boldsymbol{\lambda}_t^{(k)} = (\boldsymbol{\lambda}_{i,t} : i \in \mathcal{G}_k)' \in \mathbb{R}^{N_k}$, $\mathbf{D}^{(k)} = \operatorname{diag}(d_i : i \in \mathcal{G}_k) \in \mathbb{R}^{N_k \times N_k}$, and $\mathbf{A}^{(k)} = (a_{ij} : i \in \mathcal{G}_k, 1 \leq j \leq N) \in \mathbb{R}^{N_k \times N}$. Then, the GNPAR model (2.4) can be rewritten as

$$\mathbf{Y}_t^{(k)} = \mathbf{N}_t\big(\boldsymbol{\lambda}_t^{(k)}\big), \quad \boldsymbol{\lambda}_t^{(k)} = \omega_{k0} + \alpha_{k0}\mathbf{Y}_{t-1}^{(k)} + \rho_{k0}(\mathbf{D}^{(k)})^{-1}\mathbf{A}^{(k)}\mathbf{Y}_{t-1} + \beta_{k0}\boldsymbol{\lambda}_{t-1}^{(k)}, \quad (3.1)$$

for $1 \leq k \leq K$. Under this setting, the true parameter $\boldsymbol{\theta}_{k0} = (\omega_{k0}, \alpha_{k0}, \rho_{k0}, \beta_{k0})'$ can be estimated separately for each group. Without loss of generality, we consider the MLE for the kth group hereafter.

Let $\boldsymbol{\theta}_k = (\omega_k, \alpha_k, \rho_k, \beta_k)' \in \mathbb{R}_+^4$ be the parameter. The conditional likelihood function, given $\boldsymbol{\lambda}_0$, is given by

$$L(\boldsymbol{\theta}_k) = \prod_{t=1}^{T} \prod_{i \in \mathcal{G}_k} \left\{ \frac{\lambda_{i,t}^{Y_{i,t}}(\boldsymbol{\theta}_k) \exp\left(-\lambda_{i,t}(\boldsymbol{\theta}_k)\right)}{Y_{i,t}!} \right\}, \quad (3.2)$$

and the log-likelihood function (ignoring the constant) is

$$l(\boldsymbol{\theta}_k) = \frac{1}{T}\sum_{t=1}^{T} l_t(\boldsymbol{\theta}_k), \quad l_t(\boldsymbol{\theta}_k) = \frac{1}{N_k}\sum_{i \in \mathcal{G}_k} \left(Y_{i,t} \log \lambda_{i,t}(\boldsymbol{\theta}_k) - \lambda_{i,t}(\boldsymbol{\theta}_k)\right). \quad (3.3)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

The MLE of $\boldsymbol{\theta}_{k0}$ is defined as

$$\widehat{\boldsymbol{\theta}}_k = (\widehat{\omega}_k, \widehat{\alpha}_k, \widehat{\rho}_k, \widehat{\beta}_k)' = \arg\max_{\boldsymbol{\theta}_k \in \Theta_k} l(\boldsymbol{\theta}_k). \tag{3.4}$$

Let $\boldsymbol{\Theta} := \boldsymbol{\Theta}_1 \times \cdots \times \boldsymbol{\Theta}_K \subset \mathbb{R}_+^{4K}$ be the parameter space and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Before we study the asymptotics of $\widehat{\boldsymbol{\theta}}_k$, we first give two assumptions.

Assumption 1. *The parameter space $\boldsymbol{\Theta}$ is a compact set of $\mathbb{R}_+^{4K}$, and the true parameter $\boldsymbol{\theta}_0$ is an interior point of Θ.*

Assumption 2. *$\boldsymbol{\theta}_0$ satisfies $|||\max_{1\le k\le K}(\alpha_{k0}+\beta_{k0})\mathbf{I}_N + \max_{1\le k\le K}\rho_{k0}\mathbf{D}^{-1}\mathbf{A}|||_2 < 1$.*

The following theorem states the strong consistency and asymptotic normality of the MLE $\widehat{\boldsymbol{\theta}}_k$.

Theorem 1. *If Assumptions 1–2 hold, then there exists an open neighborhood, say, $O(\boldsymbol{\theta}_{k0}) = \{\boldsymbol{\theta}_k : \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k0}\|_2 < \delta\}$, of $\boldsymbol{\theta}_{k0}$ such that, with probability tending to one as $T \to \infty$, the equation $S_T(\boldsymbol{\theta}_k) = 0$ has a unique solution, denoted by $\widehat{\boldsymbol{\theta}}_k$. Furthermore, $\widehat{\boldsymbol{\theta}}_k$ is strongly consistent, that is, $\widehat{\boldsymbol{\theta}}_k \to \boldsymbol{\theta}_{k0}$ a.s., and is asymptotically normal, that is, $\sqrt{N_k T}\big(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{k0}\big) \xrightarrow{\mathrm{d}} N\left(0, \mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}\right)$, as $T \to \infty$, where "$\xrightarrow{\mathrm{d}}$" stands for convergence in distribution. The matrices $\mathbf{G} := \mathbf{G}(\boldsymbol{\theta}_{k0})$ and $\mathbf{H} := \mathbf{H}(\boldsymbol{\theta}_{k0})$ are defined by*

$$\mathbf{G}(\boldsymbol{\theta}_{k0}) = \frac{1}{N_k}\sum_{i\in\mathcal{G}_k}\sum_{j\in\mathcal{G}_k}\mathrm{E}\left\{\frac{1}{\lambda_{i,t}(\boldsymbol{\theta}_{k0})\lambda_{j,t}(\boldsymbol{\theta}_{k0})}\boldsymbol{\Sigma}_{ij,t}^{(k)}(\boldsymbol{\theta}_{k0})\frac{\partial\lambda_{i,t}(\boldsymbol{\theta}_{k0})}{\partial\boldsymbol{\theta}_k}\frac{\partial\lambda_{j,t}(\boldsymbol{\theta}_{k0})}{\partial\boldsymbol{\theta}_k'}\right\} \text{ and}$$

$$\mathbf{H}(\boldsymbol{\theta}_{k0}) = \frac{1}{N_k}\sum_{i\in\mathcal{G}_k}\mathrm{E}\left\{\frac{1}{\lambda_{i,t}(\boldsymbol{\theta}_{k0})}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta}_{k0})}{\partial\boldsymbol{\theta}_k}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta}_{k0})}{\partial\boldsymbol{\theta}_k'}\right\}, \text{ and}$$

$$\frac{\partial\boldsymbol{\lambda}_t^{(k)\prime}(\boldsymbol{\theta}_{k0})}{\partial\boldsymbol{\theta}_k} = \left(\mathbf{1}_{N_k}, \mathbf{Y}_{t-1}^{(k)}, (\mathbf{D}^{(k)})^{-1}\mathbf{A}^{(k)}\mathbf{Y}_{t-1}, \boldsymbol{\lambda}_{t-1}^{(k)}\right)' + \beta_{k0}\frac{\partial\boldsymbol{\lambda}_{t-1}^{(k)\prime}(\boldsymbol{\theta}_{k0})}{\partial\boldsymbol{\theta}_k},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

where $\mathbf{\Sigma}_t^{(k)}(\boldsymbol{\theta}_{k0})$ *is the covariance matrix of* $\mathbf{Y}_t^{(k)}$, $\mathbf{\Sigma}_{ij,t}^{(k)}(\cdot)$ *is the* (i,j)*th entry of* $\mathbf{\Sigma}_t^{(k)}(\cdot)$, *and the expectation is taken with respect to the invariant stationary distribution of* $\{\mathbf{Y}_t^{(k)}\}$.

