Journal of Machine Learning Research 22 (2021) 1-88 Submitted 3/20; Revised 5/21; Published 9/21

Context-dependent Networks in Multivariate Time Series: Models, Methods, and Risk Bounds in High Dimensions

Lili Zheng LZHENG57@WISC.EDU
Garvesh Raskutti RASKUTTI@WISC.EDU
Department of Statistics
University of Wisconsin-Madison
1300 University Avenue
Madison, WI 53706, USA

Rebecca Willett WILLETT@UCHICAGO.EDU
Departments of Statistics and Computer Science
University of Chicago
Chicago, IL 60637, USA

Benjamin Mark BMARK2@WISC.EDU
Department of Mathematics
University of Wisconsin-Madison
480 Lincoln Drive
Madison, WI 53706, USA

Editor: Ambuj Tewari

Abstract

High-dimensional autoregressive generalized linear models arise naturally for capturing how current events trigger or inhibit future events, such as activity by one member of a social network can affect the future activities of his or her neighbors. While past work has focused on estimating the underlying network structure based solely on the times at which events occur on each node of the network, this paper examines the more nuanced problem of estimating *context-dependent* networks that reflect how features associated with an event (such as the content of a social media post) modulate the strength of influences among nodes. Specifically, we leverage ideas from compositional time series and regularization methods in machine learning to conduct context-dependent network estimation for high-dimensional autoregressive time series of *annotated* event data. Two models and corresponding estimators are considered in detail: an autoregressive multinomial model suited to categorical features and a logistic-normal model suited to features with mixed membership in different categories. Importantly, the logistic-normal model leads to a convex negative log-likelihood objective and captures dependence across categories. We provide theoretical guarantees for both estimators that are supported by simulations. We further validate our methods and demonstrate the advantages and disadvantages of both approaches through two real data examples and a synthetic data-generating model. Finally, a mixture approach enjoying both approaches' merits is proposed and illustrated on synthetic and real data examples.

Keywords: autoregressive time series, high-dimensional generalized linear models, annotated event data, context-dependent network, compositional data

1. Introduction

High-dimensional auto-regressive processes arise in a broad range of applications. For instance, in a social network, we may observe a time series of members' activities, such as posts on social media where each person's post can influence their neighbors' future posts (*e.g.,* Stomakhin et al. (2011); Romero et al. (2011)). In the broadcast of social events, influential news media sources often play a crucial role and trigger other media sources to post new articles (Leskovec et al., 2009; Farajtabar et al., 2017). In electrical systems, cascading chains of power failures reveal critical information about the underlying power distribution network (Rudin et al., 2011; Ertekin et al., 2015). During epidemics, networks among computers or people are reflected by the time at which each node becomes infected (Ganesh et al., 2005; Yang et al., 2013). In biological neural networks, firing neurons can trigger or inhibit the firing of their neighbors, so that information about the network structure is embedded within spike train observations (Linderman et al., 2016; Fletcher and Rangan, 2014; Hall and Willett, 2015; Pillow et al., 2008; Gerhard et al., 2017). The above processes are *autoregressive* in that the likelihood of future events depends on past events.

In many applications, events are associated with feature vectors describing the events. For instance, interactions in a social network have accompanying text, images, or videos; and power failures are accompanied by information about current-carrying cables, cable ages, and cable types. Prior works (Hall et al., 2016; Mark et al., 2018) describe methods and theoretical guarantees for network influence estimation given multivariate event data without accounting for the type or context of the event. The contribution of this paper focuses on *estimation methods and theoretical guarantees for context-dependent network structures* which exploit features associated with events. The key idea is that different categories of events are characterized by different (albeit related) functional networks; we think of the feature vector as revealing the *context* of each event, and our task is to infer context-specific functional networks. Allowing for features provides a much richer model class that can reflect, for instance, that people interact in a social network differently when interactions are family-focused vs. work-focused vs. political (Puniyani et al., 2010; Feller et al., 2011; Williams et al., 2013). Capturing these differences reveals how the type or content of the information affects its spreading pattern, which is an interesting problem in mass communication (Lazer et al., 2018; Mihailidis and Viotty, 2017; Shu et al., 2017). Another example is the modeling of stock prices where the past stock price of a company could influence the future price of its competitors, and the accompanying business news illustrating the reasons behind price fluctuations can be useful context. Learning a context-dependent network for this example may improve the prediction accuracy significantly since it takes advantage of the important side information that is often incorporated for stock price prediction (Li et al., 2014; Chan, 2003).

Developing a statistical model for autoregressive time series of annotated event data is a non-trivial task. One particular challenge is that we usually cannot determine the exact category of the event. For example, a post on social media may exhibit membership in several topics (Blei et al., 2003); an infected patient's symptoms can be caused by different diseases (Woodbury et al., 1978); a new product released to the market could contain several features or styles. Some natural-seeming models lead to computationally-intractable

estimators, while others fail to account for ambiguity in the categories of the events. In this paper, we propose two autoregressive time series models that suit distinct scenarios:

(i) **Multinomial Model:** This model is applied when each event (*i.e.,* its feature) naturally belongs to a single category. For example, a tweet may clearly belong to a single category (*e.g.,* "political").

(ii) **Logistic-normal Model:** This model is applied when each event is a mixture of multiple categories (*i.e.,* mixed membership). For example, a news article may belong to two or more categories (*e.g.,* "political" and "finance"), and we may only have measurements of the relative extent to which it is in each category.

To the best of our knowledge, the multinomial model we consider appeared first in Tank et al. (2017), while no theoretical guarantee was provided. From both a modeling and theoretical perspective, the logistic-normal model is more nuanced. It employs the logistic-normal distribution widely used in compositional data analysis (*e.g.,* Aitchison (1982); Brunsdon and Smith (1998); Ravishanker et al. (2001)). The logistic-normal model has advantages over other mixed membership models such as the Dirichlet distribution and the more recent Gumbel soft-max distribution since it leads to a convex negative log-likelihood function and models dependence among sub-compositions of the membership vector, which will be explained in detail at the beginning of Section 2.2.1.

High-dimensional setting: Throughout this paper, we focus on the *high-dimensional* setting, where the number of nodes in the network is large and grows with sample size. We assume the number of edges within the huge network to be *sparse*: each node should only be influenced by a limited number of other nodes. We state this condition more formally in Section 3.

1.1 Contributions

Our contributions are summarized as follows:

- For both models, we present *estimation algorithms* based on minimizing a convex loss function using a negative log-likelihood loss or a squared loss, plus a regularization term that accounts for the sparsity of networks but shared network structure between models corresponding to different categories. See Section 2.1-2.2.

- Meanwhile, we establish *risk bounds* that characterize the error decay rate as a function of network size, sparsity, shared structure, and the number of observations, and these bounds are illustrated with a variety of simulation studies. See Section 3-5.2.

- Since our network parameters in the two models have different interpretations and cannot be compared when applied to the same dataset, we introduce a *novel concept of variable importance network* based on prediction, which shares the same interpretation across different time series models. Its shared interpretation facilitates the comparison between our models in numerical experiments on both synthetic data and real datasets. In addition, the formulation of the variable importance network is applicable to interpret any autoregressive time series models with one-step-ahead

predictions, which could be of independent interest. We also provide estimation error bounds for the variable importance networks under our proposed models. See Section 4 and Section 3.

- Furthermore, we validate the following hypothesis through experimental results on a synthetic data-generating model and real data from two datasets: *the logistic-normal method is more suitable for mixed membership settings while the multinomial method is more suitable for settings with a clear dominant category.* The synthetic data model is based on a noisy logistic-normal distribution with some nodes having events with a single dominant category and other nodes following a mixed membership setting. The multinomial model tends to correctly detect the edges between nodes with a single dominant category, while the logistic-normal approach tends to correctly detect the edges corresponding to nodes with mixed membership categories. We further provide evidence for the hypothesis with two datasets: (1) a political tweets dataset focusing on the network that varies according to political leanings of tweets and (2) online media dataset where the network depends on topics of memes. The networks detected for both datasets tend to support the above hypothesis. See Section 5.3 and Section 6.

- Finally, inspired by our synthetic data-generating model, we *develop a mixture approach* including two main steps: (i) testing which model suits each node better in a network using a log-likelihood ratio test, and (ii) fitting a mixture model (some nodes following the multinomial model while the others following the logistic-normal model) based on the test results. This mixture approach works reasonably well for both the synthetic data example and two real datasets. See Section 5.3 and Section 6.

1.2 Related Work

There has been substantial literature on recovering network structure using time series of event data in recent years, including continuous-time approaches (Zhou et al., 2013; Yang et al., 2017) based on Hawkes process (Hawkes, 1971) and discrete-time approaches (Linderman et al., 2016; Fletcher and Rangan, 2014; Hall et al., 2016; Mark et al., 2018). Our work follows the line of works (discrete-time approaches): Hall et al. (2016); Mark et al. (2018), but with the additional challenge of incorporating the context information of events. Tank et al. (2017) considers the multinomial model with exact categorical information of events but provides no theoretical guarantees.

Another popular approach aiming to recover the text-dependent network structure in social media is the cascade analysis (Lerman and Ghosh, 2010; Yu et al., 2017b, 2018), which focuses on the diffusion of information, *e.g.,* retweeting or sharing the same hyperlink. However, it is also possible for users to interact in social media by posting about similar topics (*e.g.,* showing condolence for shooting events) or arguing about opposite opinions (*e.g.,* tweets sent by presidential candidates) without sharing exactly the same text. This kind of interaction is captured by our approach but not by the cascade analysis. Due to the nature of our models, we can also study time series of event data with any categorical features (either exact or with uncertainty/mixed membership), without diffusion of information involved. Examples include the stock price changes with corresponding business news as

side information. We can also analyze multi-node compositional time series (Brunsdon and Smith (1998); Ravishanker et al. (2001) are existing works on single-node compositional time series) if we consider a special case of the logistic-normal model (7) and (8) with $q = 1$. Our work also incorporates proof techniques from the high-dimensional statistics literature (*e.g.,* Bickel et al. (2009); Raskutti et al. (2010)) whilst incorporating the nuances of temporal dependence, non-linearity, and context-based information not captured in prior works.

We will further elaborate on the connection of our models to prior work (point process and compositional time series literature) in Section 2.3 after introducing detailed formulations of our models in Section 2.1 and Section 2.2.

The remainder of this paper is organized as follows: we elaborate on our problem formulations and corresponding estimators in Section 2; theoretical guarantees on estimation errors are provided in Section 3; we present simulation results on synthetic data and our synthetic model example in Section 5, which also introduces our mixture method inspired by the synthetic model example; the real data experiments are included in Section 6.

2. Problem Formulation and Estimators

We begin by introducing basic notations. For any integer $p > 0$, $\Delta^p = \{v \in \mathbb{R}^{p+1} : \sum_{i=1}^{p+1} v_i = 1; \forall i, v_i \geqslant 0, \}$ is the p-dimensional simplex. For any tensor or matrix $A \in \mathbb{R}^{p_1 \times \cdots \times p_k}$ and $1 \leqslant l \leqslant k$, let $A_{i_1,\ldots,i_l} \in \mathbb{R}^{p_{l+1} \times \cdots \times p_k}$ be determined by fixing the first l dimensions of A to be $i_1, \ldots, i_l$. We will also use $A_{:,\ldots,:,i_m,:,\ldots,:} \in \mathbb{R}^{p_1 \times \cdots \times p_{m-1} \times p_{m+1} \times \cdots \times p_k}$ to denote the tensor that fixes the mth dimension of A to be i_m. For any two tensors A and B of the same dimension, let $\langle A, B \rangle$ denote the Euclidean inner product of A and B. If $A \in \mathbb{R}^{p_1 \times \cdots \times p_k}$ and $B \in \mathbb{R}^{p_l \times \cdots \times p_k}$ for some $1 < l \leqslant k$, let $\langle A, B \rangle$ be of dimension $p_1 \times \cdots \times p_{l-1}$, and $(\langle A, B \rangle)_{i_1,\ldots,i_{l-1}} = \langle A_{i_1,\ldots,i_{l-1}}, B \rangle$. Also, define the Frobenius norm of tensor A as $\|A\|_F = \langle A, A \rangle^{\frac{1}{2}}$. In addition, for any $A \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_k}, B \in \mathbb{R}^{p_1' \times p_2 \times \cdots \times p_k}$, we use $[A, B] \in \mathbb{R}^{(p_1+p_1') \times p_2 \times \cdots \times p_k}$ to denote concatenation of A and B in the first dimension: $([A, B])_{1:p_1} = A$ and $([A, B])_{(p_1+1):(p_1+p_1')} = B$.

For any 3rd-order tensor $A \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, define the regularization norm $\|A\|_R$ as

$$\|A\|_R = \sum_{m=1}^{n_2} \|A_{:,m,:}\|_F. \tag{1}$$

For any $0 \leqslant \alpha \leqslant 1$, 3rd-order tensor $A \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and matrix $B \in \mathbb{R}^{n_2 \times n_3}$, define $R_\alpha(A, B)$ as

$$R_\alpha(A, B) = \sum_m \left(\alpha \|A_{:,m,:}\|_F^2 + (1-\alpha)\|B_{m,:}\|_2^2\right)^{\frac{1}{2}}. \tag{2}$$

With a little abuse of notation, for any 4th-order tensor $A \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$, let $\|A\|_R = \sum_{m_1,m_2} \|A_{m_1,:,m_2,:}\|_F$. For any matrix A, we let $\lambda_{\min}(A)$ denote the smallest eigenvalue of A. Let

$$\mathbb{1}_{\{E\}} = \begin{cases} 1, & \text{if } E \text{ true} \\ 0, & \text{else} \end{cases}$$

be the indicator function.

Let M refer to the number of nodes (multiple time series) and let $X^t \in \mathbb{R}^{M\times K}$ be the observed data at time t for $t = 0, 1, ..., T$, where K is the number of categories of events. We assume there is either zero or one event for each node and time point. For each $1 \leqslant m \leqslant M$, $0 \leqslant t \leqslant T$, if there is no event, $X_m^t \in \mathbb{R}^K$ is a zero vector. For times and nodes with events, we consider two different observation models:

- The first model is the multinomial model (Section 2.1) corresponding to the setting in which each event only belongs to a single category. In this case, if the event at time t and node m is in category $k \in \{1, \ldots, K\}$, then we let $X_m^t = e_k$, where e_k is the kth vector in the canonical basis of $\mathbb{R}^K$.
- The second is the logistic-normal model (Section 2.2) corresponding to the setting in which each event has mixed category membership, and that membership is potentially observed with noise. In this case, we let X_m^t be a vector on the simplex $\triangle^{K-1}$, with non-negative elements summing up to one.

The following two sections address these two cases separately, where the corresponding estimator for each model is also discussed.

2.1 Multinomial Model

When each event belongs to a single category, the distribution of $\{X_m^{t+1}\}_{m=1}^M$ conditioned on the past data X^t can be modeled as independent multinomial random vectors. As mentioned earlier, we assume that there is either one or zero event for each node at each time, and hence $X_m^t \in \{0, 1\}^K$ under this multinomial model. Specifically, let tensor $A^{\text{MN}} \in \mathbb{R}^{M\times K\times M\times K}$ encode the context-dependent network, and each entry $A^{\text{MN}}_{m,k,m',k'}$ is the influence exerted upon {node m, category k} by {node m', category k'}. We will refer to this influence as *absolute* influence, contrasted with the relative influence and overall influence in the logistic-normal model introduced later. That is, an event from node m' in category k' may increase or decrease the likelihood of a future event by node m in category k, and $A^{\text{MN}}_{m,k,m',k'}$ parameterizes that change in likelihood. Further define $\nu^{\text{MN}} \in \mathbb{R}^{M\times K}$ as the intercept term where each entry $\nu^{\text{MN}}_{m,k}$ determines the event likelihood of {node m, category k} when there are no past stimuli. For any $p > 0$, $A \in \mathbb{R}^{p\times M\times K}$, $\nu \in \mathbb{R}^p$, let

$$\mu^{t+1}(A, \nu) = \langle A, X^t\rangle + \nu = \sum_{m',k'} A_{:,m',k'} X^t_{m',k'} + \nu \in \mathbb{R}^p. \tag{3}$$

Then $\mu^{t+1}(A^{\text{MN}}_m, \nu^{\text{MN}}_m) \in \mathbb{R}^K$ and we use $\mu^{t+1}_k(A^{\text{MN}}_m, \nu^{\text{MN}}_m)$ to denote the likelihood parameter of {node m, category k} at time $t + 1$ given the past. Then the conditional distribution of X_m^{t+1} is

$$\begin{aligned}
\mathbb{P}(X_m^{t+1} = e_k|X^t) =& \frac{e^{\mu_k^{t+1}(A^{\text{MN}}_m,\nu^{\text{MN}}_m)}}{1 + \sum_{k'=1}^K e^{\mu_{k'}^{t+1}(A^{\text{MN}}_m,\nu^{\text{MN}}_m)}}, \ 1 \leqslant k \leqslant K \\
\mathbb{P}(X_m^{t+1} = 0|X^t) =& \frac{1}{1 + \sum_{k'=1}^K e^{\mu_{k'}^{t+1}(A^{\text{MN}}_m,\nu^{\text{MN}}_m)}}.
\end{aligned} \tag{4}$$

This is also the multinomial logistic transition distribution (mLTD) model considered in Tank et al. (2017). See Figure 1(a) for a visualization of the multinomial model.

For simplicity of exposition, we assume the offset parameter ν^{MN} to be known[1] and only estimate the parameter $A^{\mathrm{MN}} \in \mathbb{R}^{M\times K\times M\times K}$. One straightforward estimation method is to find the minimizer of the penalized negative log-likelihood:

$$\widehat{A}_m^{\mathrm{MN}} = \underset{A\in\mathbb{R}^{K\times M\times K}}{\arg\min} \frac{1}{T}\sum_{t=0}^{T-1} \ell^{\mathrm{MN}}(A; X^t, X_m^{t+1}, \nu_m^{\mathrm{MN}}) + \lambda\|A\|_R, \tag{5}$$

where

$$\ell^{\mathrm{MN}}(A; X^t, X_m^{t+1}, \nu_m^{\mathrm{MN}}) = f_1(\mu_k^{t+1}(A, \nu_m^{\mathrm{MN}})) - \left\langle \mu_k^{t+1}(A, \nu_m^{\mathrm{MN}}), X_m^{t+1}\right\rangle, \tag{6}$$

and $f_1 : \mathbb{R}^K \to \mathbb{R}$ is defined by $f_1(x) = \log\left(\sum_{i=1}^K e^{x_i} + 1\right)$. Note that $\|A\|_R$ is the group sparsity penalty defined in (1).

2.2 Logistic-normal Model

When there is mixed membership, for each $0 \leqslant t \leqslant T, 1 \leqslant m \leqslant M$, the $K \times 1$ vector X_m^t is either the zero vector or a vector on the simplex corresponding to the mixed membership probability of categories. Thus we need to address the distribution in two parts: the probability mass of $\mathbb{1}_{\{X_m^{t+1}\neq 0\}}$ and the distribution of X_m^{t+1} given $X_m^{t+1} \neq 0$.

Let $Z_m^{t+1} \in \triangle^{K-1}$ be a random vector on the simplex with a distribution to be specified shortly. We model the distribution of $\{X_m^{t+1}\}_{m=1}^M$ conditioned on the past as:

$$X_m^{t+1} = \begin{cases} Z_m^{t+1}, & \text{with probability } q_m^{t+1}, \\ 0_K, & \text{with probability 1-}q_m^{t+1}, \end{cases} \tag{7}$$

and further assume conditional independence of entries for $\{X_m^{t+1}\}_{m=1}^M$. For $q^{t+1} \in [0,1]^M$, each element is the probability that an event occurs at the corresponding node and time $t+1$. We specify how q^{t+1} is modeled later.

2.2.1 Modeling Z^t

Z_m^{t+1} may be modeled by two kinds of distributions widely used for compositional data: the Dirichlet distribution (Bacon-Shone, 2011) and the logistic-normal distribution (Aitchison, 1982). The Dirichlet model gains its popularity in Bayesian statistics but makes the limiting assumption that the sub-compositions are independent. More specifically, for any r.v. $X \in \mathbb{R}^K \sim \mathrm{Dir}(\alpha)$,

$$\left(\frac{X_1}{\sum_{i=1}^k X_i}, \ldots, \frac{X_k}{\sum_{i=1}^k X_i}\right) \quad \text{and} \quad \left(\frac{X_{k+1}}{\sum_{i=k+1}^K X_i}, \ldots, \frac{X_K}{\sum_{i=k+1}^K X_i}\right)$$

are independent for any $1 \leqslant k \leqslant K-1$. Another difficulty associated with the Dirichlet modeling is the non-convexity of the negative log-likelihood objective, which presents challenges both in terms of run-time and from a statistical perspective.

1. It is straightforward to estimate ν^{MN} together with A^{MN}, and the theoretical guarantees would still hold. For synthetic mixture model experiments in Section 5.3 and real data applications in Section 6, we will always estimate the offset parameter and network parameter together for all models.

Hence we employ the logistic-normal distribution, which (i) has log-concave density function and thus facilitates fast convergence to global optimizers and more tractable theoretical analysis; (ii) incorporates the potential dependence among sub-compositions in different categories by introducing dependent Gaussian noise in the log-ratio (Atchison and Shen, 1980; Blei and Lafferty, 2006). The logistic-normal distribution is also related to the Gumbel-Softmax distribution (Jang et al., 2016), which has gained popularity in approximating a categorical distribution using a continuous one. The difference is that the logistic-normal distribution assumes the noise to be Gaussian and is thus more amenable to statistical analysis, whereas the Gumbel-Softmax employs the Gumbel distribution.

Specifically, we let the Kth category be the baseline category[2], so that we could transform $Z_m^{t+1} \in \triangle^{K-1}$ to log-ratios $\log \frac{Z_{m,k}^{t+1}}{Z_{m,K}^{t+1}}, k = 1, \ldots, K-1$, which take values on the entire $\mathbb{R}^{K-1}$ and can be modeled by a multivariate normal distribution. To model the conditional expectation of $\log \frac{Z_{m,k}^{t+1}}{Z_{m,K}^{t+1}}$ give X^t, let $A^{\mathrm{LN}} \in \mathbb{R}^{M\times(K-1)\times M\times K}$ encode the network, where $A^{\mathrm{LN}}_{m,k,m',k'}$ is the relative influence exerted upon {node m, category k} relative to {node m, category K} by {node m', category k'}. In addition, we let $\nu^{\mathrm{LN}} \in \mathbb{R}^{M\times(K-1)}$ be the corresponding intercept term. Then we use $\mu_k^{t+1}(A_m^{\mathrm{LN}}, \nu_m^{\mathrm{LN}})$ to model the conditional expectation of $\log \frac{Z_{m,k}^{t+1}}{Z_{m,K}^{t+1}}$ give X^t, where $\mu^{t+1}(\cdot,\cdot)$ is defined in (3).

More specifically, for any $t \geqslant 0, 1 \leqslant m \leqslant M$, given $\{X^{t'}\}_{t'=0}^t$,

$$\begin{aligned} &\log \frac{Z_{m,k}^{t+1}}{Z_{m,K}^{t+1}} = \mu_k^{t+1}(A_m^{\mathrm{LN}}, \nu_m^{\mathrm{LN}}) + \epsilon_{m,k}^{t+1}, \quad 1 \leqslant k \leqslant K-1, \\ &\{\epsilon_m^{t+1}\}_{t,m} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma),\ \Sigma \in \mathbb{R}^{(K-1)\times(K-1)}, \end{aligned} \tag{8}$$

where $\epsilon_m^{t+1} \in \mathbb{R}^{(K-1)}$ is a Gaussian noise vector with covariance $\Sigma \in \mathbb{R}^{(K-1)\times(K-1)}$.

See Figure 1(b) for a visualization of the logistic-normal model.

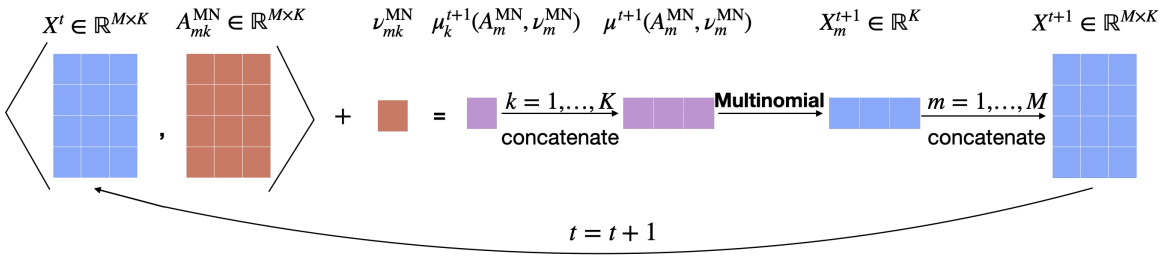
2.2.2 Modeling q^t

We now discuss models for the event probability q^{t+1} in (7) and study the following two cases: (a) q^{t+1} is a constant vector across t, which can be specified by $q \in \mathbb{R}^M$; (b) q^{t+1} depends on the past X^t.

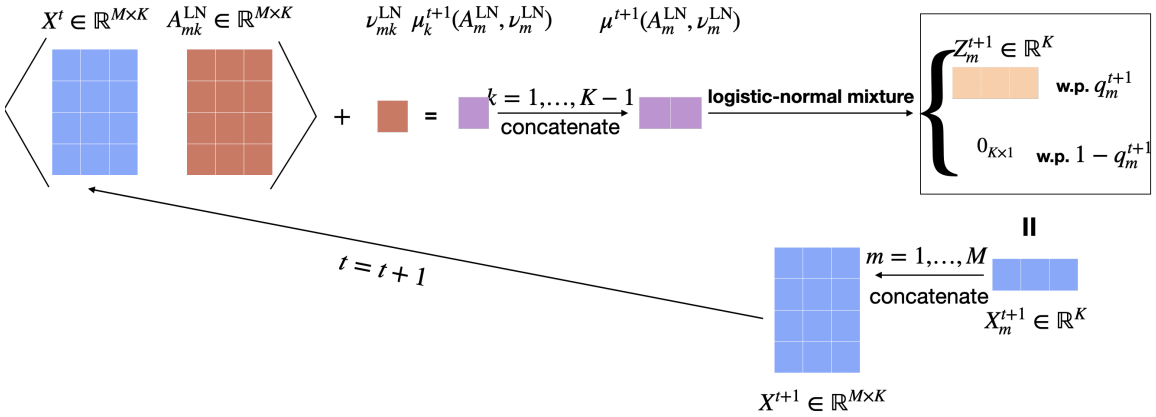
Constant $q^t = q$: This model is reasonable if we consider event rates that are constant over time or multi-node compositional time series. For example, users on social media may have constant activity levels; compositional data (*e.g.,* labor/expenditure statistics) for each node (*e.g.,* state/country) may be released on a regular schedule. The latter case can be thought of as a special case with $q = 1$ [3].

2. Using a different baseline category would not change our model form, but only lead to reparameterization. More details on this is included in Appendix C.1.

3. In this case, all X_m^t are non-zero and constrained in the $(K-1)$-dimensional simplex $\triangle^{K-1}$, so for identifiability, we have to take $X^t_{:,1,:,(K-1)} \in \mathbb{R}^{M\times(K-1)}$ instead of $X^t \in \mathbb{R}^{M\times K}$ as the covariate for predicting X^{t+1} and thus assume $A^{\mathrm{LN}} \in \mathbb{R}^{M\times(K-1)\times M\times(K-1)}$.



(a) Multinomial model



(b) Logistic-normal model

Figure 1: Visualizations for the multinomial model (4) and logistic-normal model (8), with $M = 4$, $K = 3$. Each X^{t+1}_m is independently generated from the corresponding distributions, with parameter $\mu^{t+1}(A^{\mathrm{MN}}_m, \nu^{\mathrm{MN}}_m)$ or $\mu^{t+1}(A^{\mathrm{LN}}_m, \nu^{\mathrm{LN}}_m)$ determined by the past data X^t and model parameters.

Estimator: In the case of constant q^t, we only estimate $A^{\mathrm{LN}} \in \mathbb{R}^{M\times(K-1)\times M\times K}$ and assume ν^{LN} to be known for ease of exposition, while q and the covariance matrix Σ are unknown nuisance parameters. We define the estimator as the minimizer of a penalized squared error loss:

$$\widehat{A}^{\mathrm{LN}}_m = \underset{A\in\mathbb{R}^{(K-1)\times M\times K}}{\arg\min} \frac{1}{T}\sum_{t=0}^{T-1} \ell^{\mathrm{LN}}(A; X^t, X^{t+1}_m, \nu^{\mathrm{LN}}_m) + \lambda\|A\|_R, \tag{9}$$

where

$$\begin{aligned} \ell^{\mathrm{LN}}(A; X^t, X^{t+1}_m, \nu^{\mathrm{LN}}_m) =& \frac{1}{2}\mathbb{1}_{\{X^{t+1}_m\neq 0\}}\|Y^{t+1}_m - \mu^{t+1}(A, \nu^{\mathrm{LN}}_m)\|_2^2, \\ Y^t_{m,k} =& \begin{cases} \log(X^t_{m,k}/X^t_{m,K}), & X^t_m \neq 0 \\ 0, & X^t_m = 0 \end{cases}, \qquad 1\leqslant k\leqslant K-1 \end{aligned} \tag{10}$$

Note that if $\Sigma = I_{K-1}$, the squared loss is exactly the negative log-likelihood loss, while for a general Σ, this loss is still applicable without knowing Σ. Note that q does not appear

in the objective function. This is because that the log-likelihood can be written as the summation of a function of A and a function of q, and thus we could directly minimize an objective function that does not depend on q.

q^t depends on past events: We model q^{t+1} using the logistic link: for $1 \leqslant m \leqslant M$,

$$q_m^{t+1} = \frac{\exp\{\langle B_m^{\text{Bern}}, X^t \rangle + \eta_m^{\text{Bern}}\}}{1 + \exp\{\langle B_m^{\text{Bern}}, X^t \rangle + \eta_m^{\text{Bern}}\}}, \tag{11}$$

where $B^{\text{Bern}} \in \mathbb{R}^{M \times M \times K}$, and $B^{\text{Bern}}_{m,m',k'}$ is the overall influence exerted on node m by {node m', category k'}, while $\eta^{\text{Bern}} \in \mathbb{R}^M$ is the offset parameter. If we set $B^{\text{Bern}} = 0$, this reduces to the constant $q^t = q$ case with $q_m = \left(1 + \exp\{-\nu_m^{\text{LN}}\}\right)^{-1}$. In general, our goal is to jointly estimate A^{LN} and B^{Bern}, while ν^{LN} and η^{Bern} are assumed known for ease of exposition, and the covariance matrix Σ is regarded as an unknown nuisance parameter. The loss function $\ell^{\text{LN}}(A)$ defined in (10) can still be used to estimate A^{LN}. While for B^{Bern}, we can define $\ell^{\text{Bern}}(B)$ as the log-likelihood loss of the Bernoulli distributed $\mathbb{1}_{\{X_m^t \neq 0\}}$:

$$\ell^{\text{Bern}}(B; X^t, X_m^{t+1}, \eta) = f_2(\langle B, X^t \rangle + \eta) - (\langle B, X^t \rangle + \eta)\mathbb{1}_{\{X_m^{t+1} \neq 0\}}, \tag{12}$$

where $f_2 : \mathbb{R} \to \mathbb{R}$ is defined by $f_2(x) = \log(e^x + 1)$. To exploit the sparsity structure shared by A^{LN} and B^{Bern}, we pool the two losses together and add a group sparsity penalty on A^{LN} and B^{Bern}. To account for various noise levels Σ, we put different weights on the two losses, and intuitively the weight on $L^{\text{LN}}(A)$ should be smaller if Σ is large. Formally, for any $1 \leqslant m \leqslant M$,

$$\begin{aligned} &(\widehat{A}_m^{\text{LN}}, \widehat{B}_m^{\text{Bern}}) \\ = &\underset{\substack{A \in \mathbb{R}^{(K-1)\times M \times K} \\ B \in \mathbb{R}^{M \times K}}}{\arg\min} \left[\frac{\alpha}{T} \sum_{t=0}^{T-1} \ell^{\text{LN}}(A; X^t, X_m^{t+1}, \nu_m^{\text{LN}}) \right. \\ &\left. + \frac{1-\alpha}{T} \sum_{t=0}^{T-1} \ell^{\text{Bern}}(B; X^t, X_m^{t+1}, \eta_m^{\text{Bern}}) + \lambda R_\alpha(A, B) \right], \end{aligned} \tag{13}$$

where the penalty term $R_\alpha(A, B)$ is defined in (2). If we let $\alpha = 0.5$, this type of estimator has been widely seen in the literature of multi-task learning (Zhang and Yang, 2017; Obozinski et al., 2006; Lounici et al., 2009). When $\alpha = 0$ or 1, we are estimating A^{LN} or B^{Bern} only with the penalty term $\lambda\|A\|_R$ or $\lambda\|B\|_R$, respectively.

2.3 Connection of Our Models to Prior Work

After presenting the detailed formulations of our models in previous sections, in this section, we discuss connections between our models and existing approaches in the literature.

Connection to Point Process Literature: Our work is most closely related to Hall et al. (2016); Mark et al. (2018), which discuss a discrete-time modeling approach for point process data. As illustrated by Mark et al. (2018), the multivariate Hawkes process (Hawkes, 1971; Daley and Vere-Jones, 2003; Yang et al., 2017) can be discretized and represented as a Poisson generalized linear ARMA model. Considering a discrete approach improves the

computational efficiency and can also deal with real-world data that is collected at discrete time points. More specifically, Hall et al. (2016) investigates the following high-dimensional generalized linear autoregressive process:

$$X^{t+1}|X^t \sim P(\nu + A^* X^t), \tag{14}$$

where $\{X_t \in \mathbb{R}^M\}_{t=0}^T$ is the observed data, $\nu \in \mathbb{R}^M$ is a known offset parameter, and $A^* \in \mathbb{R}^{M\times M}$ is the network parameter of interest. Hall et al. (2016) specifies P to be the product measure of independent Poisson or Bernoulli distributions. For a Bernoulli autoregressive process, the model is:

$$\mathbb{P}(X^{t+1}|X^t) = \prod_{m=1}^{M} \frac{\exp\{(\nu_m + \langle A_m^*, X^t\rangle)X_m^{t+1}\}}{1+\exp\{\nu_m + \langle A_m^*, X^t\rangle\}}. \tag{15}$$

This model ignores the context/categorical information of the events, which is what our methods aim to capture.

When each event corresponds to one exact category, the multinomial model (4) can capture the category-dependent network as a natural extension from Bernoulli autoregressive process. This model can also be seen as a multi-variate ($M > 1$) version of the categorical time series (Fokianos et al., 2003). By considering the multi-variate version, our model reflects not only an autoregressive model for each node independently but also the autoregressive model of *interactions* between them. However, when the event presents imprecise mixed membership in multiple categories, there is no established model that can be directly applied or naturally extended for this type of data. Our logistic-normal approach (7), (8) combines ideas from compositional time series and autoregressive process framework.

Connection to Compositional Time Series: Compositional time series arise from the study of labor statistics (Brunsdon and Smith, 1998), expenditure shares (Mills, 2010) and industrial production (Kynčlová et al., 2015). In a classical setup, one would observe a time series $\{X^t\}_{t=0}^T$ where $X^t \in \mathbb{R}^K$ lies on a simplex $\triangle^{K-1}$, representing the composition of a quantity of interest (*i.e.,* proportion belonging to each category). Directly modeling compositional time series data is difficult because the observations are all constrained on the simplex. This challenge can be avoided by modeling the data after transforming the data via taking the log of ratios between each category and some baseline category as discussed earlier. In classical compositional time series analysis, we might use an ARMA model to describe the transformed data.

Our *logistic-normal* model is closely connected to the compositional time series models but deviates from this classical setting in two ways. On the one hand, even when we consider the special case where event probability $q = 1$, we have a multi-variate compositional time series (one for each node in our network), and so our model reflects the interactions between the nodes. The number of nodes can be large, making this a high-dimensional problem. A more significant difference is that we consider the scenario where there is *no event* during a time period t for node m, meaning $X_m^t = 0_K$ instead of lying on the simplex. This presents a significant methodological challenge, as discussed earlier, and we cannot simply apply the log-ratio transformations to all X_m^t. Hence we introduce a latent variable Z_m^t lying on the simplex to address this issue: we only apply the log-ratio transformation on Z_m^t when modeling the conditional distribution of Z_m^t given X^{t-1}, and with probability q_m^t we observe $X_m^t = Z_m^t$, otherwise $X_m^t = 0_K$.

3. Theoretical Guarantees

In this section, we derive the estimation error bounds for the estimators defined in Section 2.1, Section 2.2, alongside the error bounds for the variable importance parameter estimators defined in Section 4, under their corresponding model set-ups. We first introduce sparsity and boundedness notions that will appear in the theoretical results. In particular, for the multinomial model (4), we define the following notions:

(i) **Group sparsity parameters**: For $1 \leqslant m \leqslant M$, let $S_m^{\mathrm{MN}} := \{m' : \|A^{\mathrm{MN}}_{m,:,m',:}\|_F > 0\}$ be the set of nodes that have an influence on node m in any category, sparsity $\rho_m^{\mathrm{MN}} := |S_m^{\mathrm{MN}}|$, and $\rho^{\mathrm{MN}} := \max_{1\leqslant m\leqslant M} \rho_m^{\mathrm{MN}}$. Further let $s^{\mathrm{MN}} := \sum_{m=1}^{M} \rho_m^{\mathrm{MN}}$.

(ii) **Boundedness parameters**: Let $R_{\max}^{\mathrm{MN}} := \|A^{\mathrm{MN}}\|_{\infty,\infty,1,\infty} = \max_{m,k} \sum_{m'} \max_{k'} |A^{\mathrm{MN}}_{m,k,m',k'}|$.

For the logistic-normal model with constant event probability ((7), (8) with $q^t = q$), we can define S_m^{LN}, ρ_m^{LN}, ρ^{LN}, s^{LN}, and $R_{\max}^{\mathrm{LN}}$ similarly from above, except that we substitute A^{MN} by A^{LN}. While for the logistic-normal model with event probability depending on the past ((7), (8), (15)), we assume shared sparsity in A^{LN} and B^{Bern} among nodes, and both of them need to be bounded. Thus under this model, we define $S_m^{\mathrm{LN,Bern}}$, $\rho_m^{\mathrm{LN,Bern}}$, $\rho^{\mathrm{LN,Bern}}$, $s^{\mathrm{LN,Bern}}$ and $R_{\max}^{\mathrm{LN,Bern}}$, similarly to the above, except that we substitute A^{MN} by the concatenated tensor $(A^{\mathrm{LN}}, B^{\mathrm{Bern}}) \in \mathbb{R}^{M\times K\times M\times K}$ (concatenated in the second dimension).

3.1 Multinomial Model

Theorem 1 *Consider the generation process* (4) *and estimator* (5). *If* $\lambda = CK\sqrt{\frac{\log M}{T}}$, $K \leqslant M$, *and* $T \geqslant ce^{2C_1}(1 + Ke^{C_1})^2K^4(\rho^{\mathrm{MN}})^2 \log M$, *then with probability at least* $1 - 3\exp\{-c\log M\}$,

$$\left\|\widehat{A}^{\mathrm{MN}} - A^{\mathrm{MN}}\right\|_F^2 \leqslant Ce^{4C_1}(CKe^{C_1} + 1)^6K^2\frac{s^{\mathrm{MN}}\log M}{T},$$
$$\left\|\widehat{A}^{\mathrm{MN}} - A^{\mathrm{MN}}\right\|_R \leqslant Ce^{2C_1}(CKe^{C_1} + 1)^3Ks^{\mathrm{MN}}\sqrt{\frac{\log M}{T}},$$

where $C_1 = R_{\max}^{\mathrm{MN}} + \|\nu^{\mathrm{MN}}\|_\infty$, $c, C > 0$ *are universal constants.*

The proof can be found in Section 7.1.

