Poisson-Randomised DirBN: Large mutation is needed in Dirichlet belief networks

Xuhui Fan [1] **Bin Li** [2] **Yaqiong Li** [3] **Scott A. Sisson** [1]

Abstract

The Dirichlet Belief Network (DirBN) was recently proposed as a promising deep generative model to learn interpretable deep latent distributions for objects. However, its current representation capability is limited since its latent distributions across different layers is prone to form similar patterns and can thus hardly use multi-layer structure to form flexible distributions. In this work, we propose Poisson-randomised Dirichlet Belief Networks (Pois-DirBN), which allows large mutations for the latent distributions across layers to enlarge the representation capability. Based on our key idea of inserting Poisson random variables in the layer-wise connection, Pois-DirBN first introduces a component-wise propagation mechanism to enable latent distributions to have large variations across different layers. Then, we develop a layer-wise Gibbs sampling algorithm to infer the latent distributions, leading to a larger number of effective layers compared to DirBN. In addition, we integrate out latent distributions and form a multi-stochastic deep integer network, which provides an alternative view on Pois-DirBN. We apply Pois-DirBN to relational modelling and validate its effectiveness through improved link prediction performance and more interpretable latent distribution visualisations. The code can be downloaded at `https://github.com/xuhuifan/Pois_DirBN`.

[1]UNSW Data Science Hub, and School of Mathematics and Statistics, University of New South Wales [2]Shanghai Key Laboratory of IIP, School of Computer Science, Fudan University [3]Australian Artificial Intelligence Institute, University of Technology, Sydney. Correspondence to: Scott A. Sisson <scott.sisson@unsw.edu.au>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

1. Introduction

The Dirichlet Belief Network (DirBN) (Zhao et al., 2018) is a promising approach for learning easily interpretable and meaningful deep latent distributions for objects. In comparison to existing deep generative models (e.g., Variational AutoEncoders (VAE) (Kingma & Welling, 2013) and Generative Adversarial Networks (GAN) (Goodfellow et al., 2014)), which usually use parameterised functions to build deep architectures, the DirBN uses a multi-stochastic-layers (Zhou et al., 2016) structure to generate deep latent distributions. Each node is affiliated with a latent distribution at each layer, with these latent distributions again generated through Dirichlet distributions, forming Dirichlet-Dirichlet connections. The DirBN is claimed to be interpretable as these latent distributions can be regarded as categorical distributions over latent components. The DirBN has been used in two scenarios: topic modelling (Zhao et al., 2018), in which the latent distributions correspond to topic-word distributions; and relational modelling (Fan et al., 2019; Li et al., 2020), in which nodes' membership distributions over communities are modelled through these latent distributions.

However, the current modelling capability of the DirBN, especially in the relational modelling setting, is limited since the latent distributions share similar patterns across different layers. Figure 1 shows an example visualisation of the latent distributions at layer 2 and layer 3 of the SDREM (Fan et al., 2019) (a DirBN based relational model) on the PPI dataset (Zitnik & Leskove, 2017). *There are no obvious variations in the patterns of the latent distributions from layer* 2 *to layer* 3. This phenomenon may be due to the fact that the DirBN uses scaled latent distributions to constitute the concentration parameters of the Dirichlet distribution for the next layer. Accordingly, the generated latent distributions at the next layer will have similar expectations to those at the current layer, and so it may be difficult to generate new latent distribution patterns. As it is constructed using similar patterns, the application of DirBN into relational modelling therefore appears unable to make full use of deep structure to form flexible distributions and enhanced modelling capability.

In this work, we propose a Poisson-randomised Dirichlet

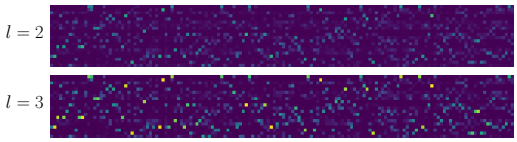

Figure 1. Visualisations of latent distributions on 150 nodes from layers 2 and 3 in a SDREM (Fan et al., 2019), with $K = 20$. Columns in each panel represent nodes' latent distributions. The entries are larger when their colours move from blue to yellow.

Belief Network (Pois-DirBN) to address the aforementioned issue and thereby enhance the modeling capabilities of the DirBN. We first insert auxiliary Poisson counting variables into the DirBN's layer-wise connections. This replaces the previous Dirichlet-Dirichlet connection with a newly formulated Dirichlet-Poisson-Dirichlet connection. These Poisson counting variables can be organized to approximate the latent distributions, which helps to maintain the existing benefits of the DirBN. However, more importantly, they introduce a *new component-wise counting variable* into the layer-wise connection. In contrast to the previously fixed components' ratios in the concentration parameters of the Dirichlet distribution, due to the scaled effect on the latent distributions, our proposed component-wise propagation can flexibly adjust the components' ratios in the concentration parameters of the Dirichlet distribution. The modelling capability is thereby enhanced by allowing for larger mutations in these latent distributions across different layers.

For model inference, we introduce auxiliary variables to augment the Poisson likelihood and construct an efficient layer-wise Gibbs sampling algorithm. This algorithm can circumvent the previously complicated strategy of upward propagating latent counts and then downward sampling random variables (Zhou et al., 2016; Zhao et al., 2018; Fan et al., 2019; Li et al., 2020). As the observations can be propagated to deeper layers based on the proposed sampling algorithm, we can thus set arbitrary number of layers in the deep architecture of the DirBN.

By integrating out the latent distributions, the Pois-DirBN can alternatively be reformulated as a multi-stochastic deep integer network, in which each layer is composed of counting variables. This integer network may benefit from small storage and low-memory requirements. We demonstrate the modelling advantages of the Pois-DirBN for relational modelling, by improved link prediction performance over the state-of-the-art models, and more interpretable visualisations on the latent distributions.

2. DirBN preliminaries

The modelling strategy of the DirBN is to construct a *multi-stochastic* layered architecture to represent interpretable hierarchical latent distributions for objects. In general, the DirBN constructs L layers of K-length latent distributions $\boldsymbol{\pi}_i = \{\boldsymbol{\pi}_i^{(l)}\}_{l=1}^{L}$ for each object i. The generative process of propagating the latent distributions $\{\boldsymbol{\pi}_j^{(l-1)}\}_j$ at layer $(l-1)$ to $\boldsymbol{\pi}_i^{(l)}$ at layer l can be constructed as

$$\omega_{ji}^{(l)} \sim \text{Gam}(c_j, d), \quad \boldsymbol{\pi}_i^{(l)} \sim \text{Dir}(\sum_j \omega_{ji}^{(l)} \boldsymbol{\pi}_j^{(l-1)}) \qquad (1)$$

where $\text{Gam}(c, d)$ is the Gamma distribution with mean c/d and variance c/d^2, $\omega_{ji}^{(l)}$ represents the information propagation coefficient from node j at layer $l-1$ to node i at layer l, and c_j, d are hyper-parameters.

Particular observation types can then be generated by specifying suitable probabilistic distributions, with nodes' latent distributions at layer L constituting the distribution's parameters. For example, a Bernoulli distribution can be used to generate the binary relation value between nodes i and j, with the probability being a combination of their latent distributions at layer L. The DirBN has found promising application in topic modelling and relational modelling.