The proof of Theorem 1 is postponed to the Supplementary Material. Note that when the components of the process $\left\{\mathbf{Y}_t^{(k)}\right\}$ are uncorrelated, we have $\mathbf{G} = \mathbf{H}$, and thus the asymptotic covariance matrix reduces to the standard one for the ordinary MLE. In practice, the above quantities can all be consistently estimated using their respective sample counterparts, for example,

$$\widehat{\mathbf{H}} = \frac{1}{N_k T} \sum_{i \in \mathcal{G}_k} \sum_{t=1}^{T} \left\{ \frac{1}{\lambda_{i,t}(\widehat{\boldsymbol{\theta}}_k)} \frac{\partial \lambda_{i,t}(\widehat{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\theta}_k} \frac{\partial \lambda_{i,t}(\widehat{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\theta}_k'} \right\}.$$

Remark 3. From Theorem 1, the convergence rate depends on both N_k and T, although N_k is fixed. The network structure characterized by $\mathbf{A}$ and the number of groups K are fixed in our model setting. There are no additional assumptions on the network structure. Because the parameters $\boldsymbol{\theta}_k$ from different groups are uncorrelated in the asymptotic covariance matrix, the MLE of $\boldsymbol{\theta}_{k0}$ for each group can be conducted separately. Thus, the fixed group number K does not affect the convergence rate in each group's estimation, whereas a larger value may consume more computational time, because more unknown parameters are involved. Furthermore, if N_k is diverging, Theorem 1 may break. In this case, we need to impose some connectivity and uniformity assumptions on the network structure and some regularity assumptions

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

on the structure of the dependence between the errors; see, for example, Zhu et al. (2017) and Armillotta and Fokianos (2021). This is left to future research.

3.2 Estimation with unknown group labels

When the group labels are unknown, the estimation includes the latent variables. A common method for dealing with such mixture models is the EM algorithm. Recall that $z_{ik} \in \{0, 1\}$ indicates whether object i belongs to the kth group. The full likelihood function is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left\{ \gamma_k \prod_{t=1}^{T} \frac{\lambda_{i,t}^{Y_{i,t}}(\boldsymbol{\theta}_k) \exp\left(-\lambda_{i,t}(\boldsymbol{\theta}_k)\right)}{Y_{i,t}!} \right\}^{z_{ik}}. \tag{3.5}$$

The EM algorithm consists of two steps: an expectation step, and a maximization step. First, we set initial values for the parameters $\widehat{\boldsymbol{\theta}}^{(0)}$ and $\widehat{\boldsymbol{\gamma}}^{(0)}$, and follow the procedure described below. Specifically, in the mth ($m \geq 1$) iteration, the estimation procedure is as follows:

- E-STEP. Estimate z_{ik} using its posterior mean $z_{ik}^{(m)}$. Here,

$$z_{ik}^{(m)} = \mathrm{E}\big(z_{ik} \big| \widehat{\boldsymbol{\theta}}^{(m-1)}\big) = \frac{\widehat{\gamma}_k^{(m-1)} \prod_{t=1}^{T} \widehat{\Delta}_{it,k}^{(m-1)}}{\sum_{j=1}^{K} \widehat{\gamma}_j^{(m-1)} \prod_{t=1}^{T} \widehat{\Delta}_{it,j}^{(m-1)}}, \tag{3.6}$$

  where $\widehat{\Delta}_{it,k}^{(m-1)} = \lambda_{i,t}^{Y_{i,t}}\big(\widehat{\boldsymbol{\theta}}_k^{(m-1)}\big) \exp\big(-\lambda_{i,t}\big(\widehat{\boldsymbol{\theta}}_k^{(m-1)}\big)\big)$ (omitting the constant term), and $\widehat{\boldsymbol{\theta}}_k^{(m-1)}$ is an estimate in the $(m-1)$th iteration.

- M-STEP. Given an estimate $z_{ik}^{(m)}$, we maximize the following Q-function with

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

respect to $\boldsymbol{\theta}_k$ and γ_k (ignoring the constant term):

$$\begin{aligned} Q\big(\boldsymbol{\theta}\big|\widehat{\boldsymbol{\theta}}^{(m-1)}\big) &= \mathrm{E}\left\{\log L(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{z})\big|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(m-1)}\right\} \\ &= \mathrm{E}\left\{\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\Big(\log\gamma_k + \sum_{t=1}^{T}\left(Y_{i,t}\log\lambda_{i,t}(\boldsymbol{\theta}_k) - \lambda_{i,t}(\boldsymbol{\theta}_k)\right)\Big)\Big|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(m-1)}\right\} \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}^{(m)}\Big(\log\gamma_k + \sum_{t=1}^{T}\left(Y_{i,t}\log\lambda_{i,t}(\boldsymbol{\theta}_k) - \lambda_{i,t}(\boldsymbol{\theta}_k)\right)\Big). \end{aligned}$$

Thus, we have

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_k^{(m)} &= \arg\max_{\boldsymbol{\theta}_k\in\Theta_k}\left\{\sum_{i=1}^{N} z_{ik}^{(m)}\sum_{t=1}^{T}\left(Y_{i,t}\log\lambda_{i,t}(\boldsymbol{\theta}_k) - \lambda_{i,t}(\boldsymbol{\theta}_k)\right)\right\}, \\ \widehat{\gamma}_k^{(m)} &= \frac{1}{N}\sum_{i=1}^{N} z_{ik}^{(m)}. \end{aligned} \tag{3.7}$$

Repeat steps (3.6)–(3.7) until the EM algorithm converges, yielding the EM estimates $\widehat{\boldsymbol{\theta}}_k$ and $\widehat{\gamma}_k$, for $1 \le k \le K$. Note that the EM estimator $\widehat{\boldsymbol{\theta}}_k$ given in (3.7) can be viewed as a weighted MLE estimator in (3.4), with the latent group variables z_{ik} as the weights.

