# Gamma Belief Networks

Mingyuan Zhou[†][*], Yulai Cong[‡] and Bo Chen[‡]
[†]The University of Texas at Austin and [‡]Xidian University

December 9, 2015

**Abstract**

To infer multilayer deep representations of high-dimensional discrete and nonnegative real vectors, we propose the gamma belief network (GBN) that factorizes each of its hidden layers into the product of a sparse connection weight matrix and the nonnegative real hidden units of the next layer. The GBN's hidden layers are jointly trained with an upward-downward Gibbs sampler that solves each layer with the same subroutine. The gamma-negative binomial process combined with a layer-wise training strategy allows inferring the width of each layer given a fixed budget on the width of the first layer. Example results illustrate interesting relationships between the width of the first layer and the inferred network structure, and demonstrate that the GBN can add more layers to improve its performance in both unsupervisedly extracting features and predicting heldout data. For exploratory data analysis, we extract trees and subnetworks from the learned deep network to visualize how the very specific factors discovered at the first hidden layer and the increasingly more general factors discovered at deeper hidden layers are related to each other, and we generate synthetic data by propagating random variables through the deep network from the top hidden layer back to the bottom data layer.

**Keywords:** Bayesian Nonparametrics, Deep Learning, Multilayer Representation, Poisson Factor Analysis, Unsupervised Learning.

## 1 Introduction

There has been significant recent interest in deep learning. Despite its tremendous success in supervised learning, inferring a multilayer data representation in an unsupervised manner

[*] Address for correspondence: 2110 Speedway Stop B6500, IROM Dept., Austin, TX 78712, USA.

remains a challenging problem (Bengio and LeCun, 2007, Bengio et al., 2015, Ranzato et al., 2007). To represent the data with a deep network, it is often unclear how to set the structure of the network, including the depth (number of layers) of the network and the width (number of hidden units) of each layer. Moreover, for tractable inference, the hidden units are often restricted to be binary. For example, the sigmoid belief network (SBN), which connects the binary units of adjacent layers via the sigmoid functions, infers a deep representation of multivariate binary vectors (Neal, 1992, Saul et al., 1996). The deep belief network (DBN) (Hinton et al., 2006) is a SBN whose top hidden layer is replaced by the restricted Boltzmann machine (RBM) (Hinton, 2002) that is undirected. The deep Boltzmann machine (DBM) is an undirected deep network that connects the binary units of adjacent layers using the RBMs (Salakhutdinov and Hinton, 2009). All these commonly used deep networks are designed to model binary observations, without principled ways to set the network structure. Although one may modify the bottom layer to model Gaussian and multinomial observations, the hidden units of these networks are still typically restricted to be binary (Larochelle and Lauly, 2012, Salakhutdinov and Hinton, 2009, Salakhutdinov et al., 2013). One may further consider the exponential family harmoniums (Welling et al., 2004, Xing et al., 2005) to construct more general networks with non-binary hidden units, but often at the expense of noticeably increased complexity in training and data fitting.

Moving beyond conventional deep networks using binary hidden units and setting the network structure in a heuristic manner, we construct deep networks using gamma distributed nonnegative real hidden units, and combine the gamma-negative binomial process (Zhou and Carin, 2015, Zhou et al., 2015b) with a greedy-layer wise training strategy to automatically infer the network structure. The proposed model is called the gamma belief network (GBN), which factorizes the observed or latent count vectors under the Poisson likelihood into the product of a factor loading matrix and the gamma distributed hidden units (factor scores) of layer one; and further factorizes the shape parameters of the gamma hidden units of each layer into the product of a connection weight matrix and the gamma hidden units of the next layer. The GBN together with Poisson factor analysis can unsupervisedly infer a multilayer representation from multivariate count vectors, with a simple but powerful mechanism to capture the correlations between the visible/hidden features across all layers and handle highly overdispersed counts. With the Bernoulli-Poisson link function (Zhou, 2015), the GBN is further applied to high-dimensional sparse binary vectors by truncating latent counts, and with a Poisson randomized gamma distribution, the GBN is further applied to high-dimensional sparse nonnegative real data by randomizing the gamma shape parameters with latent counts.

Distinct from previous deep networks that often require tuning both the width (number of hidden units) of each layer and the network depth (number of layers), the GBN employs nonnegative real hidden units and automatically infers the widths of subsequent layers given a fixed budget on the width of its first layer. Note that the budget could be infinite and hence the whole network can grow without bound as more data are being observed. When the budget is finite and hence the ultimate capacity of the network is limited, we find that the GBN equipped with a narrower first layer could increase its depth to match or even outperform a shallower network with a substantially wider first layer.