Remark 3.1 *The upper bounds in Theorem 1 grow with* K, $R_{\max}^{\mathrm{MN}}$, *and* $\|\nu^{\mathrm{MN}}\|_\infty$. *To understand this phenomenon, we notice that when these three quantities increase,* $\left|\langle A^{\mathrm{MN}}, X^t\rangle + \nu^{\mathrm{MN}}\right|$ *may also increase, and hence the event probability in each category becomes more extreme (closer to* 0 *or* 1*). Extreme probabilities would then cause the decrease of both the curvature of the loss function and the eigenvalues of* $\mathrm{Cov}(X^t|X^{t-1})$.

This type of estimation error bound is widely seen in the high-dimensional statistics literature (see *e.g.,* Bickel et al. (2009); Zhang and Yang (2017)). As in Hall et al. (2016) and Mark et al. (2018), a martingale concentration inequality is applied to adapt to the time

series setting, and the major difference in this proof from past work includes lower bounds on the strong convexity parameter for our multinomial loss function, and the eigenvalues of covariance matrices of multinomial random vectors. One may be curious about why the singular values of A^{MN} are not required to be bounded by 1, similarly to the linear VAR model studied in prior works (Basu et al., 2015; Han et al., 2015). In fact, linear VAR models require this condition since it is necessary to ensure that the magnitude of the time series data would not get larger and larger ("explode") when t increases. In contrast, under our set-up, each entry of the data $X^t_{m,k}$ is bounded, and thus this would not be a concern.

3.2 Logistic-normal Model with Constant $q^t = q$

Theorem 2 *Consider the generation process* (7), (8) *with* $q^t = q$ *and estimator* (9). *If* $K \leqslant M$, $T \geqslant \frac{CK^4}{q_m^2\gamma_1^2}(\rho_m^{\mathrm{LN}})^2 \log M$, *and* $\lambda = CK \max_k \Sigma_{k,k}\sqrt{\frac{\max_m T_m \log M}{T^2}}$, *where* $T_m = \sum_{t=1}^T \mathbb{1}_{\{X_m^t \neq 0\}}$, *then with probability at least* $1 - 3\exp\{-c\log M\}$,

$$\begin{aligned}
\|\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}\|_F^2 \leqslant & \frac{C\max_k \Sigma_{k,k} K^2}{\gamma_1^2}\frac{s^{\mathrm{LN}}\max_m q_m \log M}{\min_m q_m^2 T}, \\
\|\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}\|_R \leqslant & \frac{C\sqrt{\max_k \Sigma_{k,k}}K}{\gamma_1} s^{\mathrm{LN}}\sqrt{\frac{\max_m q_m \log M}{\min_m q_m^2 T}}.
\end{aligned} \tag{16}$$

Here $c, C > 0$ *are universal constants,*

$$\begin{aligned}
\gamma_1 =& \min\left\{\frac{\min_j q_j\beta_1}{4K+1}, \frac{\min_j q_j(1-q_j)}{4K}\right\}, \\
\beta_1 =& \frac{2(e-1)^2}{e^6(2\pi)^{\frac{K-1}{2}}K^2|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{(K-1)(R_{\max}^{\mathrm{LN}} + \|\nu^{\mathrm{LN}}\|_\infty + 2)^2}{2\lambda_{\min}(\Sigma)}\right\}.
\end{aligned}$$

The proof is provided in Section 7.2.

Remark 3.2 *In the error bounds* (16), γ_1 *encodes the curvature of the loss (the larger, the better), while* $K\max_k \Sigma_{k,k}$ *reveals the effect of model parameters on the deviation bound* $\|\nabla L^{\mathrm{LN}}(A^{\mathrm{LN}})\|_R$. *Larger* K *and* Σ *lead to larger deviations. The curvature term* γ_1 *is smaller if there are more categories (larger* K*), more extreme event probability* q *(close to* 0 *or* 1*), and larger parameter values (larger* $R_{\max}^{\mathrm{LN}} + \|\nu^{\mathrm{LN}}\|_\infty$*). A more extreme event probability* q_m *leads to a lower variance of* $\mathbb{1}_{\{X_m^{t+1}\neq 0\}}$ *conditioning on* X^t, *while larger* $R_{\max}^{\mathrm{LN}} + \|\nu^{\mathrm{LN}}\|_\infty$ *leads to smaller covariances (controlled by* β_1*) of logistic-normal vectors* $Z_m^{t+1}, 1 \leqslant m \leqslant M$.

Remark 3.3 *One challenge for lower bounding the curvature term* γ_1 *lies in lower bounding the smallest eigenvalue of* $\mathrm{Cov}(Z_m^t|X^{t-1}) \in \mathbb{R}^{(K-1)\times(K-1)}$ *for* $1 \leqslant m \leqslant M$, *which does not have a closed analytical form for each entry. Instead, for any vector* $u \in \mathbb{R}^{K-1}$, *we lower bound* $\mathrm{Var}(u^\top Z_m^t|X^{t-1})$ *by showing that* $\mathbb{P}(u^\top Z_m^t \leqslant -c\|u\|_2|X^{t-1}), \mathbb{P}(u^\top Z_m^t \geqslant c\|u\|_2|X^{t-1}) \geqslant p$ *for some* $0 < p < 1$, *which implies that* $\mathrm{Var}(u^\top Z_m^t|X^{t-1}) \geqslant 2c^2p\|u\|_2^2$. *More details are presented in the proof for Lemma 6 in Section 7.*

The error bounds in Theorem 2 have an extra factor depending on q. If $q_m = q_0$ for $1 \leqslant m \leqslant M$ and some $0 < q_0 < 1$, then this factor becomes $\frac{1}{q_0}$. If q_m's differ too much from each other, a better choice is to use specific $\lambda_m = C\sqrt{\frac{T_m \log M}{T^2}}$ for the estimation of each A_m^{LN}, which would lead to a term $\frac{1}{q_m}$ instead of $\frac{\max_{m'} q_{m'}}{q_m^2}$ in the error bounds. This extra factor can be understood as follows: under the multinomial model (4), the number of samples for estimating A_m^{MN} is T, while in this section, the expected number of samples is $q_m T$ for estimating A_m^{LN}.

Different from Theorem 2, the estimation error rates for the other two models (the multinomial model and the logistic-normal model with q^t depending on the past) presented in Theorem 1 and Theorem 3 do not explicitly depend on $\mathbb{P}(X_m^t \neq 0_{K\times 1})$. (Recall that $\mathbb{P}(X_m^t \neq 0_{K\times 1}) = q_m$ in Theorem 2.) This is because, under the models of Theorems 1 and 3, the event probability $\mathbb{P}(X_m^t \neq 0_{K\times 1})$ depends on the model parameters A^{MN} and ν^{MN}, or B^{Bern} and η^{Bern}. In these cases, $\mathbb{P}(X_m^t \neq 0_{K\times 1})$ can be lower bounded by a function of $R_{\max}^{\mathrm{MN}}, \|\nu^{\mathrm{MN}}\|_\infty$, or a function of $R_{\max}^{\mathrm{LN,Bern}}, \|\eta^{\mathrm{Bern}}\|_\infty$. Therefore, the influence of event probability upon estimation accuracy is absorbed in $(1 + CKe^{R_{\max}^{\mathrm{MN}} + \|\nu^{\mathrm{MN}}\|_\infty})^6$ in Theorem 1 and $(1 + e^{R_{\max}^{\mathrm{LN,Bern}} + \|\eta^{\mathrm{Bern}}\|_\infty + 1})^6$ in Theorem 3.

3.3 Logistic-normal Model with q^t Depending on the Past

Theorem 3 *Consider the generation process* (7), (8), (11) *and estimator* (13) *for some* $0 \leqslant \alpha < 1$.[4] *If* $\lambda = C(\alpha)K\sqrt{\frac{\log M}{T}}$, $K \leqslant M$, *and* $T \geqslant \frac{CC(\alpha)^2K^4}{(1-\alpha)\gamma_2^2}(\rho^{\mathrm{LN,Bern}})^2 \log M$, *then with probability at least* $1 - 3\exp\{-c\log M\}$,

$$\alpha\|\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}\|_F^2 + (1-\alpha)\|\widehat{B}^{\mathrm{Bern}} - B^{\mathrm{Bern}}\|_F^2 \leqslant \frac{9C(\alpha)^2K^2}{\gamma_2^2}\frac{s^{\mathrm{LN,Bern}}\log M}{T},$$
$$R_\alpha(\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}, \widehat{B}^{\mathrm{Bern}} - B^{\mathrm{Bern}}) \leqslant \frac{12C(\alpha)K}{\gamma_2}s^{\mathrm{LN,Bern}}\sqrt{\frac{\log M}{T}},$$

where $R_\alpha(\cdot,\cdot)$ *is defined in* (2), $C(\alpha) = [C\max_k \Sigma_{k,k}\alpha + C'(1-\alpha)]^{\frac{1}{2}}$,

$$\begin{aligned}
\gamma_2 =& \frac{e^{C_2+1}}{2\left(1+e^{C_2+1}\right)^3}\min\left\{\frac{\beta_2}{4K+1}, \frac{e^{C_2}}{4K(1+e^{C_2})}\right\},\\
\beta_2 =& \frac{2(e-1)^2}{e^6(2\pi)^{\frac{K-1}{2}}K^2|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{(K-1)(C_3+2)^2}{2\lambda_{\min}(\Sigma)}\right\},\\
C_2 =& R_{\max}^{\mathrm{LN,Bern}} + \|\eta^{\mathrm{Bern}}\|_\infty, C_3 = R_{\max}^{\mathrm{LN,Bern}} + \|\nu^{\mathrm{LN}}\|_\infty,
\end{aligned} \tag{17}$$

and $c, C, C' > 0$ *are universal constants.*

The proof can be found in Section 7.3.

Similar to Theorem 1 and Theorem 2, the curvature term γ_2 is smaller if there are more categories and larger parameter values (larger K, $R_{\max}^{\mathrm{LN,Bern}}$, $\|\nu^{\mathrm{LN}}\|_\infty$, $\|\eta^{\mathrm{Bern}}\|_\infty$). Larger $C_2 =$

4. Although Theorem 3 is only stated for $0 \leqslant \alpha < 1$, our proof also leads to the same estimation error bound if $\alpha = 1$, for $T \geqslant C(\rho^{\mathrm{LN,Bern}})^2 \log M$ instead of $T \geqslant \frac{CK^4}{\gamma_2^2}(\rho_m^{\mathrm{LN,Bern}})^2 \log M$.

$R_{\max}^{\mathrm{LN,Bern}}+\|\eta^{\mathrm{Bern}}\|_\infty$ leads to more extreme event probability q_m^t, while larger $C_3 = R_{\max}^{\mathrm{LN,Bern}}+\|\nu^{\mathrm{LN}}\|_\infty$ could cause more extreme means of log-ratios $\mu^t(A_m^{\mathrm{LN}}, \nu_m^{\mathrm{LN}}) = \mathbb{E}\left(\log \frac{Z_{m,1:(K-1)}^t}{Z_{m,K}^t}|X^{t-1}\right)$, and both extreme q_m^t and $\mu^t(A_m^{\mathrm{LN}}, \nu_m^{\mathrm{LN}})$ contribute to smaller covariance of X_m^t.

When $0 < \alpha < 1$, the estimation errors for A^{LN} and B^{Bern} are implied directly, although they may be loose in their dependence on α. It's difficult to determine an optimal α for estimation based on the theoretical result. Intuitively we need α to be away from 0 and 1 so that we boost the estimation performance by pooling the two estimation tasks together. We will demonstrate the interplay between α and the noise level Σ in terms of estimation errors in the numerical results in Section 5.2.3.

Remark 3.4 *Here we explain some connections between Theorem 2 and Theorem 3. If we let* $\alpha = \frac{C'}{C\max_k \Sigma_{k,k}+C'}$ *for estimating the logistic-normal model with time-varying* q^t*, then Theorem 3 implies that*

$$\begin{aligned}\|\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}\|_F^2 \leqslant & \frac{9C(\alpha)^2K^2}{\alpha\gamma_2^2}\frac{s^{\mathrm{LN,Bern}}\log M}{T} \\ \leqslant & \frac{C\max_k \Sigma_{k,k}K^2}{\gamma_2^2}\frac{s^{\mathrm{LN,Bern}}\log M}{T}.\end{aligned}$$

Compared to Theorem 2, we can see that the only difference in the upper bounds for $\|\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}\|_F^2$ *is that the event probability term* q_m *in Theorem 2 changes to* $\frac{e^{C_2}}{1+e^{C_2}}$ *or* $\frac{1}{1+e^{C_2}}$ *in Theorem 3. This is because that we have* $\frac{1}{1+e^{C_2}} \leqslant q_m^t \leqslant \frac{e^{C_2}}{1+e^{C_2}}$ *for any* t, m*, under the logistic-normal model with time-varying* q^t*.*

4. Post Hoc Signed Variable Importance Network

So far, we have defined an absolute network parameter A^{MN} for the multinomial model, a relative network parameter A^{LN} and overall network parameter B^{Bern} for the logistic-normal model, which have different interpretations due to the nature of the corresponding models. However, in real applications, influence networks that share the same meaning across different models are usually desired to facilitate comparison.

Therefore, in this section, we consider a more model-agnostic approach to determine edge presence and weights by calling on the recent literature on *variable importance and post hoc interpretation methods* (see e.g. Breiman et al. (2001); Strobl et al. (2008); Grömping (2009); Féraud and Clérot (2002)). This allows us to develop a post hoc signed variable importance network for any multivariate autoregressive model, which focuses on the models' ability to predict future data and is therefore not as sensitive to the specific choice of model parameterization. First, we revisit the literature on variable importance and post hoc interpretation methods, which serve as the inspiration for our approach.

Past literature on variable importance and post hoc interpretations: The variable importance or predictive importance measure has been widely studied in random forests (Breiman et al., 2001; Strobl et al., 2008; Grömping, 2009) and neural networks (Féraud and Clérot, 2002; Lundberg and Lee, 2017),where the key idea is to measure the effect of each predictor on the prediction results.

Among these past works, our approach is most closely related to the post hoc interpretation methods proposed for neural networks. For example, suppose that there are d predictors $X_1, \ldots, X_d$ used for predicting the response, and $f(X_1, \ldots, X_d)$ is the fitted prediction function based on the complete data with all predictors. Given the fitted function f, Féraud and Clérot (2002) measures the variable importance of the predictor of interest, say X_1, by looking into the average change in $f(X_1, \ldots, X_d)$ when X_1 changes by some value δ that follows certain distributions. More specifically, they consider the following quantity:

$$\int_{\delta} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(\delta | x_{i,1}) \left(f(x_{i,1}, \ldots, x_{i,d}) - f(x_{i,1} + \delta, \ldots, x_{i,d}) \right) \right| d\delta, \tag{18}$$

where $x_i, i = 1, \ldots, n$ are sample versions of X. Here $\mathbb{P}(\delta | x_{i,1})$ should characterize how likely we would encounter an observation $X_1 = x_{i,1} + \delta$, and thus (18) shows how the prediction result would be affected if X_1 changes to some comparison baseline that likely appears in data.

General idea of our approach and connection to past work: In this paper, we consider a similar strategy to the approach of Féraud and Clérot (2002) discussed above, with some modifications suited for our needs. First, we describe our strategy under the same context as the aforementioned neural network example and then extend it to any multivariate autoregressive models.

As mentioned earlier, one understanding of (18) is that it reflects the change in prediction when X_1 changes to some baseline that is likely to appear in data. Based on this idea and for simplicity, we use a fixed comparison baseline that is representative of the distribution of X_1, and one natural choice would be the sample average $\bar{x}_1 = \frac{1}{n} \sum_{i'=1}^{n} x_{i',1}$. Hence we consider the following quantity as the *post hoc variable importance* of X_1:

$$\frac{1}{n} \sum_{i=1}^{n} \left| f(x_{i,1}, \ldots, x_{i,d}) - f(\bar{x}_1, \ldots, x_{i,d}) \right|, \tag{19}$$

where each term inside the summation in (19) quantifies how much the deviation of $x_{i,1}$ from its average value influences the prediction. In fact, (19) is a special case of (18) under a particular $\mathbb{P}(\delta | x_{i,1})$, and a detailed explanation for this connection is presented in Appendix B.4.

Furthermore, we want our variable importance parameter to reflect whether the influence of X_1 upon the prediction is stimulatory or inhibitory. The influence is stimulatory if the increase (decrease) of $x_{i,1}$ from its average $\bar{x}_1$ leads to the increase (decrease) of $f(x_{i,1}, \ldots, x_{i,d})$ from $f(\bar{x}_1, \ldots, x_{i,d})$, while it is inhibitory if the increase (decrease) of the former leads to the decrease (increase) of the latter. Therefore, we consider the *post hoc signed variable importance* of X_1 defined as follows:

$$\frac{1}{n} \sum_{i=1}^{n} \operatorname{sgn}(x_{i,1} - \bar{x}_1) \left(f(x_{i,1}, \ldots, x_{i,d}) - f(\bar{x}_1, \ldots, x_{i,d}) \right), \tag{20}$$

which is positive for stimulatory effect but negative for inhibitory effect. We will illustrate how to extend the definition (20) to the post hoc signed variable importance network notion under the autoregressive model setting in the following.

Post hoc signed variable importance network definition: We now define the variable importance network parameter $V \in \mathbb{R}^{M\times K\times M\times K}$ for any multivariate autoregressive model with data $\{X^t \in \mathbb{R}^{M\times K}\}_{t=0}^{T-1}$. Our goal is to let each entry V_{m_1,k_1,m_2,k_2} reflect the signed variable importance of $X^t_{m_2,k_2}$ for predicting $X^{t+1}_{m_1,k_1}$. Similar to (20), where we consider how the prediction function $\hat{f}$ changes as $x_{i,1}$ deviates from the average value $\bar{x}_1$, now we define V_{m_1,k_1,m_2,k_2} by how much the prediction function $\mathbb{E}(X^{t+1}_{m_1,k_1}|X^t)$ changes as $X^t_{m_2,k_2}$ deviates from $\bar{X}_{m_2,k_2} = \frac{1}{T}\sum_{t'=0}^{T-1} X^{t'}_{m_2,k_2}$. Define $\bar{X}^t(m_2,k_2) \in \mathbb{R}^{M\times K}$ as the comparison baseline for X^t:

$$(\bar{X}^t(m_2,k_2))_{m,k} = \begin{cases} X^t_{m,k}, & (m,k) \neq (m_2,k_2), \\ \bar{X}_{m,k}, & (m,k) = (m_2,k_2), \end{cases}$$

which equals X^t at all entries other than (m_2,k_2) and takes the value of $\bar{X}_{m,k}$ at entry (m_2,k_2). The variable importance of $X^t_{m_2,k_2}$ for predicting $X^{t+1}_{m_1,k_1}$ can then be measured for any model as follows:

$$\begin{aligned} &V_{m_1,k_1,m_2,k_2} \\ :=&\frac{1}{T}\sum_{t=0}^{T-1} \operatorname{sgn}(X^t_{m_2,k_2} - \bar{X}_{m,k}) \left(\mathbb{E}(X^{t+1}_{m_1,k_1}|X^t) - \mathbb{E}(X^{t+1}_{m_1,k_1}|\bar{X}^t(m_2,k_2))\right), \end{aligned} \tag{21}$$

which resembles (20), and whose sign suggests whether the influence is stimulatory or inhibitory. Here the expectation function $\mathbb{E}(X^{t+1}_{m_1,k_1}|X^t)$ depends on the ground truth, and thus the variable importance parameter V defined here is an unknown population quantity. However, for any method that performs one-step-ahead prediction, one can simply substitute $\mathbb{E}(X^{t+1}_{m_1,k_1}|X^t)$ in (21) with the prediction output by the method. Specifically, for our three models, we define the ground truth variable importance network parameters V^{MN}, V^{LN}, $V^{\mathrm{LN,Bern}} \in \mathbb{R}^{M\times K\times M\times K}$ and their estimates $\widehat{V}^{\mathrm{MN}}$, $\widehat{V}^{\mathrm{LN}}$, $\widehat{V}^{\mathrm{LN,Bern}}$ as follows:

1. The multinomial model:

$$\begin{aligned} V^{\mathrm{MN}}_{m_1,k_1,m_2,k_2} =&\frac{1}{T}\sum_{t=0}^{T-1} \operatorname{sgn}(X^t_{m_2,k_2} - \bar{X}_{m_2,k_2}) \\ &\cdot \left[\mathbb{E}_{\mathrm{MN}}(X^{t+1}_{m_1,k_1}|X^t) - \mathbb{E}_{\mathrm{MN}}(X^{t+1}_{m_1,k_1}|\bar{X}^t(m_2,k_2))\right], \end{aligned} \tag{22}$$

   where the subscript MN means that we take the conditional expectation under the multinomial model, and

$$\mathbb{E}_{\mathrm{MN}}(X^{t+1}_{m_1,k_1}|X^t) = \frac{e^{\langle A^{\mathrm{MN}}_{m_1,k_1,:,:}, X^t\rangle + \nu^{\mathrm{MN}}_{m_1,k_1}}}{1+\sum_{k=1}^{K} e^{\langle A^{\mathrm{MN}}_{m_1,k,:,:}, X^t\rangle + \nu^{\mathrm{MN}}_{m_1,k}}}. \tag{23}$$

   To estimate V^{MN} that depends on A^{MN}, we can substitute A^{MN} in (23) by $\widehat{A}^{\mathrm{MN}}$ defined in (5) and obtain $\widehat{V}^{\mathrm{MN}} \in \mathbb{R}^{M\times K\times M\times K}$.

2. The logistic-normal model with constant q^t:
   Since there is no closed-form expression for calculating $\mathbb{E}_{\mathrm{LN}}(X^{t+1}_m|X^t)$ due to the

nature of logistic-normal distribution, we consider the following alternative as the prediction:
$\mathbb{E}_{\mathrm{LN}}(X_m^{t+1}|X^t, \epsilon^{t+1}=0)$. This is the conditional expectation of X_m^{t+1} if there is no Gaussian noise ϵ_m^{t+1} in (8). Then we can define $V^{\mathrm{LN}}_{m_1,k_1,m_2,k_2}$ by

$$\begin{aligned} &V^{\mathrm{LN}}_{m_1,k_1,m_2,k_2} \\ =&\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{sgn}(X^t_{m_2,k_2}-\bar{X}_{m_2,k_2}) \\ &\cdot\left[\mathbb{E}_{\mathrm{LN}}(X^{t+1}_{m_1,k_1}|X^t,\epsilon^{t+1}=0)-\mathbb{E}_{\mathrm{LN}}(X^{t+1}_{m_1,k_1}|\bar{X}^t(m_2,k_2),\epsilon^{t+1}=0)\right], \end{aligned} \tag{24}$$

where

$$\begin{aligned} \mathbb{E}_{\mathrm{LN}}(X^{t+1}_{m_1,k_1}|X^t,\epsilon^{t+1}=0) =&q_{m_1}\frac{e^{\langle A^{\mathrm{LN}}_{m_1,k_1,:,:},X^t\rangle+\nu^{\mathrm{LN}}_{m_1,k_1}}}{1+\sum_{k=1}^{K-1}e^{\langle A^{\mathrm{LN}}_{m_1,k,:,:},X^t\rangle+\nu^{\mathrm{LN}}_{m_1,k}}}, \quad k_1<K \\ \mathbb{E}_{\mathrm{LN}}(X^{t+1}_{m_1,K}|X^t,\epsilon^{t+1}=0) =&q_{m_1}\frac{1}{1+\sum_{k=1}^{K-1}e^{\langle A^{\mathrm{LN}}_{m_1,k,:,:},X^t\rangle+\nu^{\mathrm{LN}}_{m_1,k}}}. \end{aligned} \tag{25}$$

Similarly, we can estimate V^{LN} by substituting A^{LN} in (25) by $\widehat{A}^{\mathrm{LN}}$ defined in (9) and obtain $\widehat{V}^{\mathrm{LN}} \in \mathbb{R}^{M\times K\times M\times K}$.

3. The logistic-normal model with q^t depending on the past: The definition for $V^{\mathrm{LN,Bern}}$ is basically the same as that of V^{LN}, except that the expectation $\mathbb{E}_{\mathrm{LN,Bern}}(\cdot)$ would take a different form:

$$\begin{aligned} &V^{\mathrm{LN,Bern}}_{m_1,k_1,m_2,k_2} \\ =&\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{sgn}(X^t_{m_2,k_2}-\bar{X}_{m_2,k_2}) \\ &\cdot\left[\mathbb{E}_{\mathrm{LN,Bern}}(X^{t+1}_{m_1,k_1}|X^t,\epsilon^{t+1}=0)-\mathbb{E}_{\mathrm{LN,Bern}}(X^{t+1}_{m_1,k_1}|\bar{X}^t(m_2,k_2),\epsilon^{t+1}=0)\right], \end{aligned} \tag{26}$$

where

$$\begin{aligned} &\mathbb{E}_{\mathrm{LN,Bern}}(X^{t+1}_{m_1,k_1}|X^t,\epsilon^{t+1}=0) \\ =&\frac{e^{\langle B^{\mathrm{Bern}}_{m_1,:,:},X^t\rangle+\eta^{\mathrm{Bern}}_{m_1}}}{1+e^{\langle B^{\mathrm{Bern}}_{m_1,:,:},X^t\rangle+\eta^{\mathrm{Bern}}_{m_1}}}\frac{e^{\langle A^{\mathrm{LN}}_{m_1,k_1,:,:},X^t\rangle+\nu^{\mathrm{LN}}_{m_1,k_1}}}{1+\sum_{k=1}^{K-1}e^{\langle A^{\mathrm{LN}}_{m_1,k,:,:},X^t\rangle+\nu^{\mathrm{LN}}_{m_1,k}}}, \quad k_1<K \\ &\mathbb{E}_{\mathrm{LN,Bern}}(X^{t+1}_{m_1,K}|X^t,\epsilon^{t+1}=0) \\ =&\frac{e^{\langle B^{\mathrm{Bern}}_{m_1,:,:},X^t\rangle+\eta^{\mathrm{Bern}}_{m_1}}}{1+e^{\langle B^{\mathrm{Bern}}_{m_1,:,:},X^t\rangle+\eta^{\mathrm{Bern}}_{m_1}}}\frac{1}{1+\sum_{k=1}^{K-1}e^{\langle A^{\mathrm{LN}}_{m_1,k,:,:},X^t\rangle+\nu^{\mathrm{LN}}_{m_1,k}}}. \end{aligned} \tag{27}$$

We will substitute A^{LN}, B^{Bern} by $\widehat{A}^{\mathrm{LN}}$, $\widehat{B}^{\mathrm{Bern}}$ for calculating (27) and obtain $\widehat{V}^{\mathrm{LN,Bern}} \in \mathbb{R}^{M\times K\times M\times K}$.

Given the definitions above, for any time series dataset $\{X^t \in \mathbb{R}^{M\times K}\}_{t=0}^T$ and any chosen model (the three models proposed in this paper or any model that can be estimated

and perform one-step-ahead prediction), we are now able to obtain an estimated post hoc signed variable importance network. The estimated variable importance network can then be visualized and provide insights into the influence patterns among nodes. Moreover, we can compare the estimated variable importance networks generated under different models to understand the advantages and disadvantages of these modeling approaches. This is more reasonable than directly comparing the estimated model parameters ($\widehat{A}^{\mathrm{MN}}$, $\widehat{A}^{\mathrm{LN}}$, and $\widehat{B}^{\mathrm{Bern}}$), which have different interpretations, as mentioned at the beginning of Section 4.

Estimation error bounds: Based on the estimation error bounds for $\widehat{A}^{\mathrm{MN}}$, $\widehat{A}^{\mathrm{LN}}$ and $\widehat{A}^{\mathrm{LN,Bern}}$ in Section 3, we can also prove the following error bounds on variable importance parameters $\widehat{V}^{\mathrm{MN}}$, $\widehat{V}^{\mathrm{LN}}$ and $\widehat{V}^{\mathrm{LN,Bern}}$ under each corresponding model.

Proposition 1 *1. Under the same conditions as Theorem 1, with probability at least* $1-3\exp\{-c\log M\}$,

$$\left\|\widehat{V}^{\mathrm{MN}}-V^{\mathrm{MN}}\right\|_F^2 \leqslant Ce^{4C_1}(CKe^{C_1}+1)^6K^3(\rho^{\mathrm{MN}})^2\frac{s^{\mathrm{MN}}\log M}{T},$$

where $c, C>0$ are universal constants and C_1 is as defined in Theorem 1.

2. Under the same conditions as Theorem 2, with probability at least $1-3\exp\{-c\log M\}$,

$$\left\|\widehat{V}^{\mathrm{LN}}-V^{\mathrm{LN}}\right\|_F^2 \leqslant \frac{C\max_k\Sigma_{k,k}^2K^3(\rho^{\mathrm{LN}})^2}{\gamma_1^2}\frac{s^{\mathrm{LN}}\max_m q_m\log M}{\min_m q_m^2 T},$$

where $c, C>0$ are universal constants and γ_1 is as defined in Theorem 2.

3. Under the same conditions as Theorem 3, with probability at least $1-3\exp\{-c\log M\}$,

$$\left\|\widehat{V}^{\mathrm{LN,Bern}}-V^{\mathrm{LN,Bern}}\right\|_F^2 \leqslant \frac{CC(\alpha)^2K^3(\rho^{\mathrm{LN,Bern}})^2}{\gamma_2^2\min\{\alpha,1-\alpha\}}\frac{s^{\mathrm{LN,Bern}}\log M}{T},$$

where $c, C>0$ are universal constants, and γ_2 is defined in Theorem 3.

The proof can be found in Section 7.4.

Compared to the error bounds for $\|\widehat{A}^{\mathrm{MN}}-A^{\mathrm{MN}}\|_F^2$, $\|\widehat{A}^{\mathrm{LN}}-A^{\mathrm{LN}}\|_F^2$ and $\|\widehat{A}^{\mathrm{LN,Bern}}-A^{\mathrm{LN,Bern}}\|_F^2$ in Section 3, Proposition 1 has an additional term $K(\rho^{\mathrm{MN}})^2$ (or $K(\rho^{\mathrm{LN}})^2$, $K(\rho^{\mathrm{LN,Bern}})^2$). This is because that each entry of V^{MN}, V^{LN} and $V^{\mathrm{LN,Bern}}$ involves MK entries of A^{MN}, A^{LN} and $A^{\mathrm{LN,Bern}}$, and we have taken the sparsity into account.

5. Synthetic Data Simulation

In this section, we conduct simulation studies to validate our approaches in two ways: we first use synthetic data generated according to the three aforementioned models to validate our theoretical results on the rates of estimation error; and then test our method(s) on data generated from a synthetic mixture model, which is a hybrid of the multinomial model (4) and the logistic-normal model with q depending on the past ((7), (8) and (11)). The latter experiment is inspired by real applications and illustrates the advantages and disadvantages of our multinomial and logistic-normal methods. Moreover, it inspires us

to develop a mixture approach based on a testing procedure, where one can make a data-dependent choice over the multinomial and logistic-normal methods for each node and then fit a mixture model based on the node-wise choices. The mixture approach also provides a unified view of the underlying network and enjoys promising performance compared to the multinomial and logistic-normal approaches.

5.1 Numerical Details

Optimization algorithm: For all numerical experiments, we use the standard proximal gradient descent algorithm with a group sparsity penalty (Wright et al., 2009) to solve the optimization problems. In particular, for solving (5) and (9), we reparameterize $\{A_m\}_{m=1}^M$ to vectors in $\mathbb{R}^{MK^2}$ and $\mathbb{R}^{MK(K-1)}$ with group sizes K^2 and $K(K-1)$, respectively. To solve (13), we reparameterize $\{(\sqrt{\alpha}A_m, \sqrt{1-\alpha}B_m)\}_{m=1}^M$ to vectors in $\mathbb{R}^{MK^2}$ with group size K^2. A vector soft-threshold method can then be applied in each iteration.

Choices for tuning parameters: Across the experiments in Section 5.2, we use penalty parameter $\lambda = C_\lambda K\sqrt{\frac{\log M}{T}}$, where the constant C_λ is selected for each model via cross-validation. The detailed cross-validation procedure is included in Appendix D.1. Since the purpose of Section 5.2 is to validate our theoretical error rates, we do not tune α for the logistic-normal model with q^t depending on the past: we either use a fixed α (Figures 6, 7, 8) or run experiments for each α from a list (Figure 9).

While for the synthetic mixture experiments in Section 5.3.2 and real data applications in Section 6, cross-validation is done for each model and each data set separately. In addition, in Section 5.3.2 and Section 6, both α and λ need to be tuned for the logistic-normal model with event probability depending on the past.

5.2 Estimation Error Rates

For each of the three generation processes defined in Section 2, we investigate the performance of the corresponding estimators (5),(9), and (13). For all the figures in this section, the averages of 50 trials are shown, and error bars are the standard deviations.

5.2.1 Multinomial Model

The synthetic data is generated according to (4) (initial data $\{X_m^0\}_{m=1}^M$ are i.i.d. multinomial random vectors), and A^{MN} is estimated by (5). Under all settings, for each m, the $\rho_m^{\mathrm{MN}} = \frac{s^{\mathrm{MN}}}{M}$ non-zero slices $A^{\mathrm{MN}}_{m,:,m',:}$ are sampled uniformly from $1 \leqslant m' \leqslant M$. We set $K = 2$, and given that $A^{\mathrm{MN}}_{m,:,m',:}$ is non-zero, each of its K^2 entries is sampled independently from $U(-2, 2)$. To ensure the same baseline event rate under the three generation processes, which is set as 0.8, we let $\nu^{\mathrm{MN}} = (\log \frac{4}{K})_{M\times K}$. The tuning parameter $\lambda = 0.12 \times K\sqrt{\frac{\log M}{T}}$ where 0.12 arises from cross-validation, as explained in Section 5.1. The scaling of mean squared error $\|\widehat{A}^{\mathrm{MN}} - A^{\mathrm{MN}}\|_F^2$, $\|\widehat{V}^{\mathrm{MN}} - V^{\mathrm{MN}}\|_F^2$ with respect to sparsity s^{MN}, dimension M and sample size T are shown in Figures 2, 3.

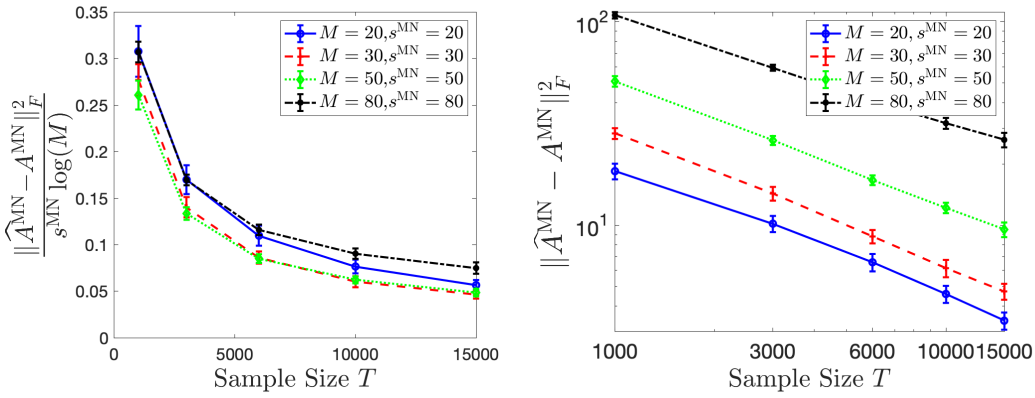


Figure 2: $\frac{\|\widehat{A}^{\mathrm{MN}}-A^{\mathrm{MN}}\|_F^2}{s^{\mathrm{MN}}\log M}$, $\|\widehat{A}^{\mathrm{MN}}-A^{\mathrm{MN}}\|_F^2$ vs. T under the multinomial data generation process and estimator (5), where the second plot has a log-scale. The scaling of $\|\widehat{A}^{\mathrm{MN}}-A^{\mathrm{MN}}\|_F^2$ with respect to $s^{\mathrm{MN}}\log M$ is similar to the theoretical bound. Its scaling w.r.t. T is a little larger than $\frac{1}{T}$ since the multinomial log-likelihood loss has a low curvature under our set-up of A.

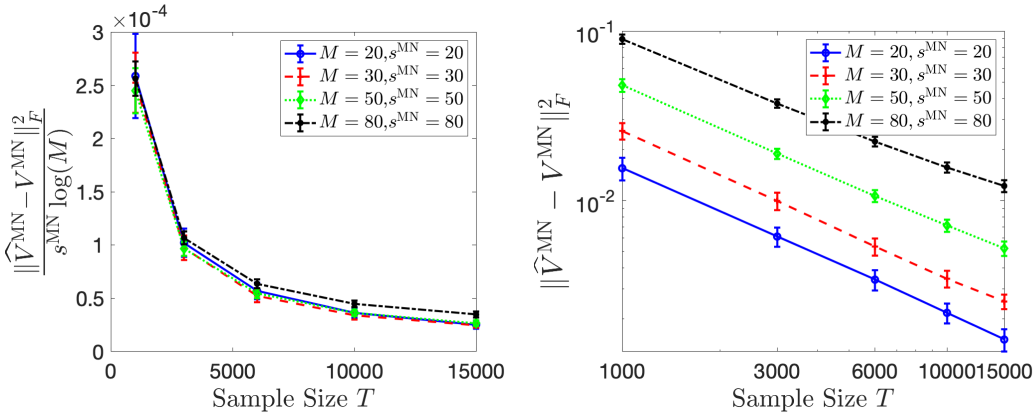


Figure 3: $\frac{\|\widehat{V}^{\mathrm{MN}}-V^{\mathrm{MN}}\|_F^2}{s^{\mathrm{MN}}\log M}$ and $\|\widehat{V}^{\mathrm{MN}}-V^{\mathrm{MN}}\|_F^2$ v.s. sample size T under the multinomial data generation process and estimator (5), where the second plot has a log-scale. The scaling of $\|\widehat{V}^{\mathrm{MN}}-V^{\mathrm{MN}}\|_F^2$ seems similar to that of $\|\widehat{A}^{\mathrm{MN}}-A^{\mathrm{MN}}\|_F^2$ in Figure 2.

5.2.2 Logistic-normal Model with $q^t = q$

Here the data is generated under (7) (initial data $\{X_m^0\}_{m=1}^M$ are i.i.d. multinomial random vectors) and (8) with constant vector $q = (0.8)^{M\times 1}$, and the estimator is as specified in (9). We set $K = 2$, the covariance $\Sigma = I_{(K-1)\times(K-1)}$ and intercept $\nu^{(\mathrm{MM})} = 0^{M\times(K-1)}$. $A^{\mathrm{LN}} \in \mathbb{R}^{M\times(K-1)\times M\times K}$ is generated in the same way as in Section 5.2.1, except that the dimension is different. The penalty parameter λ is set as $0.13\times K\sqrt{\frac{\log M}{T}}$, where 0.13 arises from cross-validation. The scaling of the mean squared error $\|\widehat{A}^{\mathrm{LN}}-A^{\mathrm{LN}}\|_F^2$, $\|\widehat{V}^{\mathrm{LN}}-V^{\mathrm{LN}}\|_F^2$

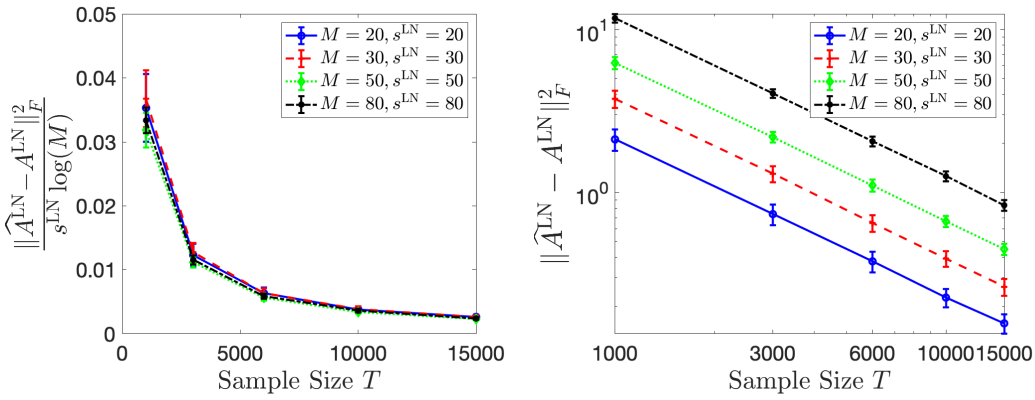


Figure 4: $\frac{\|\widehat{A}^{\mathrm{LN}}-A^{\mathrm{LN}}\|_F^2}{s^{\mathrm{LN}}\log M}$, $\|\widehat{A}^{\mathrm{LN}}-A^{\mathrm{LN}}\|_F^2$ vs. T under the logistic-normal data generation process with constant q^t and estimator (9), where the second figure is under log-scale. The scaling of MSE aligns well with Theorem 2 in s^{LN}, M, and T.