Remark 4. In practice, the computation of the E-Step (3.6) may be unstable and sensitive to initial values, especially when the sample size T is large, which leads to unsatisfactory performance of the estimator $\widehat{\boldsymbol{\gamma}}$. To address this problem, we adopt the two-step (TS) estimation method introduced by Zhu and Pan (2020) to set the initial value. Specifically, we first estimate the coefficient parameter $\boldsymbol{\theta}$ at the nodal level and obtain N sets of MLE $\widehat{\boldsymbol{\theta}}_k$, for $1 \le k \le N$. Next, we apply some cluster

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

algorithm (e.g., k-means clustering) to partition these N sets of estimates into K groups. Let $\widehat{\mathcal{G}_k}$ be the corresponding members in group k, and $\widehat{N}_k := |\widehat{\mathcal{G}_k}|$ be the cardinality. Then, the initial value of the group proportion γ_k can be estimated as $\widehat{\gamma}_k^{(0)} = \widehat{N}_k/N$, for all $1 \leq k \leq K$. Finally, given the group information $\widehat{\mathcal{G}}$, we estimate $\boldsymbol{\theta}$ using the MLE (3.4), and set the estimate as the initial value of $\boldsymbol{\theta}$, that is, $\widehat{\boldsymbol{\theta}}^{(0)}$.

Remark 5. How to select a reasonable number of groups K is a long-standing problem. Here, we recommend two procedures to determine K.

The first is from the perspective of the model setting. Because nodes in the same group are characterized by the same set of parameters $\boldsymbol{\theta}_k$, we can adopt the TS estimation method introduced in **Remark 4** and classify the estimated parameters using classical cluster algorithms. Specifically, we first estimate the coefficient parameter $\boldsymbol{\theta}$ at the nodal level and obtain N sets of ML estimates $\widehat{\boldsymbol{\theta}}_k$, for $1 \leq k \leq N$. Then, we apply k-means clustering to partition these N sets of estimates into K groups, and select an optimal K based on the elbow plot, silhouette coefficient, or gap statistic.

The second procedure is from the perspective of model fitting. As discussed in Section 4.3, we can try model fitting with different numbers of groups, say $K = 1, \ldots, 5$, then compare their out-of-sample predicted RMSEs in (4.2) among candidate models, and choose a reasonable K from the RMSEs. The in-sample fitted RMSEs in (4.1) can also be used as an auxiliary measure to select K.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

4. Simulation Studies

In this section, we investigate the finite-sample performance of the proposed model and estimation procedure. We first explore the performance of the MLE of $\boldsymbol{\theta}$ when the group labels are unknown. Then, we evaluate the model estimation and prediction accuracy when the number of groups K is misspecified. The performance of the MLE of $\boldsymbol{\theta}$ when the group labels are known is reported in the Supplementary Material S2.1.

4.1 Simulated data

We first generate the adjacency matrix $\mathbf{A}$ using two mechanisms: the Erdös–Rényi model, and the stochastic blockmodel. These mechanisms are chosen to illustrate the performance of our model under different network structures $\mathbf{A}$, and are independent of the membership-generating mechanism. Note that directed graphs are considered here, which include undirected graphs.

Case 1: The Erdös–Rényi model

The Erdös–Rényi model (Erdös and Rényi, 1960) is the most thoroughly studied network model in the literature. It assumes that given a number of vertices N_v, all edges are independent with a given probability $p \in (0, 1)$. The Erdös–Rényi model has the property that for large N_v, the degree distribution of the graph is approximately Poisson distributed with mean $p(N_v - 1)$. Here, we set $p = 3/N_v$. Loops are not allowed. A visualization of the network structure and the histogram

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

of the degree distributions of one realization are shown in Fig. 1, with $N_v = 50$.

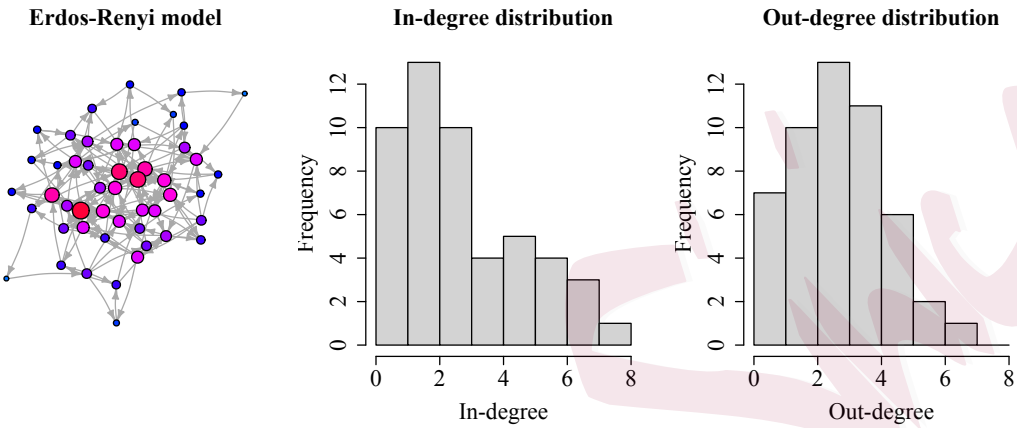


Figure 1: The network visualization and histogram of the degree for the Erdös–Rényi model, with $N_v = 50$ and $p = 0.06$.

Case 2: Stochastic Blockmodel

The stochastic blockmodel (Holland, Laskey, and Leinhardt, 1983) is another popular network topology. It assumes that the nodes in the same block are more likely to be connected to each other than they are to those from different blocks. Here, we set $K_v = \{3, 5, 10\}$ as the total number of blocks, and $N_v = \{20, 50, 100\}$ as the total number of nodes, with each block having N_v/K_v nodes. We assume there is a directed edge to every pair of vertices with probability $3K_v/N_v$ if they belong to the same community, and $0.3/N_v$ for those in different communities. A network

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

visualization and histogram of the degree with $N_v = 50$ and $K_v = 5$ are shown in Fig. 2.

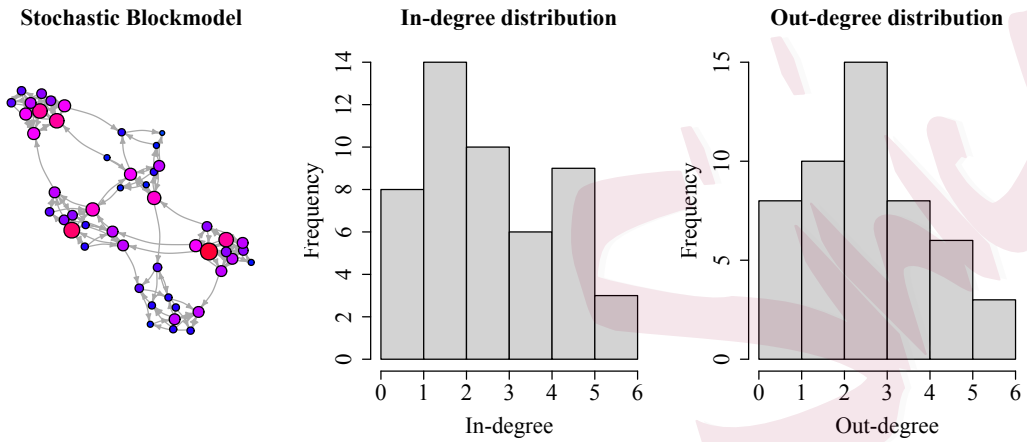


Figure 2: The network visualization and histogram of the degree for the stochastic blockmodel, with $N_v = 50$ vertices and $K_v = 5$ blocks.