The gamma distribution density function has the highly desired strong non-linearity for deep learning, but the existence of neither a conjugate prior nor a closed-form maximum likelihood estimate (Choi and Wette, 1969) for its shape parameter makes a deep network with gamma hidden units appear unattractive. Despite seemingly difficult, we discover that, by generalizing the data augmentation and marginalization techniques for discrete data (Zhou and Carin, 2015), one may propagate latent counts one layer at a time from the bottom data layer to the top hidden layer, with which one may derive an efficient upward-downward Gibbs sampler that, one layer at a time in each iteration, upward samples Dirichlet distributed connection weight vectors and then downward samples gamma distributed hidden units, with the latent parameters of each layer solved with the same subroutine.

With extensive experiments in text and image analysis, we demonstrate that the deep GBN with two or more hidden layers clearly outperforms the shallow GBN with a single hidden layer in both unsupervisedly extracting latent features for classification and predicting heldout data. Moreover, we demonstrate the excellent ability of the GBN in exploratory data analysis: by extracting trees and subnetworks from the learned deep network, we can follow the paths of each tree to visualize various aspects of the data, from very general to very specific and understand how they are related to each other.

In addition to constructing a new deep network that well fits high-dimensional sparse binary, count, and nonnegative real data, developing an efficient upward-downward Gibbs sampler, and applying the learned deep network for exploratory data analysis, other contributions of the paper include: 1) proposing novel link functions, 2) combining the gamma-negative binomial process (Zhou and Carin, 2015, Zhou et al., 2015b) with a layer-wise training strategy to automatically infer the network structure; 3) revealing the relationship between the upper bound imposed on the width of the first layer and the inferred widths of subsequent layers; 4) revealing the relationship between the depth of the network and the model's ability to model overdispersed counts; and 5) generating multivariate high-dimensional discrete or nonnegative real vectors, whose distributions are governed by the GBN, by propagating the gamma hidden units of the top hidden layer back to the bottom data layer. We note this paper significantly extends our recent conference publication (Zhou et al., 2015a) that proposes the Poisson GBN.

# 2 Distributions for Count, Binary, and Nonnegative Real Data

## 2.1 Useful count distributions and their relationships

Let the Chinese restaurant table (CRT) distribution $l \sim \mathrm{CRT}(n, r)$ represent the random number of tables seated by $n$ customers in a Chinese restaurant process (Aldous, 1985, Antoniak, 1974, Blackwell and MacQueen, 1973, Pitman, 2006) with concentration parameter $r$. Its probability mass function (PMF) can be expressed as

$$P(l \,|\, n, r) = \frac{\Gamma(r) r^l}{\Gamma(n+r)} |s(n, l)|,$$

where $l \in \mathbb{Z}$, $\mathbb{Z} := \{0, 1, \ldots, n\}$, and $|s(n, l)|$ are unsigned Stirling numbers of the first kind. A CRT distributed sample can be generated by taking the summation of $n$ independent Bernoulli random variables as

$$l = \sum_{i=1}^{n} b_i, \;\; b_i \sim \mathrm{Bernoulli}\left[r/(r+i-1)\right].$$

Let $u \sim \mathrm{Log}(p)$ denote the logarithmic distribution (Anscombe, 1950, Fisher et al., 1943, Johnson et al., 1997) with PMF

$$P(u \,|\, p) = \frac{1}{-\ln(1-p)} \frac{p^u}{u},$$

where $u \in \{1, 2, \ldots\}$, and let $n \sim \mathrm{NB}(r, p)$ denote the negative binomial (NB) distribution with PMF

$$P(n \,|\, r, p) = \frac{\Gamma(n+r)}{n!\Gamma(r)} p^n (1-p)^r,$$

where $n \in \mathbb{Z}$. The NB distribution $n \sim \mathrm{NB}(r, p)$ can be generated as a gamma mixed Poisson distribution as

$$n \sim \mathrm{Pois}(\lambda), \;\; \lambda \sim \mathrm{Gam}\left[r, p/(1-p)\right],$$

where $p/(1-p)$ is the gamma scale parameter.