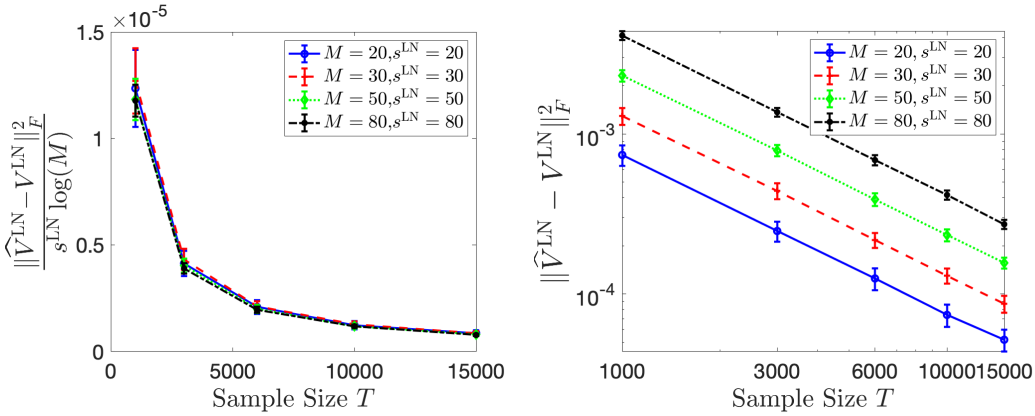


Figure 5: $\frac{\|\widehat{V}^{\mathrm{LN}}-V^{\mathrm{LN}}\|_F^2}{s^{\mathrm{LN}}\log M}$ and $\|\widehat{V}^{\mathrm{LN}}-V^{\mathrm{LN}}\|_F^2$ v.s. sample size T under the logistic-normal data generation process with constant q^t and estimator (9), where the second plot has a log-scale. The scaling of $\|\widehat{V}^{\mathrm{LN}}-V^{\mathrm{LN}}\|_F^2$ seems similar to that of $\|\widehat{A}^{\mathrm{LN}}-A^{\mathrm{LN}}\|_F^2$ in Figure 4.

with respect to sparsity s^{LN}, dimension M, and sample size T are shown in Figure 4 and Figure 5.

5.2.3 Logistic-normal Model with q^t Depending on the Past

We generate data according to (7), (8), and (11) (initial data $\{X_m^0\}_{m=1}^M$ are i.i.d. multinomial random vectors) and estimate A^{LN} and B^{Bern} using (13). For each $1 \leqslant m \leqslant M$, we sample the support set S_m uniformly from $[M] = \{1, \ldots, M\}$. Given that $A^{\mathrm{LN}}_{m,:,m',:}$ or $B^{\mathrm{Bern}}_{m,m',:}$ is non-zero, each entry is sampled independently from $U(-2, 2)$. We set $K = 2$, the covariance $\Sigma = I_{(K-1)\times(K-1)}$, intercept $\nu^{\mathrm{LN}} = (0)^{M\times(K-1)}$, and $\eta^{\mathrm{Bern}} = (\log 4)^{M\times 1}$ to ensure a base

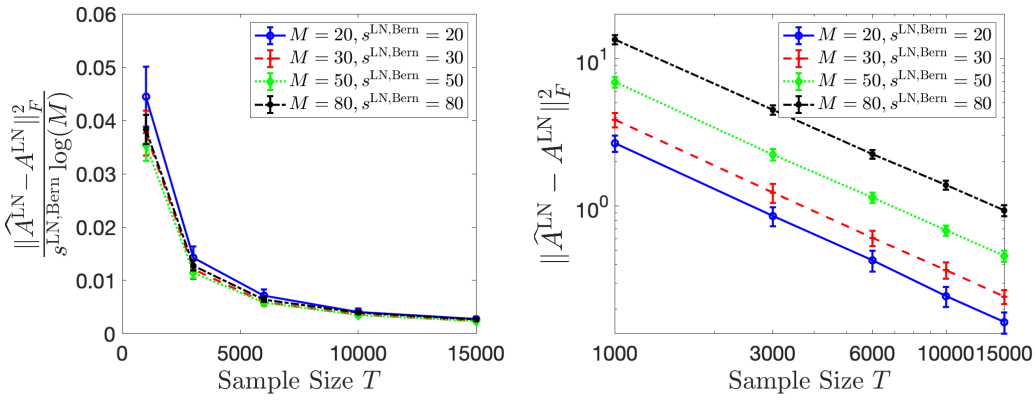


Figure 6: $\frac{\|\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}\|_F^2}{s^{\mathrm{LN,Bern}} \log M}, \|\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}\|_F^2$ vs. T under the logistic-normal data generation process with q^t depending on the past and estimator (13). The second plot is under log-scale. The scaling of $\|\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}\|_F^2$ aligns well with Theorem 3 in $s^{\mathrm{LN,Bern}}$, M and T.

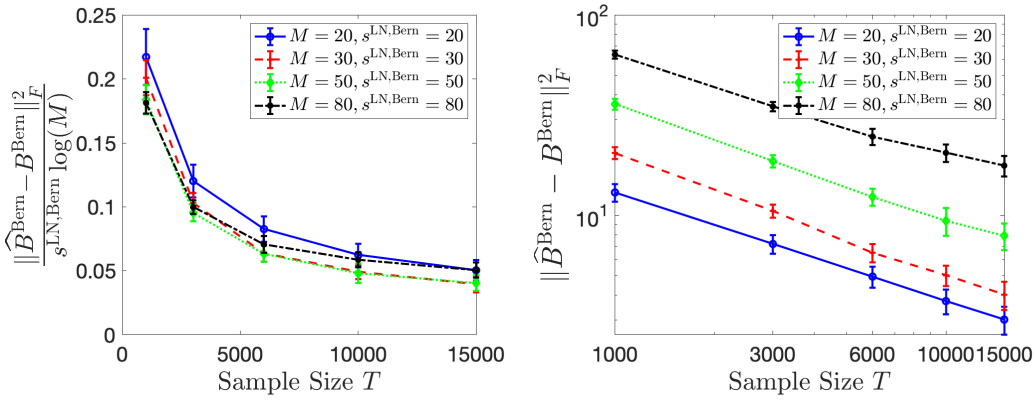


Figure 7: $\frac{\|\widehat{B}^{\mathrm{Bern}} - B^{\mathrm{Bern}}\|_F^2}{s^{\mathrm{LN,Bern}} \log M}, \|\widehat{B}^{\mathrm{Bern}} - B^{\mathrm{Bern}}\|_F^2$ vs. T under the logistic-normal data generation process with q^t depending on the past and estimator (13). The scaling of $\|\widehat{B}^{\mathrm{Bern}} - B^{\mathrm{Bern}}\|_F^2$ w.r.t. $s^{\mathrm{LN,Bern}} \log M$ is similar to the theoretical bound in Theorem 3. The second plot is under log-scale, and the scaling of $\|\widehat{B}^{\mathrm{Bern}} - B^{\mathrm{Bern}}\|_F^2$ w.r.t. T is a little larger than $\frac{1}{T}$ since the Bernoulli log-likelihood loss has a low curvature under our set-up of A.

probability of 0.8. The penalty parameter $\lambda = 0.08 \times K\sqrt{\frac{\log M}{T}}$ where 0.08 arises from cross-validation and $\alpha = 0.4$. We present the scaling of mean squared errors $\|\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}\|_F^2$, $\|\widehat{B}^{\mathrm{Bern}} - B^{\mathrm{Bern}}\|_F^2$, and $\|\widehat{V}^{\mathrm{LN,Bern}} - V^{\mathrm{LN,Bern}}\|_F^2$ in Figures 6, 7, and 8.

We also check the influence of α on the estimation error when the noise covariance Σ of the logistic-normal distribution varies. We consider the setting where $M = 20$, $s^{\mathrm{LN,Bern}} = 20$, $K = 2$, $T = 1000$, and each non-zero entry of A^{LN}, B^{Bern} is sampled from $U(-1, 1)$. We

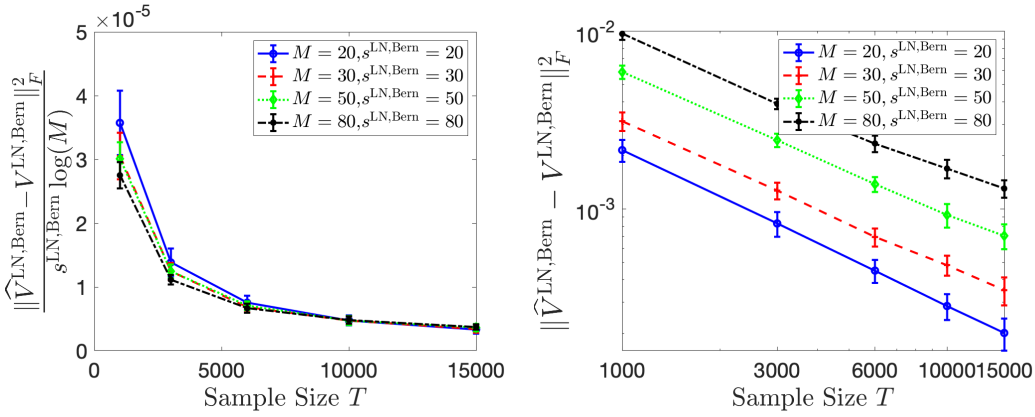


Figure 8: $\frac{\|\widehat{V}^{\mathrm{LN,Bern}} - V^{\mathrm{LN,Bern}}\|_F^2}{s^{\mathrm{LN,Bern}} \log M}$ and $\|\widehat{V}^{\mathrm{LN,Bern}} - V^{\mathrm{LN,Bern}}\|_F^2$ v.s. sample size T under the logistic-normal data generation process with q^t depending on the past and estimator (13), where the second plot has a log-scale. The scaling of $\|\widehat{V}^{\mathrm{LN,Bern}} - V^{\mathrm{LN,Bern}}\|_F^2$ also seems similar to that of $\|\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}\|_F^2$ in Figure 6.

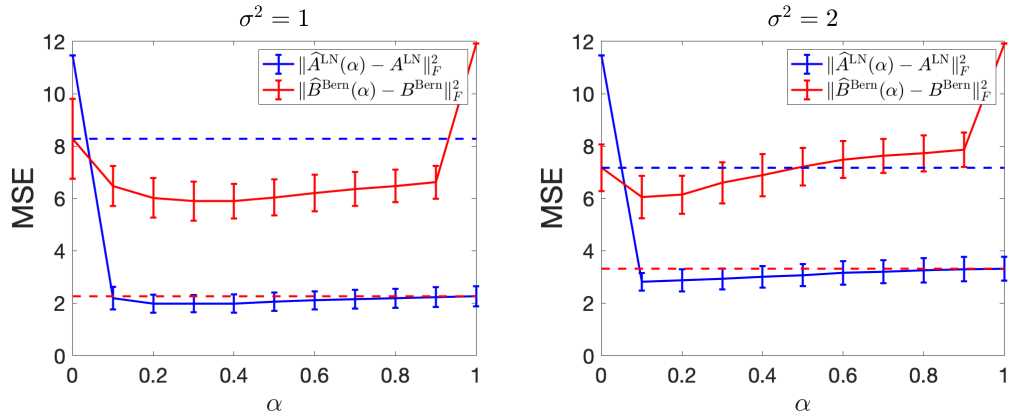


Figure 9: $\|\widehat{A}^{\mathrm{LN}}(\alpha) - A^{\mathrm{LN}}\|_F^2$ and $\|\widehat{B}^{\mathrm{Bern}}(\alpha) - B^{\mathrm{Bern}}\|_F^2$ v.s. α. The first figure shows the results when $\sigma^2 = 1$, while the second one is when $\sigma^2 = 2$. The dashed lines are $\|\widehat{A}^{\mathrm{LN}}(1) - A^{\mathrm{LN}}\|_F^2$ and $\|\widehat{B}^{\mathrm{Bern}}(0) - B^{\mathrm{Bern}}\|_F^2$. When $\alpha = 0$ or 1, $\widehat{A}^{\mathrm{LN}}$ or $\widehat{B}^{\mathrm{Bern}}$ would stay at the initializers (set as zeros tensors), while $\|\widehat{A}^{\mathrm{LN}}(1) - A^{\mathrm{LN}}\|_F^2$, $\|\widehat{B}^{\mathrm{Bern}}(0) - B^{\mathrm{Bern}}\|_F^2$ would be the estimation error of separate estimations. When α moves from the extremes (0 or 1) to the middle, the estimation errors of both are lower. When variance $\sigma^2 = 1$, choosing α around 0.4 would make $\|\widehat{A}^{\mathrm{LN}}(\alpha) - A^{\mathrm{LN}}\|_F^2$ and $\|\widehat{B}^{\mathrm{Bern}}(\alpha) - B^{\mathrm{Bern}}\|_F^2$ both lower than separate estimation. When $\sigma^2 = 2$, the figure suggests choosing a smaller α.

run 20 replicates for each α in $\{0, 0.1, 0.2, \ldots, 1\}$, and for each replicate, cross-validation is used for choosing λ. We set $\Sigma = \sigma^2 I_{(K-1)\times(K-1)}$ where $\sigma^2 = 1$ or 2, and Figure 9 shows that α should be smaller when σ^2 increases.

5.3 Synthetic Mixture Model

The simulation study in Section 5.2 shows that the three methods all perform well when data is generated from the models these methods are proposed for. However, in reality, and as we will see with our real data examples, for each event, we may always observe positive membership weights in multiple categories while we cannot directly tell if it is the *noisy observation for a single category* or it indeed represents true mixed membership. Meanwhile, data from real applications is unlikely to match a true model. In particular, one might expect that: (i) some nodes' events have *mixed memberships in different categories*, (ii) while other nodes in the network only focus on one particular category of events, and thus each of their events *falls in one category*. This is inspired by a news media example where some media sources cover multiple topics, and others focus primarily on one topic.

One key question is, how would our approaches work under this complicated real-world situation? In this section, we design a *contaminated mixture model* to mimic this situation and provide numerical evidence for our central hypothesis: *The logistic-normal approach will be more effective at estimating edges among nodes whose events exhibit mixed memberships in multiple categories; while for a node more likely to have events mainly in a single category, the multinomial approach will be more effective.* The contaminated mixture model also inspires us to propose a new *mixture approach* that leverages both the logistic-normal and multinomial models in settings with uncertainty about node type. Specifically, this section is organized as follows:

(i) We first introduce a *contaminated mixture model* with some nodes following the multinomial distribution while the others follow the logistic-normal distribution, and the non-zero multinomial vectors are contaminated to have a positive weight in each category. We also propose an estimation algorithm (Algorithm 1) that assumes knowing the type of each node. See Section 5.3.1.

(ii) We simulate a synthetic network under this contaminated mixture model to *explore the central hypothesis* articulated above. See Section 5.3.2.

(iii) Furthermore, under the contaminated model, we develop a *test procedure* based on a likelihood ratio test to estimate the type of each node. Data with unknown node types can be analyzed by computing these estimates and then performing the mixture model estimation discussed in Section 5.3.1, as illustrated on the simulated network defined in Section 5.3.2. See Section 5.3.3.

The hypothesis and the mixture approach (testing and then estimation) will also be supported by our real data experiments in Section 6.

5.3.1 Contaminated Mixture Model

Here we propose a contaminated mixture model that mimics the real data behavior mentioned at the beginning of Section 5.3. Specifically, consider a collection of M nodes that are divided into two distinct sets: $\mathcal{N}_1 \cup \mathcal{N}_2 = [M]$ with $\mathcal{N}_1 \cap \mathcal{N}_2 = \varnothing$. Under the mixture model, each event associated with a node in $\mathcal{N}_1$ only belongs to a single category and its distribution can be captured by the multinomial model, while each event associated with a node in $\mathcal{N}_2$ has mixed category membership and thus can be modeled using the logistic-normal model.

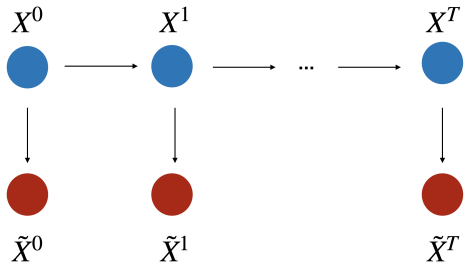


Figure 10: An illustration of the contaminated mixture model. The hidden $\{X^t\}_{t=0}^T$ are generated from the mixture model while $\{\widetilde{X}^t\}_{t=0}^T$ are observed contaminated data.

The multinomial and logistic-normal models introduced in Sections 2.1 and 2.2 correspond to two special cases of this mixture model: $\mathcal{N}_1 = [M]$, $\mathcal{N}_2 = \varnothing$, and $\mathcal{N}_1 = \varnothing$, $\mathcal{N}_2 = [M]$, respectively. The data generated from the mixture model is denoted by $\{X^t\}_{t=0}^T$, and we assume it to be contaminated to $\{\widetilde{X}^t\}_{t=0}^T$. This contamination is constructed as follows: if X_m^t is a non-zero multinomial vector, it would be contaminated to $\widetilde{X}_m^t$ that has positive weights in all categories; otherwise, we observe the true data $\widetilde{X}_m^t = X_m^t$.

Formally, let a 4th order tensor $A^{\text{mix}} \in \mathbb{R}^{M\times K\times M\times K}$ encode the context-dependent network and define $\nu^{\text{mix}} \in \mathbb{R}^{M\times K}$ as the offset parameter. Also, define $V^{\text{mix}} \in \mathbb{R}^{M\times K\times M\times K}$ as the variable importance parameter. Conditioning on all past data $(X^0, \widetilde{X}^0, \ldots, X^t, \widetilde{X}^t)$, we assume that X^{t+1} only depends on X^t through a mixture of the multinomial and logistic-normal models, while $\widetilde{X}_m^{t+1}$ only depends on X_m^{t+1} given all past data and X^{t+1} (this can be viewed as a hidden Markov model, see Figure 10). Given X^t, the future data $X_1^{t+1}, \ldots, X_M^{t+1}$ are conditionally independent. The conditional distributions of each X_m^{t+1} and $\widetilde{X}_m^{t+1}$ are specified as follows:

1. When $m \in \mathcal{N}_1$, the distribution of X_m^{t+1} given X^t and the definition of V_m^{mix} are the same as the multinomial model, defined in (3), (4), and (22), except that we substitute A_m^{MN} and ν_m^{MN} by A_m^{mix} and ν_m^{mix}.

   While for the conditional distribution of the observed data $\widetilde{X}_m^{t+1}$, we assume $\widetilde{X}_m^{t+1} = X_m^{t+1}$ if $X_m^{t+1} = 0_{K\times 1}$ which corresponds to the "no event" case; otherwise, we assume that $\widetilde{X}_m^{t+1}$ follows a logistic-normal distribution with parameters depending on X_m^{t+1}, and hence each event is observed with positive membership weights in all categories. The detailed logistic-normal distribution for the contaminated data $\widetilde{X}_m^{t+1}$ given X_m^{t+1} and and its motivation is included in Appendix D.2.

2. When $m \in \mathcal{N}_2$, the conditional distribution of X_m^{t+1} given X^t and the definition of V_m^{mix} are the same as the logistic-normal model with q^t depending on the past, defined in (7), (8), (11), and (26), except that we substitute $A_m^{\text{LN}}, B_m^{\text{Bern}}$, ν_m^{LN}, η_m^{Bern} by $A_{m,1:(K-1),:,:}^{\text{mix}}$, $A_{m,K,:,:}^{\text{mix}}$, $\nu_{m,1:(K-1)}^{\text{mix}}$, and $\nu_{m,K}^{\text{mix}}$. The covariance matrix of Gaussian noise ϵ_m^{t+1} is still denoted by Σ. We assume that the observed data $\widetilde{X}_m^{t+1} = X_m^{t+1}$ in this case.

Under this observational model, the *key challenge is to come up with estimates* $\widehat{\mathcal{N}}_1$, $\widehat{\mathcal{N}}_2$ of the node sets $\mathcal{N}_1$, $\mathcal{N}_2$. If given $\widehat{\mathcal{N}}_1$, $\widehat{\mathcal{N}}_2$, we can round $\widetilde{X}_m^t$ to be an estimate of X_m^t for $m \in \widehat{\mathcal{N}}_1$ and treat $\widetilde{X}_m^t$ as equivalent to X_m^t for $m \in \widehat{\mathcal{N}}_2$; then we can still apply the estimation methods proposed in Section 2 for each node separately, upon the estimated data. The detailed procedure is summarized in Algorithm 1. In the following section, we will consider

Algorithm 1 Network Estimation with Contaminated Data

Input: Contaminated data $\{\widetilde{X}^t\}_{t=0}^T$, number of nodes M, number of categories K, node sets $\widehat{\mathcal{N}}_1, \widehat{\mathcal{N}}_2$, tuning parameters $\lambda^{\mathrm{MN}}, \lambda^{\mathrm{LN}} > 0$, $\alpha \in (0,1)$

for $t = 0, \ldots, T$ **do**
  $\widehat{X}^{\mathrm{mix},t} = 0_{M\times K}$
  for $m = 1, \ldots, M$ **do**
    if $m \in \widehat{\mathcal{N}}_1$ and $\widetilde{X}_m^t \neq 0_{K\times 1}$ **then**
      $\widehat{X}_m^{\mathrm{mix},t} = e_k$ where $k = \arg\max_i \widetilde{X}_{m,i}^t$
    else
      $\widehat{X}_m^{\mathrm{mix},t} = \widetilde{X}_m^t$
    end if
  end for
end for
$\widehat{A}^{\mathrm{mix}} = 0_{M\times K\times M\times K}$, $\widehat{\nu}^{\mathrm{mix}} = 0_{M\times K}$, $\widehat{V}^{\mathrm{mix}} = 0_{M\times K\times M\times K}$
for $m = 1, \ldots, M$ **do**
  if $m \in \widehat{\mathcal{N}}_1$ **then**

$$(\widehat{A}_m^{\mathrm{mix}}, \widehat{\nu}_m^{\mathrm{mix}}) = \underset{A\in\mathbb{R}^{K\times M\times K}, \nu\in\mathbb{R}^K}{\arg\min} \frac{1}{T}\sum_{t=0}^{T-1} \ell^{\mathrm{MN}}(A; \widehat{X}^{\mathrm{mix},t}, \widehat{X}_m^{\mathrm{mix},t+1}, \nu) + \lambda^{\mathrm{MN}}\|A\|_R$$

    Obtain $\widehat{V}_m^{\mathrm{mix}}$ by plugging in $A_m^{\mathrm{MN}} = \widehat{A}_m^{\mathrm{mix}}$, $\nu_m^{\mathrm{MN}} = \widehat{\nu}_m^{\mathrm{mix}}$ to (22)
  else

$$\begin{aligned}(\widehat{A}_m^{\mathrm{mix}}, \widehat{\nu}_m^{\mathrm{mix}}) = \underset{A\in\mathbb{R}^{K\times M\times K}, \nu\in\mathbb{R}^K}{\arg\min} & \frac{\alpha}{T}\sum_{t=0}^{T-1} \ell^{\mathrm{LN}}(A_{1:(K-1),:,:}; \widehat{X}^{\mathrm{mix},t}, \widehat{X}_m^{\mathrm{mix},t+1}, \nu_{1:(K-1)}) \\ & + \frac{1-\alpha}{T}\sum_{t=0}^{T-1} \ell^{\mathrm{Bern}}(A_{K,:,:}; \widehat{X}^{\mathrm{mix},t}, \widehat{X}_m^{\mathrm{mix},t+1}, \nu_K) \\ & + \lambda^{\mathrm{LN}} R_\alpha(A_{1:(K-1),:,:}, A_{K,:,:})\end{aligned}$$

    Obtain $\widehat{V}_m^{\mathrm{mix}}$ by plugging in $A_m^{\mathrm{LN}} = \widehat{A}_{m,1:(K-1),:,:}^{\mathrm{mix}}$, $B_m^{\mathrm{Bern}} = \widehat{A}_{m,K,:,:}^{\mathrm{mix}}$, $\nu_m^{\mathrm{LN}} = \widehat{\nu}_{m,1:(K-1)}^{\mathrm{mix}}$, $\eta_m^{\mathrm{Bern}} = \widehat{\nu}_{m,K}^{\mathrm{mix}}$ to (24)
  end if
end for

Output: $\widehat{A}^{\mathrm{mix}}$, $\widehat{\nu}^{\mathrm{mix}}$ and $\widehat{V}^{\mathrm{mix}}$

two naive estimates for the node sets $\mathcal{N}_1$ and $\mathcal{N}_2$: (i) $\widehat{\mathcal{N}}_1 = [M]$, $\widehat{\mathcal{N}}_2 = \varnothing$, and (ii) $\widehat{\mathcal{N}}_1 = \varnothing$, $\widehat{\mathcal{N}}_2 = [M]$. Algorithm 1 with these two estimates correspond to the multinomial and logistic-normal approaches. Then we will propose data-dependent estimators for $\mathcal{N}_1$ and $\mathcal{N}_2$ and investigate the performance of Algorithm 1 with these estimators in Section 5.3.3.

5.3.2 Synthetic Example for Validating the Hypothesis

To investigate how the multinomial and logistic-normal approaches work under this contaminated mixture model, or to explore our hypothesis mentioned before Section 5.3.1, we simulate a toy example under this model. The detailed model parameters of the simulated example are deferred to Appendix D.3. Under this model set-up, we generate time series $\{X^t\}_{t=0}^T$ and $\{\widetilde{X}^t\}_{t=0}^T$ with $T = 10000$, where the initial data $\{X_m^0\}_{m=1}^M$ are i.i.d. multinomial random vectors. The true variable importance parameter V^{mix} of the hidden mixture model (calculated from $\{X^t\}_{t=0}^T$ and true model parameters) is visualized in Figure 11(a): there are 17 nodes ($M = 17$) with 5 categories of events ($K = 5$) in total: "blue", "black", "red", "green", and "yellow" events. Only influences within each category exist ($A^{\text{mix}}_{:,k,:,k'} = 0$ if $k \neq k'$), and the edge colors indicate the categories of the influence[5]. Purple nodes (nodes 1-5) belong to $\mathcal{N}_1$, while nodes 6-17 are from $\mathcal{N}_2$.

After applying the multinomial (Algorithm 1 with $\widehat{\mathcal{N}}_1 = [M]$, $\widehat{\mathcal{N}}_2 = \varnothing$) and logistic-normal (Algorithm 1 with $\widehat{\mathcal{N}}_1 = \varnothing$, $\widehat{\mathcal{N}}_2 = [M]$) approaches upon the generated data $\{\widetilde{X}^t\}_{t=0}^T$, the estimated variable importance networks are presented in Figure 11(b),(c). We can see from Figure 11 that the multinomial approach mainly picks edges correctly among nodes in $\mathcal{N}_1$, while the logistic-normal approach works better for nodes in $\mathcal{N}_2$, which validates our hypothesis mentioned at the beginning of Section 5.3.

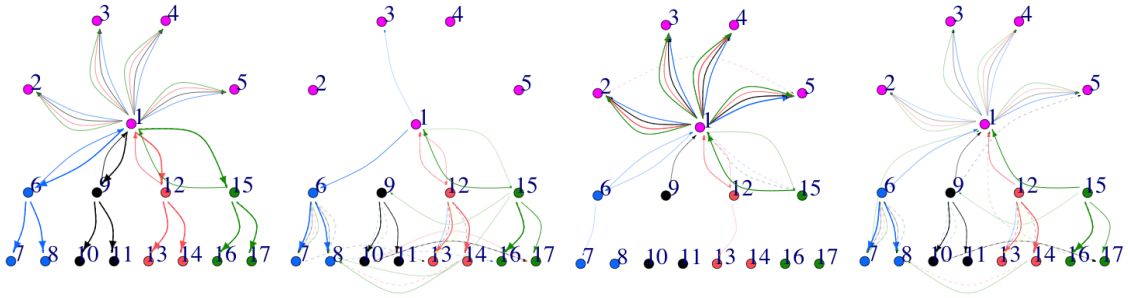
Notation	Description
$X^t \in \mathbb{R}^{M\times K}$	Hidden data at time t generated from the mixture model, defined in Section 5.3.1
$\widetilde{X}^t \in \mathbb{R}^{M\times K}$	Observed data in the contaminated mixture model, defined in Section 5.3.1; $\widetilde{X}^t$ is the contaminated version of X^t
$\widehat{X}^t_m \in \mathbb{R}^K$	Rounded data for node m at time t given the contaminated data $\widetilde{X}^t_m$, defined in Algorithm 2
$\widehat{X}^{\text{mix},t} \in \mathbb{R}^{M\times K}$	Estimated data at time t given contaminated data $\widetilde{X}^t$ and estimated node sets $\widehat{\mathcal{N}}_1$, $\widehat{\mathcal{N}}_2$, defined in Algorithm 1

Table 1: Notations for data in the contaminated mixture model

5.3.3 Mixture Approach Based on a Testing Procedure

Based on the findings in the previous section, naive estimates for the node sets, $\widehat{\mathcal{N}}_1 = [M]$ (multinomial approach) or $\widehat{\mathcal{N}}_2 = [M]$ (logistic-normal approach), may not be the best choices. In this section, we propose a heuristic test based on the idea of likelihood ratio tests, which takes the contaminated data $\{\widetilde{X}^t\}_{t=0}^T$ as input and outputs $\widehat{\mathcal{N}}_1, \widehat{\mathcal{N}}_2 \subset [M]$

5. We set no influence in the "yellow" category (no yellow edge), which can be set as a natural baseline in the logistic-normal model.



(a) True Network (b) Estimated Multinomial Network (c) Estimated Logistic-normal Network (d) Estimated Mixture Network

Figure 11: True variable importance network and estimated variable importance networks by the multinomial, logistic-normal, and mixture approaches. Solid edges are stimulatory while dashed ones are inhibitory, and edge colors indicate the categories of the influence. After normalizing the maximal absolute value of network parameters to 1 for each network, edges are only visualized if their corresponding parameters have larger absolute values than 0.15, and edge width is proportional to these values. *We can see that the multinomial approach is more likely to underestimate the edges connecting purple nodes (nodes in $\mathcal{N}_1$) compared to the nodes 6-17 (nodes in $\mathcal{N}_2$), while the logistic-normal approach is more likely to ignore edges connecting nodes in $\mathcal{N}_2$. As a comparison, the mixture approach performs reasonably well for both types of nodes (details of the mixture approach are presented in Section 5.3.3).*

as estimates of $\mathcal{N}_1, \mathcal{N}_2$. Once we obtain $\widehat{\mathcal{N}}_1, \widehat{\mathcal{N}}_2$, Algorithm 1 can be applied to estimate the underlying mixture model, and we will refer to this procedure, including testing and estimation, as the *mixture approach*. The pseudocode of our testing procedure is given in Algorithm 2, and we will validate its performance using the aforementioned synthetic example at the end of this section. In particular, this testing procedure calculates the log likelihood-ratio statistic $\widehat{R}_m$ for each node m and uses it to determine whether $m \in \widehat{\mathcal{N}}_1$ or $\widehat{\mathcal{N}}_2$. A detailed explanation of the testing procedure is presented in the following.

Likelihood functions: For any $1 \leqslant m \leqslant M$, first define the negative log-likelihood function for $\widetilde{X}_m^{t+1}$ (observed data from the synthetic model) given X^t (true unknown data) by $\widetilde{\ell}^{\mathrm{MN}}(A_m^{\mathrm{mix}}, \nu_m^{\mathrm{mix}}, a, (\sigma^{\mathrm{MN}})^2; X^t, \widetilde{X}_m^{t+1})$ if $m \in \mathcal{N}_1$; by $\widetilde{\ell}^{\mathrm{LN}}(A_m^{\mathrm{mix}}, \nu_m^{\mathrm{mix}}, \Sigma; X^t, \widetilde{X}_m^{t+1})$ if $m \in \mathcal{N}_2$. The detailed forms of $\widetilde{\ell}^{\mathrm{MN}}$ and $\widetilde{\ell}^{\mathrm{LN}}$ are included in Appendix C.2 (see (86) and (87)). Here $a, (\sigma^{\mathrm{MN}})^2$ are the parameters for multinomial nodes in the contaminated mixture model proposed in Section 5.3.2 (details presented in Appendix D.2), and Σ is the noise covariance matrix for logistic-normal nodes. We aim at finding estimators for A_m^{mix}, ν_m^{mix}, a, $(\sigma^{\mathrm{MN}})^2$, Σ under each model and then derive a log likelihood-ratio statistic.

Algorithm 2 Node Type Testing

Input: Contaminated data $\{\widetilde{X}^t\}_{t=0}^T$, tuning parameters $\lambda^{\mathrm{MN}}, \lambda^{\mathrm{LN}} > 0$, $\alpha \in (0,1)$

$\widehat{\mathcal{N}}_1 = \varnothing, \widehat{\mathcal{N}}_2 = \varnothing$
for $m = 1, \ldots, M$ **do**
  for $t = 1, \ldots, T$ **do**
    if $\widetilde{X}_m^t \neq 0_{K\times 1}$ **then**
      $\widehat{X}_m^t = e_k$ where $k = \arg\max_i \widetilde{X}_{m,i}^t$
    else
      $\widehat{X}_m^t = \widetilde{X}_m^t$
    end if
  end for
end for
for $m = 1, \ldots, M$ **do**
  Obtain $\widehat{A}_m^{\mathrm{MN}}$, $\widehat{\nu}_m^{\mathrm{MN}}$, $\widehat{a}_m$, $(\widehat{\sigma}_m^{\mathrm{MN}})^2$ by Algorithm 3 with input $\{\widetilde{X}^t\}_{t=0}^T$, $\{\widehat{X}_m^t\}_{t=1}^T$, λ^{MN}
  Obtain $\widehat{A}_m^{\mathrm{LN}}$, $\widehat{B}_m^{\mathrm{Bern}}$, $\widehat{\nu}_m^{\mathrm{LN}}$, $\widehat{\eta}_m^{\mathrm{Bern}}$, $\widehat{\Sigma}_m$ by Algorithm 4 with input $\{\widetilde{X}^t\}_{t=0}^T$, m, λ^{LN} and α
  Calculate test statistic:

$$\begin{aligned}\widehat{R}_m =& \frac{1}{T}\sum_{t=0}^{T-1} \widetilde{\ell}^{\mathrm{LN}}([\widehat{A}_m^{\mathrm{LN}}, \widehat{B}_m^{\mathrm{Bern}}], [\widehat{\nu}_m^{\mathrm{LN}}, \widehat{\eta}_m^{\mathrm{Bern}}], \widehat{\Sigma}_m; \widetilde{X}^t, \widetilde{X}_m^{t+1}) \\ & - \widetilde{\ell}^{\mathrm{MN}}(\widehat{A}_m^{\mathrm{MN}}, \widehat{\nu}_m^{\mathrm{MN}}, \widehat{a}_m, (\widehat{\sigma}_m^{\mathrm{MN}})^2; \widetilde{X}^t, \widetilde{X}_m^{t+1})\end{aligned}$$

  where $\widetilde{\ell}^{\mathrm{LN}}$ and $\widetilde{\ell}^{\mathrm{MN}}$ are defined in (87) and (86)
  if $\widehat{R}_m > -\infty$ **then**
    $\widehat{\mathcal{N}}_1 = \widehat{\mathcal{N}}_1 \cup \{m\}$
  else
    $\widehat{\mathcal{N}}_2 = \widehat{\mathcal{N}}_2 \cup \{m\}$
  end if
end for

Output: $\widehat{\mathcal{N}}_1, \widehat{\mathcal{N}}_2$

Approximation and estimation: Since X^t is unknown, we first substitute X^t in $\widetilde{\ell}^{\mathrm{MN}}$ and $\widetilde{\ell}^{\mathrm{LN}}$ by $\widetilde{X}^t$ as an approximation[6]. For each node m, we propose estimators for the model parameters under each model separately:

(i) Under the multinomial model, we regress rounded data $\{\widehat{X}_m^t\}_{t=1}^T$ (defined in Algorithm 2) upon past observed data $\{\widetilde{X}^t\}_{t=0}^{T-1}$ with the multinomial loss and regularization parameter $\lambda^{\mathrm{MN}} > 0$. Here we regress future rounded data upon past observed data instead of past rounded data $\{\widehat{X}^t\}_{t=0}^{T-1}$, since we don't assume the types of other nodes when testing node m, and hence we should not round the data associated with all nodes. Then we obtain estimators $\widehat{A}_m^{\mathrm{MN}}$, $\widehat{\nu}_m^{\mathrm{MN}}$ for A_m^{mix}, ν_m^{mix}; a and σ^{MN} are estimated by $\widehat{a}_m$, $\widehat{\sigma}_m^{\mathrm{MN}}$ using the method of moments; The detailed estimation procedure is summarized in Algorithm 3 in Appendix D.4.

(ii) Under the logistic-normal model, we regress observed data $\{\widetilde{X}_m^t\}_{t=1}^T$ upon past observed data $\{\widetilde{X}^t\}_{t=0}^{T-1}$ with the logistic-normal loss and tuning parameters λ^{LN}, $\alpha > 0$. This leads to estimators $\widehat{A}_m^{\mathrm{LN}}$, $\widehat{B}_m^{\mathrm{Bern}}$, $\widehat{\nu}_m^{\mathrm{LN}}$, $\widehat{\eta}_m^{\mathrm{Bern}}$ for $A_{m,1:(K-1),:,:}^{\mathrm{mix}}$, $A_{m,K,:,:}^{\mathrm{mix}}$, $\nu_{m,1:(K-1)}^{\mathrm{mix}}$, $\nu_{m,K}^{\mathrm{mix}}$. The covariance parameter Σ is then estimated by the MLE $\widehat{\Sigma}_m$ given other parameter estimates. The detailed estimation procedure is summarized in Algorithm 4 in Appendix D.4.

In practice, the tuning parameters λ^{MN}, λ^{LN} and α can be chosen by cross-validation. The motivation and formal definition for these estimators are also included in Appendix D.4. Based on these estimates under each model, our test statistic for node m is defined as

$$\begin{aligned}\widehat{R}_m =& \frac{1}{T}\sum_{t=0}^{T-1} \widetilde{\ell}^{\mathrm{LN}}([\widehat{A}_m^{\mathrm{LN}}, \widehat{B}_m^{\mathrm{Bern}}], [\widehat{\nu}_m^{\mathrm{LN}}, \widehat{\eta}_m^{\mathrm{Bern}}], \widehat{\Sigma}_m; \widetilde{X}^t, \widetilde{X}^{t+1}) \\ & - \widetilde{\ell}^{\mathrm{MN}}(\widehat{A}_m^{\mathrm{MN}}, \widehat{\nu}_m^{\mathrm{MN}}, \widehat{a}_m, (\widehat{\sigma}_m^{\mathrm{MN}})^2; \widetilde{X}^t, \widetilde{X}^{t+1}),\end{aligned} \tag{28}$$

where $\widetilde{\ell}^{\mathrm{MN}}$ and $\widetilde{\ell}^{\mathrm{LN}}$ are the negative log-likelihood functions mentioned in the beginning of Section 5.3.3. Here recall that $[A, B]$ denotes concatenation of A and B in the first dimension for any $A \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_k}, B \in \mathbb{R}^{p_1' \times p_2 \times \cdots \times p_k}$. Larger $\widehat{R}_m$ suggests a better suit for the multinomial model, while lower $\widehat{R}_m$ suggests a better suit for the logistic-normal model.