We generate the adjacency matrix $\mathbf{A}$ once, after which it is fixed throughout the remaining simulation studies. We set the number of groups to $K = 3$. To characterize different nodal behaviors, we set different parameters $\boldsymbol{\theta}_k = (\omega_k, \alpha_k, \rho_k, \beta_k)'$ for each group, as listed in Table 1. Group 1 has relatively low regression coefficients on past observations and the past intensity process (i.e., α and β), but a relatively high network effect (i.e., ρ), implying that the behavior of most objects is affected by the objects they follow. In contrast, Group 3 has a lower ρ and a higher ω, α, and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

β, representing influential nodes that are more likely to be self-driven and rely less on others. They account for 20% of the objects. Group 2 has medium coefficients and a medium group size. Note that the parameters satisfy Assumptions 1–2. We randomly assign each node to the K groups based on the group proportion γ_k.

Table 1: True parameters in model (2.3) for each group, with $K = 3$.

	ω	α	ρ	β	γ
Group 1	0.2	0.1	0.3	0.2	0.5
Group 2	0.5	0.2	0.2	0.3	0.3
Group 3	1	0.3	0.1	0.4	0.2

Given an initial value $\boldsymbol{\lambda}_0 = \mathbf{4}$, the observed time series $\mathbf{Y}_0$ are generated as an N-dimensional count time series with intensity $\boldsymbol{\lambda}_0$, and $\{\mathbf{Y}_t, \boldsymbol{\lambda}_t; t \geq 1\}$ are simulated consecutively, conditioned on the previous information, using the GNPAR model (2.3). The first 50 samples are discarded to eliminate the effect of initial values. As mentioned before, to establish a well-defined joint distribution of multivariate count data with a marginal Poisson distribution, we use the copula-based data-generating process introduced in Fokianos et al. (2020). This process has the advantage that the copula is defined uniquely for continuous multivariate random variables, and it imposes arbitrary dependence among the marginal Poisson components. Denote $\{\mathbf{N}_t\}$ as a sequence of independent N-dimensional copula-Poisson processes. The

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

data-generating process is given below. Suppose that $\boldsymbol{\lambda}_0 = (\lambda_{1,0}, \ldots, \lambda_{p,0})'$ is an initial value.

(1) Let $\mathbf{U}_l = (U_{1,l}, \ldots, U_{N,l})'$, for $l = 1, \ldots, S$, be a sample from an N-dimensional copula $C(u_1, \ldots, u_N)$, where $U_{i,l}$, for $l = 1, \ldots, S$, follows the uniform distribution $U(0, 1)$, for $i = 1, \ldots, N$.

(2) Consider the transformation $X_{i,l} = -(\log U_{i,l})/\lambda_{i,0}$, for $i = 1, \ldots, N$. Then, $X_{i,l}$, for $l = 1, \ldots, S$, follows an exponential distribution with parameter $\lambda_{i,0}$, for $i = 1, \ldots, N$.

(3) Define $Y_{i,0} = \max\left\{0 \leq s \leq S : \sum_{l=1}^{s} X_{i,l} \leq 1\right\}$, for $i = 1, \ldots, N$, by taking S large enough. Then, $\mathbf{Y}_0 = (Y_{1,0}, \ldots, Y_{N,0})'$ is a set of marginal Poisson processes with parameter $\boldsymbol{\lambda}_0$.

(4) Use model (2.3) to obtain $\boldsymbol{\lambda}_1$, return to step (1) to obtain $\mathbf{Y}_1$, and so on.

In practice, the sample size S should be large, say $S = 1000$. The copula $C(\cdot)$ can be chosen as the Gaussian or the Clayton copula, and the unknown parameter of the copula, say ϕ, needs to be determined based on the contemporaneous correlation among the random variables. A parametric bootstrap-based algorithm can be used to identify the copula structure and unknown parameter; see S-7 in Fokianos et al. (2020). In this section, we employ the Gaussian copula with parameter $\phi = 0.5$, allowing for arbitrary dependence among the marginal Poisson components.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

4.2 Simulation results when group labels are unknown

To assess the finite-sample performance of the MLE when the group labels are unknown, we apply the EM algorithm to estimate $\boldsymbol{\theta}_0$ and $\boldsymbol{\gamma}_0$ simultaneously. The initial value is set using the TS estimation method described in **Remark 4**. Two types of network structures are considered, each with combinations of network size (i.e., $N = 20, 50, 100$) and sample size (i.e., $T = 100, 200, 400$). Each case is randomly simulated with $R = 1000$ replicates. Denote the estimates obtained in the rth simulation as $\widehat{\boldsymbol{\theta}}^{(r)} = (\widehat{\boldsymbol{\omega}}^{(r)}, \widehat{\boldsymbol{\alpha}}^{(r)}, \widehat{\boldsymbol{\rho}}^{(r)}, \widehat{\boldsymbol{\beta}}^{(r)})'$ and $\widehat{\boldsymbol{\gamma}}^{(r)}$, where $1 \leq r \leq R$. Moreover, the group label for each node is estimated as $\widehat{z}_i^{(r)} = \arg\max_k\{\widehat{z}_{ik}^{(r)}\}$. The simulation results are summarized in Tables 2–3 for the Erdös–Rényi model and stochastic blockmodel, respectively.

First, the RMSE is calculated for each estimator. Here, we report the average RMSE taken over all groups. For example, for the network effect coefficient $\boldsymbol{\rho}$, $\text{RMSE}_\rho = \{(KR)^{-1}\sum_{k=1}^{K}\sum_{r=1}^{R}(\widehat{\rho}_k^{(r)} - \rho_k)^2\}^{1/2}$. For the group ratio $\boldsymbol{\gamma}$, $\text{RMSE}_\gamma = \{(KR)^{-1}\sum_{k=1}^{K}\sum_{r=1}^{R}(\widehat{\gamma}_k^{(r)} - \gamma_k)^2\}^{1/2}$. Next, we employ the misclassification rate (MCR) to evaluate the accuracy of the estimated group label. Specifically, $\text{MCR} = (NR)^{-1}\sum_{r=1}^{R}\sum_{i=1}^{N} I(\widehat{z}_i^{(r)} \neq z_i)$, where z_i is the true group label of object i. The last column calculates the network density, which is defined as $\{N(N-1)\}^{-1}\sum_{i,j} a_{ij}$.