As shown in (Zhou and Carin, 2015), the joint distribution of $n$ and $l$ given $r$ and $p$ in

$$l \sim \mathrm{CRT}(n, r), \;\; n \sim \mathrm{NB}(r, p),$$

where $l \in \{0, \ldots, n\}$ and $n \in \mathbb{Z}$, is the same as that in

$$n = \textstyle\sum_{t=1}^{l} u_t, \; u_t \sim \text{Log}(p), \; l \sim \text{Pois}[-r \ln(1-p)], \tag{1}$$

which is called the Poisson-logarithmic bivariate distribution, with PMF

$$P(n, l \,|\, r, p) = \frac{|s(n, l)| r^l}{n!} p^n (1-p)^r.$$

We will exploit these relationships to derive efficient inference for the proposed GBN.

## 2.2 Bernoulli-Poisson link and truncated Poisson distribution

As in Zhou (2015), the Bernoulli-Poisson (BerPo) link thresholds a random count at one to obtain a binary variable as

$$b = \mathbf{1}(m \geq 1), \; m \sim \text{Pois}(\lambda), \tag{2}$$

where $b = 1$ if $m \geq 1$ and $b = 0$ if $m = 0$. If $m$ is marginalized out from (2), then given $\lambda$, one obtains a Bernoulli random variable as

$$b \sim \text{Ber}\left(1 - e^{-\lambda}\right).$$

The conditional posterior of $m$ can be expressed as

$$(m \,|\, b, \lambda) \sim b \cdot \text{Pois}_+(\lambda),$$

where $x \sim \text{Pois}_+(\lambda)$ follows a truncated Poisson distribution, with $P(x = k) = (1 - e^{-\lambda})^{-1} \lambda^k e^{-\lambda}/k!$ for $k \in \{1, 2, \ldots\}$. Thus if $b = 0$, then $m = 0$ almost surely (a.s.), and if $b = 1$, then $m \sim \text{Pois}_+(\lambda)$, which can be simulated with a rejection sampler that has a minimal acceptance rate of 63.2% at $\lambda = 1$ (Zhou, 2015). Given the latent count $m$ and a gamma prior on $\lambda$, one can then update $\lambda$ using the gamma-Poisson conjugacy. The BerPo link shares some similarities with the probit link that thresholds a normal random variable at zero, and the logistic link that lets $b \sim \text{Ber}[e^x/(1 + e^x)]$. We advocate the BerPo link as an alternative to the probit and logistic links since if $b = 0$, then $m = 0$ a.s., which could lead to significant computational savings if the binary vectors are sparse. In addition, the conjugacy between the gamma and Poisson distributions makes it convenient to construct hierarchical Bayesian models amenable to posterior simulation.

## 2.3 Poisson randomized gamma and truncated Bessel distributions

To model nonnegative data that include both zeros and positive observations, we introduce the Poisson randomized gamma (PRG) distribution as

$$x \sim \mathrm{PRG}(\lambda, c),$$

whose distribution has a point mass at $x = 0$ and is continuous for $x > 0$. The PRG distribution is generated as a Poisson mixed gamma distribution as

$$x \sim \mathrm{Gam}(n, 1/c),\ \ n \sim \mathrm{Pois}(\lambda), \tag{3}$$

in which we define $\mathrm{Gam}(0, 1/c) = 0$ a.s. and hence $x = 0$ if and only $n = 0$. Thus the PMF of $x \sim \mathrm{PRG}(\lambda, c)$ can be expressed as

$$\begin{aligned} f_X(x \,|\, \lambda, c) &= \sum_{n=0}^{\infty} \mathrm{Gam}(x; n, 1/c)\mathrm{Pois}(n; \lambda) \\ &= \left(e^{-\lambda}\right)^{\mathbf{1}(x=0)} \left[ e^{-\lambda - cx} \sqrt{\frac{\lambda c}{x}}\ I_{-1}\left(2\sqrt{\lambda c x}\right) \right]^{\mathbf{1}(x>0)} \end{aligned} \tag{4}$$

where

$$I_{-1}(\alpha) = \left(\frac{\alpha}{2}\right)^{-1} \sum_{n=1}^{\infty} \frac{\left(\frac{\alpha^2}{4}\right)^n}{n!\Gamma(n)}, \quad \alpha > 0 \tag{5}$$

is the modified Bessel function of the first kind $I_\nu(\alpha)$ with $\nu$ fixed at $-1$. Using the laws of total expectation and total variance, or using the PMF directly, one may show that

$$\mathbb{E}[x \,|\, \lambda, c] = \lambda/c, \quad \mathrm{Var}[x \,|\, \lambda, c] = 2\lambda/c^2. \tag{6}$$

Thus the variance to mean ratio of the PRG distribution is $2/c$, as controlled by $c$.