Classification criterion: The test statistic $\widehat{R}_m$ for each node m in the synthetic example discussed in Section 5.3.2 is plotted in Figure 12, and we notice that if and only if $m \in \mathcal{N}_2$, the test statistic $\widehat{R}_m = -\infty$ due to a zero estimate $\widehat{\sigma}_m^2$. Furthermore, as long as the data is generated from the synthetic mixture model and $m \in \mathcal{N}_1$, $\widehat{\sigma}_m^2$ should be a good estimate for σ^2 with sufficiently many samples (large T), and thus $\widehat{\sigma}_m^2 = 0$ is a strong indicator for $m \in \mathcal{N}_2$. Therefore, we propose estimators $\widehat{\mathcal{N}}_1 = \{m : \widehat{R}_m > -\infty\}$ and $\widehat{\mathcal{N}}_2 = \{m : \widehat{R}_m = -\infty\}$ for the true node sets $\mathcal{N}_1$ (multinomial nodes), $\mathcal{N}_2$ (logistic-normal nodes). Although this criterion seems a bit conservative for classifying the logistic-normal nodes ($\widehat{\mathcal{N}}_2$), both our synthetic experiment and real data experiments in Section 6 suggest that it has promising performances. Finding a data-dependent threshold for $\widehat{R}_m$ is an interesting but very challenging open problem.

6. We don't calculate the exact log-likelihood function given $\widetilde{X}^t$ since it is computationally heavy, involving summation over all potential hidden variables X^t and 2^{M-1} combinations of node types of the rest $M-1$ nodes.

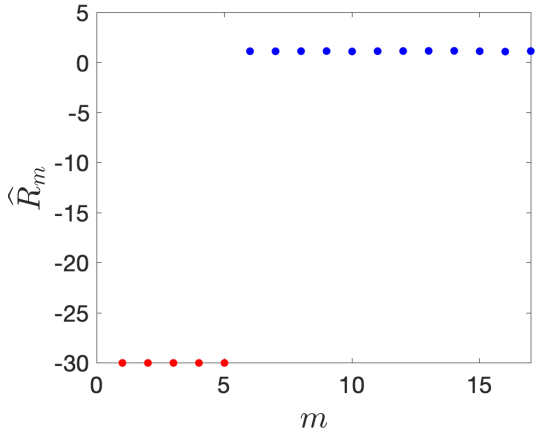


Figure 12: The test statistic $\widehat{R}_m$ for $m = 1, \ldots, 17$ in our synthetic example. Blue nodes are multinomial nodes ($\mathcal{N}_1$), while red nodes are logistic-normal nodes ($\mathcal{N}_2$). We can see that $\widehat{R}_m = -\infty$ if and only if $m \in \mathcal{N}_2$.

Figure 11(d) visualizes the estimated variable importance network $\widehat{V}^{\text{mix}}$ for the synthetic example using the mixture approach (Algorithm 2 followed by Algorithm 1), which recovers the edges for both the multinomial and logistic-normal nodes. The real data experiments in Section 6 would also demonstrate the effectiveness of this mixture approach.

6. Real Data Examples

We evaluate the multinomial, logistic-normal, mixture approaches and the main hypothesis on a political tweets data set (Littman et al., 2016), and a MemeTracker data set[7] (Leskovec et al., 2009). For both data sets, we apply some NLP methods to extract a membership vector for each post, organize them into time series $\{X^t \in \mathbb{R}^{M \times K}\}_{t=0}^{T}$ (details provided later), which are then fitted by the three approaches. These two data sets demonstrate the advantages of our methodology over existed approaches and provide some evidence for our hypothesis that for one type of nodes, the multinomial approach is better than the logistic-normal while the converse holds for the other type. Furthermore, the mixture approach shows promising performance for both types of nodes. In the following, we first elaborate on the evaluation procedures of our real data experiments, and then we discuss each example in detail (Sections 6.1 and 6.2).

One of the major challenges for network estimation is performance evaluation since there is no obvious ground truth. For both applications, we provide two evaluations: (1) prediction error performance that demonstrates the advantage of allowing influence to depend on categories; (2) a subset of directed edges are supported by external knowledge (political tweets example) or information extracted from a cascade data set (MemeTracker example),

7. Data available at `http://www.memetracker.org/data.html`

which further supports the hypothesis from the synthetic model in the previous section, and also demonstrate the usefulness of our mixture approach.

Comparison of estimates: Since the three approaches take different data as input (rounded data for the multinomial method, unrounded for the logistic-normal method, and a mixture of both for the mixture method), we also use their corresponding test data to measure the prediction errors; thus they are not directly comparable. The detailed procedure for calculating prediction errors is deferred to Appendix D.6. To investigate the benefit of learning different networks for different categories, we compare the prediction errors of the three methods relative to fitting (1) a context-independent network model where the influences among nodes do not depend on categories[8] , and (2) a constant process where the network parameters are all zeros (no influence from the past)[9]. For comparing network estimates, we visualize the variable importance parameter $\widehat{V}^{\mathrm{MN}}$, $\widehat{V}^{\mathrm{LN}}$ defined in Section 4 and $\widehat{V}^{\mathrm{mix}}$ defined in Section 5.3.1.

6.1 Political tweets data

A central question in political science and mass communication is how politicians influence each other. Here we measure influence using the time series of their posts on Twitter. While constructing an adjacency matrix for this network (*e.g.,* by looking at who follows whom) is a simple task, it does not reveal how the level of influences among politicians varies as a function of political tendencies of posts (*i.e.,* left-wing or right-wing). To address this challenge, we use a collection of tweets from the 2016 United States Presidential Election Tweets Data set (Littman et al., 2016), collected from Jan 1, 2016, to Nov 11, 2016. The collection includes $83,459$ tweets sent by 23 Twitter accounts ($M = 23$): 17 presidential candidates' accounts and the House, Senate, party accounts for each party (Democrats and Republicans). We consider two categories of tweets: left-leaning and right-leaning ($K = 2$), and we aim to learn the influence network among the 23 Twitter accounts that depend on the ideologies of tweets.

Due to the lack of a pre-trained NLP model for identifying political tendencies of tweets given their contents, we use the tweets from the first half of the time period (55,859 tweets from Jan 1, 2016, to Jun 6, 2016) to train a neural network for categorizing tweets into two political tendencies (left- and right-leaning) and apply it on the tweets from the second half of the time period. The detailed procedure for training the neural network and how we obtain the data $\{X^t\}_{t=0}^T$ with $T = 999$ is contained in Appendix D.5. Under this pre-processing procedure, each X^t spans a time interval of 3.7 hours.

Figure 13 shows the histogram of the unrounded $\{X^t_{m,2} : X^t_m \neq 0\}$, the right-leaning weights of all tweets (averaged for multiple tweets from the same user and time window). Since the sum of the left-leaning weight and right-leaning weight of any tweet equals 1, it suffices to present only one of them. One important thing to note is that there are two

8. This is equivalent to assuming a Bernoulli auto-regressive (BAR) model (Hall et al., 2016) for whether events occur, and each node's events membership in categories follow the same multinomial/logistic-normal distribution over time. We use ℓ_1 penalized MLE for estimating the BAR parameter and MLE for estimating the multinomial/logistic-normal distribution parameter.
9. MLE is used for estimating the constant process parameter.

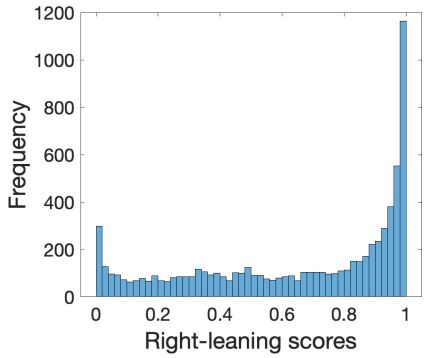


Figure 13: The histogram of right-leaning weights of all tweets (averaged for multiple tweets from the same user and time window) in political tweets example. The peaks in frequency at 0 and 1 suggest that the political tendencies of these tweets contain little ambiguity.

peaks in frequency centred at 0 and 1, which suggests many clearly left-leaning tweets (0 score) or right-leaning tweets (1 score).

Prediction performance: We apply the three methods (multinomial, logistic-normal, and mixture approaches) using the first 70% of the input data (from Jun 7 to Sept 25, 2016) and test their prediction performance on the latter 30% (from Sept 26 to Nov 11, 2016). Similar to Section 5.3.2, 5.3.3, we round the data accordingly for the multinomial and mixture methods. Meanwhile, the right-leaning category is set as the baseline for the logistic-normal approach and also for the logistic-normal nodes in the mixture approach: as discussed in Section 2.2.1, all choices of the baseline category lead to equivalent models. The prediction errors of the three fitted models and that of their corresponding fitted sub-models (defined before Section 6.1) are presented in Table 2-4 respectively. The prediction error tables show that both the multinomial and mixture approaches take advantage of the context information since our context-dependent model yields a slightly lower prediction error, but the logistic-normal approach doesn't since the context-independent approach out-performs our approach.

Method	Constant Process	Context-independent Network Model	Multinomial (Our Model)
Prediction Error	0.30580	0.25520	0.25200

Table 2: The prediction errors of the fitted multinomial model (full model), and that of its two sub-models: fitted constant multinomial process and context-independent network model under multinomial framework, evaluated on the latter 30% of the data set. We can see that the prediction error of the context-dependent network (full model) is lower than that of the context-independent one, which *illustrates the potential benefit of incorporating context information when using the multinomial method.*

Method	Constant Process	Context-independent Network Model	Logistic-normal (Our Model)
Prediction Error	0.15800	0.14373	0.14442

Table 3: The prediction errors of the fitted logistic-normal model (full model), and that of its two sub-models: fitted constant logistic-normal process and context-independent network model under logistic-normal modeling framework, evaluated on the latter 30% of the data set. The prediction error of the fitted logistic-normal model (full model) is slightly larger than that of the context-independent network model, suggesting that the *logistic-normal approach does not capture the contextual information well.*

Method	Constant Process	Context-independent Network Model	Mixture (Our Model)
Prediction Error	0.29355	0.22921	0.22913

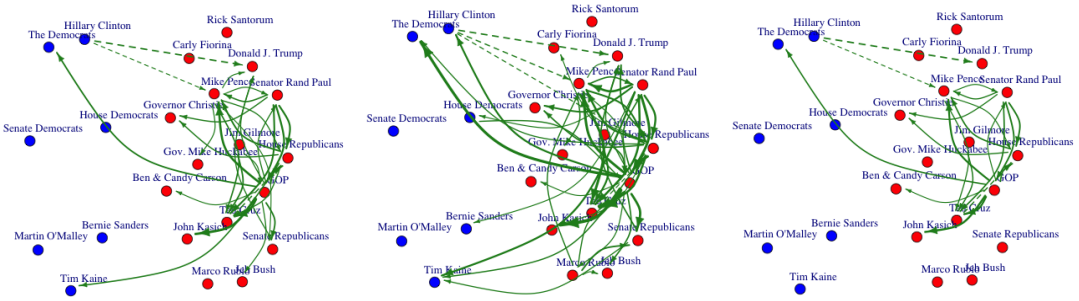
Table 4: The prediction errors of the fitted mixture model (full model), and that of its two sub-models: fitted constant mixture process and context-independent network model under mixture modeling framework, evaluated on the latter 30% of the data set. The prediction error of the fitted mixture model (full model) is slightly lower than that of the context-independent network model, suggesting that the *mixture approach probably captures the contextual information.*

Network estimates: We apply the three methods on the whole data set with the same tuning parameters used in the prediction task, obtaining the estimated variable importance networks. Although there is no notion of ground truth, we treat the following plausible hypothesis as external knowledge: Republicans' right-leaning tweets tend to have more influence than their left-leaning tweets, encouraging other Republicans' right-leaning tweets and vice versa for Democrats and their left-leaning tweets. We only present the estimated variable importance edges that are from right-leaning to right-leaning in Figure 14 since the other three types of edges (left-leaning→left-leaning, etc.) look very similar across the three methods and all align well with our external knowledge. The visualization for the other three types of edges is deferred to Appendix D.7.

In Figure 14, we can see that the multinomial and mixture networks include fewer right→right edges from Republicans (red) to Democrats (blue) than the logistic-normal network, which may suggest an improvement of these two approaches over the logistic-normal approach. Since the tweets of all political candidates have extreme ideologies, as shown in Figure 13, this is consistent with our hypothesis from the previous section. Our test classified most users (17 out of 23 total users) as multinomial nodes, which may explain why the mixture approach also has good performance.

6.2 MemeTracker Data Set

In this section, we consider the question of how past posts sent by one online media source influence another media source in posting new articles; and how this influence network depends on the topics of articles. To answer this question, we apply our methods upon



(a) Multinomial (b) Logistic-normal (c) Mixture

Figure 14: Estimated variable importance networks by the three approaches for the tweets example, including edges from right-leaning tweets to right-leaning tweets. The largest absolute entry of each of the three variable importance parameters is normalized to one, and each visualized edge width is proportional to the normalized absolute value of its corresponding parameter. For clarity, only the edges with absolute parameters larger than 0.3 are shown for each network, and blue nodes are Democrats, red nodes are Republicans. Solid edges are positive influences (stimulatory), while dashed edges are negative influences (inhibitory).

the "Raw phrases data" in the MemeTracker data set (Leskovec et al., 2009). This data set consists of news stories and blog posts from 1 million online sources (including mass media sources and personal blogs) over the time period from August 2008 to April 2009. For each news or blog item, only its phrases/quotes that have variants frequently occurring across the entire online news corpus are recorded in the data set, and we use them as the approximate content of the post. Note that most news media sources cover multiple topics (although not with the same amount of coverage), so we don't have labels for each news article and thus cannot use supervised learning as we did for the tweet example to obtain the membership vectors. Instead, we use topic modeling (Latent Dirichlet Allocation proposed in Blei et al. (2003)) for extracting mixed membership vectors, and we set the number of topics as $K = 5$. Based on the top keywords generated from topic modeling for each topic (shown in Table 14 in Appendix D.5), we choose the topic names as "Sports", "International Affairs", "Lifestyle", "Finance" and "Health". For simplicity and interpretability, we also filter out $M = 58$ media sources based on their languages, frequencies, and topic coverage. A 1-hour discretization is adopted (also applied in the prior work Mark et al. (2018)) and leads to $T+1 = 5807$ time intervals. The detailed pre-processing of the data (how we obtain $\{X^t\}_{t=0}^T$) is contained in Appendix D.5.

Prediction performance: We fit three models (multinomial, logistic-normal, and mixture models) using the first 70% of the data (from Sept 1, 2008, to Feb 16, 2009), and test their prediction performance on the latter 30% (from Feb 17 to Apr 30, 2009). The topic "Health" is set as the baseline for the logistic-normal approach and also for the logistic-normal nodes in the mixture approach.

Method	Constant Process	Context-independent Network Model	Multinomial (Our Model)
Prediction Error	0.49741	0.45062	0.43351

Table 5: The prediction errors of the fitted multinomial model (full model), and that of its two sub-models: fitted constant multinomial process and context-independent network model under multinomial framework, evaluated on the latter 30% of the data set. The prediction error of the full model is lower than the error of the context-independent one, *showing potential benefit of incorporating context information using the multinomial approach.*

Method	Constant Process	Context-independent Network Model	Logistic-normal (Our Model)
Prediction Error	0.11269	0.10809	0.10229

Table 6: The prediction errors of the fitted logistic-normal model (full model), and that of its two sub-models: fitted constant logistic-normal process and context-independent network model under logistic-normal framework, evaluated on the latter 30% of the data set. The prediction error of the full model is smaller than that of the context-independent one, *showing potential benefit of incorporating context information using the logistic-normal approach.*

As explained before Section 6.1, we calculate the prediction errors of the three fitted models and that of their corresponding fitted sub-models, which are presented in Table 5-7. All three approaches demonstrate some advantage of estimating context-dependent networks since the context-dependent networks give lower prediction errors in all cases.

Method	Constant Process	Context-independent Network Model	Mixture (Our Model)
Prediction Error	0.37158	0.31140	0.29930

Table 7: The prediction errors of the fitted mixture model (full model), and that of its two sub-models: fitted constant mixture process and context-independent network model under mixture framework, evaluated on the latter 30% of the data set. The prediction error of the full model is smaller than that of the context-independent one, *showing potential benefit of incorporating context information using the mixture approach.*

Network estimates: We apply the three approaches on the whole data set, with the same tuning parameters as those used in the prediction task. For each media source, we visualize the influences it receives (a star-shaped sub-network with this media source being the center), encoded by the estimated variable importance parameters from the three methods. Since all sub-networks look very similar across the methods except for a few edges, we defer the visualizations and details of the visualization procedures to Appendix D.8.

To compare the performance of these methods, we extract some supporting evidence from a cascade data set: the *"Phrase cluster data"* (more details included in Appendix D.9)from

Edges \ Networks	MN	LN	mix
alertnet → *reuters*	**I***	S, **I***, L, F	**I***
uk.reuters → *reuters*	**I, F**	**S, I, L, F**	**S, I, F**
canadianbusiness → *wral*	**F***	S, **I**	**F***
breitbart → *wral*	**S, F**	**S, I, L, F**	**S, I L, F**

Table 8: **Edge topics** suggested by the estimated variable importance networks (column 2-4) for edges in column 1. Here we use "S", "I", "L", and "F" as abbreviations for the topics "Sports", "International Affairs", "Lifestyle", and "Finance". In the first row of the table, "LN', "MN", "mix" refer to the logistic-normal, multinomial, and mixture approaches. We will present supporting evidence for the estimated edge topics marked in bold. Our evidence also suggests the edge topics with "*" are likely to be dominant. We can see that *the multinomial and mixture approaches may work better for the first and third edges, while the logistic-normal and mixture approaches seem to work better for the other two edges.*

Aug 2008 to Jan 2009 in the MemeTracker data set, which is also used in Yu et al. (2017a) for studying influences among media sources. In Table 8, we present four edges (*alertnet*, *uk.reuters*→*reuters*, *canadianbusiness*, *breitbart*→*wral*) with supporting evidence suggesting that some methods may do better than the others. From Table 8, we can see that the multinomial approach is likely to work better for the first and third edges since "International Affairs" and "Finance" seem to be the dominant topics in these two edges. However, for the other two edges, multiple topics are plausible, and none is dominant. Hence the logistic-normal approach may work better for the second and fourth edges. Meanwhile, the mixture approach seems to work well for all four edges. For most other edges estimated differently by the three methods, we don't have evidence suggesting which one may be better. The detailed supporting evidence and arguments for comparing these four estimated edges are included in Appendix D.9.

Hypothesis support based on validated edges: Table 8 suggests that the logistic-normal and mixture methods estimate edges better if they connect *uk.reuters* and *breitbart*, while the multinomial and mixture methods estimate edges better if they connect *alertnet* and *canadianbusiness*. The first two media sources tend to cover multiple topics, while the latter two media sources tend to be primarily about one topic. To further emphasize this *mixed membership* or *single category* behavior, we consider the *top* topic weights of averaged posts sent by each media source within each time interval, and then we take an average over all time intervals when each media source posts. We present the average *top* topic weights of these four media sources in Table 9. A higher top topic weight suggests less mixed membership. We can see that posts sent by *uk.reuters*, *breitbart* within the same time units are more mixed in topics, while those by *alertnet*, *canadianbusiness* are more

Media sources	uk.reuters	breitbart	alertnet	canadian-business
Top topic weight	0.4400	0.4061	0.5539	0.5694
% of media sources with lower top weights	27.59%	8.62%	74.14%	84.48%
$\widehat{R}_m$	$-\infty$	$-\infty$	-1.2601	-1.4343

Table 9: Top topic weights of the averaged posts within each time unit sent by the four media sources, averaged over time. The four media sources all have edges estimated well by multinomial or logistic-normal method but not the other. The third row is the percentage of all 58 media sources that have lower top topic weights than the media source in the first row. A higher top topic weight and percentage suggest that the posts sent by the media source are more likely to fall in one topic, while a lower top topic weight suggests that its posts are more likely to have mixed membership. The fourth row presents the test statistic $\widehat{R}_m$, indicating that *our test identifies the types of these nodes correctly, which explains why the mixture approach based on the test works well for all four nodes.*

exclusively about one topic. This finding further validates our main hypothesis from the previous section. In addition, the test statistic $\widehat{R}_m$ for each node is presented in Table 9, suggesting that both *uk.reuters* and *breitbart* are classified as logistic-normal nodes by our test, which explains why our mixture approach would work well.

6.3 Summary of findings

Since real data validation is quite involved, we briefly summarize the key findings in Table 10, which provides further evidence for the hypothesis that the logistic-normal approach will be more effective at estimating influences among nodes whose events exhibit mixed memberships in multiple categories; while for a node more likely to have events in one category than others and thus each of its events falls in that category, the multinomial approach will be more effective. Furthermore, the mixture approach works reasonably well for both types of nodes.

7. Proofs

In this section we provide proofs for Theorems 1, 2, 3, Propositions 1, and Lemma 6. Proofs for other supporting lemmas are deferred to the appendix.

7.1 Proof of Theorem 1

We prove the error bounds for arbitrary $1 \leqslant m \leqslant M$ and then take a union bound. Let $\Delta_m \in \mathbb{R}^{K\times M\times K}$, and define

$$F(\Delta_m) = L_m^{\mathrm{MN}}(A_m^{\mathrm{MN}} + \Delta_m) - L_m^{\mathrm{MN}}(A_m^{\mathrm{MN}}) + \lambda\|A_m^{\mathrm{MN}} + \Delta_m\|_R - \lambda\|A_m^{\mathrm{MN}}\|_R, \tag{29}$$

Examples	Prediction	Network estimates	Mixed membership v.s. single category
Political tweets	MN, mix are better	MN, mix are better	Each Twitter user has one ideology tendency
MemeTracker	All methods work well	LN, mix better for *uk.reuters* → *reuters* and *breitbart* → *wral*	*uk.reuters*, *breitbart* and *breitbart* cover multiple topics
		MN, mix better for *alertnet* → *reuters* and *canadianbusiness* → *wral*	*alertnet* and *bizjournals* are primarily about one topic

Table 10: Summary of the comparison among the three methods in the two real data examples. "MN", "LN", "mix" refer to the multinomial, logistic-normal, and mixture methods. The last column shows whether nodes exhibit mixed membership in multiple categories or falls mainly in single categories and further validates our main hypothesis.

where

$$L_m^{\mathrm{MN}}(A_m) = \frac{1}{T}\sum_{t=0}^{T-1}\left[f(\langle A_m, X^t\rangle + \nu_m^{\mathrm{MN}}) - \sum_{k=1}^{K}\langle A_{m,k}, X^t\rangle X_{m,k}^{t+1}\right], \quad f(x) = \log\left(\sum_{i=1}^{K} e^{x_i} + 1\right),$$

and

$$\|A_m\|_R = \sum_{m'=1}^{M} \|A_{m,:,m',:}\|_F.$$

Our goal is to show that if $F(\Delta_m) \leqslant 0$, the following holds with high probability:

$$\|\Delta_m\|_F^2 \leqslant \frac{C\rho_m^{\mathrm{MN}} \log M}{T}, \quad \|\Delta_m\|_R \leqslant C\rho_m^{\mathrm{MN}}\sqrt{\frac{\log M}{T}}. \tag{30}$$

The following lemma shows that we only need to prove the claim above for $\|\Delta_m\|_R \leqslant C$.

Lemma 2 *For any convex function g and norm $\|\cdot\|$, if $g(0) = 0$, $g(x) > 0$ as long as $\|x\| = C$, then $g(x) \leqslant 0$ implies $\|x\| < C$.*

Since $F(\cdot)$ is convex, we only need to show that $F(\Delta_m) \leqslant 0$ and $\|\Delta_m\|_R \leqslant C$ imply the error bounds (30). This is because that the error bounds suggest $\|\Delta_m\|_R \leqslant C\rho_m^{\mathrm{MN}}\sqrt{\frac{\log M}{T}} < C$, thus the condition in Lemma 2 holds.

Denote the Bregman divergence induced by any function g as $D_g(\cdot,\cdot)$, then if $F(\Delta_m) \leqslant 0$,

$$D_{L_m^{\mathrm{MN}}}(A_m^{\mathrm{MN}} + \Delta_m, A_m^{\mathrm{MN}}) \leqslant -\langle \nabla L_m^{\mathrm{MN}}(A_m^{\mathrm{MN}}), \Delta_m\rangle + \lambda\|A_m^{\mathrm{MN}}\|_R - \lambda\|A_m^{\mathrm{MN}} + \Delta_m\|_R, \tag{31}$$

The following lemma provide an upper bound for the R.H.S.

Lemma 3 *Under the model generation process* (4)*, if* $K \leqslant M$*, then with probability at least* $1 - \exp\{-c \log M\}$*,*

$$\left\| L_m^{\mathrm{MN}}(A_m^{\mathrm{MN}}) \right\|_{R^*} < CK\sqrt{\frac{\log M}{T}} \leqslant \frac{\lambda}{2},$$

where $c, C > 0$ *are universal constants.*

Thus we can bound the R.H.S. of (31) by

$$\frac{\lambda}{2}\|\Delta_m\|_R + \lambda\|\Delta_{m,:,S_m^{\mathrm{MN}},:}\|_R - \lambda\|\Delta_{m,:,(S_m^{\mathrm{MN}})^c,:}\|_R \leqslant \frac{3\lambda}{2}\|\Delta_{m,:,S_m^{\mathrm{MN}},:}\|_R - \frac{\lambda}{2}\|\Delta_{m,:,(S_m^{\mathrm{MN}})^c,:}\|_R.$$

By the definition of L_m^{MN},

$$\begin{aligned} D_{L_m^{\mathrm{MN}}}(A_m^{\mathrm{MN}} + \Delta_m, A_m^{\mathrm{MN}}) =& \frac{1}{T}\sum_{t=0}^{T-1} D_f(\langle A_m^{\mathrm{MN}}, X^t\rangle + \nu_m^{\mathrm{MN}} + \langle \Delta_m, X^t\rangle, \langle A_m^{\mathrm{MN}}, X^t\rangle + \nu_m^{\mathrm{MN}}) \\ \geqslant& \frac{1}{T}\sum_{t=0}^{T-1} \frac{\lambda_{\min}(\nabla^2 f(\xi^t))}{2}\|\langle \Delta_m, X^t\rangle\|_2^2, \end{aligned}$$

where $\xi^t \in \mathbb{R}^K$ is some point lying between $\langle A_m^{\mathrm{MN}} + \Delta_m, X^t\rangle + \nu_m^{\mathrm{MN}}$ and $\langle A_m^{\mathrm{MN}}, X^t\rangle + \nu_m^{\mathrm{MN}}$. Since we have assumed

$$\|A^{\mathrm{MN}}\|_{\infty,\infty,1,\infty} \leqslant R_{\max}^{\mathrm{MN}}, \quad \|\Delta_m\|_R \leqslant C,$$

we know $\langle A_m^{\mathrm{MN}} + \Delta_m, X^t\rangle + \nu_m^{\mathrm{MN}}, \langle A_m^{\mathrm{MN}}, X^t\rangle + \nu_m^{\mathrm{MN}} \in [-C_1 - C, C_1 + C]^K$, and thus $\xi^t \in [-C_1 - C, C_1 + C]^K$, where $C_1 = R_{\max}^{\mathrm{MN}} + \|\nu^{\mathrm{MN}}\|_\infty$.

The next step is to lower bound $\lambda_{\min}(\nabla^2 f(\xi^t))$. First we calculate the Hessian matrix of f:

$$\left(\nabla^2 f(x)\right)_{i,j} = -\frac{e^{x_i + x_j}}{\left(\sum_{k=1}^K e^{x_k} + 1\right)^2} + \frac{e^{x_i}\mathbb{1}_{\{i=j\}}}{\sum_{k=1}^K e^{x_k} + 1},$$

then for any $u \in \mathbb{R}^K$,

$$\begin{aligned} u^\top \nabla^2 f(x) u =& \sum_{i,j} u_i u_j \left(\nabla^2 f(x)\right)_{i,j} \\ =& \left(\sum_{k=1}^K e^{x_k} + 1\right)^{-2} \left\{ -\left(\sum_{i=1}^K u_i e^{x_i}\right)^2 + \left(\sum_{i=1}^K u_i^2 e^{x_i}\right)\left(\sum_{i=1}^K e^{x_i} + 1\right)\right\} \\ \geqslant& \left(\sum_{k=1}^K e^{x_k} + 1\right)^{-2} \left(\sum_{i=1}^K u_i^2 e^{x_i}\right) \\ \geqslant& \|u\|_2^2 \min_i e^{x_i} \left(\sum_{k=1}^K e^{x_k} + 1\right)^{-2}. \end{aligned}$$

The third line is due to Cauchey-Schwartz inequality:

$$\left(\sum_{i=1}^K u_i e^{x_i}\right)^2 = \left(\sum_{i=1}^K u_i e^{\frac{x_i}{2}} e^{\frac{x_i}{2}}\right)^2 \leqslant \left(\sum_{i=1}^K u_i^2 e^{x_i}\right)\left(\sum_{i=1}^K e^{x_i}\right).$$

Therefore, $\lambda_{\min}(\nabla^2 f(\xi^t)) \geqslant \frac{e^{-(C_1+C)}}{\left(Ke^{C_1+C}+1\right)^2} > 0$. Combining this with (31), we know that

$$\left\|(\Delta_m)_{:,(S_m^{\mathrm{MN}})^c,:}\right\|_R \leqslant 3\left\|(\Delta_m)_{:,S_m^{\mathrm{MN}},:}\right\|_R, \tag{32}$$

Now we would like to lower bound $\frac{1}{T}\sum_{t=0}^{T-1}\left\|\langle \Delta_m, X^t\rangle\right\|_2^2$ with the following restricted eigenvalue condition. First we define set $\mathcal{C}(S,\kappa)$ of $K \times M \times K$ tensors, for any set $S \subseteq \{1,\ldots,M\}$, and constant $\kappa > 0$:

$$\mathcal{C}(S,\kappa) = \{U \in \mathbb{R}^{K\times M\times K} : \|U_{:,S^c,:}\|_R \leqslant \kappa \|U_{:,S,:}\|_R\}.$$

Lemma 4 *Under the model generation process* (4)*, if* $K \leqslant M$ *and* $T \geqslant Ce^{2C_1}(1+Ke_1^C)^2K^4 \cdot(\rho^{\mathrm{MN}})^2 \log M$*, then with probability at least* $1-2\exp\{-c\log M\}$*,*

$$\inf_{U\in\mathcal{C}(S_m^{\mathrm{MN}},3)} \frac{1}{T}\sum_{t=0}^{T-1}\frac{\left\|\langle U, X^t\rangle\right\|_2^2}{\|U\|_F^2} \geqslant \frac{1}{2}e^{-C_1}(1+Ke_1^C)^{-1},$$

where $C_1 = R_{\max}^{\mathrm{MN}} + \|\nu^{\mathrm{MN}}\|_\infty$*,* $c, C > 0$ *are universal constants.*

By (32), $\Delta_m \in \mathcal{C}(S_m^{\mathrm{MN}}, 3)$. Therefore, with probability at least $1-3\exp\{-c\log M\}$,

$$\frac{ce^{-2C_1}}{4(CKe_1^C+1)^3}\|\Delta_m\|_F^2 \leqslant \frac{3\lambda}{2}\left\|(\Delta_m)_{:,S_m^{\mathrm{MN}},:}\right\|_R \leqslant \frac{3\sqrt{\rho_m^{\mathrm{MN}}}\lambda}{2}\|(\Delta_m)\|_F$$

which further implies,

$$\|\Delta_m\|_F \leqslant CKe^{2C_1}(CKe^{C_1}+1)^3\sqrt{\frac{\rho_m^{\mathrm{MN}}\log M}{T}},$$

and

$$\begin{aligned}\|\Delta_m\|_R \leqslant& 4\left\|(\Delta_m)_{:,S_m^{\mathrm{MN}},:}\right\|_R\\ \leqslant& 4\sqrt{\rho_m^{\mathrm{MN}}}\|\Delta_m\|_F\\ \leqslant& CKe^{2C_1}(CKe^{C_1}+1)^3\rho_m^{\mathrm{MN}}\sqrt{\frac{\log M}{T}}.\end{aligned}$$

where $C_1 = R_{\max}^{\mathrm{MN}} + \|\nu^{\mathrm{MN}}\|_\infty, C > 0$ is universal constants.

7.2 Proof of Theorem 2

We follow similar steps from the proof of Theorem 1. Here for any $\Delta_m \in \mathbb{R}^{(K-1)\times M\times K}$ we define $F(\Delta_m)$ as

$$F(\Delta_m) = L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}}+\Delta_m) - L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}}) + \lambda\|A_m^{\mathrm{LN}}+\Delta_m\|_R - \lambda\|A_m^{\mathrm{LN}}\|_R, \tag{33}$$

where $L_m^{\mathrm{LN}}(A_m) = \frac{1}{2T}\sum_{t\in\mathcal{T}_m}\|Y_m^{t+1} - \mu^{t+1}(A_m, \nu_m^{\mathrm{LN}})\|_2^2$ and $\mathcal{T}_m = \{t : X_m^{t+1} \neq 0_{K\times 1}\}$. We will prove that $F(\Delta_m) \leqslant 0$ implies the error bounds for Δ_m. We start with the standard equations

$$D_{L_m^{\mathrm{LN}}}(A_m^{\mathrm{LN}} + \Delta_m, A_m^{\mathrm{LN}}) \leqslant -\langle \nabla L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}}), \Delta_m\rangle + \lambda\|A_m^{\mathrm{LN}}\|_R - \lambda\|A_m^{\mathrm{LN}} + \Delta_m\|_R. \tag{34}$$

Lemma 5 (Deviation Bound) *Under the data generation process* (7) *and* (8) *with* $q^t = q$*,*

$$\left\|\nabla L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}})\right\|_{R^*} \leqslant CK\sqrt{\max_k \Sigma_{k,k}}\sqrt{\frac{\log M|\mathcal{T}_m|}{T^2}} \leqslant \frac{\lambda}{2}.$$

With probability at least $1 - \exp(-c\log M)$*, for universal constants* $c, C > 0$*.*

Similarly we can also write

$$-\langle \nabla L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}}), \Delta_m\rangle + \lambda\|A_m^{\mathrm{LN}}\|_R - \lambda\|A_m^{\mathrm{LN}} + \Delta_m\|_R \leqslant \frac{3\lambda}{2}\|\Delta_{m,:,S_m^{\mathrm{LN}},:}\|_R - \frac{\lambda}{2}\|\Delta_{m,:,(S_m^{\mathrm{LN}})^c,:}\|_R,$$

and thus $\|\Delta_{m,:,(S_m^{\mathrm{LN}})^c,:}\|_R \leqslant 3\|\Delta_{m,:,S_m^{\mathrm{LN}},:}\|_R$. By the definition of L_m^{LN}, $D_{L_m^{\mathrm{LN}}}(A_m^{\mathrm{LN}}+\Delta_m, A_m^{\mathrm{LN}}) = \frac{1}{2T}\sum_{t\in\mathcal{T}_m}\|\langle\Delta_m, X^t\rangle\|_2^2$, and it can be lower bounded based on the following Lemma that holds for $\mathcal{C}(S_m^{\mathrm{LN}}, 3) = \{U \in \mathbb{R}^{(K-1)\times M\times K} : \left\|U_{:,(S_m^{\mathrm{LN}})^c,:}\right\|_R \leqslant \kappa\left\|U_{:,S_m^{\mathrm{LN}},:}\right\|_R\}$,

Lemma 6 (Restricted Eigenvalue Condition) *Under the data generation process* (7) *and* (8) *with* $q^t = q$*, if* $K \leqslant M$ *and* $T \geqslant \frac{CK^4(\rho_m^{\mathrm{LN}})^2}{q_m^2\gamma_1^2}\log M$*, then*

$$\inf_{U\in\mathcal{C}(S_m^{\mathrm{LN}},3)}\frac{1}{2T\|U\|_F^2}\sum_{t\in\mathcal{T}_m}\|\langle U, X^t\rangle\|_2^2 \geqslant \frac{q_m\gamma_1}{4}$$

with probability at least $1 - 3\exp\{-c\log M\}$*, where*

$$\gamma_1 = \min\{\frac{\min_j q_j\beta_1}{4K+1}, \frac{\min_j q_j(1-q_j)}{4K}\},$$
$$\beta_1 = \frac{2(e-1)^2}{e^6(2\pi)^{\frac{K-1}{2}}K^2|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{(K-1)(R_{\max}^{\mathrm{LN}} + \|\nu^{\mathrm{LN}}\|_\infty + 2)^2}{2\lambda_{\min}(\Sigma)}\right\},$$

and $c, C > 0$ *are universal constants.*

Due to Lemma 6, with probability at least $1 - 2\exp\{-c\log M\}$,

$$\begin{aligned}\|\Delta_m\|_F^2 \leqslant& \frac{6\lambda}{q_m\gamma_1}\|\Delta_{m,S_m^{\mathrm{LN}}}\|_R\\ \leqslant& \frac{CK\sqrt{\max_k\Sigma_{k,k}}}{q_m\gamma_1}\sqrt{\frac{\rho_m^{\mathrm{LN}}\log M\max_{m'}|\mathcal{T}_{m'}|}{T^2}}\|\Delta_m\|_F,\\ \|\Delta_m\|_R \leqslant& \frac{CK\sqrt{\max_k\Sigma_{k,k}}\rho_m^{\mathrm{LN}}}{q_m\gamma_1}\sqrt{\frac{\log M\max_{m'}|\mathcal{T}_{m'}|}{T^2}}.\end{aligned} \tag{35}$$

The following lemma provides an upper bound for $T_m = |\mathcal{T}_m|$:

Lemma 7

$$\mathbb{P}\left(|\mathcal{T}_m| > 2q_m T\right) \leqslant \exp\{-2q_m^2 T\}.$$

Note that $\gamma_1 \leqslant 1$, thus if $T \geqslant \frac{C\rho^{\mathrm{LN}}(K^2+\log M)}{q_m^2\gamma_1^2}$, $\exp\{-2q_m^2 T\} \leqslant \exp\{-c\log M\}$. Therefore, with probability at least $1 - C\exp\{-c\log M\}$,

$$\begin{aligned}
\|\Delta_m\|_F^2 \leqslant & \frac{CK^2 \max_k \Sigma_{k,k}^2 \max_{m'} q_{m'}}{\gamma_1^2 q_m^2} \frac{\rho_m^{\mathrm{LN}} \log M}{T}, \\
\|\Delta_m\|_R \leqslant & \frac{CK \max_k \Sigma_{k,k} \rho_m^{\mathrm{LN}}}{\gamma_1} \sqrt{\frac{\max_{m'} q_{m'} \log M}{q_m^2 T}},
\end{aligned} \tag{36}$$

holds for $1 \leqslant m \leqslant M$, which implies

$$\begin{aligned}
\|\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}\|_F^2 \leqslant & \frac{CK^2 \max_k \Sigma_{k,k}^2 \max_m q_{m'}}{\gamma_1^2 \min_m q_m^2} \frac{s^{\mathrm{LN}} \log M}{T}, \\
\|\widehat{A}^{\mathrm{LN}} - A^{\mathrm{LN}}\|_R \leqslant & \frac{CK \max_k \Sigma_{k,k} s^{\mathrm{LN}}}{\gamma_1} \sqrt{\frac{\max_m q_m \log M}{\min_m q_m^2 T}}.
\end{aligned} \tag{37}$$

7.3 Proof of Theorem 3

Similarly from the previous proofs, we only prove the error bounds for an arbitrary m first. Let $\Delta_m^A \in \mathbb{R}^{(K-1)\times M\times K}$, $\Delta_m^B \in \mathbb{R}^{M\times K}$, and $\Delta_m(\alpha) \in \mathbb{R}^{K\times M\times K}$ be concatenated by $\sqrt{\alpha}\Delta_m^A$ and $\sqrt{1-\alpha}\Delta_m^B$ in the first dimension. Formally, $\Delta_{m,1:(K-1),:,:}(\alpha) = \sqrt{\alpha}\Delta_m^A$, $\Delta_{m,K,:,:}(\alpha) = \sqrt{1-\alpha}\Delta_m^B$. For simplicity, we will omit $\Delta_m(\alpha)$ to Δ_m. Define

$$\begin{aligned}
F(\Delta_m) = & \alpha L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}} + \Delta_m^A) + (1-\alpha) L_m^{\mathrm{Bern}}(B_m^{\mathrm{Bern}} + \Delta_m^B) + \lambda R_\alpha(A_m^{\mathrm{LN}} + \Delta_m^A, B_m^{\mathrm{Bern}} + \Delta_m^B) \\
& - \alpha L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}}) - (1-\alpha) L_m^{\mathrm{Bern}}(B_m^{\mathrm{Bern}}) - \lambda R_\alpha(A_m^{\mathrm{LN}}, B_m^{\mathrm{Bern}}).
\end{aligned} \tag{38}$$

Our goal is to show that if $F(\Delta_m) \leqslant 0$, the following holds with high probability:

$$\begin{aligned}
\|\Delta_m\|_F^2 = \alpha\|\Delta_m^A\|_F^2 + (1-\alpha)\|\Delta_m^B\|_F^2 \leqslant & \frac{C\rho_m^{\mathrm{LN,Bern}} \log M}{T}, \\
\|\Delta_m\|_R = R_\alpha(\Delta_m^A, \Delta_m^B) \leqslant & C\rho_m^{\mathrm{LN,Bern}} \sqrt{\frac{\log M}{T}}.
\end{aligned} \tag{39}$$

Given Lemma 2, we only need to show that $F(\Delta_m) \leqslant 0$ and $\|\Delta_m\|_R \leqslant \sqrt{1-\alpha}$ imply the error bounds (39). This is because that the error bounds suggest $\|\Delta_m\|_R \leqslant \frac{12C(\alpha)K}{\gamma_2}\rho_m^{\mathrm{LN,Bern}}\sqrt{\frac{\log M}{T}} < \sqrt{1-\alpha}$, thus the condition in Lemma 2 holds.