Tables 2–3 show that the RMSEs are all very small for the estimators $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\rho}}, \widehat{\boldsymbol{\beta}}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

and $\widehat{\boldsymbol{\gamma}}$. For the baseline effect estimator $\widehat{\omega}_k$, the RMSEs are relatively large. As the network dimension N and sample size T increase, the RMSEs of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\gamma}}$ decrease toward zero, which implies more accurate estimates and smaller standard deviations. Moreover, the misclassification rates of the group labels are quite small, and decrease rapidly as the network size and sample size increase. These facts indicate the good performance of the MLE and the effectiveness of the EM algorithm.

Table 2: Simulation results for the Erdös–Rényi model. The RMSEs ($\times 10^2$) are reported with the misclassification rate (%) and the network density (%).

N	T	$\boldsymbol{\omega}$	$\boldsymbol{\alpha}$	$\boldsymbol{\rho}$	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	MCR	Network Density
20	100	39.09	7.24	9.70	18.19	5.28	7.39	13.16%
	200	23.93	5.14	7.00	11.33	3.90	3.37	
	400	16.98	4.00	5.73	8.12	3.33	2.95	
50	100	27.44	4.72	6.51	14.99	2.85	2.49	7.10%
	200	19.59	3.37	4.72	8.87	2.51	0.96	
	400	14.53	2.63	3.21	6.25	2.72	1.07	
100	100	21.86	3.96	4.53	12.66	2.45	1.45	3.10%
	200	14.84	2.60	2.90	7.56	1.80	0.37	
	400	9.13	1.81	2.05	5.20	0.96	0.19	

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

Table 3: Simulation results for the stochastic blockmodel. The RMSEs ($\times 10^2$) for each estimator are reported with the misclassification rate (%) and the network density (%).

N	T	$\boldsymbol{\omega}$	$\boldsymbol{\alpha}$	$\boldsymbol{\rho}$	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	MCR	Network Density
20	100	43.03	7.00	7.22	16.89	5.42	8.98	13.16%
	200	28.10	4.60	4.76	10.16	4.27	3.03	
	400	15.32	3.39	3.61	6.84	2.87	2.02	
50	100	26.56	5.12	6.69	15.12	3.46	3.30	6.24%
	200	15.80	3.11	4.39	8.36	2.29	0.60	
	400	11.70	2.49	3.22	6.15	2.20	0.73	
100	100	22.86	4.03	4.38	12.55	2.20	1.31	2.92%
	200	15.63	2.74	2.78	7.59	2.39	0.62	
	400	10.58	1.91	2.03	5.40	1.58	0.25	

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

4.3 Model performance when the number of groups K is misspecified

Thus far, we have set the group number as $K = 3$. However, in reality, the true number of groups is unknown and could be specified incorrectly. In this subsection, we study the effects of such a misspecification on the model estimation and prediction accuracy.

The data are generated under the stochastic blockmodel, and the true number of groups is $K = 3$, with the same parameters as those in Table 1. We choose $K = 1, 2, 4, 5$ as the misspecified number of groups. The network size is $N = 20, 50, 100$, and the sample size is $T = 100, 200, 400$, each with $R = 1000$ replicates. The total period of the generated data is $T + 20$, where the first T periods are used for the parameter estimation, and the remaining 20 periods are used for the prediction.

For each selected number of groups K, denote $\widehat{\mathbf{Y}}_t$ as the fitting response for $t = 1, \ldots, T$, and the predicted value for $t = T+1, \ldots, T+20$. Because the parameter estimation error cannot be defined naturally when the number of groups is incorrect, we employ the estimation error of the response instead to compare the performance of the model with different K. The in-sample RMSE for the fitted value is defined as

$$\text{RMSE}_{esti} = \left\{ (NT)^{-1} \sum_{t=1}^{T} ||\widehat{\mathbf{Y}}_t - \text{E}(\mathbf{Y}_t|\mathcal{F}_{t-1}, \mathbf{Z})||^2 \right\}^{1/2}, \tag{4.1}$$

where $\text{E}(\mathbf{Y}_t|\mathcal{F}_{t-1}, \mathbf{Z})$ is the conditional expectation of the response $\mathbf{Y}_t$ based on the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

historical and group information, which is equal to $\boldsymbol{\lambda}_t$ in our model. The out-of-sample predictive RMSE is defined as

$$\text{RMSE}_{pred} = \Big\{(20N)^{-1} \sum_{t=T+1}^{T+20} ||\widehat{\mathbf{Y}}_t - \mathbf{Y}_t||^2\Big\}^{1/2}. \tag{4.2}$$

The mean values of these statistics are summarized in Table 4. We can see that both the estimation errors and the prediction errors shrink sharply from $K \leq 2$ to the true value $K = 3$ in all scenarios, and decrease smoothly as K increases. In particular, the prediction errors remain steady for $K \geq 3$. Therefore, in practice, we could try model fitting with different numbers of groups, say $K = 1, \ldots, 5$, compare the prediction errors among the candidate models, and then select a reasonable number of groups K. This confirms the effectiveness of the second K-selection method in **Remark 5**. We verify the performance of the first method using simulations; see the Supplementary Material S2.2.

5. Case study: Chicago Police Department Investigatory Stop Report (ISR) data

5.1 Data description

Here, we apply the proposed methodology to crime data from Chicago. Chicago is one of the most racially and socio-economically segregated cities in America, and its crime rate remains high, even by worldwide standards. We use data from the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

Table 4: Simulation results for different numbers of groups K in stochastic block-models with 500 replicates. The RMSE$_{esti}$($\times 10^2$) and the RMSE$_{pred}$ are reported.

N	T	Estimation					Prediction				
		K=1	K=2	K=3	K=4	K=5	K=1	K=2	K=3	K=4	K=5
20	100	50.68	31.30	19.81	20.04	20.97	1.50	1.45	1.43	1.43	1.44
	200	49.94	29.08	13.49	14.18	14.97	1.49	1.44	1.42	1.42	1.42
	400	49.54	28.02	10.56	10.69	11.34	1.50	1.45	1.42	1.42	1.42
50	100	45.19	27.87	14.33	14.44	15.21	1.38	1.34	1.32	1.32	1.32
	200	44.51	26.04	9.56	10.00	10.60	1.39	1.34	1.32	1.32	1.32
	400	44.22	25.39	6.68	6.80	7.21	1.39	1.34	1.32	1.32	1.32
100	100	46.55	26.78	12.79	13.15	13.70	1.42	1.37	1.35	1.35	1.35
	200	45.90	26.68	8.52	8.96	9.41	1.40	1.36	1.34	1.34	1.34
	400	45.59	25.52	5.94	6.21	6.52	1.42	1.37	1.35	1.35	1.35