However, it seems inefficient that the concentration parameter vector of the Dirichlet distribution in Eq. (1) only uses the previous layers' scaled latent distributions. Under this setting, the components' weight ratio may be highly similar to the expected components' weight ratio for the latent distributions at the previous layer. Because the latent distributions at the current and previous layers will then share similar patterns, it might be difficult for the Dir-BN to be sufficiently flexible to model complex real-world data structures, and efficiently utilize deep structure to enhance modelling capability.

Another disadvantage of the DirBN is that little information can be propagated to higher layers via the sampling algorithm. As observed in (Zhou et al., 2016; Fan et al., 2019), the DirBN uses the CRT distribution to upward propagate counts to higher layers. As a result, the amount of information arriving at the higher layer scales $\mathcal{O}(\log)$ to the information at the previous layer (Zhou et al., 2016). For a deep architecture with L layers at the output layer, the expected amount of information will be reduced with a function of $\mathcal{O}(\log^L(\cdot))$, which is quite small when L is large. As higher layers can only receive little information, the effectiveness of the DirBN is *limited to a few layers*.

3. Poisson-Randomised Dirichlet Belief Networks

We introduce the Poisson-randomised Dirichlet Belief Network (Pois-DirBN) to generate L-layered hierarchical latent distributions for N nodes, in which we use $\boldsymbol{\pi}_i^{(l)}$ to denote node i's latent distribution at layer l. The Pois-DirBN is able to promote large mutations in latent distributions across different layers, and build the deep architecture with no re-

strictions on the number of layers.

3.1. Dirichlet-Poisson-Dirichlet connection

In the Pois-DirBN, we insert Poisson counting variables into the DirBN's Dirichlet-Dirichlet layer-wise connection (i.e. Eq. (1)) to form a new Dirichlet-Poisson-Dirichlet layer-wise connection. These Poisson variables can be categorized into two groups: $C_{iki'}^{(l+\frac{1}{2})}$, which denotes the counting information of node i at layer l propagated to node i' at layer $(l+1)$ within component k; and $D_{ikk'}^{(l+\frac{1}{2})}$, which represents the information of node i in component k at layer l propagated to its component k' at layer $l+1$, where the superscript of $(l+\frac{1}{2})$ denotes that these counting variables are located between layer l and layer $(l+1)$. The new layer-wise connections can be expressed as:

$$
\begin{aligned}
C_{iki'}^{(l+\frac{1}{2})} &\sim \text{Poisson}(M_c \pi_{ik}^{(l)} \omega_{ii'}^{(l)}),\\
D_{ikk'}^{(l+\frac{1}{2})} &\sim \text{Poisson}(M_d \pi_{ik}^{(l)} \phi_{kk'}^{(l)}),\\
\boldsymbol{\pi}_i^{(l+1)} &\sim \text{Dirichlet}(\alpha_k^{(\pi)} + \sum_{i'} \boldsymbol{C}_{i'\cdot i}^{(l+\frac{1}{2})} + \sum_{k'} \boldsymbol{D}_{ik'\cdot}^{(l+\frac{1}{2})}), \quad (2)
\end{aligned}
$$

where M_c, M_d are hyper-parameters to control the scale of the variables $\boldsymbol{C}^{(l+\frac{1}{2})}, \boldsymbol{D}^{(l+\frac{1}{2})}$; $\omega_{ii'}^{(l)}$ regulates the information propagation of the latent distribution of node i at layer l to node i' at layer $(l+1)$; $\phi_{kk'}^{(l)}$ regulates the propagation of component k at layer l to component k' at layer $(l+1)$ within the same node; and $\alpha_k^{(\pi)}$ is the offset parameter to circumvent the case of 0 incoming counts. The ith row $\boldsymbol{\omega}_i^{(l)}$ in $\boldsymbol{\omega}^{(l)}$ and the kth row $\boldsymbol{\phi}_k^{(l)}$ in $\boldsymbol{\phi}^{(l)}$ are restricted to have an L_1 norm for scale identifiability and inferential convenience. We let $\boldsymbol{\omega}_i^{(l)}, \boldsymbol{\phi}_k^{(l)}$ follow a Dirichlet distribution: $\boldsymbol{\omega}_i^{(l)} \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(\omega)}), \boldsymbol{\phi}_k^{(l)} \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(\phi)})$, where $\boldsymbol{\alpha}^{(\omega)}, \boldsymbol{\alpha}^{(\phi)} \in [\mathbb{R}^+]^{1\times K}$ are the concentration parameters.

Due to the Poisson-Multinomial equivalence (Dunson & Herring, 2005) and the fact that $\sum_k \pi_{ik} = 1, \sum_{i'} \omega_{ii'}^{(l)} = 1, \sum_{k'} \phi_{kk'}^{(l)} = 1$, we can observe the following for the summary statistics of the counting variables $\boldsymbol{C}, \boldsymbol{D}$:

$$
\begin{aligned}
&M_{i,C}^{(l+\frac{1}{2})}, M_{k,C}^{(l+\frac{1}{2})} \sim \text{Poisson}(M_c),\\
&M_{i,D}^{(l+\frac{1}{2})}, M_{k,D}^{(l+\frac{1}{2})} \sim \text{Poisson}(M_d),\\
&(\sum_{i'} C_{i1i'}^{(l+\frac{1}{2})}, \ldots, \sum_{i'} C_{iKi'}^{(l+\frac{1}{2})}) \sim \text{Multinomial}\ (M_{i,C}^{(l+\frac{1}{2})}; \boldsymbol{\pi}_i^{(l)}),\\
&(\sum_{k} C_{ik1}^{(l+\frac{1}{2})}, \ldots, \sum_{k} C_{ikN}^{(l+\frac{1}{2})}) \sim \text{Multinomial}\ (M_{k,C}^{(l+\frac{1}{2})}; \boldsymbol{\omega}_i^{(l)}),\\
&(\sum_{k'} D_{i1k'}^{(l+\frac{1}{2})}, \ldots, \sum_{k'} D_{iKk'}^{(l+\frac{1}{2})}) \sim \text{Multinomial}(M_{i,D}^{(l+\frac{1}{2})}; \boldsymbol{\pi}_i^{(l)}),\\
&(\sum_{k} D_{ik1}^{(l+\frac{1}{2})}, \ldots, \sum_{k} D_{ikK}^{(l+\frac{1}{2})}) \sim \text{Multinomial}(M_{k,D}^{(l+\frac{1}{2})}; \boldsymbol{\phi}_k^{(l)}).
\end{aligned}
$$