If $F(\Delta_m) \leqslant 0$,

$$\begin{aligned}
& \alpha D_{L_m^{\mathrm{LN}}}(A_m^{\mathrm{LN}} + \Delta_m^A, A_m^{\mathrm{LN}}) + (1-\alpha) D_{L_m^{\mathrm{Bern}}}(B_m^{\mathrm{Bern}} + \Delta_m^B, B_m^{\mathrm{Bern}}) \\
\leqslant & -\alpha\langle \nabla L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}}), \Delta_m^A\rangle - (1-\alpha)\langle \nabla L_m^{\mathrm{Bern}}(B_m^{\mathrm{Bern}}), \Delta_m^B\rangle \\
& + \lambda R_\alpha(A_m^{\mathrm{LN}}, B_m^{\mathrm{Bern}}) - \lambda R_\alpha(A_m^{\mathrm{LN}} + \Delta_m^A, B_m^{\mathrm{Bern}} + \Delta_m^B).
\end{aligned} \tag{40}$$

The following lemmas provide an upper bound for the R.H.S.

Lemma 8 (Deviation bound for continuous error) *Under the data generation process* (7), (8), (11), *with probability at least* $1-\exp(-c\log(M))$,

$$\left\|\nabla L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}})\right\|_\infty \leqslant C\max_k\sqrt{\Sigma_{k,k}}\sqrt{\frac{\log(M)}{T}},$$

for universal constants $c, C>0$.

Lemma 9 (Deviation bound for discrete error) *Under the data generation process* (7), (8), (11), *with probability at least* $1-\exp(-c\log M)$,

$$\left\|\nabla L_m^{\mathrm{Bern}}(B_m^{\mathrm{Bern}})\right\|_\infty \leqslant C\sqrt{\frac{\log(M)}{T}},$$

for universal constants $c, C>0$.

By Lemma 8 and Lemma 9, with probability at least $1-\exp\{-c\log M\}$,

$$\begin{aligned}
&-\alpha\left\langle\nabla L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}}),\Delta_m^A\right\rangle-(1-\alpha)\left\langle\nabla L_m^{\mathrm{Bern}}(B_m^{\mathrm{Bern}}),\Delta_m^B\right\rangle\\
=&-\sum_{m'=1}^M\left\langle\sqrt{\alpha}(\nabla L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}}))_{:,m',:},\sqrt{\alpha}\Delta^A_{m,:,m',:}\right\rangle+\left\langle\sqrt{1-\alpha}(\nabla L_m^{\mathrm{LN}}(B_m^{\mathrm{Bern}}))_{m',:},\sqrt{1-\alpha}\Delta^B_{m,m',:}\right\rangle\\
\leqslant&\sum_{m'=1}^M\left(\alpha\|\nabla L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}})_{:,m',:}\|_F^2+(1-\alpha)\|\nabla L_m^{\mathrm{Bern}}(B_m^{\mathrm{Bern}})_{m',:}\|_2^2\right)^{\frac12}\|\Delta_{m,:,m',:}(\alpha)\|_F\\
\leqslant&\left(C\alpha(K-1)K\Sigma_{k,k}\frac{\log M}{T}+C'(1-\alpha)K\frac{\log M}{T}\right)^{\frac12}\|\Delta_m\|_R\\
\leqslant&\left(C(K-1)\max_k\Sigma_{k,k}\alpha+C'(1-\alpha)\right)^{\frac12}\sqrt{\frac{K\log M}{T}}\|\Delta_m\|_R.
\end{aligned}$$

Setting $\lambda=C(\alpha)K\sqrt{\frac{\log M}{T}}$, where $C(\alpha)=[C\max_k\Sigma_{k,k}\alpha+C'(1-\alpha)]^{\frac12}$ for some universal constants $C, C'>0$. Then we have

$$-\alpha\left\langle\nabla L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}}),\Delta_m^A\right\rangle-(1-\alpha)\left\langle\nabla L_m^{\mathrm{Bern}}(B_m^{\mathrm{Bern}}),\Delta_m^B\right\rangle\leqslant\frac{\lambda}{2}\|\Delta_m\|_R.$$

Let $S_m^{\mathrm{LN,Bern}}=\{(i,j,k):\alpha\|A^{\mathrm{LN}}_{m,:,j,:}\|_F^2+(1-\alpha)\|B^{\mathrm{Bern}}_{m,:,j,:}\|_F^2>0\}$ be the support set of A_m^{LN} and B_m^{Bern}, then we can write

$$\begin{aligned}
&R_\alpha(A_m^{\mathrm{LN}},B_m^{\mathrm{Bern}})-R_\alpha(\widehat{A}_m^{\mathrm{LN}},\widehat{B}_m^{\mathrm{Bern}})\\
=&R_\alpha(A^{\mathrm{LN}}_{m,S_m^{\mathrm{LN,Bern}}},B^{\mathrm{Bern}}_{m,S_m^{\mathrm{LN,Bern}}})-R_\alpha(\widehat{A}^{\mathrm{LN}}_{m,S_m^{\mathrm{LN,Bern}}},\widehat{B}^{\mathrm{Bern}}_{m,S_m^{\mathrm{LN,Bern}}})-R_\alpha(\widehat{A}^{\mathrm{LN}}_{m,(S_m^{\mathrm{LN,Bern}})^c},\widehat{B}^{\mathrm{Bern}}_{m,(S_m^{\mathrm{LN,Bern}})^c})\\
\leqslant&R_\alpha(\Delta^A_{m,S_m^{\mathrm{LN,Bern}}},\Delta^B_{m,S_m^{\mathrm{LN,Bern}}})-R_\alpha(\Delta^A_{m,S_m^{\mathrm{LN,Bernc}}},\Delta^B_{m,(S_m^{\mathrm{LN,Bern}})^c})\\
=&\|\Delta_{m,S_m^{\mathrm{LN,Bern}}}\|_R-\|\Delta_{m,(S_m^{\mathrm{LN,Bern}})^c}\|_R
\end{aligned}$$

Therefore, the R.H.S of (40) is bounded by $\frac{3\lambda}{2}\|\Delta_{m,S_m^{\text{LN,Bern}}}\|_R - \frac{\lambda}{2}\|\Delta_{m,(S_m^{\text{LN,Bern}})^c}\|_R$. Since L_m^{LN} and L_m^{Bern} are both convex, the L.H.S. of (40) is non-negative. Thus $\|\Delta_{m,(S_m^{\text{LN,Bern}})^c}\|_R \leqslant 3\|\Delta_{m,S_m^{\text{LN,Bern}}}\|_R$. Define set $\mathcal{C}(S_m^{\text{LN,Bern}},\kappa)$ of $K\times M\times K$ tensors for any $\kappa>0$ as follows:

$$\mathcal{C}(S_m^{\text{LN,Bern}},\kappa)=\{U\in\mathbb{R}^{K\times M\times K}:\left\|U_{(S_m^{\text{LN,Bern}})^c}\right\|_R\leqslant\kappa\left\|U_{S_m^{\text{LN,Bern}}}\right\|_R\}, \tag{41}$$

then $\Delta_m\in\mathcal{C}(S_m^{\text{LN,Bern}},3)$.

Now we would like to show the strong convexity of L_m^{LN} and L_m^{Bern} as a function of $\langle A_m,X^t\rangle$ and $\langle B_m,X^t\rangle$. As shown in the proof of Theorem 2,

$$D_{L_m^{\text{LN}}}(A_m^{\text{LN}}+\Delta_m^A,A_m^{\text{LN}})=\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{1}_{\{X_m^{t-1}\neq 0\}}\|\langle\Delta_m^A,X^t\rangle\|_2^2. \tag{42}$$

Meanwhile, $\|\Delta_m^B\|_{1,\infty}\leqslant\|\Delta_m^B\|_{1,2}\leqslant\frac{\|\Delta_m\|_R}{\sqrt{1-\alpha}}$, thus the strong convexity of L_m^{Bern} is guaranteed by the following lemma:

Lemma 10 (Strong convexity (L_m^{Bern})) *Define $\sigma_C\triangleq\frac{e^{C_2+1}}{(1+e^{C_2+1})^2}$, where $C_2=R_{\max}^{\text{LN,Bern}}+\|\eta^{\text{Bern}}\|_\infty$, then we have*

$$D_{L_m^{\text{Bern}}}(B_m^{\text{Bern}}+\Delta_m^B,B_m^{\text{Bern}})\geqslant\frac{\sigma_C}{2T}\sum_{t=0}^{T-1}\langle\Delta_m^B,X^t\rangle^2.$$

The following Lemma provides a lower bound for

$$\frac{\alpha}{2T}\sum_{t=0}^{T-1}\mathbb{1}_{\{X_m^{t-1}\neq 0\}}\|\langle\Delta_m^A,X^t\rangle\|_2^2+\frac{(1-\alpha)\sigma_C}{2T}\sum_{t=0}^{T-1}\langle\Delta_m^B,X^t\rangle^2$$

in terms of $\|\Delta_m\|_F^2$.

Lemma 11 (Restricted Eigenvalue Condition) *For any $U\in\mathbb{R}^{K\times M\times K}$, let $U^{(1)}=U_{1:(K-1),:,:}$ and $U^{(2)}=U_{K,:,:}$. There exists a constant c_1, such that if $K\leqslant M$ and $T\geqslant\frac{CK^4(\rho_m^{\text{LN,Bern}})^2}{\gamma_2^2}\log M$,*

$$\inf_{U\in\mathcal{C}(S_m^{\text{LN,Bern}},3)\cap B_F(1)}\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{1}_{\{\mathbf{X}_m^{t-1}\neq 0\}}\|\langle U^{(1)},X^t\rangle\|_2^2+\frac{\sigma_C}{2T}\sum_{t=0}^{T-1}\langle U^{(2)},X^t\rangle^2\geqslant\frac{\gamma_2}{2},$$

with probability at least $1-2\exp\{-c\log M\}$. Here $c,C>0$ are universal constants, and

$$\begin{aligned}\gamma_2=&\frac{e^{C_2+1}}{2\left(1+e^{C_2+1}\right)^3}\min\left\{\frac{\beta_2}{4K+1},\frac{e^{C_2}}{4K(1+e^{C_2})}\right\},\\ C_2=&R_{\max}^{\text{LN,Bern}}+\|\eta^{\text{Bern}}\|_\infty,C_3=R_{\max}^{\text{LN,Bern}}+\|\nu^{\text{LN}}\|_\infty\\ \beta_2=&\frac{2(e-1)^2}{e^6(2\pi)^{\frac{K-1}{2}}K^2|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{(K-1)(C_3+2)^2}{2\lambda_{\min}(\Sigma)}\right\}.\end{aligned} \tag{43}$$

Therefore, combining (40), (42), Lemma 10 and Lemma 11 leads us to

$$\begin{aligned}\|\Delta_m\|_F^2 \leqslant & \frac{9C(\alpha)^2K^2}{\gamma_2^2}\frac{\rho_m^{\mathrm{LN,Bern}}\log M}{T},\\ \|\Delta_m\|_R \leqslant & 4\sqrt{\rho_m^{\mathrm{LN,Bern}}}\|\Delta_m\|_F \leqslant \frac{12C(\alpha)K}{\gamma_2}\rho_m^{\mathrm{LN,Bern}}\sqrt{\frac{\log M}{T}},\end{aligned}$$

with probability at least $1-3\exp\{-c\log M\}$. Taking a union bound over $1 \leqslant m \leqslant M$ gives us the final result.

7.4 Proof of Proposition 1

For notational convenience, under the logistic-normal model with $q^t = q$, we let $\eta_m^{\mathrm{Bern}} = \log(q_m/(1-q_m))$ in this proof. First we define the prediction functions $g^{\mathrm{MN}} : \mathbb{R}^K \to \mathbb{R}^K$ and $g^{\mathrm{LN}} : \mathbb{R}^K \to \mathbb{R}^K$ as follows:

$$\begin{aligned}g_k^{\mathrm{MN}}(u) =& \frac{e^{u_k}}{1+\sum_{l=1}^K e^{u_l}}, 1 \leqslant k \leqslant K\\ g_k^{\mathrm{LN}}(u) =& \frac{e^{u_K}}{1+e^{u_K}}\frac{e^{u_k}}{1+\sum_{l=1}^{K-1}e^{u_l}}, 1 \leqslant k \leqslant K-1\\ g_K^{\mathrm{LN}}(u) =& \frac{e^{u_K}}{1+e^{u_K}}\frac{1}{1+\sum_{l=1}^{K-1}e^{u_l}}.\end{aligned} \tag{44}$$

Let $E^{\mathrm{MN}}_{t,m_1,m_2,k_2}, E^{\mathrm{LN}}_{t,m_1,m_2,k_2}, E^{\mathrm{LN,Bern}} \in \mathbb{R}$ be defined as follows:

$$\begin{aligned}E^{\mathrm{MN}}_{t,m_1,m_2,k_2} =& \|g^{\mathrm{MN}}(\langle \widehat{A}^{\mathrm{MN}}_{m_1}, X^t\rangle + \nu^{\mathrm{MN}}_{m_1}) - g^{\mathrm{MN}}(\langle \widehat{A}^{\mathrm{MN}}_{m_1}, \bar{X}^t(m_2,k_2)\rangle + \nu^{\mathrm{MN}}_{m_1})\\ & - g^{\mathrm{MN}}(\langle A^{\mathrm{MN}}_{m_1}, X^t\rangle + \nu^{\mathrm{MN}}_{m_1}) + g^{\mathrm{MN}}(\langle A^{\mathrm{MN}}_{m_1}, \bar{X}^t(m_2,k_2)\rangle + \nu^{\mathrm{MN}}_{m_1})\|_2,\end{aligned}$$

$$\begin{aligned}E^{\mathrm{LN}}_{t,m_1,m_2,k_2} =& \|g^{\mathrm{LN}}(\langle \widehat{A}^{\mathrm{LN}}_{m_1}, X^t\rangle + \nu^{\mathrm{LN}}_{m_1}, \eta^{\mathrm{Bern}}_{m_1})\\ & - g^{\mathrm{LN}}(\langle \widehat{A}^{\mathrm{LN}}_{m_1}, \bar{X}^t(m_2,k_2)\rangle + \nu^{\mathrm{LN}}_{m_1}, \eta^{\mathrm{Bern}}_{m_1})\\ & - g^{\mathrm{LN}}(\langle A^{\mathrm{LN}}_{m_1}, X^t\rangle + \nu^{\mathrm{LN}}_{m_1}, \eta^{\mathrm{Bern}}_{m_1})\\ & + g^{\mathrm{LN}}(\langle A^{\mathrm{LN}}_{m_1}, \bar{X}^t(m_2,k_2)\rangle + \nu^{\mathrm{LN}}_{m_1}, \eta^{\mathrm{Bern}}_{m_1})\|_2,\end{aligned}$$

$$\begin{aligned}E^{\mathrm{LN,Bern}}_{t,m_1,m_2,k_2} =& \|g^{\mathrm{LN,Bern}}(\langle \widehat{A}^{\mathrm{LN}}_{m_1}, X^t\rangle + \nu^{\mathrm{LN}}_{m_1}, \langle \widehat{B}^{\mathrm{Bern}}_{m_1}, X^t\rangle + \eta^{\mathrm{Bern}}_{m_1})\\ & - g^{\mathrm{LN}}(\langle \widehat{A}^{\mathrm{LN}}_{m_1}, \bar{X}^t(m_2,k_2)\rangle + \nu^{\mathrm{LN}}_{m_1}, \langle \widehat{B}^{\mathrm{Bern}}_{m_1}, \bar{X}^t(m_2,k_2)\rangle + \eta^{\mathrm{Bern}}_{m_1})\\ & - g^{\mathrm{LN}}(\langle A^{\mathrm{LN}}_{m_1}, X^t\rangle + \nu^{\mathrm{LN}}_{m_1}, \langle B^{\mathrm{Bern}}_{m_1}, X^t\rangle + \eta^{\mathrm{Bern}}_{m_1})\\ & + g^{\mathrm{LN}}(\langle A^{\mathrm{LN}}_{m_1}, \bar{X}^t(m_2,k_2)\rangle + \nu^{\mathrm{LN}}_{m_1}, \langle B^{\mathrm{Bern}}_{m_1}, \bar{X}^t(m_2,k_2)\rangle + \eta^{\mathrm{Bern}}_{m_1})\|_2.\end{aligned}$$

Then by definition and Jensen's inequality,

$$\left\|\widehat{V}^{\mathrm{MN}} - V^{\mathrm{MN}}\right\|_F \leqslant \frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{m_1,m_2,k_2}(E^{\mathrm{MN}}_{t,m_1,m_2,k_2})^2\right)^{\frac{1}{2}},$$

$$\left\|\widehat{V}^{\mathrm{LN}} - V^{\mathrm{LN}}\right\|_F \leqslant \frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{m_1,m_2,k_2}(E^{\mathrm{LN}}_{t,m_1,m_2,k_2})^2\right)^{\frac{1}{2}},$$

$$\left\|\widehat{V}^{\mathrm{LN,Bern}} - V^{\mathrm{LN,Bern}}\right\|_F \leqslant \frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{m_1,m_2,k_2}(E^{\mathrm{LN,Bern}}_{t,m_1,m_2,k_2})^2\right)^{\frac{1}{2}},$$

In order to bound $E^{\mathrm{MN}}_{t,m_1,m_2,k_2}, E^{\mathrm{LN}}_{t,m_1,m_2,k_2}$ and $E^{\mathrm{LN,Bern}}_{t,m_1,m_2,k_2}$, first we would like to upper bound the largest singular value of $\nabla g^{\mathrm{MN}}(u), \nabla g^{\mathrm{LN}}(u) \in \mathbb{R}^{K\times K}$. For any $k > 0$, define $M^{(k)} : \mathbb{R}^k \to \mathbb{R}^{k\times k}$ satisfying

$$(M^{(k)}(u))_{i,j} = \frac{e^{u_i}}{1+\sum_{l=1}^k e^{u_l}}\left(\mathbb{1}_{\{i=j\}} - \frac{e^{u_j}}{1+\sum_{l=1}^k e^{u_l}}\right),$$

and $g^{(k)} : \mathbb{R}^k \to \mathbb{R}^k$ satisfying

$$(g^{(k)}(u))_i = \frac{e^{u_i}}{1+\sum_{l=1}^k e^{u_l}},$$

then some calculation shows that $\nabla g^{\mathrm{MN}}(u) = M_K(u)$ and

$$\nabla g^{\mathrm{LN}}(u) = \begin{pmatrix} \frac{e^{u_K}}{1+e^{u_K}}M^{(K-1)}(u_{1:(K-1)}) & \frac{e^{u_K}}{(1+e^{u_K})^2}g^{(K-1)}(u_{1:(K-1)}) \\ -\frac{e^{u_K}g^{(K-1)}(u_{1:(K-1)})}{(1+e^{u_K})(1+\sum_{l=1}^{K-1}e^{u_l})} & \frac{e^{u_K}}{(1+e^{u_K})^2(1+\sum_{l=1}^{K-1}e^{u_l})}. \end{pmatrix}$$

Note that $M^{(k)}(u)$ is symmetric, and for any $v \in \mathbb{R}^K$,

$$\begin{aligned} v^\top M^{(k)}(u)v &= \sum_{j=1}^k g_j^{(k)}(u)v_j^2 - \left(\sum_{j=1}^k g_j^{(k)}v_j\right)^2 \\ &\geqslant \sum_{j=1}^k g_j^{(k)}(u)v_j^2 - \sum_{j=1}^k g_j^{(k)}v_j^2\sum_{j=1}^k g_j^{(k)} \\ &\geqslant \sum_{j=1}^k g_j^{(k)}(u)v_j^2\left(1-\sum_{j=1}^k g_j^{(k)}\right) \\ &\geqslant 0, \end{aligned} \tag{45}$$

$$v^\top M^{(k)}(u)v \leqslant \sum_{j=1}^k g_j^{(k)}(u)v_j^2 \leqslant \|v\|_2^2,$$

which implies that $\|M^{(k)}(u)\| \leqslant 1$ and thus $\|\nabla g^{\mathrm{MN}}(u)\| \leqslant 1$. Here we have applied Cauchey-Schwartz inequality on the second line of (45). Meanwhile, for any $v \in \mathbb{R}^K$,

$$\begin{aligned}
&\left\| \frac{e^{u_K}}{1+e^{u_K}} M^{(K-1)}(u_{1:(K-1)}) v_{1:(K-1)} + \frac{e^{u_K} v_K}{(1+e^{u_K})^2} g^{(K-1)}(u_{1:(K-1)}) \right\|_2 \\
\leqslant & \frac{e^{u_K}}{1+e^{u_K}} \left(\|v_{1:(K-1)}\|_2 + \frac{|v_K|}{1+e^{u_K}} \right) \\
\leqslant & \frac{e^{u_K}}{1+e^{u_K}} \sqrt{1+(1+e^{u_K})^{-2}} \|v\|_2,
\end{aligned} \tag{46}$$

and

$$\begin{aligned}
&\left| -\frac{e^{u_K} g^{(K-1)\top}(u_{1:(K-1)}) v_{1:(K-1)}}{(1+e^{u_K})(1+\sum_{l=1}^{K-1} e^{u_l})} + \frac{e^{u_K} v_K}{(1+e^{u_K})^2 (1+\sum_{l=1}^{K-1} e^{u_l})} \right| \\
\leqslant & \frac{e^{u_K}}{1+e^{u_K}} \left(\frac{\|v_{1:(K-1)}\|_2}{1+\sum_{l=1}^{K-1} e^{u_l}} + \frac{|v_K|}{(1+e^{u_K})(1+\sum_{l=1}^{K-1} e^{u_l})} \right) \\
\leqslant & \frac{e^{u_K}}{(1+e^{u_K})(1+\sum_{l=1}^{K-1} e^{u_l})} \sqrt{1+(1+e^{u_K})^{-2}} \|v\|_2,
\end{aligned} \tag{47}$$

where we have applied Cauchey-Schwarz inequality on the last line of both inequalities above. Thus we have

$$\begin{aligned}
\left\| \nabla g^{\mathrm{LN}}(u) v \right\|_2 \leqslant & \frac{e^{u_K}}{1+e^{u_K}} \sqrt{\left(1 + (1 + \sum_{l=1}^{K-1} e^{u_l})^{-2} \right) (1 + (1+e^{u_K})^{-2}) \|v\|_2} \\
\leqslant & \sqrt{2} \|v\|_2,
\end{aligned} \tag{48}$$

where the last line is due to that $(1+e^{u_K})^{-2} \leqslant e^{-u_K}$, which implies that

$$\left(1 + (1+e^{u_K})^{-2}\right) \frac{e^{u_K}}{1+e^{u_K}} \leqslant 1.$$

Therefore, $g^{\mathrm{MN}}(u)$ is 1-Lipschitz while g^{LN} is $\sqrt{2}$-Lipschitz. Now we are ready to prove the error bounds for $\widehat{V}^{\mathrm{MN}}$, $\widehat{V}^{\mathrm{LN}}$, and $\widehat{V}^{\mathrm{LN,Bern}}$.

1. Upper bounding $\|\widehat{V}^{\mathrm{MN}} - V^{\mathrm{MN}}\|_F$ under the conditions in Theorem 1
   In the following we discuss upper bounds for each error term $E^{\mathrm{MN}}_{t,m_1,m_2,k_2}$ for two cases: $m_2 \in S^{\mathrm{MN}}_{m_1} = \{m' : A^{\mathrm{MN}}_{m_1,:,m',:} \neq 0\}$ and $m_2 \notin S^{\mathrm{MN}}_{m_1}$.

(a) $m_2 \in S^{\mathrm{MN}}_{m_1}$
Under this case, note that

$$\begin{aligned}
&\left\|\langle \widehat{A}^{\mathrm{MN}}_{m_1}, X^t\rangle - \langle A^{\mathrm{MN}}_{m_1}, X^t\rangle\right\|_2 \\
=&\left(\sum_{k_1=1}^{K} \langle \widehat{A}^{\mathrm{MN}}_{m_1,k_1} - A^{\mathrm{MN}}_{m_1,k_1}, X^t\rangle^2\right)^{\frac{1}{2}} \\
\leqslant&\left(\sum_{k_1=1}^{K}\left(\sum_{m'}\max_{k'}\left|\widehat{A}^{\mathrm{MN}}_{m_1,k_1,m',k'} - A^{\mathrm{MN}}_{m_1,k_1,m',k'}\right|\right)^2\right)^{\frac{1}{2}} \\
\leqslant&\left(\sum_{k_1=1}^{K}\left(\sum_{m'}\left\|\widehat{A}^{\mathrm{MN}}_{m_1,k_1,m',:} - A^{\mathrm{MN}}_{m_1,k_1,m',:}\right\|_2\right)^2\right)^{\frac{1}{2}} \\
\leqslant&\sum_{m'}\left\|\widehat{A}^{\mathrm{MN}}_{m_1,:,m',:} - A^{\mathrm{MN}}_{m_1,:,m',:}\right\|_F \\
=&\|\widehat{A}^{\mathrm{MN}}_{m_1} - A^{\mathrm{MN}}_{m_1}\|_R
\end{aligned} \tag{49}$$

where we have applied Minkowski's inequality on the 5th line of the above inequality. Similarly we also have

$$\left\|\langle \widehat{A}^{\mathrm{MN}}_{m_1}, \bar{X}^t(m_2,k_2)\rangle - \langle A^{\mathrm{MN}}_{m_1}, \bar{X}^t(m_2,k_2)\rangle\right\|_2 \leqslant \|\widehat{A}^{\mathrm{MN}}_{m_1} - A^{\mathrm{MN}}_{m_1}\|_R.$$

Thus one can show that

$$E^{\mathrm{MN}}_{t,m_1,m_2,k_2} \leqslant 2\|\widehat{A}^{\mathrm{MN}}_{m_1} - A^{\mathrm{MN}}_{m_1}\|_R.$$

(b) $m_2 \notin S^{\mathrm{MN}}_{m_1}$
Since $A^{\mathrm{MN}}_{m_1,:,m_2,:} = 0$, $\langle A^{\mathrm{MN}}_{m_1}, X^t\rangle = \langle A^{\mathrm{MN}}_{m_1}, \bar{X}^t(m_2,k_2)\rangle$, and

$$\begin{aligned}
\left\|\langle \widehat{A}^{\mathrm{MN}}_{m_1}, X^t\rangle - \langle \widehat{A}^{\mathrm{MN}}_{m_1}, \bar{X}^t(m_2,k_2)\rangle\right\|_2 &= \left\|\widehat{A}^{\mathrm{MN}}_{m_1,:,m_2,k_2}(X^t_{m_2,k_2} - \bar{X}_{m_2,k_2})\right\|_2 \\
&\leqslant \left\|\widehat{A}^{\mathrm{MN}}_{m_1,:,m_2,k_2} - A^{\mathrm{MN}}_{m_1,:,m_2,k_2}\right\|_2.
\end{aligned}$$

Since g^{MN} is 1-Lipschitz, this implies that

$$E^{\mathrm{MN}}_{t,m_1,m_2,k_2} \leqslant \left\|\widehat{A}^{\mathrm{MN}}_{m_1,:,m_2,k_2} - A^{\mathrm{MN}}_{m_1,:,m_2,k_2}\right\|_2. \tag{50}$$

Combining the two cases together, we know that

$$\begin{aligned}
&\left\|\widehat{V}^{\mathrm{MN}}-V^{\mathrm{MN}}\right\|_F \\
\leqslant&\frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{m_1}\left(\sum_{m_2\in S_{m_1}^{\mathrm{MN}}}\sum_{k_2=1}^{K}(E_{t,m_1,m_2,k_2}^{\mathrm{MN}})^2+\sum_{m_2\notin S_{m_1}^{\mathrm{MN}}}\sum_{k_2=1}^{K}(E_{t,m_1,m_2,k_2}^{\mathrm{MN}})^2\right)\right)^{\frac{1}{2}} \\
\leqslant&\left(\sum_{m_1}\left(4K\rho_{m_1}^{\mathrm{MN}}\|\widehat{A}_{m_1}^{\mathrm{MN}}-A_{m_1}^{\mathrm{MN}}\|_R^2+\|\widehat{A}_{m_1}^{\mathrm{MN}}-A_{m_1}^{\mathrm{MN}}\|_F^2\right)\right)^{\frac{1}{2}} \\
\leqslant&\left(\sum_{m_1}CK^3e^{4C_1}(CKe^{C_1}+1)^6\frac{\rho_{m_1}^{\mathrm{MN3}}\log M}{T}\right)^{\frac{1}{2}} \\
\leqslant&CK^{\frac{3}{2}}e^{2C_1}(CKe^{C_1}+1)^3\rho^{\mathrm{MN}}\sqrt{\frac{s^{\mathrm{MN}}\log M}{T}},
\end{aligned} \tag{51}$$

where the third line is due to that

$$\begin{aligned}
\|\widehat{A}_{m_1}^{\mathrm{MN}}-A_{m_1}^{\mathrm{MN}}\|_R\leqslant&CKe^{2C_1}(CKe^{C_1}+1)^3\rho_{m_1}^{\mathrm{MN}}\sqrt{\frac{\log M}{T}} \\
\|\widehat{A}_{m_1}^{\mathrm{MN}}-A_{m_1}^{\mathrm{MN}}\|_F\leqslant&CK^2e^{4C_1}(CKe^{C_1}+1)^6\frac{\rho_{m_1}^{\mathrm{MN}}\log M}{T},
\end{aligned}$$

which has been shown in the proof of Theorem 1, and C_1 is as defined in Theorem 1.

2. Upper bounding $\|\widehat{V}^{\mathrm{LN}}-V^{\mathrm{LN}}\|_F$ under the conditions in Theorem 2
Following similar arguments for bounding $E_{t,m_1,m_2,k_2}^{\mathrm{MN}}$, we can also show that under the logistic-normal model with constant q^t,

$$E_{t,m_1,m_2,k_2}^{\mathrm{LN}}\leqslant 2\sqrt{2}\left\|\widehat{A}_{m_1}^{\mathrm{LN}}-A_{m_1}^{\mathrm{LN}}\right\|_R,$$

for $m_2\in S_{m_1}^{\mathrm{LN}}$, and

$$E_{t,m_1,m_2,k_2}^{\mathrm{LN}}\leqslant\sqrt{2}\left\|\widehat{A}_{m_1,:,m_2,k_2}^{\mathrm{LN}}-A_{m_1,:,m_2,k_2}^{\mathrm{LN}}\right\|_2, \tag{52}$$

which implies that

$$\begin{aligned}
\left\|\widehat{V}^{\mathrm{LN}}-V^{\mathrm{LN}}\right\|_F\leqslant&\left(\sum_{m_1}\left(8K\rho_{m_1}^{\mathrm{LN}}\left\|\widehat{A}_{m_1}^{\mathrm{LN}}-A_{m_1}^{\mathrm{LN}}\right\|_R\right.\right. \\
&\left.\left.+2\|\widehat{A}_{m_1}^{\mathrm{LN}}-A_{m_1}^{\mathrm{LN}}\|_F^2\right)\right)^{\frac{1}{2}} \\
\leqslant&\left(\sum_{m_1}\frac{CK^3\max_k\Sigma_{k,k}^2}{\gamma_1^2}\frac{\rho_{m_1}^{\mathrm{LN3}}\max_m q_m\log M}{\min_m q_m^2T}\right)^{\frac{1}{2}} \\
\leqslant&\frac{CK^{\frac{3}{2}}\max_k\Sigma_{k,k}\rho^{\mathrm{LN}}}{\gamma_1}\sqrt{\frac{s^{\mathrm{LN}}\max_m q_m\log M}{\min_m q_m^2T}},
\end{aligned} \tag{53}$$

where the third line is due to that

$$\|\widehat{A}^{\mathrm{LN}}_{m_1} - A^{\mathrm{LN}}_{m_1}\|_R \leqslant \frac{CK \max_k \Sigma_{k,k}}{\gamma_1} \rho^{\mathrm{LN}}_{m_1} \sqrt{\frac{\max_m q_m \log M}{\min_m q_m^2 T}}$$
$$\|\widehat{A}^{\mathrm{LN}}_{m_1} - A^{\mathrm{LN}}_{m_1}\|_F \leqslant \frac{CK^2 \max_k \Sigma^2_{k,k}}{\gamma_1^2} \frac{\rho^{\mathrm{LN}}_{m_1} \max_m q_m \log M}{\min_m q_m^2 T},$$

which has been shown in the proof of Theorem 2.

3. Upper bounding $\|\widehat{V}^{\mathrm{LN,Bern}} - V^{\mathrm{LN,Bern}}\|_F$ under the conditions in Theorem 3 While for the case where q^t depending on the past,

$$E^{\mathrm{LN,Bern}}_{t,m_1,m_2,k_2} \leqslant \frac{2\sqrt{2}}{\min\{\sqrt{\alpha}, \sqrt{1-\alpha}\}} R_\alpha(\widehat{A}^{\mathrm{LN}}_{m_1} - A^{\mathrm{LN}}_{m_1}, \widehat{B}^{\mathrm{Bern}}_{m_1} - B^{\mathrm{Bern}}_{m_1}),$$

for $m_2 \in S^{\mathrm{LN}}_{m_1}$, and

$$E^{\mathrm{LN,Bern}}_{t,m_1,m_2,k_2} \leqslant \sqrt{2} \left(\left\| \widehat{A}^{\mathrm{LN}}_{m_1,:,m_2,k_2} - A^{\mathrm{LN}}_{m_1,:,m_2,k_2} \right\|_2^2 + \left(\widehat{B}^{\mathrm{Bern}}_{m_1,m_2,k_2} - B^{\mathrm{Bern}}_{m_1,m_2,k_2} \right)^2 \right)^{\frac{1}{2}}. \quad (54)$$

which implies that

$$\begin{aligned} \left\| \widehat{V}^{\mathrm{LN,Bern}} - V^{\mathrm{LN,Bern}} \right\|_F \leqslant & \left(\sum_{m_1} \left(\frac{8K\rho^{\mathrm{LN}}_{m_1}}{\min\{\alpha, 1-\alpha\}} R^2_\alpha(\widehat{A}^{\mathrm{LN}}_{m_1} - A^{\mathrm{LN}}_{m_1}, \widehat{B}^{\mathrm{Bern}}_{m_1} - B^{\mathrm{Bern}}_{m_1}) \right. \right. \\ & \left. \left. + 2\|\widehat{A}^{\mathrm{LN}}_{m_1} - A^{\mathrm{LN}}_{m_1}\|_F^2 + 2\|\widehat{B}^{\mathrm{Bern}}_{m_1} - B^{\mathrm{Bern}}_{m_1}\|_2^2 \right) \right)^{\frac{1}{2}} \\ \leqslant & \left(\sum_{m_1} \frac{CC(\alpha)^2 K^3}{\gamma_2^2 \min\{\alpha, 1-\alpha\}} \frac{\rho^{\mathrm{LN,Bern3}}_{m_1} \log M}{T} \right)^{\frac{1}{2}} \\ \leqslant & \frac{CC(\alpha) K^{\frac{3}{2}} \rho^{\mathrm{LN,Bern}}}{\gamma_2 \min\{\sqrt{\alpha}, \sqrt{1-\alpha}\}} \sqrt{\frac{s^{\mathrm{LN,Bern}} \log M}{T}}, \end{aligned} \quad (55)$$

where the third line is due to that

$$R_\alpha(\widehat{A}^{\mathrm{LN}}_{m_1} - A^{\mathrm{LN}}_{m_1}, \widehat{B}^{\mathrm{Bern}}_{m_1} - B^{\mathrm{Bern}}_{m_1}) \leqslant \frac{12C(\alpha)K}{\gamma_2} \rho^{\mathrm{LN,Bern}}_{m_1} \sqrt{\frac{\log M}{T}}$$
$$\|\widehat{A}^{\mathrm{LN}}_{m_1} - A^{\mathrm{LN}}_{m_1}\|_F^2 + \|\widehat{B}^{\mathrm{Bern}}_{m_1} - B^{\mathrm{Bern}}_{m_1}\|_2^2 \leqslant \frac{9C(\alpha)^2 K^2}{\gamma_2^2 \min\{\alpha, 1-\alpha\}} \frac{\rho^{\mathrm{LN}}_{m_1} \log M}{T},$$

which has been shown in the proof of Theorem 3.