The conditional posterior of $n$ given $x$, $\lambda$, and $c$ can be expressed as

$$\begin{aligned} f_N(n \,|\, x, \lambda, c) &= \frac{\mathrm{Gam}(x; n, 1/c)\mathrm{Pois}(n; \lambda)}{\mathrm{PRG}(x; \lambda, c)} \\ &= \mathbf{1}(x = 0)\delta_0 + \mathbf{1}(x > 0) \sum_{n=1}^{\infty} \mathrm{Bessel}_{-1}(n; 2\sqrt{cx\lambda})\delta_n\,, \end{aligned} \tag{7}$$

where we define $n \sim \mathrm{Bessel}_{-1}(\alpha)$ as the truncated Bessel distribution, with PMF

$$\mathrm{Bessel}_{-1}(n;\alpha)=\frac{\left(\frac{\alpha}{2}\right)^{2n-1}}{I_{-1}(\alpha)n!\Gamma(n)},\quad n\in\{1,2,\ldots\}. \tag{8}$$

Thus $n=0$ if and only if $x=0$, and $n$ is a positive integer drawn from a truncated Bessel distribution if $x>0$.

Related to our work, Yuan and Kalbfleisch (2000) proposed the randomized gamma distribution to generate a random positive real number as

$$x\,|\,n,\nu\sim\mathrm{Gam}(n+\nu+1,1/c),\ n\sim\mathrm{Pois}(\lambda), \tag{9}$$

where $\nu>-1$ and $c>0$. As in Yuan and Kalbfleisch (2000), the conditional posterior of $n$ can be expressed as

$$(n\,|\,x,\nu,\alpha)\sim\mathrm{Bessel}_\nu(2\sqrt{cx\lambda}) \tag{10}$$

where we denote $n\sim\mathrm{Bessel}_\nu(\alpha)$ as the Bessel distribution with parameters $\nu>-1$ and $\alpha>0$, with PMF

$$\mathrm{Bessel}_\nu(n;\alpha)=\frac{\left(\frac{\alpha}{2}\right)^{2n+\nu}}{I_\nu(\alpha)n!\Gamma(n+\nu+1)},\quad n\in\{0,1,2,\ldots\}. \tag{11}$$

Algorithms to draw Bessel random variables can be found in Devroye (2002).

The proposed PRG is different from the randomized gamma distribution of Yuan and Kalbfleisch (2000) in that it models both positive real numbers and exact zeros, and the proposed truncated Bessel distribution $n\sim\mathrm{Bessel}_{-1}(\alpha)$ is different from the Bessel distribution $n\sim\mathrm{Bessel}_\nu(\alpha)$, where $\nu>-1$, in that it is defined only on positive integers.

# 3 Gamma Belief Networks

Denoting $\boldsymbol{\theta}_j^{(t)}\in\mathbb{R}_+^{K_t}$ as the $K_t$ hidden units of sample $j$ at layer $t$, where $\mathbb{R}_+=\{x:x\geq 0\}$, the generative model of the gamma belief network (GBN) with $T$ hidden layers, from top to bottom, is expressed as

$$\begin{aligned}
\boldsymbol{\theta}_j^{(T)}&\sim\mathrm{Gam}\left(\boldsymbol{r},1/c_j^{(T+1)}\right),\\
&\vdots\\
\boldsymbol{\theta}_j^{(t)}&\sim\mathrm{Gam}\left(\boldsymbol{\Phi}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)},1/c_j^{(t+1)}\right),\\
&\vdots\\
\boldsymbol{\theta}_j^{(1)}&\sim\mathrm{Gam}\left(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)},p_j^{(2)}/\big(1-p_j^{(2)}\big)\right).
\end{aligned} \tag{12}$$

For $t = 1, 2, \ldots, T-1$, the GBN factorizes the shape parameters of the gamma distributed hidden units $\boldsymbol{\theta}_j^{(t)} \in \mathbb{R}_+^{K_t}$ of layer $t$ into the product of the connection weight matrix $\boldsymbol{\Phi}^{(t+1)} \in \mathbb{R}_+^{K_t \times K_{t+1}}$ and the hidden units $\boldsymbol{\theta}_j^{(t+1)} \in \mathbb{R}_+^{K_{t+1}}$ of layer $t+1$; the top layer's hidden units $\boldsymbol{\theta}_j^{(T)}$ share the same vector $\boldsymbol{r} = (r_1, \ldots, r_{K^{(T)}})'$ as their gamma shape parameters; and the $p_j^{(2)}$ are probability parameters and $\{1/c^{(t)}\}_{3,T+1}$ are gamma scale parameters, with $c_j^{(2)} := \big(1 - p_j^{(2)}\big)/p_j^{(2)}$. We will discuss later how to measure the connection strengths between the nodes of adjacent layers and the overall popularity of a factor at a particular hidden layer.