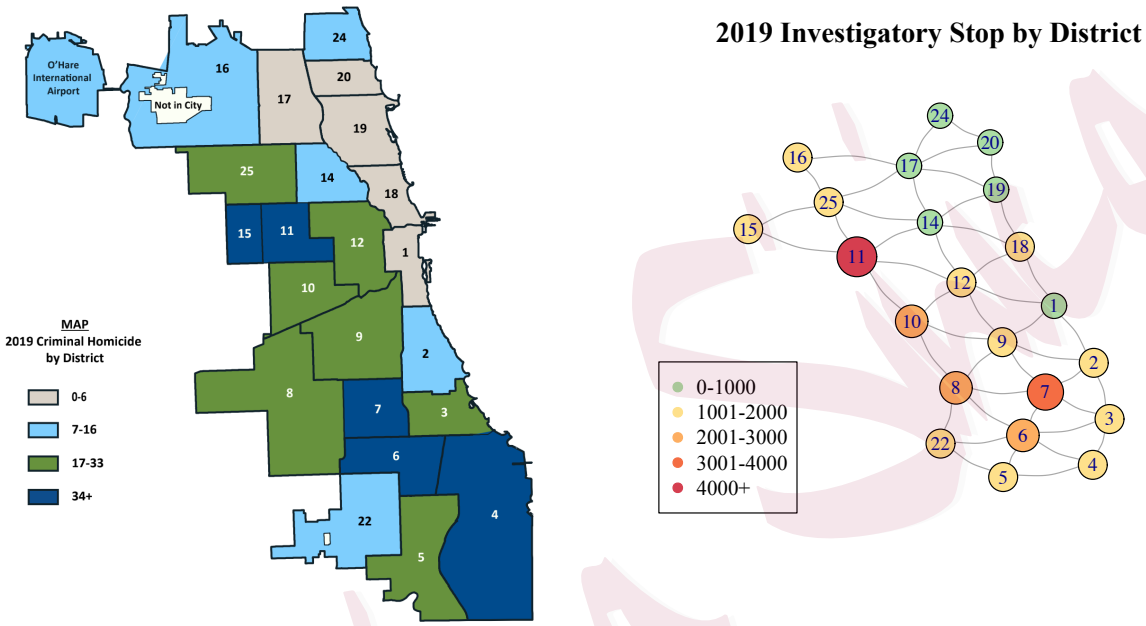
Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

Chicago Police Department Investigatory Stop Report (ISR). A police officer can perform an investigatory stop if there are specific and articulable facts leading to a suspicion of criminal activity. Thus, the number of investigatory stops in which any enforcement action was taken can be viewed as a measure of the crime index in an area. Here, we study the dynamic and spatial patterns of investigatory stops and how the crime numbers from different districts interact with each other, which will be helpful in crime prevention and policy making.

We consider the number of daily investigatory stops that involve an enforcement action (arrest, personal service citation, etc.) in each district in 2019 ($T = 365$) as the response Y_{it}. The data set is taken from the public data of the Chicago Police Department, named "ISR Data 2019" (https://home.chicagopolice.org/statistics-data/isr-data/). The Chicago Police Department divides the city into $N = 22$ districts, as shown in Fig. 3 (a). The figure shows the criminal homicide distribution by district in 2019, where darker colors represent districts with relatively more criminal homicide cases. We see that the crime rate is high in the middle and southern part of Chicago, and relatively low in the northeast areas. Fig. 3 (b) shows the construction of the symmetric adjacency matrix, which is based on the spatial distribution of the districts, that is, there is an edge between district i and j if the two districts share a border. The network density is 19.9%. Larger nodes indicate more investigatory stops in these districts, and smaller nodes denote fewer stops. We can see that the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

distributions of the investigatory stops and criminal homicide are very similar.



(a) (b)

Figure 3: (a) District division in Chicago (2019 Criminal Homicide by District). Note that Nos.13, 21, and 23 are unused. The figure is from the 2019 annual report of the Chicago Police Department (https://home.chicagopolice.org/wp-content/uploads/2020/09/19AR.pdf). (b) The constructed network structure, where the node's color and size denote the level of 2019 yearly investigatory stops (involving enforcement action). Bigger nodes with deeper colors indicate that a greater number of investigatory stops occurred in this district.

The time series of the number of daily investigatory stops for Districts 1 and 6 are

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

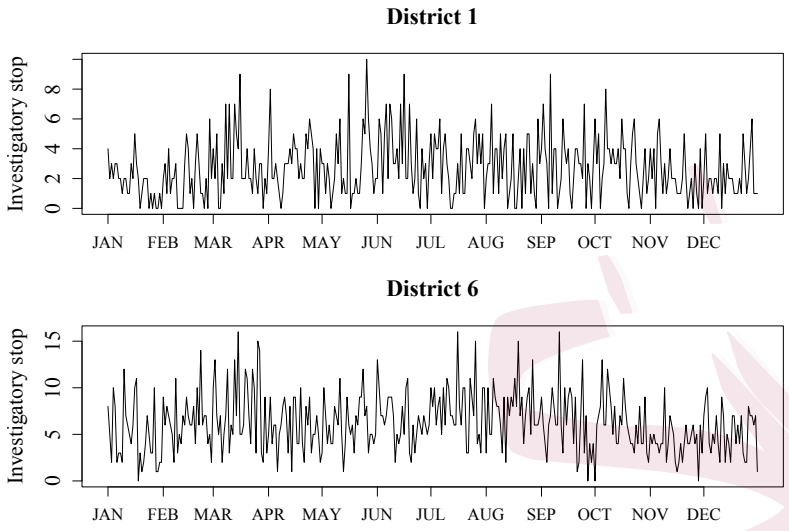


Figure 4: Number of daily investigatory stops in Districts 1 and 6.

plotted in Fig. 4, for illustration. There exists a dependency within the individual series. The average variance-to-mean ratio for each district is 1.85, and the overall variance-to-mean ratio is 3.16, which imply over-dispersion in the data.

5.2 Model estimation and interpretation

To determine the number of groups, we follow the two approaches described in **Remark 5**. First, we fit the data using different numbers of groups, say $K = 1, \ldots, 5$, and calculate the RMSEs. The first 11 months are employed for model training, and the last month is used for prediction evaluation. The in-sample RMSEs, defined as $\text{RMSE}_{esti} = \{(N(T-31))^{-1} \sum_{t=1}^{T-31} ||\widehat{\mathbf{Y}}_t - \mathbf{Y}_t||^2\}^{1/2}$, are $2.81, 2.78, 2.77, 2.77$, and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

2.76 for each K, respectively, and the out-of-sample RMSEs, defined as $\text{RMSE}_{pred} = \{(31N)^{-1}\sum_{t=T-30}^{T}||\widehat{\mathbf{Y}}_t - \mathbf{Y}_t||^2\}^{1/2}$, are $2.50, 2.51, 2.51, 2.51$, and 2.50, respectively. It appears that $3, 4$, and 5 are reasonable candidates for K. We then try the clustering method, that is, we estimate the coefficient parameter $\boldsymbol{\theta}$ at the nodal level, and apply k-means clustering to partition these N sets of estimates into K groups. Fig. 5 illustrates the selection of the optimal number of groups based on the elbow plot, silhouette coefficient, and gap statistic. All measures recommend $K = 3$. Thus, $K = 3$ is chosen in the following analysis.