That is, the component-wise vector of sums over the counts from node i to other nodes follows a Multinomial distribution with $\boldsymbol{\pi}_i$ as event probabilities, the node-wise vector of sums over the counts on all the components from node i to other nodes follows a Multinomial distribution with $\boldsymbol{\omega}_i^{(l)}$ as event probabilities, and the component-wise vector of sums over the counts from component k to all other components follows a Multinomial distribution with $\boldsymbol{\phi}_k^{(l)}$ as event probabilities. It is easy to see that $\frac{1}{M_{i,i,C}^{(l)}}\mathbb{E}[\sum_{i'} \boldsymbol{C}_{i\cdot i'}^{(l+\frac{1}{2})}] = \frac{1}{M_{i,k,D}^{(l)}}\mathbb{E}[\sum_{k'} \boldsymbol{D}_{i\cdot k'}^{(l+\frac{1}{2})}] = \boldsymbol{\pi}_i$, $\frac{1}{M_{i,k,C}^{(l)}}\mathbb{E}[\sum_{k} \boldsymbol{C}_{ik\cdot}^{(l+\frac{1}{2})}] = \boldsymbol{\omega}_i^{(l)}$, $\frac{1}{M_{i,k,D}^{(l)}}\mathbb{E}[\sum_{k} \boldsymbol{D}_{ik\cdot}^{(l+\frac{1}{2})}] = \boldsymbol{\phi}_k^{(l)}$, which shows that these summary statistics of $\boldsymbol{C}, \boldsymbol{D}$ can be seen as a finite proxy of the distributions $\boldsymbol{\pi}, \boldsymbol{\omega}^{(l)}, \boldsymbol{\phi}^{(l)}$. It is obvious that larger values of $M_{i,i,C}^{(l)}, M_{i,i,D}^{(l)}, M_{i,k,C}^{(l)}, M_{i,k,D}^{(l)}$ would result in closer approximations.

After obtaining these counting variables, we use the Dirichlet distribution (Eq. (2)) to generate the new latent distribution at layer $(l+1)$, with the concentration parameter summarizing all the counts $\boldsymbol{C}$ that propagated to node i and all the counts $\boldsymbol{D}$ that propagated to the components within the node (see Figure 2 for an graphical illustration). Due to the structure of these distributions, we refer to such layer-wise connections as Dirichlet-Poisson-Dirichlet connections.

It is easy to see that the generation of the Poisson counting variables $\boldsymbol{C}, \boldsymbol{D}$ makes the prior and posterior distribution of $\boldsymbol{\pi}_i^{(l)}$ conjugate (i.e. Dirichlet distributions). We do not need to use the usual upward propagation of latent counts to form the "pseudo" counts for $\boldsymbol{\pi}_i^{(l)}$, and can thereby circumvent the issue of "insufficient counts" for higher layers. Also, as the Poisson counting variables form finite approximations to the $\boldsymbol{\pi}_i^{(l)}$, the primary structure of the DirBN is maintained, and so the Pois-DirBN retains its original properties.

Component-wise propagation through $\boldsymbol{D}$: Note that using the counting variable $\boldsymbol{C}$ produces similar benefits as the Dirichlet-Dirichlet connections in the DirBN (Zhao et al., 2018; Fan et al., 2019; Li et al., 2020), which incorporate the sum of scaled latent distributions into the concentration parameters of the next layer's latent distributions. While this node-wise propagation helps the sharing of latent distributions between interacting nodes, the component-wise propagation, which may promote cross-component weights transfer within the same node, has never been explored before.

The counting variable $\boldsymbol{D}$ is used to model the cross-component information propagation. The expectation of each $\boldsymbol{D}$'s entry is $\mathbb{E}[D_{ikk'}^{(l+\frac{1}{2})}] = M_d \pi_{ik}^{(l)} \phi_{kk'}^{(l)}$, representing that node i propagates its component k to its component k' through the coefficient $\phi_{kk'}^{(l)}$. As with the above, the usage of $\boldsymbol{C}, \boldsymbol{D}$ produces similar effects with the generation $\boldsymbol{\pi}_i^{(l+1)} \sim \text{Dirichlet}(\sum_{i'} \omega_{i'i}^{(l-1)} \boldsymbol{\pi}_{i'}^{(l-1)} + \boldsymbol{\pi}_i^{(l)} \boldsymbol{\phi}^{(l)})$ as for the DirBN. However, the usage of $\boldsymbol{\phi}^{(l)}$ changes the component

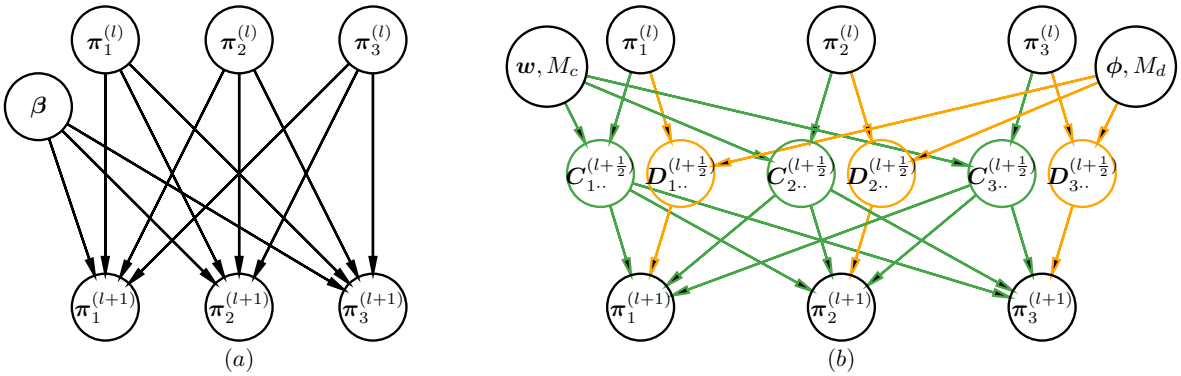

Figure 2. Graphical illustrations of the layer-wise connection for the Dirichlet Belief Network (DirBN) (left panel (a)) and the Poisson-randomised Dirichlet Belief Network (Pois-DirBN) (right panel (b)). In the right panel, we use green (or orange) nodes and arrows to denote the node-wise Poisson counting variables $\boldsymbol{C}$ (or component-wise Poisson counting variables $\boldsymbol{D}$) and their related connections.

ratio and thus violates the implicit condition mentioned in Section 2. It is thereby difficult to use the algorithm of first upward propagating latent counts and then downward sampling random variables (Zhou et al., 2016; Zhao et al., 2018) to infer the latent distribution $\boldsymbol{\pi}_i^{(l)}$.

This component-wise propagation provides convenient ways to set flexible concentration parameters. Under the previous node-wise propagation methods, it would be difficult to set larger weights on features when these features are unimportant in other latent distributions. Through the component-wise propagation, the weights of these features can be enlarged by propagating other features into it. More importantly, we can have different components' ratios for the same node's latent distributions across consecutive layers, which promotes larger mutations than those without component-wise propagation.

It should be noted that, we cannot directly sampling each entries of the propagation variable $\boldsymbol{D}$ when applying $\boldsymbol{D}$ in the topic modelling setting, which regards topic-word distributions as the latent distributions. As the vocabulary size is usually large, it is usually impossible to infer each $\boldsymbol{D}$ entry. Instead, we can manually set the sparsity of $\boldsymbol{D}$ by observing the co-occurence of words in the same document. For example, we may only sample the entries when the two corresponding vocabularies have appeared in the same document at least two times.