For scale identifiability and ease of inference and interpretation, each column of $\boldsymbol{\Phi}^{(t)} \in \mathbb{R}_+^{K_{t-1} \times K_t}$ is restricted to have a unit $L_1$ norm and hence $0 \leq \boldsymbol{\Phi}^{(t)}(k', k) \leq 1$. To complete the hierarchical model, for $t \in \{1, \ldots, T-1\}$, we let

$$\boldsymbol{\phi}_k^{(t)} \sim \mathrm{Dir}\big(\eta^{(t)}, \ldots, \eta^{(t)}\big), \quad r_k \sim \mathrm{Gam}\big(\gamma_0/K_T, 1/c_0\big) \tag{13}$$

where $\boldsymbol{\phi}_k^{(t)} \in \mathbb{R}_+^{K_{t-1}}$ is the $k$th column of $\boldsymbol{\Phi}^{(t)}$; we impose $c_0 \sim \mathrm{Gam}(e_0, 1/f_0)$ and $\gamma_0 \sim \mathrm{Gam}(a_0, 1/b_0)$; and for $t \in \{3, \ldots, T+1\}$, we let

$$p_j^{(2)} \sim \mathrm{Beta}(a_0, b_0), \quad c_j^{(t)} \sim \mathrm{Gam}(e_0, 1/f_0). \tag{14}$$

We expect the correlations between the $K_t$ rows (latent features) of $(\boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_J^{(t)})$ to be captured by the columns of $\boldsymbol{\Phi}^{(t+1)}$. Even if $\boldsymbol{\Phi}^{(t)}$ for $t \geq 2$ are all identity matrices, indicating no correlations between the latent features to be captured, our analysis in Section 4.2 will show that a deep structure with $T \geq 2$ could still benefit data fitting by better modeling the variability of the latent features $\boldsymbol{\theta}_j^{(1)}$.

## 3.1 Link functions for three different types of observations

If the observations are multivariate count vectors $\boldsymbol{x}_j^{(1)} \in \mathbb{Z}^V$, where $V := K_0$, then we link the integer-valued visible units to the nonnegative real hidden units at layer one using Poisson factor analysis (PFA) as

$$\boldsymbol{x}_j^{(1)} \sim \mathrm{Pois}\left(\boldsymbol{\Phi}^{(1)} \boldsymbol{\theta}_j^{(1)}\right). \tag{15}$$

Under this construction, the correlations between the $K_0$ rows (features) of $(\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_J^{(1)})$ are captured by the columns of $\boldsymbol{\Phi}^{(1)}$. Detailed descriptions on how PFA is related to a wide variety of discrete latent variable models, including nonnegative matrix factorization (Lee and Seung, 2001), latent Dirichlet allocation (Blei et al., 2003), the gamma-Poisson model (Canny, 2004), discrete Principal component analysis (Buntine and Jakulin, 2006), and the focused topic model (Williamson et al., 2010), can be found in Zhou et al. (2012) and Zhou and Carin (2015).

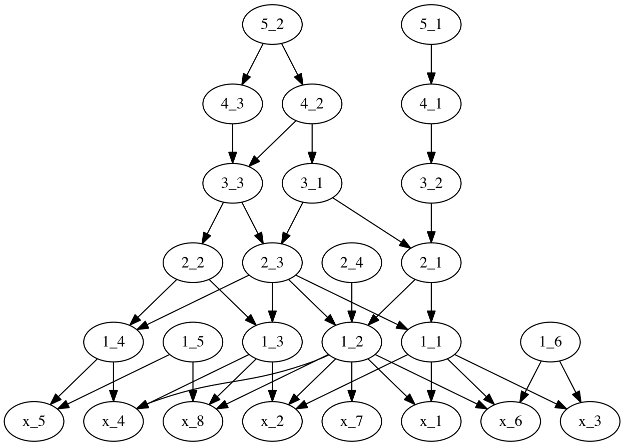


Figure 1: An example directed network of five hidden layers, with $K_0 = 8$ visible units, $[K_1, K_2, K_3, K_4, K_5] = [6, 4, 3, 3, 2]$, and sparse connections between the units of adjacent layers.