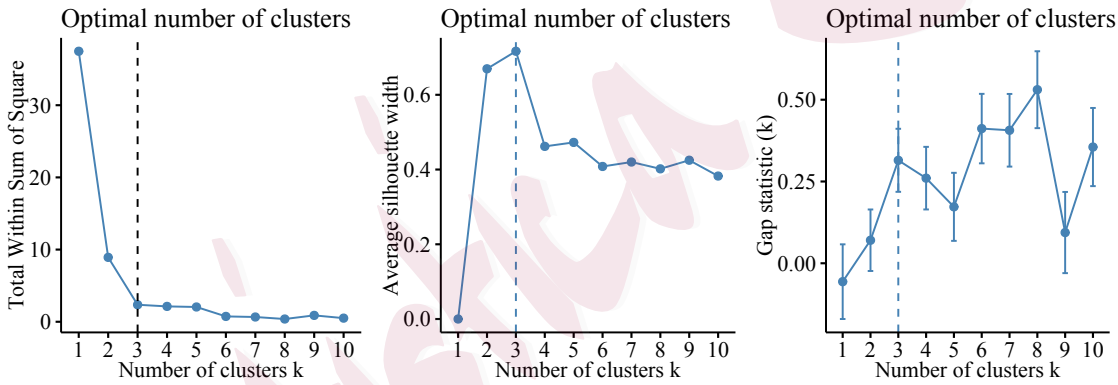


Figure 5: The selection of optimal number of groups based on the elbow plot, silhouette coefficient, and gap statistic, respectively.

We fit the GNPAR model (2.3) to the data set with $K = 3$. The results are summarized in Table 5. For all groups, the estimated regression coefficient on the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

past intensity process $\widehat{\beta}$ appears to be much bigger than the regression coefficients on the past observations $\widehat{\alpha}$ and the network effect $\widehat{\rho}$, implying that districts with a large (small) number of investigatory stops are more likely to have a large (small) number of investigatory stops in the future.

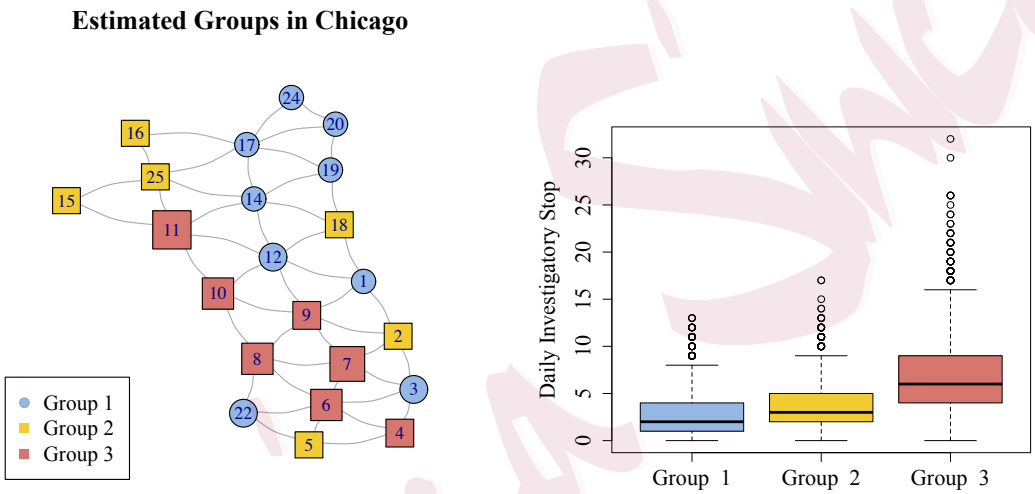
Table 5: Estimation results for ISR data using model (2.3), with $K = 3$.

	$\widehat{\omega}$	$\widehat{\alpha}$	$\widehat{\rho}$	$\widehat{\beta}$	$\widehat{\gamma}$
Group 1	0.0503	0.0931	0.0140	0.8650	0.4125
Group 2	1.4464	0.1109	0.1124	0.3281	0.2689
Group 3	0.0527	0.0659	1.9e-04	0.9259	0.3186

Fig. 6 (a) plots the estimated group labels for each district, and Fig. 6 (b) displays a boxplot for the number of daily investigatory stops Y_{it} in a grouped manner. The proportion of districts in each group is 0.41, 0.27, and 0.32, respectively.

The three groups show distinct numbers of stops and patterns of dependence. The districts in Group 3 are mainly in the southwest part of the city, which coincide with those areas with the highest level of crime risk in Fig. 3. Group 3 has the highest number of stops, and the intensity of the count does not depend on its surroundings, but mostly on its past intensity. Group 1 contains the safest areas, and also has a very small network effect, indicating that it is less likely to be affected by the surrounding areas. The future stops for Groups 1 and 3 can be predicted reliably

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040



(a) (b)

Figure 6: (a) Estimated group label for each district, marked in different colors and shapes. The size of each node denotes the level of yearly investigatory stops, with bigger nodes representing more crimes occurred in that district. (b) Box plot for the number of daily investigatory stops in a grouped manner.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

using historical information. In contrast, Group 2 has a moderate crime level and the districts surround high-risk areas. Group 2 has a relatively large network effect, implying that the intensity of the districts in Group 2 tend to be affected by events in their neighborhood.

The above observations also imply that although the network structure is symmetric, the mutual network effects between each pair of nodes could be different. The districts in Group 3 that are connected to districts in Group 2 are little affected by their neighbors, whereas those in Group 2 are more likely to be affected by events in their neighborhood. We guess that the latent flow network of the population of Chicago is directed and asymmetric, but more data are needed to verify this conjecture.

We also fit the NPAR model (2.2) that does not involve a group structure on the same data set for comparison. The estimation results are summarized in Table 6. For each estimator, the standard deviations are computed as $\mathbf{H}_T(\hat{\theta})^{-1}\mathbf{G}_T(\hat{\theta})\mathbf{H}_T(\hat{\theta})^{-1}$, where $\mathbf{H}_T$ and $\mathbf{G}_T$ are given in (S1.3) and (S1.4), respectively, in the Supplementary Material. All estimates are statistically significant at the 1% level. Still, the momentum effect is much greater than the network effect. The results show that the group-wise information provided by the GNPAR model provides greater insight into the real data and exhibits better interpretability. The AIC values for the GNPAR and NPAR models are 37340.57 and 37500.59, respectively, which suggests that the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

Table 6: Estimation results for ISR data using NPAR model (2.2). The estimates, estimated standard deviations, and p-values for each estimator are summarized.

	$\widehat{\omega}$	$\widehat{\alpha}$	$\widehat{\rho}$	$\widehat{\beta}$
Estimate	0.0128	0.0770	0.0033	0.9159
$\widehat{SE}$	0.0014	0.0015	7.1e-05	0.0017
p-value	< 0.001	< 0.001	< 0.001	< 0.001

GNPAR model fits the data better.