3.2. Marginal deep stochastic integer networks of $\boldsymbol{C}, \boldsymbol{D}$

We can integrate out the latent distributions $\{\boldsymbol{\pi}_i^{(l)}\}_{i,l}$ in the Pois-DirBN to obtain a counting variables $\boldsymbol{C}, \boldsymbol{D}$ composed deep neural network. Let $\psi_{ik}^{(l-\frac{1}{2})} = \sum_{i'} C_{i'ki}^{(l-\frac{1}{2})} + \sum_{k'} D_{ik'k}^{(l-\frac{1}{2})}$, which summarizes the counts from layer $(l-\frac{1}{2})$ into the component k of node i at layer $(l+\frac{1}{2})$. The counting variables $\boldsymbol{C}_{ik\cdot}^{(l+\frac{1}{2})}, \boldsymbol{D}_{ik\cdot}^{(l+\frac{1}{2})}$ for layer $(l+\frac{1}{2})$ can be directly generated as follows:

1. $M_{i,C}^{(l+\frac{1}{2})} \sim \text{Poisson}(M_c), M_{i,D}^{(l+\frac{1}{2})} \sim \text{Poisson}(M_d)$;
2. $P(\{M_{ik,C/D}^{(l+\frac{1}{2})}\}_{k=1}^K) \propto \prod_k \binom{\alpha_k^{(\pi)}+\psi_{ik}^{(l-\frac{1}{2})}+M_{ik,C/D}^{(l+\frac{1}{2})}}{M_{ik,C/D}^{(l+\frac{1}{2})}}$, such that $\sum_k M_{ik,C}^{(l+\frac{1}{2})} = M_{i,C}^{(l+\frac{1}{2})}, \sum_k M_{ik,D}^{(l+\frac{1}{2})} = M_{i,D}^{(l+\frac{1}{2})}$;
3. $(C_{ik1}^{(l+\frac{1}{2})}, \dots, C_{ikN}^{(l+\frac{1}{2})}) \sim \text{Multinomial}(M_{ik,C}^{(l+\frac{1}{2})}; \boldsymbol{\omega}_i^{(l)})$;
4. $(D_{ik1}^{(l+\frac{1}{2})}, \dots, D_{ikK}^{(l+\frac{1}{2})}) \sim \text{Multinomial}(M_{ik,D}^{(l+\frac{1}{2})}; \boldsymbol{\phi}_k^{(l)})$,

where we use $M_{ik,C/D}^{(l+\frac{1}{2})}$ to denote either $M_{ik,C}^{(l+\frac{1}{2})}$ or $M_{ik,D}^{(l+\frac{1}{2})}$ as their probability mass functions are the same, and where $\binom{\alpha_k^{(\pi)}+\psi_{ik}^{(l-\frac{1}{2})}+M_{ik,C/D}^{(l+\frac{1}{2})}}{M_{ik,C/D}^{(l+\frac{1}{2})}} = \frac{\prod_{v=0}^{M_{ik,C/D}^{(l+\frac{1}{2})}-1}(\alpha_k^{(\pi)}+\psi_{ik}^{(l-\frac{1}{2})}+v)}{(M_{ik,C/D}^{(l+\frac{1}{2})})!}$ is the generalized binomial coefficient (L Graham, 1994). The variables $\boldsymbol{M}_{i\cdot,C}^{(l+\frac{1}{2})}, \boldsymbol{M}_{i\cdot,D}^{(l+\frac{1}{2})}$ in step (2) can be sampled by first calculating the ratios of all their potential configurations and then sampling one particular configuration with probabilities proportional to these ratios. Note that although this generative process may require high computational cost for large values of $M_{i,C}^{(l+\frac{1}{2})}, M_{i,D}^{(l+\frac{1}{2})}, K$, the computational cost of posterior sampling, which can be found in the Supplementary Material, is the same as the Dirichlet-Poisson-Dirichlet connections.

Marginalizing out the latent distributions enables us to view the Pois-DirBN as a *multi-stochastic deep integer network*, in which the architecture is composed of $\{l+\frac{1}{2}\}_l$ integer composed layers and the latent distribution are marginalized out. The distribution in step 2. can be regarded as non-linear activation function in this deep network structure. Since the integer variables have smaller storage and low memory requirements, this deep stochastic integer network may have

unique merits in dealing with large-scale data tasks. This may be worth future exploration.

3.3. Layer-wise Gibbs sampling

The Pois-DirBN random variables include the latent distributions $\{\boldsymbol{\pi}_i^{(l)}\}_{i,l}$, counting variables $\{\boldsymbol{C}^{(l+\frac{1}{2})}, \boldsymbol{D}^{(l+\frac{1}{2})}\}_l$ and layer-wise propagation coefficients $\{\boldsymbol{\phi}^{(l)}, \boldsymbol{\omega}^{(l)}\}_l$. We develop an efficient layer-wise Gibbs sampling algorithm to infer the posterior distributions of these variables.

Sampling $\boldsymbol{C}$: Combining the prior and likelihood terms of $C_{iki'}^{(l+\frac{1}{2})}$, the conditional posterior of $C_{iki'}^{(l+\frac{1}{2})}$ is:

$$P(C_{iki'}^{(l+\frac{1}{2})}|-) \propto \frac{[M_c \pi_{ik}^{(l)} w_{ii'} \pi_{i'k}^{(l+1)}]^{C_{iki'}^{(l+\frac{1}{2})}}}{C_{iki'}^{(l+\frac{1}{2})}!} \cdot \frac{\Gamma(\sum_{k'} v_{k'})}{\Gamma(v_k)}$$

where $v_k = \alpha_k^{(\pi)} + C_{\cdot ki'}^{(l+\frac{1}{2})} + D_{i'\cdot k}^{(l+\frac{1}{2})}$ and $\Gamma(\cdot)$ is the Gamma function.

The second term (i.e. $\Gamma(\sum_{k'} v_{k'})/\Gamma(v_k)$) in the RHS of the above equation is a ratio of two Gamma functions which may impair the ability to form conjugate relations. In order to proceed with efficient Gibbs sampling for $C_{iki'}^{(l+\frac{1}{2})}$, we first fix $\alpha_k^{(\pi)} = \frac{1}{K-1}$, which makes the difference between $\sum_{k'} v_{k'}$ and v_k to be an integer. Then, we introduce two auxiliary variables $y_{iki'}^{(l+\frac{1}{2})}, z_{iki'}^{(l+\frac{1}{2})}$ to augment this Gamma function ratio and form the joint likelihood as

$$y_{i'k}^{(l+\frac{1}{2})} \sim \text{CRT}(1 + C_{\cdot\cdot i'}^{(l+\frac{1}{2})} + D_{i'\cdot\cdot}^{(l+\frac{1}{2})}, \alpha_k^{(\pi)} + C_{\cdot ki'}^{(l+\frac{1}{2})} + D_{i'\cdot k}^{(l+\frac{1}{2})}),$$
$$z_{iki'}^{(l+\frac{1}{2})} \sim \text{Binomial}(y_{i'k}^{(l+\frac{1}{2})}, C_{iki'}^{(l+\frac{1}{2})}/(\alpha_k + C_{\cdot ki'}^{(l+\frac{1}{2})} + D_{i'\cdot k}^{(l+\frac{1}{2})})),$$
$$P(C_{iki'}^{(l+\frac{1}{2})}, z_{iki'}^{(l+\frac{1}{2})}|-) \propto \frac{[\gamma_{iki'}^{(C,l+\frac{1}{2})}]^{C_{iki'}^{(l+\frac{1}{2})}}}{C_{iki'}^{(l+\frac{1}{2})}!} \cdot [C_{iki'}^{(l+\frac{1}{2})}]^{z_{iki'}^{(l+\frac{1}{2})}},$$

where $\text{CRT}(\cdot,\cdot)$ is the Chinese Restaurant Table (CRT) distribution (Zhou & Carin, 2015). Note that this CRT distribution will not lead to information loss in the Pois-DirBN. The expected value of $y_{i'k}^{(l+\frac{1}{2})}$ is on the same scale as that of the counting information, which is $\mathcal{O}(\mathbb{E}[y_{i'k}^{(l+\frac{1}{2})}]) = \mathcal{O}(\alpha_k^{(\pi)} + C_{\cdot ki'}^{(l+\frac{1}{2})} + D_{i'\cdot k}^{(l+\frac{1}{2})}) \log(1 + C_{\cdot\cdot i'}^{(l+\frac{1}{2})} + D_{i'\cdot\cdot}^{(l+\frac{1}{2})})$.