7.5 Proof of Lemma 6

First let $\mathcal{F}_t = \sigma(X^0, \ldots, X^t)$ be the σ field generated by $X^0, \ldots, X^t$. We can write

$$\begin{aligned} \frac{1}{T} \sum_{t \in \mathcal{T}_m} \|\langle U, X^t \rangle\|_2^2 = & \sum_k U_k^\top \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}(X^t X^{t\top} \mathbb{1}_{\{X^{t+1}_m \neq 0\}} | \mathcal{F}_t) U_k \\ & + \sum_k U_k^\top \frac{1}{T} \sum_{t=0}^{T-1} \left[X^t X^{t\top} \mathbb{1}_{\{X^{t+1}_m \neq 0\}} - \mathbb{E}(X^t X^{t\top} \mathbb{1}_{\{X^{t+1}_m \neq 0\}} | \mathcal{F}_t) \right] U_k \end{aligned} \quad (56)$$

(1) Bounding the eigenvalue of $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}(X^tX^{t\top}\mathbb{1}_{\{X_m^{t+1}\neq 0\}}|\mathcal{F}_t)$
We can write

$$\begin{aligned}\mathbb{E}(X^tX^{t\top}\mathbb{1}_{\{X_m^{t+1}\neq 0\}}|\mathcal{F}_{t-1}) =&\mathbb{E}(X^tX^{t\top}\mathbb{E}(\mathbb{1}_{\{X_m^{t+1}\neq 0\}}|\mathcal{F}_t)|\mathcal{F}_{t-1})\\ =&q_m\mathbb{E}(X^t|\mathcal{F}_{t-1})\mathbb{E}(X^t|\mathcal{F}_{t-1})^\top + q_m\mathrm{Cov}(X^t|\mathcal{F}_{t-1}),\end{aligned}$$

where $\mathbb{E}(X^t|\mathcal{F}_{t-1})\mathbb{E}(X^t|\mathcal{F}_{t-1})^\top$ is positive semi-definite, thus the smallest eigenvalue can be lower bounded by that of $q_m\mathrm{Cov}(X^t|\mathcal{F}_{t-1})$.

Given $\mathcal{F}_{t-1}$, $X_1^t,\ldots,X_M^t$ are all independent, which suggests $\mathrm{Cov}(X^t|\mathcal{F}_{t-1})$ to be a block diagonal matrix. We only need to lower bound the smallest eigenvalue of each $\mathrm{Cov}(X_j^t|\mathcal{F}_{t-1})$. Recall that $X_j^t = Z_j^t$ with probability q_j, where Z_j^t follows a logistic-normal distribution. Hence we have that, for any $1\leqslant j\leqslant M$,

$$\mathbb{E}(X_j^t|\mathcal{F}_{t-1}) = q_j\mathbb{E}(Z_j^t|\mathcal{F}_{t-1}),\quad \mathbb{E}(X_j^tX_j^{t\top}|\mathcal{F}_{t-1}) = q_j\mathbb{E}(Z_j^tZ_j^{t\top}), \tag{57}$$

which implies

$$\mathrm{Cov}(X_j^t|\mathcal{F}_{t-1}) = q_j\mathrm{Cov}(Z_j^t) + q_j(1-q_j)\mathbb{E}(Z_j^t)\mathbb{E}(Z_j^{t\top}). \tag{58}$$

Our goal is to show that $u^\top\mathrm{Cov}(X_j^t|\mathcal{F}_{t-1})u \geqslant \gamma_1\|u\|_2^2$ for any $u\in\mathbb{R}^K$. Let $\bar{u} = \frac{1}{K}\sum_{i=1}^K u_i$, $\widetilde{u} = u - \bar{u}1_K$, then

$$\begin{aligned}&u^\top\mathrm{Cov}(X_j^t|\mathcal{F}_{t-1})u\\ =&q_j\widetilde{u}^\top\mathrm{Cov}(Z_j^t)\widetilde{u} + q_j(1-q_j)((\bar{u}1_K+\widetilde{u})^\top\mathbb{E}(Z_j^t))^2\\ =&q_j\mathrm{Var}(\widetilde{u}^\top Z_j^t) + q_j(1-q_j)(\bar{u}+\widetilde{u}^\top\mathbb{E}(Z_j^t))^2,\end{aligned} \tag{59}$$

where we have applied the fact that $1_K^\top Z_j^t = 1$ holds deterministically. In the following, we first prove a lower bound for the variance of $\widetilde{u}^\top Z_j^t$. Let $\mathcal{P}_{\widetilde{u}} = \{1\leqslant i\leqslant K-1 : \widetilde{u}_i > 0\}$ and $\mathcal{N}_{\widetilde{u}} = \{1\leqslant i\leqslant K-1 : \widetilde{u}_i\leqslant 0\}$. For notational simplicity, we adopt $\mu_m^t = \mu^t(A_m^{\mathrm{LN}},\nu_m^{\mathrm{LN}})$ as a shorthand in the following arguments. Consider the following two events:

$$\begin{aligned}\mathcal{A}: &-\mu_{j,i}^t - 2\leqslant\epsilon_{j,i}^t\leqslant -\mu_{j,i}^t - 1, \forall i\in\mathcal{P}_{\widetilde{u}},\\ &-\mu_{j,i}^t + 1\leqslant\epsilon_{j,i}^t\leqslant -\mu_{j,i}^t + 2, \forall i\in\mathcal{N}_{\widetilde{u}}.\end{aligned} \tag{60}$$

and

$$\begin{aligned}\mathcal{B}: &-\mu_{j,i}^t - 2\leqslant\epsilon_{j,i}^t\leqslant -\mu_{j,i}^t - 1, \forall i\in\mathcal{N}_{\widetilde{u}},\\ &-\mu_{j,i}^t + 1\leqslant\epsilon_{j,i}^t\leqslant -\mu_{j,i}^t + 2, \forall i\in\mathcal{P}_{\widetilde{u}}.\end{aligned} \tag{61}$$

Then if event $\mathcal{A}$ happens,

$$\begin{aligned}
\widetilde{u}^\top Z_j^t =& \frac{\sum_{i=1}^{K-1} \widetilde{u}_i \exp\{\mu_{j,i}^t + \epsilon_{j,i}^t\} + \widetilde{u}_K}{\sum_{i=1}^{K-1} \exp\{\mu_{j,i}^t + \epsilon_{j,i}^t\} + 1} \\
=& \frac{\sum_{i=1}^{K-1} \widetilde{u}_i \left(\exp\{\mu_{j,i}^t + \epsilon_{j,i}^t\} - 1\right)}{\sum_{i=1}^{K-1} \exp\{\mu_{j,i}^t + \epsilon_{j,i}^t\} + 1} \\
\leqslant& \frac{\sum_{i\in\mathcal{P}_{\widetilde{u}}} \widetilde{u}_i \left(e^{-1} - 1\right) + \sum_{i\in\mathcal{N}_{\widetilde{u}}} \widetilde{u}_i \left(e - 1\right)}{\sum_{i=1}^{K-1} \exp\{\mu_{j,i}^t + \epsilon_{j,i}^t\} + 1} \\
\leqslant& - \frac{(1-e^{-1}) \sum_{i\in\mathcal{P}_{\widetilde{u}}} \widetilde{u}_i + (e-1) \sum_{i\in\mathcal{P}_{\widetilde{u}}} (-\widetilde{u}_i)}{(K-1)e^2 + 1} \\
\leqslant& - \frac{e-1}{Ke^3} \|\widetilde{u}\|_1.
\end{aligned} \tag{62}$$

On the other hand, if event $\mathcal{B}$ happens,

$$\begin{aligned}
\widetilde{u}^\top Z_j^t =& \frac{\sum_{i=1}^{K-1} \widetilde{u}_i \left(\exp\{\mu_{j,i}^t + \epsilon_{j,i}^t\} - 1\right)}{\sum_{i=1}^{K-1} \exp\{\mu_{j,i}^t + \epsilon_{j,i}^t\} + 1} \\
\geqslant& \frac{\sum_{i\in\mathcal{P}_{\widetilde{u}}} \widetilde{u}_i \left(e - 1\right) + \sum_{i\in\mathcal{N}_{\widetilde{u}}} (-\widetilde{u}_i) \left(1 - e^{-1}\right)}{\sum_{i=1}^{K-1} \exp\{\mu_{j,i}^t + \epsilon_{j,i}^t\} + 1} \\
\leqslant& \frac{e-1}{Ke^3} \|\widetilde{u}\|_1.
\end{aligned} \tag{63}$$

Now we are ready to lower bound the variance of $\widetilde{u}^\top Z_j^t$. Consider the following three cases:

- $\mathbb{E}(\widetilde{u}^\top Z_j^t) \leqslant -\frac{e-1}{Ke^3} \|\widetilde{u}\|_1$

$$\begin{aligned}
\mathrm{Var}(\widetilde{u}^\top Z_j^t) =& \mathbb{E}\left((\widetilde{u}^\top Z_j^t - \mathbb{E}(\widetilde{u}^\top Z_j^t))^2\right) \\
\geqslant& \mathbb{E}\left(\mathbb{1}_{\{\mathcal{B}\}}(\widetilde{u}^\top Z_j^t - \mathbb{E}(\widetilde{u}^\top Z_j^t))^2\right) \\
\geqslant& \mathbb{P}(\mathcal{B}) \frac{4(e-1)^2}{K^2 e^6} \|\widetilde{u}\|_1^2 \\
\geqslant& \frac{4(e-1)^2 \|\widetilde{u}\|_1^2}{(2\pi)^{\frac{K-1}{2}} e^6 K^2 |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{(K-1)(R_{\max}^{\mathrm{LN}} + \|\nu^{\mathrm{LN}}\|_\infty + 2)^2}{2\lambda_{\min}(\Sigma)}\right\},
\end{aligned} \tag{64}$$

  where we have applied the fact that $|\mu_{m,i}^t| \leqslant R_{\max}^{\mathrm{LN}} + \|\nu^{\mathrm{LN}}\|_\infty$ on the last line.

- $\mathbb{E}(\widetilde{u}^\top Z_j^t) \geqslant \frac{e-1}{Ke^3} \|\widetilde{u}\|_1$

$$\begin{aligned}
\mathrm{Var}(\widetilde{u}^\top Z_j^t) \geqslant& \mathbb{E}\left(\mathbb{1}_{\{\mathcal{A}\}}(\widetilde{u}^\top Z_j^t - \mathbb{E}(\widetilde{u}^\top Z_j^t))^2\right) \\
\geqslant& \mathbb{P}(\mathcal{A}) \frac{4(e-1)^2}{K^2 e^6} \|\widetilde{u}\|_1^2 \\
\geqslant& \frac{4(e-1)^2}{(2\pi)^{\frac{K-1}{2}} K^2 e^6 |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{(K-1)(R_{\max}^{\mathrm{LN}} + \|\nu^{\mathrm{LN}}\|_\infty + 2)^2}{2\lambda_{\min}(\Sigma)}\right\} \|\widetilde{u}\|_1^2.
\end{aligned} \tag{65}$$

- $-\frac{e-1}{Ke^3}\|\widetilde{u}\|_1 \leqslant \mathbb{E}(\widetilde{u}^\top Z_j^t) \leqslant \frac{e-1}{Ke^3}\|\widetilde{u}\|_1$

$$\begin{aligned}
\mathrm{Var}(\widetilde{u}^\top Z_j^t) \geqslant & \mathbb{P}(\mathcal{A})\left(\mathbb{E}(\widetilde{u}^\top Z_j^t) + \frac{e-1}{Ke^3}\|\widetilde{u}\|_1\right)^2 \\
& + \mathbb{P}(\mathcal{B})\left(\frac{e-1}{Ke^3}\|\widetilde{u}\|_1 - \mathbb{E}(\widetilde{u}^\top Z_j^t)\right)^2 \\
= & (\mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B}))\left((\mathbb{E}(\widetilde{u}^\top Z_j^t))^2 + \frac{(e-1)^2}{K^2e^6}\|\widetilde{u}\|_1^2\right) \\
& + (\mathbb{P}(\mathcal{A}) - \mathbb{P}(\mathcal{B}))\frac{2(e-1)}{Ke^3}\|\widetilde{u}\|_1\mathbb{E}(\widetilde{u}^\top Z_j^t) \\
\geqslant & \frac{(e-1)^2}{K^2e^6}\|\widetilde{u}\|_1^2\frac{4\mathbb{P}(\mathcal{A})\mathbb{P}(\mathcal{B})}{\mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B})} \\
\geqslant & \frac{2(e-1)^2}{(2\pi)^{\frac{K-1}{2}}K^2e^6|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{(K-1)(R_{\max}^{\mathrm{LN}} + \|\nu^{\mathrm{LN}}\|_\infty + 2)^2}{2\lambda_{\min}(\Sigma)}\right\}\|\widetilde{u}\|_1^2.
\end{aligned} \tag{66}$$

Therefore, it is guaranteed that

$$\begin{aligned}
\mathrm{Var}(\widetilde{u}^\top Z_j^t) \geqslant & \frac{2(e-1)^2\|\widetilde{u}\|_1^2}{(2\pi)^{\frac{K-1}{2}}K^2e^6|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{(K-1)(R_{\max}^{\mathrm{LN}} + \|\nu^{\mathrm{LN}}\|_\infty + 2)^2}{2\lambda_{\min}(\Sigma)}\right\} \\
:= & \beta_1\|\widetilde{u}\|_1^2.
\end{aligned} \tag{67}$$

By (59), if $|\bar{u}| \geqslant 2\|\widetilde{u}\|_\infty$, then

$$\begin{aligned}
& u^\top \mathrm{Cov}(X_j^t|\mathcal{F}_{t-1})u \\
\geqslant & q_j\mathrm{Var}(\widetilde{u}^\top Z_j^t) + \frac{q_j(1-q_j)}{4}\bar{u}^2 \\
\geqslant & q_j\beta_1\|\widetilde{u}\|_1^2 + \frac{q_j(1-q_j)}{4}\bar{u}^2 \\
\geqslant & \min\{q_j\beta_1, \frac{q_j(1-q_j)}{4K}\}(\|\widetilde{u}\|_2^2 + K\bar{u}^2) \\
= & \min\{q_j\beta_1, \frac{q_j(1-q_j)}{4K}\}\|u\|_2^2.
\end{aligned} \tag{68}$$

Otherwise,

$$u^\top \mathrm{Cov}(X_j^t|\mathcal{F}_{t-1})u \geqslant q_j\mathrm{Var}(\widetilde{u}^\top Z_j^t) \geqslant \frac{q_j\beta_1}{4K+1}(4K+1)\|\widetilde{u}\|_2^2 \geqslant \frac{q_j\beta_1}{4K+1}\|u\|_2^2. \tag{69}$$

Therefore, the smallest eigenvalue of $\mathrm{Cov}(X_j^t|\mathcal{F}_{t-1})$ can be lower bounded by $\min\{\frac{q_j\beta_1}{4K+1}, \frac{q_j(1-q_j)}{4K}\}$, which implies that

$$\sum_k U_k^\top \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}(X^t X^{t\top}\mathbb{1}_{\{X_m^{t+1}\neq 0\}}|\mathcal{F}_{t-1})U_k \geqslant q_m\gamma_1\|U\|_F^2,$$

where $\gamma_1 = \min\{\frac{\min_j q_j \beta_1}{4K+1}, \frac{\min_j q_j(1-q_j)}{4K}\}$, and

$$\beta_1 = \frac{2(e-1)^2}{e^6(2\pi)^{\frac{K-1}{2}} K^2 |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{(K-1)(R_{\max}^{\mathrm{LN}} + \|\nu^{\mathrm{LN}}\|_\infty + 2)^2}{2\lambda_{\min}(\Sigma)}\right\}.$$

(2) Uniform concentration of martingale sequence
Note that each element of $X^t X^{t\top} \mathbb{1}_{\{X_m^{t+1} \neq 0\}} - \mathbb{E}(X^t X^{t\top} \mathbb{1}_{\{X_m^{t+1} \neq 0\}} | \mathcal{F}_t)$ is bounded by 1, we can still use the same argument as in the proof of Lemma 4 (presented in Appendix B.1) and obtain

$$\inf_{U \in \mathcal{C}(S_m^{\mathrm{LN}}, 3)} \frac{1}{T} \sum_{t \in \mathcal{T}_m} \frac{\left\|\langle U, X^t \rangle\right\|_2^2}{\|U\|_F^2} \geqslant \frac{q_m \gamma_1}{2},$$

with probability at least $1 - 2\exp\{-c \log M\}$, if $K \leqslant M$ and $T \geqslant \frac{CK^4(\rho_m^{\mathrm{LN}})^2}{q_m^2 \gamma_1^2} \log M$.

8. Conclusion

In this paper, we develop two procedures that estimate context-dependent networks from autoregressive time series of annotated event data. The first approach is a standard regularized multinomial approach for estimating the influence between pairs of nodes (m, m') and pairs of categories (k, k') given that each event belongs to a particular category. Our second logistic-normal approach builds on ideas from compositional time series and is more nuanced since each event consists of a composition of several different topics. We extend existing compositional time series approaches by accounting for the scenario in which no event occurs in our algorithm; significantly, the logistic-normal distribution leads to a convex objective. Our theoretical guarantees show that we can achieve consistent estimation even when the number of network nodes, M, is much larger than the duration of the observation period, T.

We validate our network estimation procedures both with synthetic and two real data examples. Both the synthetic and real data examples suggest that the multinomial approach is better suited to nodes or networks where events tend to belong to a single category, whereas the logistic-normal approach is better suited to nodes in which each event tends to have mixed membership. To handle the situation when both types of nodes exist in the network while the type of each node is unknown, we also develop a mixture approach, including a test procedure for choosing the suitable model for each node and a node-wise fitting procedure, which preserves the merits of both multinomial and logistic-normal approaches in synthetic and real data examples.

Acknowledgments

LZ, GR, BM, and RW were partially supported by ARO W911NF-17-1-0357 and NGA HM0476-17-1-2003. GR was also partially supported by NSF DMS-1811767. RW was also partially supported by NSF DMS-1930049, AFOSR FA9550-18-1-0166, NSF Awards 0353079, 1447449, 1740707, and 1839338.

Appendix A. Table of Notations

Since this paper involves several different models and hence a complex notation system, we include the notations for our multinomial and logistic-normal models in Table 11, notations for theoretical results in Table 12 and notations for the synthetic mixture models (described in Section 5.3) in Table 13.

Notation	Description
$M \in \mathbb{Z}^+$	Number of nodes in the network
$K \in \mathbb{Z}^+$	Number of event categories
$T \in \mathbb{Z}^+$	Number of time series data points
$\triangle^p$	p-dimensional simplex
$X^t \in \mathbb{R}^{M\times K}$	Observed data matrix at time t: each row $X_m^t \in \mathbb{R}^K$ is associated with the node m
$Z^t \in \mathbb{R}^{M\times K}$	Hidden data matrix in the logistic-normal model: $Z_m^t \in \triangle^{K-1}$; X_m^t is either $0_{K\times 1}$ or Z_m^t
$A^{\mathrm{MN}} \in \mathbb{R}^{M\times K\times M\times K}$	Network parameter of the multinomial model: $A^{\mathrm{MN}}_{m_1,k_1,m_2,k_2}$ encodes the influence from (node m_2, category k_2) upon (node m_1, category k_1)
$\nu^{\mathrm{MN}} \in \mathbb{R}^{M\times K}$	Offset parameter of the multinomial model
$A^{\mathrm{LN}} \in \mathbb{R}^{M\times(K-1)\times M\times K}$	Network parameter of the logistic-normal model: $A^{\mathrm{LN}}_{m_1,k_1,m_2,k_2}$ encodes the relative influence from (node m_2, category k_2) upon (node m_1, category k_1) compared to (node m_1, category K)
$\nu^{\mathrm{LN}} \in \mathbb{R}^{M\times(K-1)}$	An offset parameter of the logistic-normal model
$\Sigma \in \mathbb{R}^{(K-1)\times(K-1)}$	Covariance matrix of the logistic-normal model
$q^t \in \mathbb{R}^M$	Event probability of the logistic-normal model: q_m^t is the event probability of node m at time t
$B^{\mathrm{Bern}} \in \mathbb{R}^{M\times M\times K}$	Network parameter of the logistic-normal model with q^t depending on the past: $B^{\mathrm{Bern}}_{m_1,m_2,k}$ encodes the overall influence from (node m_2, category k) upon node m_1
$\eta^{\mathrm{Bern}} \in \mathbb{R}^M$	An offset parameter of the logistic-normal model with q^t depending on the past
$V^{\mathrm{MN}} \in \mathbb{R}^{M\times K\times M\times K}$	Variable importance network parameter of the multinomial model
$V^{\mathrm{LN}} \in \mathbb{R}^{M\times K\times M\times K}$	Variable importance network parameter of the logistic-normal model with $q^t = q$
$V^{\mathrm{LN,Bern}} \in \mathbb{R}^{M\times K\times M\times K}$	Variable importance network parameter of the logistic-normal model with q^t depending on the past

Table 11: Notations in the multinomial and logistic-normal models

Appendix B. Proof of Supporting Lemmas

In this section, we present the proofs of the supporting Lemmas required in Section 7.

Notation	Description
S_m^{MN}	Set of nodes that influence node m in the multinomial model $\{m' : \|A^{\text{MN}}_{m,:,m',:}\|_F > 0\}$
S_m^{LN}	Set of nodes that influence node m in the logistic-normal model with $q^t = q$: $\{m' : \|A^{\text{LN}}_{m,:,m',:}\|_F > 0\}$
$S_m^{\text{LN,Bern}}$	Set of nodes that influence node m in the logistic-normal model with q^t depending on tha past: $\{m' : \|A^{\text{LN}}_{m,:,m',:}\|_F^2 + \|B^{\text{Bern}}_{m,m',:}\|_2^2 > 0\}$
$\rho_m^{\text{MN}} \in \mathbb{N}$	Indegree of node m in the multinomial model: $\|S_m^{\text{MN}}\|$
$\rho^{\text{MN}} \in \mathbb{N}$	Maximum indegree of all M nodes in the multinomial model: $\max_m \rho_m^{\text{MN}}$
s^{MN}	Network sparsity in the multinomial model: $\sum_{m=1}^M \rho_m^{\text{MN}}$
$\rho_m^{\text{LN}} \in \mathbb{N}$	Indegree of node m in the logistic-normal model with $q^t = q$: $\|S_m^{\text{LN}}\|$
$\rho^{\text{LN}} \in \mathbb{N}$	Maximum indegree of all M nodes in the logistic-normal model with $q^t = q$: $\max_m \rho_m^{\text{LN}}$
s^{LN}	Network sparsity in the logistic-normal model with $q^t = q$: $\sum_{m=1}^M \rho_m^{\text{LN}}$
$\rho_m^{\text{LN,Bern}} \in \mathbb{N}$	Indegree of node m in the logistic-normal model with q^t depending on the past: $\|S_m^{\text{LN,Bern}}\|$
$\rho^{\text{LN,Bern}} \in \mathbb{N}$	Maximum indegree of all M nodes in the logistic-normal model with q^t depending on the past: $\max_m \rho_m^{\text{LN,Bern}}$
$s^{\text{LN,Bern}}$	Network sparsity in the logistic-normal model with q^t depending on the past: $\sum_{m=1}^M \rho_m^{\text{LN,Bern}}$
$R_{\max}^{\text{MN}}$	Boundedness parameter for A^{MN} in the multinomial model: $\max_{m,k} \sum_{m'} \max_{k'} \|A^{\text{MN}}_{m,k,m',k'}\|$
$R_{\max}^{\text{LN}}$	Boundedness parameter for A^{LN} in the logistic-normal model with $q^t = q$: $\max_{m,k} \sum_{m'} \max_{k'} \|A^{\text{LN}}_{m,k,m',k'}\|$
$R_{\max}^{\text{LN,Bern}}$	Boundedness parameter for $A^{\text{LN}}, B^{\text{Bern}}$ in the logistic-normal model with q^t depending on the past: $\max_m \max\{\max_k \sum_{m'} \max_{k'} \|A^{\text{LN}}_{m,k,m',k'}\|, \sum_{m'} \max_{k'} \|B^{\text{Bern}}_{m,m',k'}\|$
$\mathcal{T}_m$	$\{t : X_m^{t+1} \neq 0_{K\times 1}\}$
T_m	Size of set $\mathcal{T}_m$: $\|\mathcal{T}_m\|$

Table 12: Notations in theoretical results

B.1 Proof of Lemmas in Section 7.1

Proof [proof of Lemma 2] We prove by contradiction. Assume that their exists $\|x\| > C$ and $g(x) \leqslant 0$, then let $\gamma = \frac{C}{\|x\|} < 1$. Due to the convexity of g,

$$g(\gamma x) = g(\gamma x + (1-\gamma) * 0) \leqslant \gamma g(x) + (1-\gamma) g(0) = \gamma g(x) \leqslant 0.$$

However, $\|\gamma x\| = C$. This contradicts with our condition, so we are forced to conclude that $\|x\| \leqslant C$ is necessary for $g(x) \leqslant 0$. ■

Notation	Description
$X^t \in \mathbb{R}^{M\times K}$	Data at time t generated from the mixture model; X^t is hidden (unobserved) under the contaminated mixture model
$\widetilde{X}^t \in \mathbb{R}^{M\times K}$	Observed data in the contaminated mixture model; $\widetilde{X}^t$ is the contaminated version of X^t
$\widehat{X}_m^t \in \mathbb{R}^K$	Rounded data for node m at time t given the contaminated data $\widetilde{X}^t$
$\widehat{X}^{\text{mix},t} \in \mathbb{R}^{M\times K}$	Estimated data at time t given contaminated data $\widetilde{X}^t$ and estimated node sets $\widehat{\mathcal{N}}_1, \widehat{\mathcal{N}}_2$
$A^{\text{mix}} \in \mathbb{R}^{M\times K\times M\times K}$	Network parameter of the mixture model and contaminated mixture model
$\nu^{\text{mix}} \in \mathbb{R}^{M\times K}$	Offset parameter of the mixture model and the contaminated mixture model
$V^{\text{mix}} \in \mathbb{R}^{M\times K\times M\times K}$	Variable importance network parameter in the mixture model and the contaminated mixture model
$\mathcal{N}_1$	Set of multinomial nodes in the mixture model
$\mathcal{N}_2$	Set of logistic-normal nodes in the mixture model
$a \in \mathbb{R}^+$	Location parameter for the contaminated non-zero multinomial vectors in the contaminated mixture model
$\sigma^{\text{MN}} \in \mathbb{R}^+$	Scale parameter for the contaminated non-zero multinomial vectors in the contaminated mixture model
$\widehat{R}_m \in \mathbb{R} \cup \{\infty\}$	Test statistic for identifying the type of node m in the contaminated mixture model

Table 13: Notations in the synthetic mixture model

Proof [Proof of Lemma 3| By the definition of L_m^{MN},

$$\nabla L_m^{\text{MN}}(A_m^{\text{MN}}) = -\frac{1}{T}\sum_{t=0}^{T-1}(X_m^{t+1} - \nabla f(\langle A_m^{\text{MN}}, X^t\rangle)) \otimes X^t.$$

Define $\epsilon_m^{t+1} := X_m^{t+1} - \mathbb{E}\left(X_m^{t+1}|\mathcal{F}_t\right)$, where $\mathcal{F}_t = \sigma(X^0, \ldots, X^t)$ is the filtration. Since

$$(\nabla f(x))_i = \frac{e^{x_i}}{\sum_{j=1}^K e^{x_j} + 1},$$

we can write $\nabla L_m^{\text{MN}}(A_m^{\text{MN}}) = -\frac{1}{T}\sum_{t=0}^{T-1}\epsilon_m^{t+1} \otimes X^t$. First note that

$$\left\|\frac{1}{T}\sum_{t=0}^{T-1}\epsilon_m^{t+1} \otimes X^t\right\|_{R*} = \max_{m'}\left\|\frac{1}{T}\sum_{t=0}^{T-1}\epsilon_m^{t+1}X_{m'}^{t\top}\right\|_F \leqslant \max_{m',k',k} K\left|\frac{1}{T}\sum_{t=0}^{T-1}\epsilon_{m,k}^{t+1}X_{m',k'}^t\right|,$$

thus we only need to look into $\frac{1}{T}\sum_{t=0}^{T-1}\epsilon_{m,k}^{t+1}X_{m',k'}^t$ for any m', k', k, and then take a union bound. Let $Y_n = \frac{1}{T}\sum_{t=0}^{n-1}\epsilon_{m,k}^{t+1}X_{m',k'}^t$, then $\{Y_n\}_{n=0}^T$ is a martingale sequence, with $Y_0 = 0$. Since

$$\xi_n \triangleq Y_n - Y_{n-1} = \frac{1}{T}\epsilon_{m,k}^n X_{m',k'}^{n-1},$$

$|\xi_n| \leqslant \frac{1}{T}$. Thus by Azuma-Hoeffding's inequality, for any $y > 0$,

$$\mathbb{P}(Y_T \geqslant y) \leqslant \exp\{-\frac{Ty^2}{2}\}.$$

Let $y = C\sqrt{\frac{\log M}{T}}$ and take a union bound over each m', k', k, we know that

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=0}^{T-1}\epsilon_m^{t+1}\otimes X^t\right\|_{R^*} \geqslant CK\sqrt{\frac{\log M}{T}}\right)\\
\leqslant& KM^2\exp\left\{-\frac{Ty^2}{2}\right\}\\
=&\exp\left\{\log K - (C^2/2-2)\log M\right\}\\
\leqslant&\exp\{-c\log M\}.
\end{aligned}$$

■

Proof [Proof of Lemma 4] For notational convenience, we view X^t as a MK-dimensional vector and U as $K \times MK$ dimensional matrix in this proof. First note that

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\left\|\langle U, X^t\rangle\right\|_2^2 = &\sum_{k=1}^{K} U_k^\top \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}(X^tX^{t\top}|\mathcal{F}_{t-1})U_k\\
&+\sum_{k=1}^{K} U_k^\top \frac{1}{T}\sum_{t=0}^{T-1}\left(X^tX^{t\top} - \mathbb{E}(X^tX^{t\top}|\mathcal{F}_{t-1})\right)U_k.
\end{aligned}$$

In the following steps we provide a lower bound for the first term, and concentrate the second term around 0.

(1) Lower bound for the first term
We can decompose the conditional expectation $\mathbb{E}(X^tX^{t\top}|\mathcal{F}_{t-1})$ as two terms:

$$\mathbb{E}(X^tX^{t\top}|\mathcal{F}_{t-1}) = \mathbb{E}(X^t)\mathbb{E}(X^{t\top}) + \mathrm{Cov}(X^t|\mathcal{F}_{t-1}),$$

where the first term is positive semi-definite, and the second term is a block diagonal matrix ($\mathrm{Cov}(X_m^t, X_{m'}^t|\mathcal{F}_{t-1}) = 0$ if $m \neq m'$). Thus we only have to lower bound the eigenvalue of the each $\mathrm{Cov}(X_m^t|\mathcal{F}_{t-1})$. Define matrix $p^t \in \mathbb{R}^{M\times(K+1)}$ as follows:

$$\begin{aligned}
p_{m,k}^t =&\mathbb{P}(X_{m,k}^t = 1|\mathcal{F}_{t-1}) = \frac{\exp\{\langle A_{m,k}^{\mathrm{MN}}, X^{t-1}\rangle\} + \nu_{m,k}^{\mathrm{MN}}}{1+\sum_{l=1}^K \exp\{\langle A_{m,l}^{\mathrm{MN}}, X^{t-1}\rangle + \nu_{m,l}^{\mathrm{MN}}\}}, \quad 1 \leqslant k \leqslant K\\
p_{m,K+1}^t =&\mathbb{P}(X_m^t = 0|\mathcal{F}_{t-1}) = \frac{1}{1+\sum_{l=1}^K \exp\{\langle A_{m,l}^{\mathrm{MN}}, X^{t-1}\rangle + \nu_{m,l}^{\mathrm{MN}}\}}.
\end{aligned}$$

Since $\|A^{\mathrm{MN}}\|_{\infty,\infty,1,\infty} \leqslant R_{\max}^{\mathrm{MN}}$, $p_{m,k}^t \geqslant \frac{e^{-C_1}}{1+Ke^{-C_1}}$ for $1 \leqslant k \leqslant K$, $p_{m,K+1}^t \geqslant \frac{1}{1+Ke^{C_1}}$. We can write

$$\mathrm{Cov}(X_m^t|\mathcal{F}_{t-1}) = \begin{pmatrix} p_{m,1}^t & 0 & \dots & 0\\ 0 & p_{m,2}^t & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & \dots & \dots & p_{m,K}^t\end{pmatrix} - \begin{pmatrix} p_{m,1}^t\\ \vdots\\ p_{m,K}^t\end{pmatrix}\begin{pmatrix} p_{m,1}^t & \dots & p_{m,K}^t\end{pmatrix},$$

For any vector $u \in \mathbb{R}^K$,

$$\begin{aligned}
u^\top \mathrm{Cov}(X_m^t|\mathcal{F}_{t-1})u &= \sum_{k=1}^K p_{m,k}^t u_k^2 - \left(\sum_{k=1}^K p_{m,k}^t u_k\right)^2 \\
&\geqslant \sum_{k=1}^K p_{m,k}^t u_k^2 - \sum_{k=1}^K p_{m,k}^t \left(\sum_{k=1}^K p_{m,k}^t u_k^2\right) \\
&= p_{m,K+1}^t \left(\sum_{k=1}^K p_{m,k}^t u_k^2\right) \\
&\geqslant p_{m,K+1}^t \min_k p_{m,k}^t \|u\|_2^2 \\
&\geqslant \frac{e^{-R_{\max}^{\mathrm{MN}} - \|\nu^{\mathrm{MN}}\|_\infty}}{1 + Ke^{R_{\max}^{\mathrm{MN}} + \|\nu^{\mathrm{MN}}\|_\infty}} \|u\|_2^2,
\end{aligned}$$

Thus the smallest eigenvalue of $\mathrm{Cov}(X^t|\mathcal{F}_{t-1})$ is lower bounded by $\gamma := \frac{e^{-C_1}}{1+Ke^{C_1}}$ where $C_1 = R_{\max}^{\mathrm{MN}} + \|\nu^{\mathrm{MN}}\|_\infty$, which implies that

$$\sum_{k=1}^K U_k^\top \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}(X^t X^{t\top}|\mathcal{F}_{t-1}) U_k \geqslant \gamma \|U\|_F^2. \tag{70}$$

(2) Concentration bound for the second term
Since $U \in \mathcal{C}(S_m^{\mathrm{MN}}, 3)$,

$$\begin{aligned}
&\left| \sum_{k=1}^K U_k^\top \frac{1}{T} \sum_{t=0}^{T-1} \left(X^t X^{t\top} - \mathbb{E}(X^t X^{t\top}|\mathcal{F}_{t-1})\right) U_k \right| \\
\leqslant &\sum_{k=1}^K \|U_k\|_1^2 \left\| \frac{1}{T} \sum_{t=0}^{T-1} \left(X^t X^{t\top} - \mathbb{E}(X^t X^{t\top}|\mathcal{F}_{t-1})\right) \right\|_\infty \\
\leqslant &16K^2 \rho_m^{\mathrm{MN}} \|U\|_F^2 \left\| \frac{1}{T} \sum_{t=0}^{T-1} \left(X^t X^{t\top} - \mathbb{E}(X^t X^{t\top}|\mathcal{F}_{t-1})\right) \right\|_\infty .
\end{aligned}$$

We can bound $\left\| \frac{1}{T} \sum_{t=0}^{T-1} \left(X^t X^{t\top} - \mathbb{E}(X^t X^{t\top}|\mathcal{F}_{t-1})\right) \right\|_\infty$ using the same argument as the proof of Lemma 3. For arbitrary $1 \leqslant m, m' \leqslant M$, $1 \leqslant k, k' \leqslant K$, let

$$Y_n := \frac{1}{T} \sum_{t=0}^{n-1} \left(X_{m,k}^t X_{m',k'}^{t\top} - \mathbb{E}(X_{m,k}^t X_{m',k'}^{t\top}|\mathcal{F}_{t-1})\right)$$

for $n \geqslant 1$, and $Y_0 = 0$, then $\{Y_n\}$ is a bounded difference martingale sequence. Since $|Y_n - Y_{n-1}| \leqslant \frac{1}{T}$, applying Azuma-Hoeffding's inequality and taking a union bound over m, m', k, k' would lead us to

$$\begin{aligned}
&\mathbb{P}\left(\left\| \frac{1}{T} \sum_{t=0}^{T-1} \left(X^t X^{t\top} - \mathbb{E}(X^t X^{t\top}|\mathcal{F}_{t-1})\right) \right\|_\infty > \frac{\gamma}{32K^2 \rho_m^{\mathrm{MN}}} \right) \\
\leqslant &2K^2 M^2 \exp\{-\frac{c\gamma^2 T}{K^4 (\rho_m^{\mathrm{MN}})^2}\} \\
\leqslant &2\exp\{-c \log M\},
\end{aligned}$$

if $K \leqslant M$ and $T \geqslant \frac{CK^4(\rho_m^{\mathrm{MN}})^2}{\gamma^2}\log M$.

Therefore, with probability at least $1 - 2\exp\{-c\log M\}$,

$$\inf_{U\in\mathcal{C}(S_m^{\mathrm{MN}},3)} \frac{1}{T}\sum_{t=0}^{T-1}\frac{\left\|\langle U, X^t\rangle\right\|_2^2}{\|U\|_F^2} \geqslant \frac{\gamma}{2},$$

where $\gamma = \frac{e^{-C_1}}{1+Ke^{C_1}}$. ■

B.2 Proof of Lemmas in Section 7.2

Proof [proof of Lemma 5] First we prove the upper bound conditioning on $\mathcal{T}_m = \{t_1, \ldots, t_{|\mathcal{T}_m|}\}$. Since $\nabla L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}}) = -\frac{1}{T}\sum_{i=1}^{|\mathcal{T}_m|}\epsilon_m^{t_i+1}\otimes X^{t_i}$, we start by bounding each entry of $\frac{1}{T}\sum_{i=1}^{|\mathcal{T}_m|}\epsilon_{m,k}^{t_i+1}X_{m',k'}^{t_i}$. Let

$$Y_n = \frac{1}{T}\sum_{i=1}^{n-1}\epsilon_{m,k}^{t_i+1}X_{m',k'}^{t_i},$$

with $Y_0 = 0$ and $Y_{|\mathcal{T}_m|} = \frac{1}{T}\sum_{i=1}^{|\mathcal{T}_m|}\epsilon_{m,k}^{t_i+1}X_{m',k'}^{t_i}$. Then $\{Y_n\}_{n=0}^{|\mathcal{T}_m|}$ is a martingale with filtrations $\mathcal{F}_n = \sigma(X^1, \ldots, X^{t_n}, \mathcal{T}_m)$. Let $\xi_n = Y_n - Y_{n-1} = -\frac{1}{T}\epsilon_{m,k}^{t_{n-1}+1}X_{m',k'}^{t_{n-1}}$ be the corresponding martingale difference sequence. The moment generating function of Y_n satisfies

$$\mathbb{E}(e^{\eta Y_n}) = \mathbb{E}[e^{\eta Y_{n-1}}\mathbb{E}(e^{\eta\xi_n}|\mathcal{F}_{n-1})], \tag{71}$$

for any η. Since $\epsilon_{m,k}^{t_{n-1}+1} \sim \mathcal{N}(0, \Sigma_{k,k})$ given $\mathcal{F}_n$, we can bound $\mathbb{E}(e^{\eta\xi_n}|\mathcal{F}_{n-1})]$ in the following:

$$\mathbb{E}(e^{\eta\xi_n}|\mathcal{F}_{n-1}) = \mathbb{E}\left(\exp\left\{\frac{\eta X_{m',k'}^{t_{n-1}}}{T}\epsilon_{m,k}^{t_{n-1}+1}\right\}|\mathcal{F}_{n-1}\right) \leqslant \exp\left\{\frac{\eta^2\Sigma_{k,k}(X_{m',k'}^{t_{n-1}})^2}{2T^2}\right\} \leqslant \exp\left\{\frac{\eta^2\Sigma_{k,k}}{2T^2}\right\}.$$

Therefore, combining this with (71) we have

$$\mathbb{E}(e^{\eta Y_T}) \leqslant e^{\frac{\eta^2\Sigma_{k,k}|\mathcal{T}_m|}{2T^2}}.$$

Applying Chernoff bound further shows that, for any $\eta > 0$,

$$\begin{aligned}\mathbb{P}(|Y_T| > r) \leqslant& e^{-\eta r}\mathbb{E}(e^{\eta Y_T} + e^{-\eta Y_T})\\ \leqslant& 2\exp\left\{\frac{\eta^2\Sigma_{k,k}|\mathcal{T}_m|}{2T^2} - \eta r\right\}.\end{aligned}$$

Let $\eta = \frac{rT^2}{\Sigma_{k,k}|\mathcal{T}_m|}$, then

$$\mathbb{P}(|Y_{T-1}| > r) \leqslant 2\exp\left\{-\frac{r^2T^2}{2\Sigma_{k,k}|\mathcal{T}_m|}\right\}.$$

Now we take a union bound for all entries of $\frac{1}{T}\sum_{t\in\mathcal{T}_m}\epsilon_m^{t+1}\otimes X^t$.