In summary, we have divided the districts in Chicago into three groups, and each group has its own spatial and dynamic patterns of investigatory stops. We find that the spatial distribution of investigatory stops with enforcement action taken largely agrees with that of actual crime that occurred, confirming the efficiency of the Investigatory Stop System.

6. Conclusion

We have proposed a GNPAR model. Compared with the traditional multivariate Poisson autoregressive model, our model has the following merits: (i) it incorporates network information to reduce the number of unknown parameters and the computational complexity; (ii) individual heterogeneity is introduced to describe different nodal behaviors for different groups, which makes the model more flexible and realis-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

tic; and (iii) the estimated group information and network effect can provide insight into real social problems and lead to better practical interpretations.

Our model can be generalized in several ways. First, we consider the linear form of the PAR, although the log-linear form of the PAR is also popular, and can be extended easily to the grouped case. Second, we assume that the network structure is fixed, but in practice, nodes may drop in and out of the model, and the association between nodes may change over time. Therefore, a time-varying network structure is worth studying. Third, additional covariates of the nodes or network structure information could be incorporated into the model for better fitting and group estimation. Lastly, in existing methods, the network dimension N is fixed, and we study the asymptotic properties with increasing time sample size T. If N is diverging, the stationarity and ergodicity of count time series are unavailable under current methods, and the estimation could become problematic, because the parameters grow quickly with the dimension of the matrix. This remains an open problem, and is left to future research.

Supplementary Material

The online Supplementary Material contains technical proofs of Proposition 1, Theorem 1, and several useful lemmas, as well as further simulation results when a group label is known and the performance of the first K-selection method in **Remark 5**.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

Acknowledgments

The research of Tao and Li was supported in part by NSFC (No.71973077 and No.11771239). The research of Niu was supported in part by NIH/NIAID R01 AI136664.

References

Ahmad, A. and Francq, C. (2016). Poisson QMLE of count time series models. *Journal of Time Series Analysis* **37**, 291–314.

Aitchison, J. and Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76**, 643–653.

Andreassen, C. M. (2013). *Models and inference for correlated count data*. Ph. D. thesis, Aaarhus University, Denmark.

Armillotta, M. and Fokianos, K. (2022). Poisson network autoregression. arXiv:2104.06296v3.

Davis, R.A., Fokianos, K., Holan, S.H., Joe, H., Livsey, J., Lund, R., Pipiras, V. and Ravishanker, N. (2021). Count time series: A methodological review. *Journal of the American Statistical Association* **116**, 1533–1547.

Davis, R.A. and Liu, H. (2016). Theory and inference for a class of nonlinear models with application to time series of counts. *Statistica Sinica* **26**, 1673–1707.

Du, J.-G. and Li, Y. (1991). The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

Analysis **12**, 129–142.

Erdös, P. and Rényi, A. (1960). On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5**, 17–61.

Ferland, R., Latour, A. and Oraichi, D. (2006). Integer-valued GARCH processes. *Journal of Time Series Analysis* **27**, 923–942.

Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association* **104**, 1430–1439.

Fokianos, K., Støve, B., Tjøstheim, D. and Doukhan, P. (2020). Multivariate count autoregressions. *Bernoulli* **26**, 471–499.

Fokianos, K. and Tjøstheim, D. (2011). Log-linear Poisson autoregression. *Journal of Multivariate Analysis* **102**, 563–578.

Fokianos, K. and Tjøstheim, D. (2012). Nonlinear Poisson autoregression. *Annals of the Institute of Statistical Mathematics* **64**, 1205–1225.

Holland, P., Laskey, K.B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5**, 109–137.

Huang, D., Wang, F., Zhu, X. and Wang, H. (2020). Two-mode network autoregressive model for large-scale networks. *Journal of Econometrics* **216**, 203–219.

Karlis, D. and Meligkotsidou, L. (2005). Multivariate Poisson regression with covariance structure. *Statistics and Computing* **15**, 255–265.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

Karlis, D. and Meligkotsidou, L. (2007). Finite mixtures of multivariate Poisson regression with application. *Journal of Statistical Planning and Inference* **137**, 1942–1960.

Latour, A. (1997). The multivariate GINAR(p) process. *Advances in Applied Probability* **29**, 228–248.

Lee, Y., Lee, S. and Tjøstheim, D. (2018). Asymptotic normality and parameter change test for bivariate Poisson INGARCH models. *TEST* **27**, 52–69.

Liu, H. (2012). *Some models for time series of counts*. Ph. D. thesis, Columbia University, USA.

Mahamunulu, D. (1967). A note on regression in the multivariate Poisson distribution. *Journal of the American Statistical Association* **62**, 251–258.

Meyn, S.P., Tweedie, R.L. (1993) *Markov Chains and Stochastic Stability*. Springer, London.

Neumann, M. (2011). Absolute regularity and ergodicity of Poisson count processes. *Bernoulli* **17**, 1268–1284.

Tjøstheim, D. (2012). Some recent theory for autoregressive count time series. *TEST* **21**, 413–438.

Pedeli, X. and Karlis, D. (2013). On composite likelihood estimation of a multivariate INAR(1) model. *Journal of Time Series Analysis* **34**, 206–220.

Wang, C., Liu, H., Yao, J.-F., Davis, R.A. and Li, W.K. (2014). Self-excited threshold Poisson autoregression. *Journal of the American Statistical Association* **109**, 777–787.

Weiß, C.H. (2018). *An Introduction to Discrete-Valued Time Series*. Wiley, Hoboken.

Zhou, J., Li, D., Pan, R. and Wang, H. (2020). Network GARCH model. *Statistica Sinica* **30**, 1–18.

Zhu, X., Pan, R., Li, G., Liu, Y. and Wang, H. (2017). Network vector autoregression. *Annals of Statistics*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0040

45, 1096–1123.

Zhu, X., Chang, X., Li, R. and Wang, H. (2019). Portal nodes screening for large scale social networks. *Journal of Econometrics* **209**, 145–157.

Zhu, X., Wang, W., Wang, H. and Härdle, W.K. (2019). Network quantile autoregression. *Journal of Econometrics* **212**, 345–358.

Zhu, X., Huang, D., Pan, R. and Wang, H. (2020). Multivariate spatial autoregressive model for large scale social networks. *Journal of Econometrics* **215**, 591–606.

Zhu X. and Pan R. (2020). Grouped network vector autoregression. *Statistica Sinica* **30**, 1437–1462.

Zhu, X., Cai, Z. and Ma, Y. (2022). Network functional varying coefficient model. *Journal of the American Statistical Association* **117**, 2074–2085.

Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, China

E-mail: taoyx19@mails.tsinghua.edu.cn

Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, China

E-mail: malidong@tsinghua.edu.cn

Department of Statistics, Pennsylvania State University, University Park, PA, USA

E-mail: xiaoyue@psu.edu