Since the distribution of $C_{iki'}^{(l+\frac{1}{2})}$ is then in the form of Touchard polynomials (Roman & Rota, 1978), we can obtain samples by using the method of (Fan et al., 2019).

Sampling $\boldsymbol{D}_{ikk'}^{(l+\frac{1}{2})}$: The conditional posterior of $D_{ikk'}^{(l+\frac{1}{2})}$ is

$$P(D_{ikk'}^{(l+\frac{1}{2})}|-) \propto \frac{[M_d \pi_{ik}^{(l)} \phi_{kk'}^{(l)} \pi_{ik'}^{(l+1)}]^{D_{ikk'}^{(l+\frac{1}{2})}}}{D_{ikk'}^{(l+\frac{1}{2})}!} \cdot \frac{\Gamma(\sum_{k'} v_{k'})}{\Gamma(v_k)},$$

where $v_k = \alpha_k^{(\pi)} + D_{i\cdot k}^{(l+\frac{1}{2})} + C_{\cdot ki}^{(l+\frac{1}{2})}$. We can sample from this distribution in a similar manner to $\boldsymbol{C}$.

Sampling $\boldsymbol{\pi}$: $\boldsymbol{\pi}_i^{(l)}$'s conditional posterior distribution is

$$\boldsymbol{\pi}_i^{(l)} \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(\pi,l)} + \boldsymbol{\psi}_i^{(l-\frac{1}{2})} + \sum_{i'} C_{i\cdot i'}^{(l+\frac{1}{2})} + \sum_{k'} D_{i\cdot k'}^{(l+\frac{1}{2})})$$

where $\boldsymbol{\psi}_i^{(l-\frac{1}{2})} = \sum_{i'} C_{i'\cdot i}^{(l-\frac{1}{2})} + \sum_{k'} D_{ik'\cdot}^{(l-\frac{1}{2})}$.

Sampling $\boldsymbol{\omega}^{(l)}, \boldsymbol{\phi}^{(l)}$: Let the prior distribution of $\boldsymbol{\omega}^{(l)}$ and $\boldsymbol{\phi}^{(l)}$ be $\boldsymbol{\omega}_i^{(l)} \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(\omega)}), \boldsymbol{\phi}_k^{(l)} \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(\phi)})$. The conditional posterior distributions of $\boldsymbol{\omega}_i^{(l)}$ and $\boldsymbol{\phi}_k^{(l)}$ are then

$$\boldsymbol{\omega}_i^{(l)} \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(\omega)} + \sum_{k'} \boldsymbol{C}_{ik'\cdot}^{(l+\frac{1}{2})}) \tag{3}$$

$$\boldsymbol{\phi}_k^{(l)} \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(\phi)} + \sum_{i} \boldsymbol{D}_{ik\cdot}^{(l+\frac{1}{2})}). \tag{4}$$

4. Related Work

In addition to the DirBN variants mentioned in the Introduction, the DirBN is closely related to the Gamma Belief Network (GBN) (Zhou et al., 2016), which is another multi-stochastic layered deep generative model. Instead of using Dirichlet distributions, the GBN uses Gamma distributions to propagate scalar variables across layers, and was the first to develop the algorithm for upward propagating latent counts and then downward sampling random variables for model inference. Applications of the GBN and its inferential algorithm have been observed in factor analysis (Wang et al., 2019), natural language modelling (Guo et al., 2020), Poisson Gamma Dynamic Systems (Schein et al., 2016; Guo et al., 2018; Yang & Koeppl, 2018) and even variational autoencoder methods (Zhang et al., 2018). The GBN does not enjoy the unique sparsity property of the Dirichlet distribution and cannot be used to model latent distributions.

The Poisson Randomised Gamma Dynamic System (PRGDS) (Schein et al., 2019) may be the closest to our approach, which inserted Poisson variables between the Gamma-Gamma connections. The Pois-DirBN differs from the PRGDS in two aspects: (1) the targets are different as the Pois-DirBN works on Dirichlet-Dirichlet connected deep generative models and aims to address the latent distribution mutation problem, whereas the PRGDS iss applied in the dynamic system setting; (2) the inference method is different. We have independently developed an efficient Gibbs sampling algorithm to sample the counting variables $\boldsymbol{C}, \boldsymbol{D}$. The PRGDS discusses the cases where $\alpha = 0$ and $\alpha > 0$ for the shape parameter α of the Gamma distribution. The Pois-DirBN discusses the case where $\alpha = \frac{1}{K-1}$ for the concentration parameter of the Dirichlet distribution. Our developed deep stochastic integer networks (Section 3.2) can be directly used to model the case of $\alpha = 0$.

5. Application of Pois-DirBN to relational modeling

5.1. Model

The performance of the Pois-DirBN is evaluated in the relational modeling setting. The relational data $\boldsymbol{R}$ are represented as a binary matrix $\boldsymbol{R} \in \{0,1\}^{N\times N}$, where N is the number of nodes and the element R_{ij} $(\forall i,j)$ indicates whether node i relates to node j ($R_{ij}=1$ if the relation exists, otherwise $R_{ij}=0$). The self-connection relation R_{ii} is not considered here. The matrix $\boldsymbol{R}$ can be symmetric (i.e. undirected) or asymmetric (i.e. directed).

Following similar settings to (Fan et al., 2019), the generative process of the Pois-DirBN in the relational modelling settings can be described as follows:

1. $r_k \sim \text{Gamma}(r_0/K, c_0)$,
2. $\Lambda_{k_1k_2} \sim \begin{cases} \text{Gamma}(\xi r_{k_1}, \eta), & \text{if } k_1 = k_2; \\ \text{Gamma}(r_{k_1} r_{k_2}, \eta), & \text{if } k_1 \neq k_2. \end{cases}$
3. $\boldsymbol{\pi}_i^{(L)} \sim \text{Pois-DirBN}(-), X_{ik} \sim \text{Poisson}(M\pi_{ik}^{(L)})$
4. $Z_{ij,k_1k_2} \sim \text{Poisson}\left(X_{ik_1}\Lambda_{k_1k_2}X_{jk_2}\right), \forall i \neq j$
5. $R_{ij} = \mathbf{1}\left(\sum_{k_1,k_2} Z_{ij,k_1k_2} > 0\right), \forall i \neq j$

where r_0, c_0, ξ, η, M are hyper-parameters.