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t\in\mathcal{T}_m}\epsilon_m^{t+1}\otimes X^t\right\|_{R*}>r\right)\leqslant&\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t\in\mathcal{T}_m}\epsilon_m^{t+1}\otimes X^t\right\|_{\infty}>\frac{r}{K}\right)\\ \leqslant&2MK^2\exp\left\{-\frac{r^2T^2}{2K^2\Sigma_{k,k}|\mathcal{T}_m|}\right\}.\end{aligned}$$

Plug in $r=CK\sqrt{\Sigma_{k,k}}\sqrt{\frac{\log M|\mathcal{T}_m|}{T^2}}\leqslant\frac{\lambda}{2}$, we obtain the final result. ■

Proof [proof for Lemma 7] Note that we can write $|\mathcal{T}_m|=\sum_{t=0}^{T-1}\mathbb{1}_{\{X_m^{t+1}\neq 0\}}$, where $\mathbb{1}_{\{X_m^{t+1}\neq 0\}}$ are i.i.d. Bernoulli r.v., with sub-Gaussian parameter bounded by $\frac{1}{2}$. Applying Hoeffding's inequality would give us

$$\mathbb{P}\left(|\mathcal{T}_m|>2q_mT\right)=\mathbb{P}\left(\sum_{t=0}^{T-1}\left(\mathbb{1}_{\{X_m^{t+1}\neq 0\}}-q_m\right)>q_mT\right)\leqslant\exp\{-2q_m^2T\}.$$

■

B.3 Proof of Lemmas in Section 7.3

Proof [proof of Lemma 8] The proof is the same as that of Lemma 5, except that we need to bound the infinity norm instead of $\|\cdot\|_R$. Using the same argument as in the proof of Lemma 5, we obtain

$$\mathbb{P}\left(\left\|\nabla L_m^{\mathrm{LN}}(A_m^{\mathrm{LN}})\right\|_{\infty}>\eta\right)\leqslant 2K^2M\exp\{-\frac{\eta^2T}{2\Sigma_{k,k}}\}.$$

Let $\eta=C\sqrt{\Sigma_{k,k}}\sqrt{\frac{\log M}{T}}$, we have the final result. ■

Proof [proof of Lemma 9] By the definition of L_m^{Bern},

$$\nabla L_m^{\mathrm{Bern}}(B_m^{\mathrm{Bern}})=-\frac{1}{T}\sum_{t=0}^{T-1}\varepsilon_m^{t+1}X^t,$$

where $\varepsilon_m^{t+1}=\mathbb{1}_{\{X_m^{t+1}\neq 0\}}-P(X_m^{t+1}\neq 0|X^t)$. Since $\mathbb{E}(\varepsilon_m^{t+1}X^t|\mathcal{F}_t)=0$ each element of $\varepsilon_m^{t+1}X^t$ is bounded by $[-1,1]$, the argument used in the proof of Lemma 3 can be directly applied here, and leads us to

$$\mathbb{P}\left(\left\|\nabla L_m^{\mathrm{Bern}}(B_m^{\mathrm{Bern}})\right\|_{\infty}>C\sqrt{\frac{\log M}{T}}\right)\leqslant\exp\{-c\log M\}.$$

■

Proof [proof of Lemma 10] Define $g(u) = \log(1 + e^u)$, and $u_m^{t*} = \langle B_m^{\text{Bern}}, X^t \rangle + \eta_m^{\text{Bern}}$, $\Delta u_m^t = \langle \Delta_m^B, X^t \rangle$, then we have

$$\begin{aligned} D_{L_m^{\text{Bern}}}(B_m^{\text{Bern}} + \Delta_m^B, B_m^{\text{Bern}}) =& \frac{1}{T} \sum_{t=0}^{T-1} \left[g(u_m^{t*} + \Delta u_m^t) - g(u_m^{t*}) - g'(u_m^{t*}) \Delta u_m^t \right] \\ =& \frac{1}{2T} \sum_{t=0}^{T-1} g''(\xi^t) (\Delta u_m^t)^2, \end{aligned}$$

where ξ^t lies between u_m^{t*} and $u_m^{t*} + \Delta u_m^t$. Since $\|B_m^{\text{Bern}}\|_{1,\infty} \leqslant R_{\max}^{\text{LN,Bern}}$, $\|\Delta_m^B\|_{1,\infty} \leqslant 1$, $u_m^{t*} \in [-R_{\max}^{\text{LN,Bern}} - \|\eta^{\text{Bern}}\|_\infty, R_{\max}^{\text{LN,Bern}} + \|\eta^{\text{Bern}}\|_\infty]$, $\Delta u_m^t \in [-1, 1]$, which implies that $|\xi^t| \leqslant R_{\max}^{\text{LN,Bern}} + \|\eta^{\text{Bern}}\|_\infty + 1 := C_2 + 1$. Therefore,

$$g''(\xi^t) = \frac{e^{-\xi^t}}{(1 + e^{-\xi^t})^2} \geqslant \exp\{C_2 + 1\}(1 + \exp\{C_2 + 1\})^{-2} = \sigma_C.$$

This implies

$$D_{L_m^{\text{Bern}}}(B_m^{\text{Bern}} + \Delta_m^B, B_m^{\text{Bern}}) \geqslant \frac{\sigma_B}{2T} \sum_{t=0}^{T-1} \langle \Delta_m^B, X^t \rangle^2.$$

■

Proof [Proof for Lemma 11] The proof is very similar to that of Lemma 6. For notational convenience, we view X^t and $U^{(2)}$ as MK-dimensional vector, $U^{(1)}$ as $(K-1) \times MK$ dimensional matrix. We can still write

$$\begin{aligned} & \frac{1}{2T} \sum_{t=0}^{T-1} \mathbb{1}_{\{X_m^{t+1} \neq 0\}} \|\langle U^{(1)}, X^t \rangle\|_2^2 + \frac{\sigma_C}{2T} \sum_{t=0}^{T-1} \langle U^{(2)}, X^t \rangle^2 \\ =& \frac{1}{2T} \sum_{t=0}^{T-1} \left\{ \sum_{k=1}^{K-1} U_k^{(1)\top} P_1^t U_k^{(1)} + \sigma_C U^{(2)\top} P_2^t U^{(2)} \right\} \\ & + \frac{1}{2T} \sum_{t=0}^{T-1} \left\{ \sum_{k=1}^{K-1} U_k^{(1)\top} E_1^t U_k^{(1)} + \sigma_C U^{(2)\top} E_2^t U^{(2)} \right\}, \end{aligned} \tag{72}$$

where

$$P_1^t = \mathbb{E}\left[X^t X^{t\top} \mathbb{1}_{\{X_m^{t+1} \neq 0\}} | \mathcal{F}_{t-1} \right], \quad P_2^t = \mathbb{E}\left[X^t X^{t\top} | \mathcal{F}_{t-1} \right]. \tag{73}$$

$$\begin{aligned} E_1^t =& X^t X^{t\top} \mathbb{1}_{\{X_m^{t+1} \neq 0\}} - \mathbb{E}\left[X^t X^{t\top} \mathbb{1}_{\{X_m^{t+1} \neq 0\}} | \mathcal{F}_{t-1} \right] \\ E_2^t =& X^t X^{t\top} - \mathbb{E}\left[X^t X^{t\top} | \mathcal{F}_{t-1} \right]. \end{aligned} \tag{74}$$

(1) Lower bounding the first term Since

$$P_1^t = \mathbb{E}\left(X^t X^{t\top} \mathbb{1}_{\{X_m^{t+1} \neq 0\}} | \mathcal{F}_{t-1} \right) = \mathbb{E}\left(X^t X^{t\top} \mathbb{P}(X_m^{t+1} \neq 0 | \mathcal{F}_t) | \mathcal{F}_{t-1} \right),$$

and

$$\begin{aligned}\mathbb{P}(X_m^{t+1} \neq 0|\mathcal{F}_t) &= \left(1+\exp\{-\langle B_m^{\text{Bern}}, X^t\rangle - \eta_m^{\text{Bern}}\}\right)^{-1} \\ &\geqslant \frac{1}{1+e^{R_{\max}^{\text{LN,Bern}}+\|\eta^{\text{Bern}}\|_\infty}},\end{aligned} \tag{75}$$

we have

$$\begin{aligned}U_k^{(1)\top} P_1^t U_k^{(1)} =& \mathbb{E}\left[\mathbb{P}(X_m^{t+1}\neq 0|\mathcal{F}_t)\left(U_k^{(1)\top}X^t\right)^2|\mathcal{F}_{t-1}\right] \\ \geqslant& \frac{1}{1+e^{R_{\max}^{\text{LN,Bern}}+\|\eta^{\text{Bern}}\|_\infty}}\mathbb{E}\left[\left(U_k^{(1)\top}X^t\right)^2|\mathcal{F}_{t-1}\right] \\ \geqslant& \frac{\lambda_{\min}(\mathbb{E}\left[X^tX^{t\top}|\mathcal{F}_{t-1}\right])}{1+e^{R_{\max}^{\text{LN,Bern}}+\|\eta^{\text{Bern}}\|_\infty}}\|U_k^{(1)}\|_2^2.\end{aligned}$$

Thus,

$$\begin{aligned}&\frac{1}{2T}\sum_{t=0}^{T-1}\left\{\sum_{k=1}^{K-1}U_k^{(1)\top}P_1^tU_k^{(1)}+\sigma_C U^{(2)\top}P_2^tU^{(2)}\right\} \\ \geqslant& \min_t \lambda_{\min}(\mathbb{E}\left[X^tX^{t\top}|\mathcal{F}_{t-1}\right])\left[\frac{\|U^{(1)}\|_F^2}{2(1+e^{R_{\max}^{\text{LN,Bern}}+\|\eta^{\text{Bern}}\|_\infty})}+\frac{\sigma_C\|U^{(2)}\|_F^2}{2}\right] \\ \geqslant& \frac{\sigma_C}{2}\min_t\lambda_{\min}(\mathbb{E}\left[X^tX^{t\top}|\mathcal{F}_{t-1}\right])\|U\|_F^2.\end{aligned}$$

To lower bound $\min_t \lambda_{\min}(\mathbb{E}\left[X^tX^{t\top}|\mathcal{F}_{t-1}\right])$, first note that

$$\begin{aligned}&\lambda_{\min}\left(\mathbb{E}(X^tX^{t\top}|\mathcal{F}_{t-1})\right) \\ \geqslant& \lambda_{\min}\left(\text{Cov}(X^t|\mathcal{F}_{t-1})\right) \\ =& \min_j \lambda_{\min}(\text{Cov}(X_j^t|\mathcal{F}_{t-1})).\end{aligned} \tag{76}$$

Following the same argument as in the proof for Lemma 6, one can show that

$$\begin{aligned}&\lambda_{\min}(\text{Cov}(X_j^t|\mathcal{F}_{t-1})) \\ \geqslant& \min\{\frac{q_j^t\beta_2}{4K+1}, \frac{q_j^t(1-q_j^t)}{4K}\} \\ \geqslant& \min\{\frac{\beta_2}{(4K+1)(1+e^{C_2})}, \frac{e^{C_2}}{4K(1+e^{C_2})^2}\},\end{aligned} \tag{77}$$

where

$$\beta_2 = \frac{2(e-1)^2}{e^6(2\pi)^{\frac{K-1}{2}}K^2|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{(K-1)(C_3+2)^2}{2\lambda_{\min}(\Sigma)}\right\} \tag{78}$$

where $C_2 = R_{\max}^{\text{LN,Bern}} + \|\eta^{\text{Bern}}\|_\infty$, $C_3 = R_{\max}^{\text{LN,Bern}} + \|\nu^{\text{LN}}\|_\infty$. Let

$$\gamma_2 = \frac{e^{C_2+1}}{2\left(1+e^{C_2+1}\right)^3}\min\left\{\frac{\beta_2}{4K+1}, \frac{e^{C_2}}{4K(1+e^{C_2})}\right\}, \tag{79}$$

then it is guaranteed that

$$\frac{1}{2T}\sum_{t=0}^{T-1}\left\{\sum_{k=1}^{K-1}U_k^{(1)\top}P_1^tU_k^{(1)}+\sigma_C U^{(2)\top}P_2^tU^{(2)}\right\}\geqslant\gamma_2\|U\|_F^2.$$

(2) Concentrating the second term uniformly Similarly from the proof for Lemma 4 and Lemma 6, one can show that for any $\varepsilon>0$, with probability at least $1-4K^2M^2\exp\{-c\varepsilon^2T\}$,

$$\left\|\frac{1}{2T}\sum_{t=0}^{T-1}E_1^t\right\|_\infty\leqslant\varepsilon,\quad\left\|\frac{1}{2T}\sum_{t=0}^{T-1}E_2^t\right\|_\infty\leqslant\varepsilon. \tag{80}$$

When the above holds,

$$\begin{aligned}&\left|\frac{1}{2T}\sum_{t=0}^{T-1}\left\{\sum_{k=1}^{K-1}U_k^{(1)\top}E_1^tU_k^{(1)}+\sigma_C U^{(2)\top}E_2^tU^{(2)}\right\}\right|\\ \leqslant&\epsilon\left(\sum_{k=1}^{K-1}\|U_k^{(1)}\|_1^2+\|U^{(2)}\|_1^2\right)\\ \leqslant&\epsilon\|U\|_1^2\\ \leqslant&16K^2\rho_m^{\mathrm{LN,Bern}}\varepsilon\|U\|_F^2.\end{aligned} \tag{81}$$

Let $\varepsilon=\frac{\gamma_2}{32K^2\rho_m^{\mathrm{LN,Bern}}}$, then

$$\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{1}_{\{X_m^{t+1}\neq0\}}\left\|\langle U^{(1)},X^t\rangle\right\|_2^2+\frac{\sigma_C}{2T}\sum_{t=0}^{T-1}\langle U^{(2)},X^t\rangle^2\geqslant\frac{\gamma_2}{2}\|U\|_F^2, \tag{82}$$

holds for any $U\in\mathcal{C}(S_m^{\mathrm{LN,Bern}},3)\cap\mathbb{B}_F(1)$, with probability at least

$$1-4K^2M^2\exp\{-\frac{c\gamma_2^2T}{K^4(\rho_m^{\mathrm{LN,Bern}})^2}\}\geqslant1-2\exp\{-c\log M\}, \tag{83}$$

as long as $T\geqslant\frac{CK^4(\rho_m^{\mathrm{LN,Bern}})^2}{\gamma_2^2}\log M$.

■

B.4 Explanation on the connection between (18) and (19)

The following lemma shows what (18) would be when we set the comparison baseline $x_{i,1}+\delta=C$ for all i. As a consequence, (18) equals (19) when the comparison baseline is set as $\frac{1}{n}\sum_{i'=1}^n x_{i',1}$.

Lemma 12 *If $\mathbb{P}(\delta|x_{i1})$ is set to ensure that $x_{i,1}+\delta=C$ for all i, then the variable importance (18) of X_1 becomes*

$$\frac{1}{n}\sum_{i=1}^n\left|f(x_{i,1},\ldots,x_{i,d})-f(C,\ldots,x_{i,d})\right|. \tag{84}$$

Proof [proof for Lemma 12] First note that setting $x_{i,1} + \delta = C$ for any C suggests that $\mathbb{P}(\delta|x_{i,1}) = \mathbb{1}_{\{\delta=C-x_{i,1}\}}$. Then some calculation shows that (18) satisfies

$$\begin{aligned}
&\int_\delta \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(\delta|x_{i,1}) \left(f(x_{i,1},\ldots,x_{i,d}) - f(x_{i,1}+\delta,\ldots,x_{i,d})\right)\right| d\delta \\
&= \int_\delta \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\delta=C-x_{i,1}\}} \left(f(x_{i,1},\ldots,x_{i,d}) - f(C,\ldots,x_{i,d})\right)\right| d\delta \\
&= \sum_{\exists 1\leqslant i\leqslant n, \delta=C-x_{i,1}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\delta=C-x_{i,1}\}} \left(f(x_{i,1},\ldots,x_{i,d}) - f(C,\ldots,x_{i,d})\right)\right| \\
&= \sum_{i=1}^{n} \left| \frac{1}{n} \left(f(x_{i,1},\ldots,x_{i,d}) - f(C,\ldots,x_{i,d})\right)\right|.
\end{aligned}$$

■

B.5 Proof for Lemma 13

By the generation scheme for $\widetilde{X}_m^{t+1}$ detailed in Secton 5.3.2, one can show that

$$\mathbb{E}\left[\log \frac{\widetilde{X}^{t+1}_{m,1:(K-1)}}{\widetilde{X}^{t+1}_{m,K}} | X_m^{t+1} = e_k^{(K)}\right] = \begin{cases} a e_k^{(K-1)}, & k < K, \\ -a 1_{K-1}, & k = K, \end{cases}$$

and

$$\mathbb{E}\left[\left\|\log \frac{\widetilde{X}^{t+1}_{m,1:(K-1)}}{\widetilde{X}^{t+1}_{m,K}}\right\|_2^2 | X_m^{t+1} = e_k^{(K)}\right] = \begin{cases} a + \sigma^2(K-1), & k < K, \\ a^2(K-1) + \sigma^2(K-1), & k = K. \end{cases}$$

Meanwhile, note that

$$\mathbb{P}(X_m^{t+1} = e_k^{(K)} | \mathcal{F}_t, X_m^{t+1} =\neq 0) = \frac{p_{m,k}^{t+1}}{\sum_{i=1}^K p_{m,i}^{t+1}},$$

where $p_{m,k}^{t+1} = \frac{e^{\langle A_{m,k}^{\text{mix}}, X^t\rangle + \nu_{m,k}^{\text{mix}}}}{1+\sum_{k'=1}^K e^{\langle A_{m,k'}^{\text{mix}}, X^t\rangle + \nu_{m,k'}^{\text{mix}}}}$. Hence we have

$$\mathbb{E}\left[\log \frac{\widetilde{X}^{t+1}_{m,1:(K-1)}}{\widetilde{X}^{t+1}_{m,K}} | \mathcal{F}_t, X_m^{t+1} \neq 0\right] = \left[\sum_{k=1}^{K-1} \frac{p_{m,k}^{t+1}}{\sum_{i=1}^K p_{m,i}^{t+1}} e_k^{(K-1)} - \frac{p_{m,K}^{t+1}}{\sum_{i=1}^K p_{m,i}^{t+1}} 1_{K-1}\right] a,$$

and

$$\begin{aligned}
&\mathbb{E}\left[\left\|\log\frac{\widetilde{X}^{t+1}_{m,1:(K-1)}}{\widetilde{X}^{t+}_{m,K}}\right\|_2^2 |\mathcal{F}_t, X^{t+1}_m \neq 0\right] \\
=&\sum_{k=1}^{K-1}\frac{p^{t+1}_{m,k}}{\sum_{i=1}^K p^{t+1}_{m,i}}(a^2+(K-1)\sigma^2)+\frac{p^{t+1}_{m,K}}{\sum_{i=1}^K p^{t+1}_{m,i}}(K-1)(a^2+\sigma^2) \\
=&\frac{\sum_{i=1}^{K-1} p^{t+1}_{m,i}+(K-1)p^{t+1}_{m,K}}{\sum_{i=1}^K p^{t+1}_{m,i}}a^2+(K-1)\sigma^2.
\end{aligned} \tag{85}$$

Appendix C. Supplementary Results

In this appendix, we provide some supplementary results to support some aforementioned arguments.

C.1 Change of Baseline Category

In this section, we illustrate that our model form would not change if using a different baseline category. Specifically, if we take a different category from the Kth category, say l, as the baseline and want to model the distribution of $\log\frac{Z^{t+1}_{m,k}}{Z^{t+1}_{m,l}}$, then the model (8) can be equivalently written as:

$$\begin{aligned}
&\log\frac{Z^{t+1}_{m,k}}{Z^{t+1}_{m,l}}=\left\langle\widetilde{A}^{\mathrm{LN}}_{m,k},X^t\right\rangle+\widetilde{\nu}^{\mathrm{LN}}_{m,k}+\widetilde{\epsilon}^{t+1}_{m,k},\quad 1\leqslant k\leqslant K, k\neq l,\\
&\{\widetilde{\epsilon}^{t+1}_m\}_{t,m}\overset{i.i.d.}{\sim}\mathcal{N}(0,\widetilde{\Sigma}),\ \widetilde{\Sigma}\in\mathbb{R}^{(K-1)\times(K-1)},
\end{aligned}$$

where $\widetilde{A}^{\mathrm{LN}}_{m,k}=A^{\mathrm{LN}}_{m,k}-A^{\mathrm{LN}}_{m,l}, k\notin\{l,K\}$, $\widetilde{A}^{\mathrm{LN}}_{m,K}=-A^{\mathrm{LN}}_{m,l}$; $\widetilde{\nu}^{\mathrm{LN}}_{m,k}=\nu^{\mathrm{LN}}_{m,k}-\nu^{\mathrm{LN}}_{m,l}, k\notin\{l,K\}$, $\widetilde{\nu}^{\mathrm{LN}}_{m,K}=-\nu^{\mathrm{LN}}_{m,l}$; $\widetilde{\epsilon}^{t+1}_m$ is transformed from ϵ^{t+1}_m through a linear full rank transformation, thus $\widetilde{\Sigma}$ is still of full rank (function of Σ).

C.2 Likelihood functions for the contaminated model

If $m\in\mathcal{N}_1$, the negative log-likelihood of $\widetilde{X}^{t+1}_m$ given X^t is

$$\begin{aligned}
&\widetilde{\ell}^{\mathrm{MN}}(A^{\mathrm{mix}}_m,\nu^{\mathrm{mix}}_m,a,\sigma^2;X^t,\widetilde{X}^{t+1}_m)\\
=&\begin{cases}\frac{\log(2\pi\sigma^2)(K-1)}{2}-\log\left[\sum_{k=1}^K p^{t+1}_{m,k}\exp\left\{-\frac{\|(E^{\mathrm{MN}})^{t+1}_{m,k}\|_2^2}{2\sigma^2}\right\}\right], & \widetilde{X}^{t+1}_m\neq 0,\\ -\log(1-\sum_{k=1}^K p^{t+1}_{m,k}), & \widetilde{X}^{t+1}_m=0,\end{cases}
\end{aligned} \tag{86}$$

where $p^{t+1}_{m,k}=\frac{e^{\langle A^{\mathrm{mix}}_{m,k},X^t\rangle+\nu^{\mathrm{mix}}_{m,k}}}{1+\sum_{i=1}^K e^{\langle A^{\mathrm{mix}}_{m,i},X^t\rangle+\nu^{\mathrm{mix}}_{m,i}}}$, and

$$(E^{\mathrm{MN}})^{t+1}_{m,k}=\begin{cases}\log\frac{\widetilde{X}^{t+1}_{m,1:(K-1)}}{\widetilde{X}^{t+1}_{m,K}}-ae^{(K-1)}_k, & 1\leqslant k\leqslant K-1\\ \log\frac{\widetilde{X}^{t+1}_{m,1:(K-1)}}{\widetilde{X}^{t+1}_{m,K}}+a1_{(K-1)\times 1}, & k=K\end{cases}$$

Otherwise, the negative log-likelihood of $\widetilde{X}_m^{t+1}$ given X^t is

$$\begin{aligned}&\widetilde{\ell}^{\mathrm{LN}}(A_m^{\mathrm{mix}}, \nu_m^{\mathrm{mix}}, \Sigma; X^t, \widetilde{X}_m^{t+1}) \\ &= \begin{cases} -\log(q_m^{t+1}) + \frac{(K-1)\log 2\pi + \log|\Sigma|}{2} + \frac{1}{2}(E^{\mathrm{LN}})_m^{t+1\top}\Sigma^{-1}(E^{\mathrm{LN}})_m^{t+1}, & \widetilde{X}_m^{t+1} \neq 0 \\ -\log(1-q_m^{t+1}), & \widetilde{X}_m^{t+1} = 0 \end{cases} \end{aligned} \tag{87}$$

where $q_m^{t+1} = \frac{e^{\langle A^{\mathrm{mix}}_{m,K,:,:}, X^t\rangle + \nu^{\mathrm{mix}}_{m,K}}}{1+e^{\langle A^{\mathrm{mix}}_{m,K,:,:}, X^t\rangle + \nu^{\mathrm{mix}}_{m,K}}}$ and

$$(E^{\mathrm{LN}})_m^{t+1} = \log \frac{\widetilde{X}^{t+1}_{m,1:(K-1)}}{\widetilde{X}_m^{t+1}} - \langle A^{\mathrm{mix}}_{m,1:(K-1),:,:}, X^t\rangle - \nu^{\mathrm{mix}}_{m,1:(K-1)}$$

Appendix D. Detailed Procedures in Numerical Experiments

In this appendix, we detail the numerical procedures for our simulations and real data experiments.

D.1 Cross-validation Procedures

We first illustrate the cross-validation procedure for the simulations in Section 5.2. As mentioned in Section 5.1, we use $\lambda = C_\lambda K\sqrt{\frac{\log M}{T}}$ across the experiments in Section 5.2, where the constant C_λ is selected for each model via cross-validation. Since the time series data is not exchangeable, we make the following modification to 5-fold cross-validation. For each model and candidate C_λ, we first generate data $\{X^t\}_{t=0}^T$ under one combination of sparsity, M, K, T, then the corresponding algorithm with $\lambda = C_\lambda K\sqrt{\frac{\log M}{T}}$ is run on 5 subsets of the data $\{X^t\}_{t=0}^T$, each including 80% of consecutive data points: $\{X^{t_i}, \ldots, X^{t_i+0.8T}\}, i = 1, \ldots, 5$, with $t_i = 0.05 \times (i-1)T$. The estimators obtained from each training subset are tested on the rest 20% of the data, and we choose the constant C_λ that results in the lowest average test loss (ℓ^{MN}, ℓ^{LN}, $\alpha\ell^{\mathrm{LN}} + (1-\alpha)\ell^{\mathrm{Bern}}$, previously defined for estimation) for the experiments in Section 5.2.

While for the experiments in Section 5.3.2 and Section 6, we run cross-validation for each model and each data set separately. In particular, for the logistic-normal model with event probability depending on the past, both α and λ need to be tuned, and hence the test loss $\alpha\ell^{\mathrm{LN}} + (1-\alpha)\ell^{\mathrm{Bern}}$ is no longer a reasonable criterion for selecting the tuning parameters. Therefore, for each pair of candidate tuning parameters (α, λ), we run 5-fold cross-validation similar to the procedure described above, except that we choose the tuning parameters resulting in the lowest prediction error on the test sets. For comparison fairness, we also use prediction error as the cross-validation criterion for other models across the experiments in Section 5.3.2 and Section 6. The detailed definition for prediction errors is included in Appendix D.6

D.2 Contaminated Mixture Model

In this contaminated model, we assume the non-zero multinomial vectors generated from the mixture model in Section 5.3.1 are contaminated to be random vectors following logistic-

normal distributions. Hence each event is observed with positive membership weights in all categories, similar to the real data sets we will discuss in Section 6.

Formally, if $m \in \mathcal{N}_1$ (set of multinomial nodes) and $X_m^{t+1} \neq 0_{K\times 1}$, the observation $\widetilde{X}_m^{t+1} \in \mathbb{R}^K$ would be a noisy version of X_m^{t+1}, following logistic-normal distribution. To define this distribution, we incorporate the following intuition: when $X_{m,k}^{t+1} = 1$ for some $1 \leqslant k \leqslant K$, $\widetilde{X}_{m,k}^{t+1}$ is likely larger than $\widetilde{X}_{m,k'}^{t+1}$ for $k' \neq k$. Hence we design the distribution of $\widetilde{X}_m^{t+1}$ so that $\mathbb{E}\left(\log \frac{\widetilde{X}_{m,k}^{t+1}}{\widetilde{X}_{m,k'}^{t+1}}\right) = a$ for some $a > 0$, if $k' \neq k$. Formally,

$$\widetilde{X}_m^{t+1} \sim \begin{cases} \mathrm{LN}(-a1_{(K-1)\times 1}, \sigma^{\mathrm{MN}}), & X_m^{t+1} = e_K^{(K)}, \\ \mathrm{LN}(ae_k^{(K-1)}, \sigma^{\mathrm{MN}}), & X_m^{t+1} = e_k^{(K)} \text{ for } k < K, \end{cases} \tag{88}$$

where $a, \sigma^{\mathrm{MN}} > 0$, $1_{(K-1)\times 1}$ is the all-ones vector in $\mathbb{R}^{K-1}$ and $e_k^{(K-1)}$ refers to the kth canonical vector in $\mathbb{R}^{K-1}$. Here, we say a vector $Y \in \mathbb{R}^K$ follows $\mathrm{LN}(\mu, \sigma^{\mathrm{MN}})$ for $\mu \in \mathbb{R}^{K-1}$ and $\sigma^{\mathrm{MN}} > 0$ if $\log(\frac{Y_{1:(K-1)}}{Y_K}) \sim \mathcal{N}(\mu, (\sigma^{\mathrm{MN}})^2 I_{(K-1)\times(K-1)})$.

D.3 Model Parameters for Synthetic Mixture Example

For the synthetic mixture network simulated in Section 4.3.2, we specify the parameters in the following. For simplicity, we assume the influence of events in one category is only imposed on future events in the same category, which is reasonable if we think of the categories as topics of news articles; also, events in the last category exerts and receives no influence, so that it can be viewed as a natural baseline. Therefore, for $m \in \mathcal{N}_1$, we set $A^{\mathrm{mix}}_{m,k,:,k'} = 0$ for $k \neq k'$ or $k' = K$; while for $m \in \mathcal{N}_2$, $A^{\mathrm{mix}}_{m,k,:,k} = A^{\mathrm{mix}}_{m,K,:,k}$, $A^{\mathrm{mix}}_{m,k,:,k'} = 0$ for $1 \leqslant k \leqslant K-1$ and $k' \neq k$.

For reproducibility, we present the non-zero parameter values here:

$$\begin{aligned} A^{\mathrm{mix}}_{1,(m-3)/3,m,(m-3)/3} = A^{\mathrm{mix}}_{1,K,m,(m-3)/3} =& \frac{1}{6}, \quad m = 6, 9, 12, 15, \\ A^{\mathrm{mix}}_{m,k,1,k} = A^{\mathrm{mix}}_{m,K,1,k} =& \frac{1}{3}, \quad 2 \leqslant m \leqslant 5, 1 \leqslant k \leqslant 4 \\ A^{\mathrm{mix}}_{m,(m-3)/3,1,(m-3)/3} =& \frac{2}{3}, \quad m = 6, 9, 12, 15, \\ A^{\mathrm{mix}}_{(m+1):(m+2),(m-3)/3,m,(m-3)/3} =& (\frac{7}{30}, \frac{7}{30})^\top, \quad m = 6, 9, 12, 15. \end{aligned} \tag{89}$$

The intercept terms ν^{mix} is defined to align with the preference of each node, so that nodes 1-5 are equally likely to have events in any of the first 4 categories, while each of nodes 6-8 (9-11, etc) is more likely to have events in one category than the others. More specifically, we set

$$\begin{aligned} \nu^{\mathrm{mix}}_{m,:} &= (1, 1, 1, 1, 0), \quad 1 \leqslant m \leqslant 5, \\ \nu^{\mathrm{mix}}_{m,:} &= \begin{cases} (0.5, 0, 0, 0, -0.5) & 6 \leqslant m \leqslant 8, \\ (0, 0.5, 0, 0, -0.5), & 9 \leqslant m \leqslant 11, \\ (0, 0, 0.5, 0, -0.5), & 12 \leqslant m \leqslant 14, \\ (0, 0, 0, 0.5, -0.5), & 15 \leqslant m \leqslant 17. \end{cases} \end{aligned} \tag{90}$$

The covariance matrix for the logistic-normal nodes is $\Sigma = I_{(K-1)\times(K-1)}$, and the noise level σ^{MN} for the contaminated multinomial vectors is set as 0.3 and a is set as 1. The comparison results can be influenced by σ^{MN}: when σ^{MN} gets too large, neither method works well and thus the performance gap between the two estimated networks on nodes 6-17 would be negligible.

D.4 Estimators in the Testing Procedure

1. Under the *multinomial* model, our estimation procedure for model parameters A_m^{mix}, ν_m^{mix}, a, and $(\sigma^{\mathrm{MN}})^2$ are summarized in Algorithm 3. First, we estimate A_m^{mix} and ν_m^{mix} using a variant of our proposed multinomial estimator (5) since it handles high-dimensionality by enforcing group sparsity. Define the rounded data $\widehat{X}_m^t$ as follows:

$$\widehat{X}_m^t = \begin{cases} 0_{K\times 1}, & \widetilde{X}_m^t = 0_{K\times 1}, \\ e_k, & \widetilde{X}_m^t \neq 0_{K\times 1}, k = \arg\max_i \widetilde{X}_{m,k}^t \end{cases} \tag{91}$$

where e_k is the kth canonical vector of $\mathbb{R}^K$. Since the loss function ℓ^{MN} in (5) is defined only for categorical response vectors, we consider regressing the rounded data $\widehat{X}_m^{t+1}$ upon the original observed data $\widetilde{X}^t$ (a surrogate for the unknown true data X^t), by solving

$$(\widehat{A}_m^{\mathrm{MN}}, \widehat{\nu}_m^{\mathrm{MN}}) = \underset{A\in\mathbb{R}^{K\times M\times K}, \nu\in\mathbb{R}^K}{\arg\min} \frac{1}{T}\sum_{t=0}^{T-1} \ell^{\mathrm{MN}}(A; \widetilde{X}^t, \widehat{X}_m^{t+1}, \nu) + \lambda^{\mathrm{MN}}\|A\|_R, \tag{92}$$

where ℓ^{MN} is defined in (6). Here we use $\ell^{\mathrm{MN}}(A; \widetilde{X}^t, \widehat{X}_m^{t+1}, \nu)$ instead of $\ell^{\mathrm{MN}}(A; \widehat{X}^t, \widehat{X}_m^{t+1}, \nu)$ since we don't assume the types of other nodes when testing node m, and hence we should not round the data associated with all nodes. Here $\widehat{A}_m^{\mathrm{MN}}$ and $\widehat{\nu}_m^{\mathrm{MN}}$ are estimators for A_m^{mix} and ν_m^{mix} under the multinomial model.

Meanwhile, noting that a and $(\sigma^{\mathrm{MN}})^2$ are Gaussian mixture parameters, we estimate them using the method of moments, which is efficient and commonly adopted for Gaussian mixture estimation (Anandkumar et al., 2012; Wu et al., 2020; Kalai et al., 2010). The following lemma provides two key moment equalities, based on which we will derive our estimators.

Lemma 13 *If node m follows the multinomial model,*

$$\begin{aligned} \mathbb{E}\left[\log \frac{\widetilde{X}_{m,1:(K-1)}^{t+1}}{\widetilde{X}_{m,K}^{t+1}} | \mathcal{F}_t, X_m^{t+1} \neq 0\right] &= a\beta_m^{t+1}, \\ \mathbb{E}\left[\left\|\log \frac{\widetilde{X}_{m,1:(K-1)}^{t+1}}{\widetilde{X}_{m,K}^{t+1}}\right\|_2^2 | \mathcal{F}_t, X_m^{t+1} \neq 0\right] &= a^2\kappa_m^{t+1} + (K-1)(\sigma^{\mathrm{MN}})^2. \end{aligned} \tag{93}$$

where $\beta_m^{t+1} = \sum_{k=1}^{K-1} \frac{p_{m,k}^{t+1}}{\sum_{i=1}^K p_{m,i}^{t+1}} e_k^{(K-1)} - \frac{p_{m,K}^{t+1}}{\sum_{i=1}^K p_{m,i}^{t+1}} 1_{(K-1)\times 1} \in \mathbb{R}^{K-1}$ with $e_k^{(K-1)}$ being the kth canonical vector, $1_{(K-1)\times 1}$ being the all-ones vector in $\mathbb{R}^{K-1}$ and $\kappa_m^{t+1} = \frac{\sum_{i=1}^{K-1} p_{m,i}^{t+1} + (K-1)p_{m,K}^{t+1}}{\sum_{i=1}^K p_{m,i}^{t+1}} \in \mathbb{R}$. Here the probability $p_{m,k}^{t+1} = \frac{e^{\langle A_{m,k}^{\mathrm{mix}}, X^t\rangle + \nu_{m,k}^{\mathrm{mix}}}}{1+\sum_{k'=1}^K e^{\langle A_{m,k'}^{\mathrm{mix}}, X^t\rangle + \nu_{m,k'}^{\mathrm{mix}}}}$.

Lemma 13 can be proved by some direct calculations, and the detailed proof is included in Appendix B.5. Given Lemma 13, we want to find estimators for a and $(\sigma^{\mathrm{MN}})^2$ by solving (93) with expectations substituted by sample average. However, since β_m^{t+1} and κ_m^{t+1} depend on the unknown $p_{m,k}^{t+1}, 1 \leqslant k \leqslant K$, we substitute it with its estimated surrogate:

$$\widehat{p}_{m,k}^{t+1} = \frac{e^{\langle \widehat{A}_{m,k}^{\mathrm{MN}}, \widetilde{X}^t \rangle + \widehat{\nu}_{m,k}^{\mathrm{MN}}}}{1 + \sum_{k'=1}^{K} e^{\langle \widehat{A}_{m,k'}^{\mathrm{MN}}, \widetilde{X}^t \rangle + \widehat{\nu}_{m,k'}^{\mathrm{MN}}}}, \tag{94}$$

and then obtain the corresponding $\widehat{\beta}_m^{t+1}$ and $\widehat{\kappa}_m^{t+1}$. Formally, we consider the following estimators for a and $(\sigma^{\mathrm{MN}})^2$:

$$\widehat{a}_m = \operatorname*{arg\,min}_{\theta \in \mathbb{R}} \sum_{t \in \mathcal{T}_m} \left\| \log \frac{\widetilde{X}_{m,1:(K-1)}^{t+1}}{\widetilde{X}_{m,K}^{t+1}} - \theta \widehat{\beta}_m^{t+1} \right\|_2^2, \tag{95}$$

and

$$(\widehat{\sigma}_m^{\mathrm{MN}})^2 = \max \left\{ \frac{1}{(K-1)|\mathcal{T}_m|} \sum_{t \in \mathcal{T}_m} \left(\left\| \log \frac{\widetilde{X}_{m,1:(K-1)}^{t+1}}{\widetilde{X}_{m,K}^{t+1}} \right\|_2^2 - \widehat{a}_m^2 \widehat{\kappa}_m^{t+1} \right), 0 \right\}, \tag{96}$$

where $\mathcal{T}_m = \{t : \widetilde{X}_m^{t+1} \neq 0\}$.