In the above generative process, steps $(1), (2)$ use the Hierarchical Gamma Process (HGaP) (Zhou, 2015a) to generate community popularity variable r_k and community compatibility value $\Lambda_{k_1k_2}$, in which larger r_k values make community k generate larger compatibility values. Λ_{k_1,k_2} is a community-versus-community compatibility parameter, where a larger value of Λ_{k_1,k_2} indicates a higher probability of generating a link between community k_1 and k_2. Step (3) uses the Pois-DirBN to generate nodes' latent distributions at layer L and then use these latent distributions to generate nodes' counting vectors $\boldsymbol{X}_i$. Similar to the discussions with $\boldsymbol{C}, \boldsymbol{D}$, $\boldsymbol{X}_i/M$ can also be regarded as a finite approximation to $\boldsymbol{\pi}_i^{(L)}$ as we have $\mathbb{E}[\boldsymbol{X}_i]/M = \boldsymbol{\pi}_i^{(L)}$. Steps $(4), (5)$ use the Poisson-Bernoulli link function (Rai et al., 2015; Zhou, 2015a) to generate the relation R_{ij}, which first generates the (k_1,k_2)-th latent integer Z_{ij,k_1k_2} and then uses the sum over all K^2 integers to determine the positiveness of R_{ij}. Steps $(4), (5)$ can be alternatively represented as $R_{ij} \sim \text{Bernoulli}(1-\exp(\sum_{k_1,k_2} X_{ik_1}\Lambda_{k_1k_2}X_{jk_2}))$ if we integrate out Z_{ij,k_1k_2}.

We follow the setting of (Fan et al., 2019) and restrict the entries of $\boldsymbol{\omega}_i^{(l)}$ to be 0 if there is no observed link from node i to the given nodes. Each relation R_{ij} is decomposed into community-to-community latent integers, and only the relations with the summation of its latent integers larger than 0 are taken as observed. In this way, the computational cost scales to the number of positive links only.

Table 1. Dataset information. N is the number of nodes, N_E is the number of positive links.

Dataset	N	N_E	Dataset	N	N_E
Citeer	3 312	4 715	Cora	2 708	5 429
Pubmed	2 000	17 522	PPI	4 000	105 775

5.2. Experiments

Dataset Information We examine four real-world datasets: three standard citation networks (*Citeer*, *Cora*, *Pubmed* (Sen et al., 2008) and one protein-to-protein interaction network (*PPI*) (Zitnik & Leskove, 2017). Summary statistics for these datasets are displayed in Table 1. For fair comparison, we use an identity matrix $I_{N\times N}$ as the feature information for all the datasets and comparison methods and do not involve any detailed feature values.

Experimental settings For hyper-parameters, we set $r_0, c_0, \xi, \eta \sim \text{Gam}(1,1), M \sim \text{Gam}(100,1)$ for all datasets. Hyper-parameters not directly related to the Pois-DirBN are specified in the supplementary material. Each run uses 2 000 MCMC iterations with the first 1 000 discarded as burn-in and the mean values of the second 1 000 posterior samples' performance score are reported. Unless specified, we are using 90% (per row) of the relational data as training data and the remaining 10% as test data. We use Area Under the ROC curve (AUC) and average precision value on the testing data to measure the link prediction performance. Unless specified otherwise, we set the number of layers $L=3$ and the number of communities $K=20$.

Comparison methods: Several Bayesian methods for relational data and two Graph Auto-Encoder models are used for comparison: the Mixed-Membership Stochastic Blockmodel (Airoldi et al., 2009), the Hierarchical Latent Feature Relational Model (HLFM) (Hu et al., 2017), the Node Attribute Relational Model (NARM) (Zhao et al., 2017), the Hierarchical Gamma Process-Edge Partition Model (HGP-EPM) (Zhou, 2015b), the graph autoencoder (GAE) and variational graph autoencoder (VGAE) (Kipf & Welling, 2016), and the Scalable Deep Relational Model (SDREM) (Fan et al., 2019).

The NARM, HGP-EPM, GAE, VGAE and SDREM methods are executed using their respective implementations from the authors, under their default settings. The MMSB and HLFM are implemented to the best of our abilities and we set the number of layers and the length of the latent binary representation in the HLFM the same as those in the Pois-DirBN. For the GAE and VGAE, the AUC and

Table 2. Links prediction performance comparison (values are displayed in percentage format). It is noted that we do not use nodes' feature information in these relational datasets.

Model	AUC (mean and standard deviation) Citeer	Cora	Pubmed	PPI	Average Precision (mean and standard deviation) Citeer	Cora	Pubmed	PPI
MMSB	69.0 ± 0.4	74.3 ± 0.7	77.4 ± 0.5	80.1 ± 0.3	66.1 ± 0.4	70.4 ± 0.5	74.2 ± 0.4	82.3 ± 0.3
NARM	75.9 ± 0.3	80.9 ± 0.3	80.8 ± 0.4	82.1 ± 0.2	78.1 ± 0.4	83.1 ± 0.4	77.1 ± 0.5	84.4 ± 0.2
HGP-EPM	76.3 ± 0.3	81.0 ± 0.3	80.3 ± 0.6	83.4 ± 0.4	77.6 ± 0.2	84.0 ± 0.3	78.6 ± 0.6	86.4 ± 0.4
HLFM	78.1 ± 1.0	82.9 ± 0.5	82.9 ± 0.5	85.6 ± 1.0	79.3 ± 0.4	84.2 ± 0.3	80.2 ± 0.3	88.3 ± 0.8
GAE	78.9 ± 0.4	84.6 ± 0.6	82.2 ± 0.4	87.4 ± 0.9	83.9 ± 0.4	88.4 ± 0.7	84.6 ± 0.4	88.9 ± 0.3
VGAE	79.0 ± 0.3	84.9 ± 0.4	82.6 ± 0.2	88.0 ± 0.7	84.6 ± 0.3	88.9 ± 0.4	85.0 ± 0.3	88.2 ± 0.4
DirBN (SDREM)	77.9 ± 0.4	83.2 ± 0.8	84.5 ± 0.8	89.2 ± 0.7	81.9 ± 0.4	87.5 ± 3.0	86.0 ± 0.7	88.4 ± 0.2
Pois-DirBN-C	78.3 ± 0.6	83.6 ± 1.1	82.4 ± 0.3	88.7 ± 0.7	82.8 ± 0.3	87.4 ± 0.6	84.3 ± 0.7	88.7 ± 0.5
Pois-DirBN-D	80.5 ± 0.8	86.1 ± 0.3	86.8 ± 0.7	92.4 ± 0.5	$\mathbf{88.2} \pm 0.3$	90.8 ± 0.3	88.7 ± 0.3	91.8 ± 0.4
Pois-DirBN-CD	$\mathbf{82.9} \pm 0.3$	88.4 ± 0.4	$\mathbf{87.8} \pm 0.2$	92.7 ± 0.3	88.1 ± 1.1	91.5 ± 0.7	$\mathbf{89.3} \pm 0.2$	$\mathbf{92.3} \pm 0.5$
Integer-DirBN	82.3 ± 1.0	$\mathbf{89.1} \pm 0.4$	87.6 ± 0.7	$\mathbf{93.1} \pm 0.7$	$\mathbf{88.2} \pm 0.4$	$\mathbf{92.2} \pm 0.6$	89.2 ± 0.7	92.8 ± 0.2

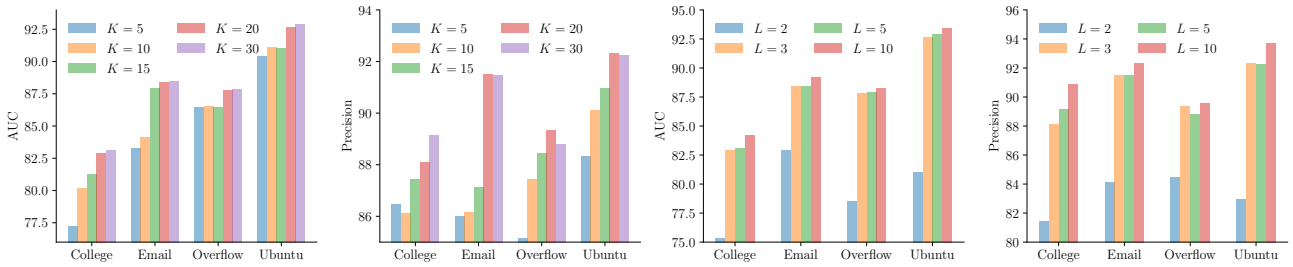

Figure 3. Barplots for AUC and average precision versus different number of communities and different number of layers.

precision values are calculated based on the pairwise similarities between the node representations. We consider the cases of introducing either $\boldsymbol{C}$, $\boldsymbol{D}$, both $\boldsymbol{C}$ and $\boldsymbol{D}$, or the integer version to verify the effectiveness of introducing the Poisson counting vectors in the Pois-DirBN, which we refer to as Pois-DirBN-C, Pois-DirBN-D, Pois-DirBN-CD and Integer-DirBN, respectively.

Link prediction performance: Table 2 displays the resulting link prediction performance on the Pois-DirBN models and other comparison methods. As can be seen, almost all deep generative models perform better than the shallow approaches, which supports the adoption of a deep structure. Among all deep generative models, the Pois-DirBN-CD usually perform the best in all four datasets. The performance of the Poisson-DirBN-C is quite comparable to the SDREM, but is worse than that of the Pois-DirBN-D and Pois-DirBN-CD, which shows that only using scaled latent distributions is not enough to fully use the deep structure. The performance of Pois-DirBN-D is comparable to the Pois-DirBN-CD, which indicates the importance of the component-wise propagation in the Pois-DirBN.

Comparisons on the model structure: We set different values for the number of layers L and the number of communities K to explore the performance of the Pois-DirBN-CD. In particular, we consider two scenarios: $K = 15$ and $L = 2, 3, 5, 10$; $L = 3$ and $K = 5, 10, 15, 20, 30$. Figure 3 shows the resulting performance. We can see that the performance of the Pois-DirBN-CD for different values of K is consistent with expectations: AUC and Average Precision increase as K increases, for all four datasets. The performance increase is not significant when K is larger than 20. This phenomenon might be due to the shrinkage property of the compatibility matrix $\boldsymbol{\Lambda}$. The performance when $L = 2$ seems be significantly worse than other cases, which may indicate that a 2-layered structure may not be deep enough. Although the performance gain is not large when L increases further, it is still the case that $L = 10$ is the best performing.

Latent distribution visualisations: Figure 4 displays visualisations of the latent distributions across different layers for the SDREM (DirBN), Pois-DirBN-C, Pois-DirBN-D and Pois-DirBN-CD models. For the SDREM, we can see that the patterns of latent distributions do not vary much across different layers. Their main difference is that the latent distribution patterns are clearer in the lower layers, which might be due to less information being propagated to the upper layers.

For the Pois-DirBN-C, the patterns are still quite similar across different layers, however, these patterns are still quite clear and distinguishable even in upper layers. For the

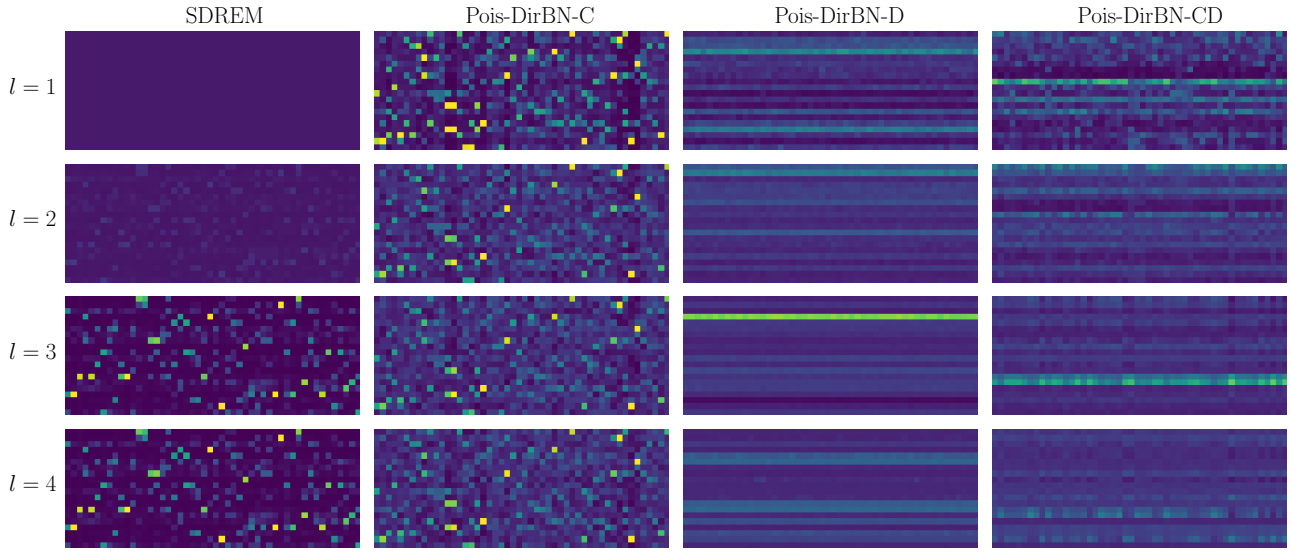

Figure 4. Visualisations of first 50 nodes' latent distributions $\{\boldsymbol{\pi}_i^{(l)}\}_l$ for the SDREM (DirBN), Pois-DirBN-C, Pois-DirBN-D and Pois-DirBN-CD models. The columns in each panel represent different latent distributions. The membership value are larger when their colours from blue to yellow.

Pois-Dir-D, the patterns are different across different layers, however, there seems to be no significant variations among the nodes' latent distributions in the same layer. This phenomenon might be related to the component-wise propagation matrix $\boldsymbol{\phi}^{(l)}$, in which the dominating entries might determine the largest-weighted component for all latent distributions. For the Pois-Dir-CD, we can see there are clear pattern changes across different layers and the patterns are varied for the nodes in the same layer, which is consistent with our expectations.