2. Under the *logistic-normal* model, the estimation procedure is summarized in Algorithm 4. Similarly to the multinomial case, we also estimate A_m^{mix} and ν_m^{mix} by a variant of (13). We consider regressing the observed data $\widetilde{X}_m^{t+1}$ upon $\widetilde{X}^t$ by solving

$$\begin{aligned} &(\widehat{A}_m^{\mathrm{LN}}, \widehat{B}_m^{\mathrm{Bern}}, \widehat{\nu}_m^{\mathrm{LN}}, \widehat{\eta}_m^{\mathrm{Bern}}) \\ &= \operatorname*{arg\,min}_{A \in \mathbb{R}^{(K-1)\times M \times K}, B \in \mathbb{R}^{M \times K}, \nu \in \mathbb{R}^{K-1}, \eta \in \mathbb{R}} \frac{\alpha}{T} \sum_{t=0}^{T-1} \ell^{\mathrm{LN}}(A; \widetilde{X}^t, \widetilde{X}_m^{t+1}, \nu) \\ &\quad + \frac{1-\alpha}{T} \sum_{t=0}^{T-1} \ell^{\mathrm{Bern}}(B; \widetilde{X}^t, \widetilde{X}_m^{t+1}, \eta) + \lambda^{\mathrm{LN}} R_\alpha(A, B), \end{aligned} \tag{97}$$

where ℓ^{LN} and ℓ^{Bern} are defined in (10) and (12). Here $\widehat{A}_m^{\mathrm{LN}}$, $\widehat{B}_m^{\mathrm{Bern}}$, $\widehat{\nu}_m^{\mathrm{LN}}$ and $\widehat{\eta}_m^{\mathrm{Bern}}$ are estimators for $A_{m,1:(K-1),:,:}^{\mathrm{mix}}$, $A_{m,K,:,:}^{\mathrm{mix}}$, $\nu_{m,1:(K-1)}^{\mathrm{mix}}$, and $\nu_{m,K}^{\mathrm{mix}}$ under the logistic-normal model.

While for estimating Σ, note that its MLE is

$$\frac{1}{|\mathcal{T}_m|} \sum_{t \in \mathcal{T}_m} \left(\log \frac{\widetilde{X}_{m,1:(K-1)}^{t+1}}{\widetilde{X}_{m,K}^{t+1}} - \mu_m^{t+1} \right) \left(\log \frac{\widetilde{X}_{m,1:(K-1)}^{t+1}}{\widetilde{X}_{m,K}^{t+1}} - \mu_m^{t+1} \right)^\top,$$

where $\mu_m^{t+1} = \langle A_{m,1:(K-1)}^{\mathrm{mix}}, X^t \rangle + \nu_{m,1:(K-1)}^{\mathrm{mix}}$ is unknown. Given the estimates $\widehat{A}_m^{\mathrm{LN}}, \widehat{\nu}_m^{\mathrm{LN}}$, we can substitute μ_m^{t+1} by $\widehat{\mu}_m^{t+1} = \langle \widehat{A}_m^{\mathrm{LN}}, \widetilde{X}^t \rangle + \widehat{\nu}_m^{\mathrm{LN}}$, then estimate Σ by

$$\widehat{\Sigma}_m = \frac{1}{|\mathcal{T}_m|} \sum_{t \in \mathcal{T}_m} \left(\log \frac{\widetilde{X}_{m,1:(K-1)}^{t+1}}{\widetilde{X}_{m,K}^{t+1}} - \widehat{\mu}_m^{t+1} \right) \left(\log \frac{\widetilde{X}_{m,1:(K-1)}^{t+1}}{\widetilde{X}_{m,K}^{t+1}} - \widehat{\mu}_m^{t+1} \right)^\top. \tag{98}$$

Algorithm 3 Estimation of Multinomial Parameters for the Test

Input: Contaminated data $\{\widetilde{X}^t\}_{t=0}^T$, rounded data for node m: $\{\widehat{X}_m^t\}_{t=1}^T$, tuning parameter $\lambda^{\mathrm{MN}} > 0$

1: Estimate network and offset parameters under the multinomial model:

$$(\widehat{A}_m^{\mathrm{MN}}, \widehat{\nu}_m^{\mathrm{MN}}) = \underset{A\in\mathbb{R}^{K\times M\times K}, \nu\in\mathbb{R}^K}{\arg\min} \frac{1}{T}\sum_{t=0}^{T-1} \ell^{\mathrm{MN}}(A; \widetilde{X}^t, \widehat{X}_m^{t+1}, \nu) + \lambda^{\mathrm{MN}}\|A\|_R$$

2: **for** $t = 0, \ldots, T-1$ **do**

3: **for** $k = 1, \ldots, K$ **do**

4:

$$\widehat{p}_{m,k}^{t+1} = \frac{e^{\langle \widehat{A}_{m,k}^{\mathrm{MN}}, \widetilde{X}^t\rangle + \widehat{\nu}_{m,k}^{\mathrm{MN}}}}{1 + \sum_{k'=1}^K e^{\langle \widehat{A}_{m,k'}^{\mathrm{MN}}, \widetilde{X}^t\rangle + \widehat{\nu}_{m,k'}^{\mathrm{MN}}}}$$

5: **end for**

6:

$$\widehat{\beta}_m^{t+1} = \sum_{k=1}^{K-1} \frac{\widehat{p}_{m,k}^{t+1}}{\sum_{i=1}^K \widehat{p}_{m,i}^{t+1}} e_k^{(K-1)} - \frac{\widehat{p}_{m,K}^{t+1}}{\sum_{i=1}^K \widehat{p}_{m,i}^{t+1}} 1_{(K-1)\times 1}$$

$$\widehat{\kappa}_m^{t+1} = \frac{\sum_{i=1}^{K-1} \widehat{p}_{m,i}^{t+1} + (K-1)\widehat{p}_{m,K}^{t+1}}{\sum_{i=1}^K \widehat{p}_{m,i}^{t+1}}$$

7: **end for**

8: Estimate $\widehat{a}_m$ and $(\widehat{\sigma}_m^{\mathrm{MN}})^2$ by the method of moments:

$$\widehat{a}_m = \underset{\theta\in\mathbb{R}}{\arg\min} \sum_{t\in\mathcal{T}_m} \left\| \log \frac{\widetilde{X}_{m,1:(K-1)}^{t+1}}{\widetilde{X}_{m,K}^{t+1}} - \theta\widehat{\beta}_m^{t+1} \right\|_2^2$$

$$(\widehat{\sigma}_m^{\mathrm{MN}})^2 = \max\left\{ \frac{1}{(K-1)|\mathcal{T}_m|} \sum_{t\in\mathcal{T}_m} \left(\left\| \log \frac{\widetilde{X}_{m,1:(K-1)}^{t+1}}{\widetilde{X}_{m,K}^{t+1}} \right\|_2^2 - \widehat{a}_m^2 \widehat{\kappa}_m^{t+1} \right), 0 \right\}$$

where $\mathcal{T}_m = \{t : \widetilde{X}_m^{t+1} \neq 0\}$

Output: $\widehat{A}_m^{\mathrm{MN}}$, $\widehat{\nu}_m^{\mathrm{MN}}$, $\widehat{a}_m$, $(\widehat{\sigma}_m^{\mathrm{MN}})^2$

Algorithm 4 Estimation of Logistic-normal Parameters for the Test

Input: Contaminated data $\{\widetilde{X}^t\}_{t=0}^T$, node index m, tuning parameters $\lambda^{\mathrm{LN}} > 0$, $\alpha \in (0,1)$

1: Estimate network and offset parameters under the logistic-normal model:

$$\begin{aligned}&(\widehat{A}_m^{\mathrm{LN}}, \widehat{B}_m^{\mathrm{Bern}}, \widehat{\nu}_m^{\mathrm{LN}}, \widehat{\eta}_m^{\mathrm{Bern}})\\ =&\underset{A\in\mathbb{R}^{(K-1)\times M\times K}, B\in\mathbb{R}^{M\times K}, \nu\in\mathbb{R}^{K-1}, \eta\in\mathbb{R}}{\arg\min} \frac{\alpha}{T}\sum_{t=0}^{T-1}\ell^{\mathrm{LN}}(A;\widetilde{X}^t,\widetilde{X}_m^{t+1},\nu)\\ &+\frac{1-\alpha}{T}\sum_{t=0}^{T-1}\ell^{\mathrm{Bern}}(B;\widetilde{X}^t,\widetilde{X}_m^{t+1},\eta)+\lambda^{\mathrm{LN}}R_\alpha(A,B)\end{aligned}$$

2: **for** $t = 0, \ldots, T-1$ **do**

3: $\quad \widehat{\mu}_m^{t+1} = \langle \widehat{A}_m^{\mathrm{LN}}, \widetilde{X}^t\rangle + \widehat{\nu}_m^{\mathrm{LN}}$

4: **end for**

5: Estimate the covariance matrix:

$$\widehat{\Sigma}_m = \frac{1}{|\mathcal{T}_m|}\sum_{t\in\mathcal{T}_m}\left(\log\frac{\widetilde{X}_{m,1:(K-1)}^{t+1}}{\widetilde{X}_{m,K}^{t+1}} - \widehat{\mu}_m^{t+1}\right)\left(\log\frac{\widetilde{X}_{m,1:(K-1)}^{t+1}}{\widetilde{X}_{m,K}^{t+1}} - \widehat{\mu}_m^{t+1}\right)^\top$$

where $\mathcal{T}_m = \{t : \widetilde{X}_m^{t+1} \neq 0\}$

Output: $\widehat{A}_m^{\mathrm{LN}}$, $\widehat{B}_m^{\mathrm{Bern}}$, $\widehat{\nu}_m^{\mathrm{LN}}$, $\widehat{\eta}_m^{\mathrm{Bern}}$, $\widehat{\Sigma}_m$.

D.5 Data Preprocessing in Section 6

Some details about how we obtain the membership vectors for each post in both examples are listed below.

1. **Identifying political tendencies of tweets:**
   We first use the tweets from the first half of the time period (55,859 tweets from Jan 1, 2016 to June 6, 2016) to train a neural network for categorizing tweets into two political tendencies (left- and right-leaning). The input feature vector of the neural network is an embedded vector of each tweet obtained by the standard pre-trained model BERT (Devlin et al., 2018; Xiao, 2018) (uncased, 24-layer); and the partisanship of the user is used as the label (tweets sent by Democrats are all labeled as "left-leaning"). The partisanship may not represent the true label, but due to the lack of human annotated labels, we believe the partisanship serves as a reasonable approximation, especially since politicians usually sent tweets with clear ideology.

   The neural network is composed of three fully connected layers (two hidden layers of 128 nodes). RELU and softmax are the activation functions of the first two layers and the last layer respectively, and the cross entropy loss is used for training.

   Since the tweets from the first half of the time period are already used for training the neural network, we don't include them in the input data set to our methods to avoid over-fitting. The trained neural network model outputs a 2-dimensional vector on the simplex for each of the 27,600 tweets from June 7, 2016 to November 11, 2016, the second half of the time period. The neural network predicts the tweet to be left-leaning if the vector has larger value in its first coordinate, and right-leaning otherwise. Therefore, we first consider this vector as the mixed membership vector of the tweet, where the first coordinate is the membership in the left-leaning category (score_L) and the second being that in the right-leaning category (score_R). Carefully examining the mixed membership vectors, we find that the some scores are very close to 0 and 1 which may lead to computational issues when calculating the log-ratios. Hence we perform the following transformation:

$$\widetilde{\text{score}}_L = \frac{1}{2} \times \frac{1}{200} + \text{score}_\text{L} \times \frac{199}{200},$$
$$\widetilde{\text{score}}_R = \frac{1}{2} \times \frac{1}{200} + \text{score}_\text{R} \times \frac{199}{200},$$

   so that both $\widetilde{\text{score}}_L$ and $\widetilde{\text{score}}_R$ lie in $[0.0025, 0.9975]$. Finally, we use $(\widetilde{\text{score}}_L, \widetilde{\text{score}}_R)$ as the mixed membership vector for each tweet.

2. **Topic membership vectors for memes in the MemeTracker example:**
   We first filter for the English media sources with high frequencies (more than 1500 posts included in the data set each month), which leads to a total of 5,684,791 posts from 101 media sources. For each post, we combine its recorded phrases/quotes together as the approximate content of the post. We then run topic modeling (Latent Dirichlet Allocation proposed in Blei et al. (2003)) on these posts, where the number of topics is set as 5 ($K = 5$), using the module `gensim` (RadimŘehůřek and Sojka, 2010) in python. For each topic, we present the top 10 keywords generated from topic

modeling in the second column of Table 14, and we choose the topic names (the first column of Table 14) based on these keywords. For each post item, topic modeling

Topics	Keywords
Sports	time, people, lot, thing, game, way, team, work, player, year
International Affairs	people, country, government, time, united_states, state, law, issue, case, work
Lifestyle	life, people, man, family, love, water, woman, world, story, music
Finance	market, company, business, economy, customer, time, service, industry, bank, product
Health	child, patient, food, health, people, drug, hospital, information, research, risk

Table 14: Keywords for the 5 topics generated from topic modeling.

also outputs a corresponding K-dimensional weight vector on the simplex, indicating its memberships in the K topics.

Using 1-hour discretizations, we obtain a sample of size $T+1 = 5807$, and if we want to learn the network among all of the 101 media sources, there would be $255,025$ $(101^2 \times 5^2)$ network parameters to estimate for both methods. Therefore for simplicity and interpretability, we select a subset of the 101 media sources and learn the network among them. To preserve a variety of topics covered in the posts, for each of the first 4 topics, we select the top 15 media sources that have the highest average topic weights in it.[10] This leads us to a list of 58 media sources ($M = 58$), due to some overlaps among top media sources in different topics, so the total number of network parameters to estimate is reduced to $84,100$.

After we get the mixed membership vector of each post for each example, the time series data $\{X^t \in \mathbb{R}^{M\times K}\}_{t=0}^{T}$ is obtained as follows. For the political tweets data, the time period is discretized into $T+1 = 1000$ intervals of length approximately 3.7 hrs, while for the MemeTracker data, we use 1-hour discretization and end up with $T+1 = 5807$. After discretizing the time period into $T+1$ time intervals, the input data $\{X_m^t \in \mathbb{R}^K, 1 \leqslant m \leqslant M, 0 \leqslant t \leqslant T\}$ (M is the number of nodes) is then constructed as follows: for each time window t, if there is no event associated with node m, let $X_m^t = 0$; otherwise, (1) for the logistic-normal approach, let $X_m^t \in \mathbb{R}^K$ be the mixed membership vector (over the categories) of the event; (2) for the multinomial approach, let $X_m^t \in \mathbb{R}^K$ be the rounded mixed membership vector, that is, $X_m^t = e_k$ if the membership vector takes the largest value in the kth category, where e_k is the kth canonical vector in $\mathbb{R}^K$. If there are multiple events associated with one node in the same time window, we average the mixed membership vector and use that as X_m^t for the logistic-normal approach, and the rounded version of that average vector as X_m^t for the multinomial approach.

10. No selected media has high weights in the topic "health", so that we have a good choice for the baseline topic, as explained shortly.

D.6 Definition of Prediction Errors in Section 6

The prediction errors for the two methods are evaluated on hold-out sets (latter 30% of each data set), after fitting the models using training sets (first 70% of each data set). The prediction error on a hold-out set is defined as follows:

- For a fitted multinomial model, given $X^{t-1} \in \mathbb{R}^{M\times K}$ (rounded data at time $t-1$ in the hold-out set), a one-step-ahead predicted probability vector $\hat{p}_m^t \in \mathbb{R}^{K+1}$ (the last dimension is the probability of no event) is output for each user m, according to (4). The prediction for X_m^t is defined as

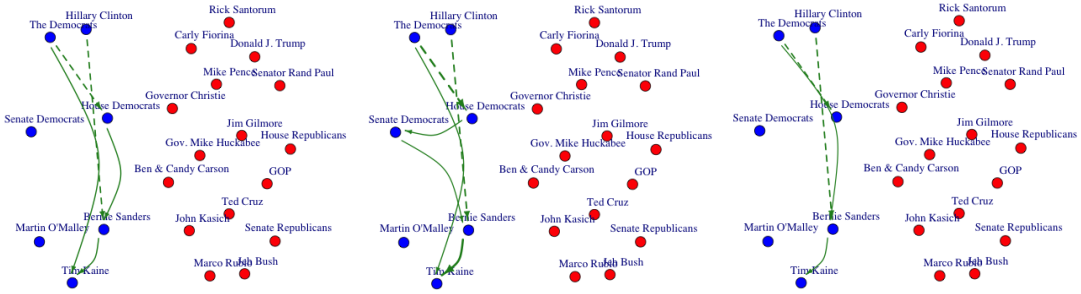
$$\widehat{X}_m^t = \begin{cases} 0, & \arg\max_{k'} \hat{p}_{m,k'}^t = K+1, \\ e_k, & \arg\max_{k'} \hat{p}_{m,k'}^t = k \leqslant K, \end{cases}$$

  and the prediction error is calculated by $\frac{1}{TM}\sum_{t,m} \|X_m^t - \widehat{X}_m^t\|_2^2$, which is the proportion of wrong predictions for all nodes and time units in the hold-out set. Here X_m^t is the observed rounded data.

- For a fitted logistic-normal model, given $X^{t-1} \in \mathbb{R}^{M\times K}$ (original, unrounded) in the hold-out set, a probability $\hat{q}_m^t$ is output for an event associated with node m to occur at time t, specified by (11); the expected log-ratios $\{\log \frac{\hat{Z}_{m,k}^t}{\hat{Z}_{m,K}^t}\}_{k=1}^{K-1}$ of the mixed membership vector $Z_m^t \in \triangle^{K-1}$ can also be specified by (8) with $\epsilon_{m,k}^t = 0$. Then we can transform the expected log-ratios back to $\widehat{Z}_m^t$ as the prediction for true mixed membership vector. Hence we define the prediction for X_m^t as $\widehat{X}_m^t = \hat{q}_m^t \widehat{Z}_m^t$, and prediction error as $\sum_{t,m} \frac{\|X_m^t - \widehat{X}_m^t\|_2^2}{TM}$ (mean squared error).

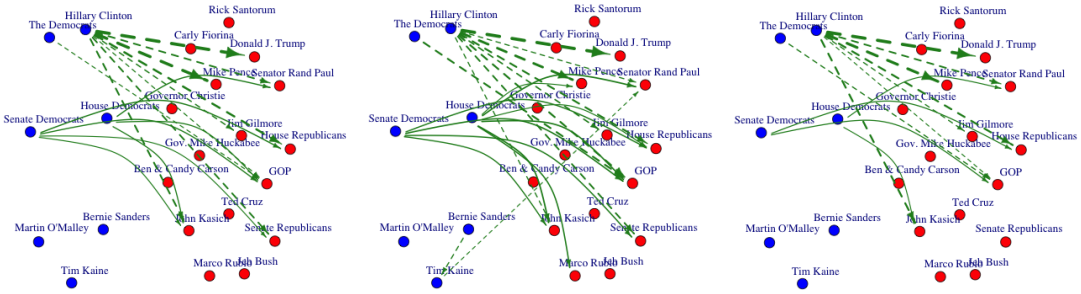
D.7 The Rest Three Sub-networks for the Political Tweets Example

Here we present the estimated variable importance edges that are left-leaning→left-leaning, left-leaning→right-leaning and right-leaning→left-leaning. As mentioned earlier in Figure 15, the largest absolute entry of each of the three variable importance parameters ($\widehat{V}^{\mathrm{MN}}$, $\widehat{V}^{\mathrm{LN}}$ and $\widehat{V}^{\mathrm{mix}}$) is normalized to one and each visualized edge width is proportional to the normalized absolute value of its corresponding parameter. For clarity, only the edges with absolute parameters larger than 0.3 are shown for each network, and blue nodes are Democrats, red nodes are Republicans. Solid edges are positive influences (stimulatory) while dashed edges are negative influences (inhibitory).



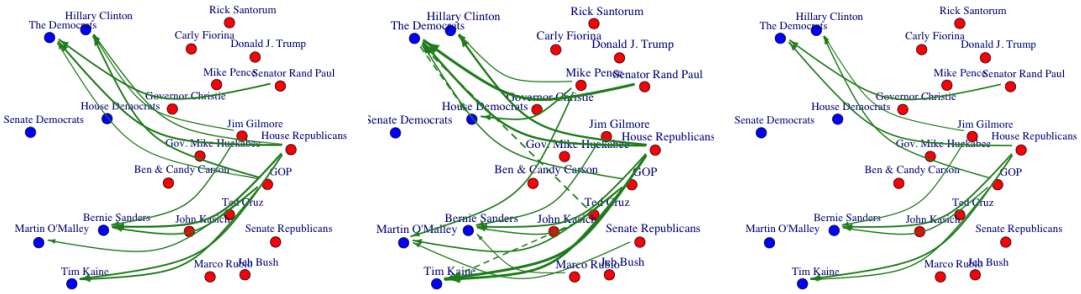
(a) Multinomial (b) Logistic-normal (c) Mixture

Figure 15: Estimated variable importance networks by the three approaches for the tweets example, including edges from left-leaning tweets to left-leaning tweets.



(a) Multinomial (b) Logistic-normal (c) Mixture

Figure 16: Estimated variable importance networks by the three approaches for the tweets example, including edges from left-leaning tweets to right-leaning tweets.

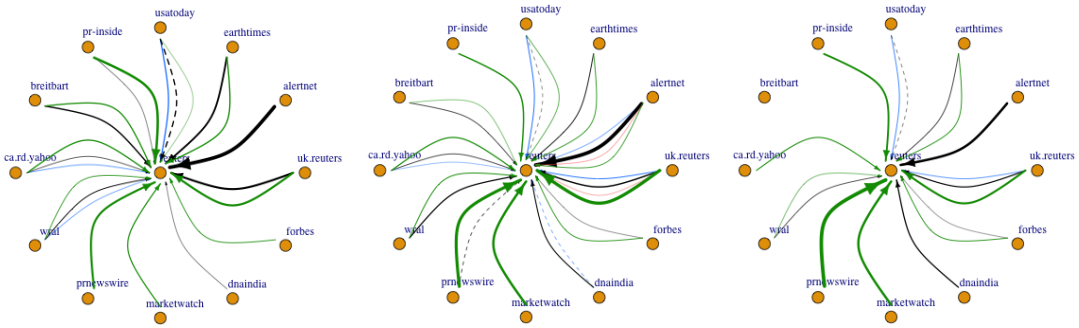


(a) Multinomial (b) Logistic-normal (c) Mixture

Figure 17: Estimated variable importance networks by the three approaches for the tweets example, including edges from right-leaning tweets to left-leaning tweets.

D.8 Neighborhood Visualization for the MemeTracker Example

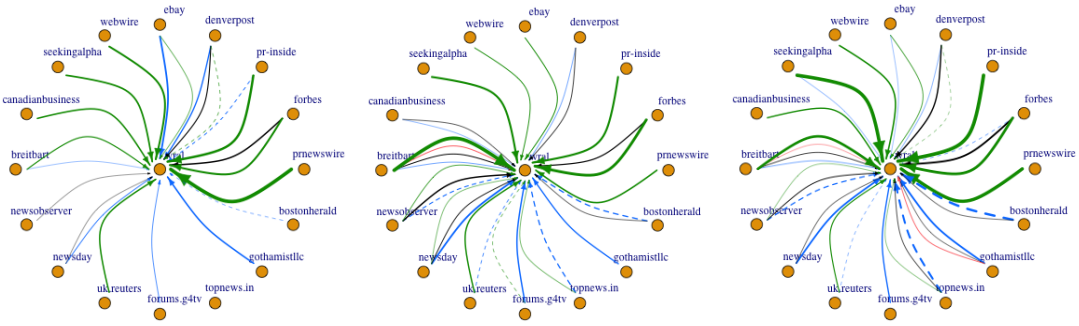
We present the neighborhood estimates around each media source, instead of the whole network estimates among 58 media sources. For each central media source, we consider the influences it receives that are between the same topic, in the estimated variable importance networks. The top 8 neighbors in any of the three estimated variable importance sub-networks are included in our visualization, and the maximum absolute entry of each variable importance parameter of the sub-network is normalized to 1. Edges with absolute parameter value higher 0.2 are visualized[11], and each visualized edge width is proportional to its absolute parameter value. Solid edges are positive influences (stimulatory) while dashed edges are negative influences (inhibitory). Figure 18 and Figure 19 present the estimated variable importance sub-networks around *reuters.com* and *wral.com*. Four edges in these two sub-networks are summarized in Table 8, since there are supporting evidence for their estimation by some approaches than the other approach.



(a) Multinomial (b) Logistic-normal (c) Mixed

Figure 18: Estimated variable importance sub-networks (influences received by *reuters.com*) by the multinomial, logistic-normal and mixture approaches.

11. We use a smaller threshold here than the political tweets example (0.2 instead of 0.3), since we present the sub-networks around each node, instead of the whole network among all nodes. Smaller threshold can still preserve clarity of presentation.



(a) Multinomial (b) Logistic-normal (c) Mixed

Figure 19: Estimated variable importance sub-networks (influences received by *wral.com*) by the multinomial, logistic-normal and mixture approaches.

D.9 Supporting Evidence for the Estimated Edges in the MemeTracker Example

Extracting evidence from the phrase cluster data: For evaluating these edges, the external knowledge used in the tweets example is not applicable here since many media sources post on multiple topics. Instead, we present supporting evidence based on a cascade data set: the *"Phrase cluster data"* from Aug 2008 to Jan 2009 in the MemeTracker data set, which is also used in Yu et al. (2017a) for studying influences among media sources. In contrast to the "Raw phrases data" used for our network estimation, where original phrases are recorded for each post, the "Phrase cluster data" collects phrase clusters consisting of variants of the same phrases, and for each phrase cluster, there are records of which media source posts variants in it and when.

For convenience, in the following, we say that a media source posts a phrase cluster if it posts a phrase in that cluster. For each phrase cluster and any pair of influencer (m) and receiver (m') media sources, if the first time m' posts the phrase cluster is within an hour after m posts it, we refer to it as an *influence-involved phrase cluster* from m to m'. Here we set the time limit as one hour since 1-hour discretization is used in the estimation task. In order to demonstrate the topics of these phrase clusters, we combine all the influence-involved phrase clusters from m to m' into one "document" and generate a word cloud and topics weights for the document. To assign topic weights, we apply the previously trained topic model (mentioned at the beginning of Section 6.2) to the document, quantifying how much the document falls in each topic. The word clouds and topic weights for the discussed four edges are presented in Figures 20-23. Details about the generation of word clouds and topic weights are included in Appendix D.10.

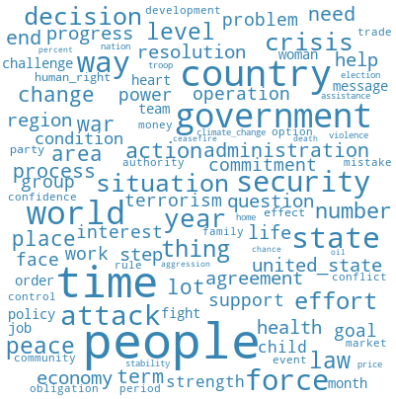
The *number* and *topics* of the influence-involved phrase clusters should reflect stimulatory influences between media sources *qualitatively*, and thus can facilitate our comparison among the proposed three approaches given that there is no ground truth. However, we don't expect this procedure based on phrase cluster data to provide us with an accurate network estimate due to the following reasons: this procedure only looks at the marginal dependence of each receiver media source on an influencer media source, instead of its con-

ditional dependence on that influencer media source given all media sources; meanwhile, the group sparsity structure is not leveraged to handle the high-dimensional problem.

Evidence for *alertnet*, *uk.reuters→reuters*: We look at the number of influence-involved phrase clusters from each media source to *reuters* and calculate the percentage of all phrase clusters that media source ever posts that are influence-involved. The results for *alertnet* and *uk.reuters* are presented in Table 15, suggesting that these two media sources are likely to be influential to *reuters*. We further investigate the topics of influence through the word clouds and topic weights in Figures 20 and 21, which support two methods than the other for each edge.

Media Source	Total Number of Posted Phrase Clusters	Influence-involved Phrase Clusters	Percent	Rank
alertnet	2552	758	29.70%	1
uk.reuters	4998	875	17.51%	2

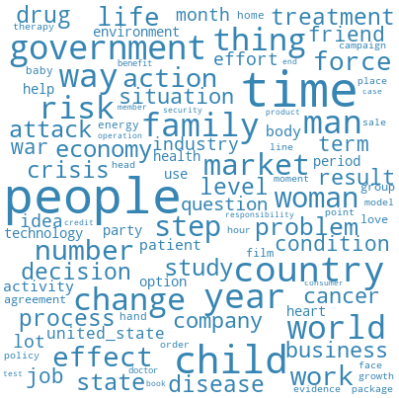
Table 15: Number of phrase clusters that are posted at least once by *alertnet* and *uk.reuters*, and the number of influence-involved phrase clusters from them to *reuters* (column 3). The third column includes the percentages of the phrase clusters these two media sources post that are influence-involved, while the last column lists the ranks of them among all the media sources in terms of these percentages.



Sports	International Affairs	Lifestyle	Finance	Health
0.1380	0.5647	0.0458	0.1370	0.1146

Figure 20: **(*alernet→ reuters*)** The word cloud and topic weights of the document consisting of influence-involved phrase clusters from *alertnet* to *reuters*. We can see from the word cloud that these phrase clusters are mostly focused on "International affairs" (*e.g.,* words like "country", "world", "people", "government"). Meanwhile, the topic weight of "International affairs" is much larger than the other topics. Since the logistic-normal method estimates edges in all of the first four topics, *the word cloud, together with the topic weights, provides evidence for the edges estimated by the multinomial method and the mixture method, other than the logistic-normal method.*

Evidence for *canadianbusiness*, *breitbart→wral*: Similarly, we also present the number, word cloud, and topic weights of influence-involved phrase clusters from *canadian-*



Sports	International Affairs	Lifestyle	Finance	Health
0.1335	0.3327	0.1143	0.1915	0.2280

Figure 21: **(*uk.reuters*→ *reuters*)** The word cloud and topic weights of the document consisting of influence-involved phrase clusters from *uk.reuters* to *reuters*. We can see from the word cloud that these phrase clusters cover multiple topics including "International affairs" (*e.g.,* words like "people", "country", "world", "government"), "Lifestyle" (*e.g.,* "child", "family"), "Finance" (*e.g.,* "market", "economy") and "Health" (*e.g.,* "drug", "treatment"). Although we can see few words clearly referring to sports, both "Sports" and "Lifestyle" have non-negligible topic weights (compared to the highest topic weight 0.3327) in the table above. We believe this is because that the topic "Sports" is not exclusively about sports although we name it so, as indicated by the key words in Table 14. Specifically, its top 10 keywords include "time", "lot", "thing", which do not clearly refer to any topic. *The word cloud, together with the topic weights, provides evidence for edges in all five topics; hence the edges estimated by the logistic-normal method and the mixture method may be more reasonable than the multinomial method.*

business and *breitbart* to *wral*, see Table 16, Figure 22, and Figure 23. Table 16 suggests that both *canadianbusiness* and *breitbart* may be influential to *wral*, while Figure 22 and Figure 23 support two methods than the other for each edge.

Media Sources	Total Number of Posted Phrase Clusters	Influence-involved Phrase Clusters	Percent	Rank
canadianbusiness	2339	252	10.77%	3
breitbart	19279	1408	7.30%	4

Table 16: Number of phrase clusters that are posted at least once by *canadianbusiness* and *breitbart* (column 2); and the number of influence-involved phrase clusters from them to *wral* (column 3). The third column includes the percentages of the phrase clusters the two media sources post that are influence-involved, while the last column lists the ranks of them among all the media sources in terms of these percentages.

Sports	International Affairs	Lifestyle	Finance	Health
0.1062	0.2153	0.0144	0.6255	0.0387

Figure 22: (***canadianbusiness→wral***) The word cloud and topic weights of the document consisting of influence-involved phrase clusters from *canadianbusiness* to *wral*. We can see from the word cloud that these phrase clusters are mostly focused on "Finance" (*e.g.,* words like "market", "economy", "company", "dollar"), and also with some coverage on "International affairs" (*e.g.,* "government", "country"). Meanwhile, the topic weight of "'Finance" is much larger than the other topics. *The word cloud, together with the topic weights, provides evidence for the edges in "International Affairs" and "Finance", and "Finance" is likely the dominant topic. Hence the edges estimated by the multinomial method and the mixture method may be better than the logistic-normal method.*

Sports	International Affairs	Lifestyle	Finance	Health
0.2412	0.3392	0.1030	0.2744	0.0423

Figure 23: (***breitbart→ wral***) The word cloud and topic weights of the document consisting of influence-involved phrase clusters from *breitbart* to *wral*. We can see from the word cloud that these phrase clusters cover multiple topics including "International affairs" (*e.g.,* words like "people", "country", "government", "taxpayer"), "Finance" (*e.g.,* "market", "business"), "Lifestyle" (*e.g.,* "life"', "family", "work"). Meanwhile, all the first four topics have non-negligible topic weights in the table above. *The word cloud, together with the topic weights, provides evidence for the edges in all the first four topics; hence the edges estimated by the logistic-normal method and the mixture method may be better than the multinomial method.*

D.10 Generation of Word Clouds and Topic Weights in the MemeTracker Example

To understand the topics of the influence, we also combine those influence-involved phrase clusters together as one document. We remove the stop words and only preserve nouns in this document, just as what we did for the pre-processing of the topic modeling. Then we generate a word cloud for this pre-processed document using the module `wordcloud`[12] in Python, which assigns larger fonts to words with higher frequencies. The top 100 words with highest frequencies are included in each word cloud. We also apply the previously trained topic model (mentioned in the beginning of Section 6.2) on the pre-processed document to obtain its topic weights, as a quantitative characterization of the influence strength in each topic.

References

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, pages 33–1, 2012.

J Atchison and Sheng M Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

John Bacon-Shone. A short history of compositional data analysis. *Compositional Data Analysis*, pages 1–11, 2011.

Sumanta Basu, George Michailidis, et al. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Teresa M Brunsdon and TMF Smith. The time series analysis of compositional data. *Journal of Official Statistics*, 14(3):237, 1998.

Wesley S Chan. Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics*, 70(2):223–260, 2003.

12. `https://github.com/amueller/word_cloud`

Daryl J Daley and David Vere-Jones. An introduction to the theory of point processes, volume 1: Elementary theory and methods. *Verlag New York Berlin Heidelberg: Springer*, 2003.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Şeyda Ertekin, Cynthia Rudin, Tyler H McCormick, et al. Reactive point processes: A new approach to predicting power failures in underground electrical systems. *The Annals of Applied Statistics*, 9(1):122–144, 2015.

Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. Fake news mitigation via point process based intervention. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1097–1106. JMLR. org, 2017.

Albert Feller, Matthias Kuhnert, Timm O Sprenger, and Isabell M Welpe. Divided they tweet: The network structure of political microbloggers and discussion topics. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

Raphael Féraud and Fabrice Clérot. A methodology to explain neural network classification. *Neural networks*, 15(2):237–246, 2002.

Alyson K Fletcher and Sundeep Rangan. Scalable inference for neuronal connectivity from calcium imaging. In *Advances in Neural Information Processing Systems*, pages 2843–2851, 2014.

Konstantinos Fokianos, Benjamin Kedem, et al. Regression theory for categorical time series. *Statistical science*, 18(3):357–376, 2003.

Ayalvadi Ganesh, Laurent Massoulié, and Don Towsley. The effect of network topology on the spread of epidemics. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 2, pages 1455–1466. IEEE, 2005.

Felipe Gerhard, Moritz Deger, and Wilson Truccolo. On the stability and dynamics of stochastic spiking neuron models: Nonlinear hawkes process and point process glms. *PLoS computational biology*, 13(2):e1005390, 2017.

Ulrike Grömping. Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4):308–319, 2009.

Eric C Hall and Rebecca M Willett. Online learning of neural network structure from spike trains. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 930–933. IEEE, 2015.

Eric C Hall, Garvesh Raskutti, and Rebecca Willett. Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*, 2016.

Fang Han, Huanran Lu, and Han Liu. A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 2015.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562, 2010.

Petra Kynčlová, Peter Filzmoser, and Karel Hron. Modeling compositional time series with vector autoregressive models. *Journal of Forecasting*, 34(4):303–314, 2015.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

Xiaodong Li, Haoran Xie, Li Chen, Jianping Wang, and Xiaotie Deng. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23, 2014.

Scott Linderman, Ryan P Adams, and Jonathan W Pillow. Bayesian latent structure discovery from multi-neuron recordings. In *Advances in neural information processing systems*, pages 2002–2010, 2016.

Justin Littman, Laura Wrubel, and Daniel Kerchner. 2016 United States Presidential Election Tweet Ids, 2016. URL `https://doi.org/10.7910/DVN/PDI7IN`.

Karim Lounici, Massimiliano Pontil, Alexandre B Tsybakov, and Sara Van De Geer. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, 2009.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

Benjamin Mark, Garvesh Raskutti, and Rebecca Willett. Network estimation from point process data. *IEEE Transactions on Information Theory*, 65(5):2953–2975, 2018.

Paul Mihailidis and Samantha Viotty. Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. *American behavioral scientist*, 61(4):441–454, 2017.

Terence C Mills. Forecasting compositional time series. *Quality & Quantity*, 44(4):673–690, 2010.

Guillaume Obozinski, Ben Taskar, and Michael Jordan. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2(2.2), 2006.

Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995, 2008.

Kriti Puniyani, Jacob Eisenstein, Shay Cohen, and Eric P Xing. Social links from latent topics in microblogs. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 19–20. Association for Computational Linguistics, 2010.

RadimŘehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259, 2010.

Nalini Ravishanker, Dipak K Dey, and Malini Iyengar. Compositional time series analysis of mortality proportions. *Communications in Statistics-Theory and Methods*, 30(11): 2281–2291, 2001.

Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. Influence and passivity in social media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 18–33. Springer, 2011.

Cynthia Rudin, David Waltz, Roger N Anderson, Albert Boulanger, Ansaf Salleb-Aouissi, Maggie Chow, Haimonti Dutta, Philip N Gross, Bert Huang, Steve Ierome, et al. Machine learning for the new york city power grid. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 34(2):328–345, 2011.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1): 22–36, 2017.

Alexey Stomakhin, Martin B Short, and Andrea L Bertozzi. Reconstruction of missing data in social networks based on temporal patterns of interactions. *Inverse Problems*, 27(11): 115013, 2011.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9 (1):307, 2008.

Alex Tank, Emily B Fox, and Ali Shojaie. Granger causality networks for categorical time series. *arXiv preprint arXiv:1706.02781*, 2017.

Timothy Peter Williams, Yew Sun Ding, Daniel Hobbs, Daniel Schmidt, and Doug Asherman. System and method determining online significance of content items and topics using social media, August 1 2013. US Patent App. 13/563,667.

Max A Woodbury, Jonathan Clive, and Arthur Garson Jr. Mathematical typology: a grade of membership technique for obtaining disease definition. *Computers and biomedical research*, 11(3):277–298, 1978.

Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

Yihong Wu, Pengkun Yang, et al. Optimal estimation of gaussian mixtures via denoised method of moments. *Annals of Statistics*, 48(4):1981–2007, 2020.

Han Xiao. bert-as-service. `https://github.com/hanxiao/bert-as-service`, 2018.

Lu-Xing Yang, Xiaofan Yang, Jiming Liu, Qingyi Zhu, and Chenquan Gan. Epidemics of computer viruses: A complex-network approach. *Applied Mathematics and Computation*, 219(16):8705–8717, 2013.

Yingxiang Yang, Jalal Etesami, Niao He, and Negar Kiyavash. Online learning for multivariate hawkes processes. In *Advances in Neural Information Processing Systems*, pages 4937–4946, 2017.

Ming Yu, Varun Gupta, and Mladen Kolar. Estimation of a low-rank topic-based model for information cascades. *arXiv preprint arXiv:1709.01919*, 2017a.

Ming Yu, Varun Gupta, and Mladen Kolar. Estimation of a low-rank topic-based model for information cascades. *arXiv preprint arXiv:1709.01919*, 2017b.

Ming Yu, Varun Gupta, and Mladen Kolar. Learning influence-receptivity network structure with guarantee. *arXiv preprint arXiv:1806.05730*, 2018.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*, 2017.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649, 2013.