Node-wise and component-wise propagation matrix: Figure 5 visualises the node-wise propagation matrix $\boldsymbol{\omega}^{(2)}$ at layer 2 and the community compatibility matrix $\boldsymbol{\Lambda}$ for the Pois-DirBN-CD and SDREM. While the values of $\boldsymbol{\omega}^{(2)}$ for the Pois-DirBN-CD do not have clear patterns, the diagonal values of $\boldsymbol{\omega}^{(2)}$ for the SDREM are significantly larger than the non-diagonal ones. This might explain the duplicate patterns of latent distributions in the SDREM. For both the Pois-DirBN-CD and SDREM, we can see that the intra-community compatibilities (diagonals) dominate the community compatibility matrix.

The left panels of Figure 6 display the component-wise propagation matrices $\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}, \boldsymbol{\phi}^{(3)}$ for the Pois-DirBN-CD on the *Citeer* dataset. It is interesting that the intra-component propagation is not significant for $\boldsymbol{\phi}^{(l)}$. Except for $\boldsymbol{\phi}^{(2)}$, which seems to propagate more information to components $14, 15$, the other matrix do not have clear patterns. The right panel of Figure 6 displays the convergence behaviour of the Pois-DirBN-CD and Integer-DirBN on the *Citeer* dataset. It is clear that the Integer-DirBN converges slightly earlier than the Pois-DirBN-CD, due to the collapsed variable effect.

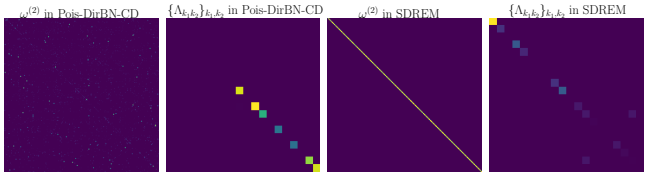

Figure 5. Visualisations of the node-wise propagation matrix $\boldsymbol{\omega}^{(2)}$ and community compatibility matrix $\boldsymbol{\Lambda}$ for both Pois-DirBN-CD and SDREM.

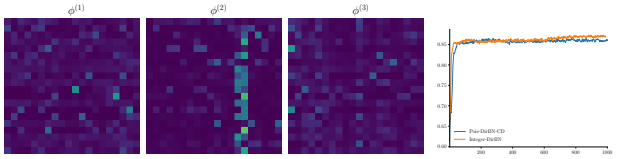

Figure 6. Left: Visualisations of the component-wise propagation matrix $\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}, \boldsymbol{\phi}^{(3)}$ for the Pois-DirBN-CD. Right: Convergence behaviour forthe Pois-DirBN-CD and Integer-DirBN.

6. Conclusion

We have proposed the Pois-DirBN to promote larger mutations for the latent distributions across different layers in Dirichlet Belief Networks, by introducing a component-wise propagation mechanism in its layer-wise connections. We introduced a Poisson counting variable between the

Dirichlet-Dirichlet layerwise-connection, forming Dirichlet-Poisson-Dirichlet connections, and developed a layer-wise Gibbs sampling method which can overcome the disadvantages of the previous methods. We also integrated out the latent distributions and formed a multi-stochastic integer network, which may be promising for reducing memory requirements and accelerating computation. The promising experimental results validate the effectiveness of the Pois-DirBN over the DirBN in terms of improved link prediction performance and more interpretable latent distribution visualisations. Using these counting variables to form a Bayesian nonparametric stick-breaking process to allow flexible model architectures and developing scalable variational amortized inference method would be worth exploring in the future.

Acknowledgements

Xuhui Fan and Scott A. Sisson are supported by the Australian Research Council (ARC) through the Australian Centre of Excellence in Mathematical and Statistical Frontiers (ACEMS, CE140100049), and Scott A. Sisson through the ARC Future Fellow Scheme (FT170100079). Bin Li is supported in part by STCSM Project (20511100400), Shanghai Municipal Science and Technology Major Projects (2018SHZDZX01, 2021SHZDZX0103), and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning. Yaqiong Li is a recipient of UTS Research Excellence Scholarship.

References

Airoldi, E., Blei, D., Fienberg, S., and Xing, E. Mixed membership stochastic blockmodels. In *NIPS*, pp. 33–40, 2009.

Dunson, D. B. and Herring, A. H. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 6(1): 11–25, 2005.

Fan, X., Li, B., Sisson, S., Li, C., and Chen, L. Scalable deep generative relational model with high-order node dependence. In *NeurIPS*, pp. 12637–12647, 2019.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Guo, D., Chen, B., Zhang, H., and Zhou, M. Deep poisson gamma dynamical systems. In *NeurIPS*, pp. 8442–8452, 2018.

Guo, D., Chen, B., Lu, R., and Zhou, M. Recurrent hierarchical topic-guided RNN for language generation. In *ICML*, pp. 3810–3821, 2020.

Hu, C., Rai, P., and Carin, L. Deep generative models for relational data with side information. In *ICML*, pp. 1578–1586, 2017.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kipf, T. N. and Welling, M. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

L Graham, R. *Concrete mathematics: a foundation for computer science*. Addison-Wesley, 1994.

Li, Y., Fan, X., Chen, L., Li, B., Yu, Z., and Sisson, S. A. Recurrent dirichlet belief networks for interpretable dynamic relational data modelling. In *IJCAI*, pp. 2470–2476, 2020.

Rai, P., Hu, C., Henao, R., and Carin, L. Large-scale bayesian multi-label learning via topic-based label embeddings. In *NIPS*, pp. 3222–3230. 2015.

Roman, S. M. and Rota, G.-C. The umbral calculus. *Advances in Mathematics*, 27(2):95 – 188, 1978.

Schein, A., Wallach, H., and Zhou, M. Poisson-gamma dynamical systems. In *NIPS*, pp. 5005–5013, 2016.

Schein, A., Linderman, S., Zhou, M., Blei, D., and Wallach, H. Poisson-randomized gamma dynamical systems. In *NeurIPS*, pp. 782–793, 2019.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. Collective classification in network data. In *AI magazine*, pp. 29–93, 2008.

Wang, C., Chen, B., Xiao, S., and Zhou, M. Convolutional poisson gamma belief network. In *ICML*, pp. 6515–6525, 2019.

Yang, S. and Koeppl, H. A poisson gamma probabilistic model for latent node-group memberships in dynamic networks. In *AAAI*, 2018.

Zhang, H., Chen, B., Guo, D., and Zhou, M. WHAI: Weibull hybrid autoencoding inference for deep topic modeling. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1cZsf-RW.

Zhao, H., Du, L., and Buntine, W. Leveraging node attributes for incomplete relational data. In *ICML*, pp. 4072–4081, 2017.

Zhao, H., Du, L., Buntine, W., and Zhou, M. Dirichlet belief networks for topic structure learning. In *NeurIPS*, pp. 7955–7966, 2018.

Zhou, M. Infinite Edge Partition Models for Overlapping Community Detection and Link Prediction. In *AISTATS*, pp. 1135–1143, 2015a.

Zhou, M. Infinite edge partition models for overlapping community detection and link prediction. In *AISTATS*, pp. 1135–1143, 2015b.

Zhou, M. and Carin, L. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2015.

Zhou, M., Cong, Y., and Chen, B. Augmentable gamma belief networks. *Journal of Machine Learning Research*, 17(163):1–44, 2016.

Zitnik, M. and Leskove, J. Predicting multicellular function through multi-layer tissue networks. In *Bioinformatics*, pp. i190–i198, 2017.