We call PFA using the GBN in (12) as the prior on its factor scores as the Poisson gamma belief network (PGBN), as proposed in Zhou et al. (2015a). The PGBN can be naturally applied to factorize the term-document frequency count matrix of a text corpus, not only extracting semantically meaningful topics at multiple layers, but also capturing the relationships between the topics of different layers using the deep network, as discussed below in both Sections 3.2 and 5.

If the observations are high-dimensional sparse binary vectors $\boldsymbol{b}_j^{(1)} \in \{0,1\}^V$, then we factorize them using Bernoulli-Poisson factor analysis (Ber-PFA) as

$$\boldsymbol{b}_j^{(1)} = \mathbf{1}\big(\boldsymbol{x}_j^{(1)} \geq 0\big),\ \boldsymbol{x}_j^{(1)} \sim \text{Pois}\left(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}\right). \tag{16}$$

We call Ber-PFA with the GBN as the prior on its factor scores $\boldsymbol{\theta}_j^{(1)}$ as the Bernoulli-Poisson gamma belief network (BerPo-GBN).

If the observations are high-dimensional sparse nonnegative real-valued vectors $\boldsymbol{y}_j^{(1)} \in \mathbb{R}_+^V$, then we factorize them using Poisson randomized gamma (PRG) factor analysis as

$$\boldsymbol{y}_j^{(1)} \sim \text{Gam}(\boldsymbol{x}_j^{(1)}, 1/a_j),\ \boldsymbol{x}_j^{(1)} \sim \text{Pois}\left(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}\right). \tag{17}$$

We call PRG factor analysis with the GBN as the prior on its factor scores $\boldsymbol{\theta}_j^{(1)}$ as the PRG gamma belief network (PRG-GBN).

We show in Figure 1 an example directed belief network of five hidden layers, with $V = 8$ visible units, 6, 4, 3, 3 and 2 hidden units for layers one, two, three, four and five, respectively, and with sparse connections between the units of adjacent layers.

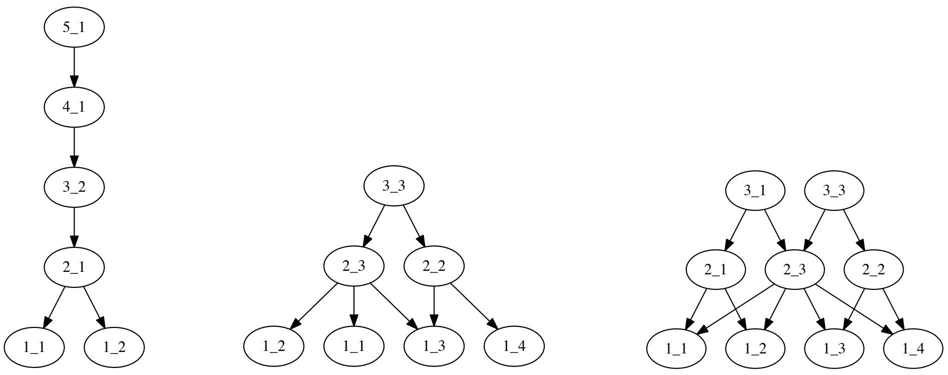


Figure 2: Extracted from the network shown in Figure 1, from left to right are a tree rooted at node 5_1, a tree rooted at node 3_3, and a subnetwork consisting of both the tree rooted at node 3_1 and the tree rooted at node 3_3.

### 3.2 Exploratory data analysis

To interpret the network structure of the GBN, we notice that

$$\mathbb{E}\big[\boldsymbol{x}_j^{(1)} \,\big|\, \boldsymbol{\theta}_j^{(t)}, \{\boldsymbol{\Phi}^{(\ell)}, c_j^{(\ell)}\}_{1,t}\big] = \left[\prod_{\ell=1}^{t} \boldsymbol{\Phi}^{(\ell)}\right] \frac{\boldsymbol{\theta}_j^{(t)}}{\prod_{\ell=2}^{t} c_j^{(\ell)}}, \tag{18}$$

and

$$\mathbb{E}\big[\boldsymbol{\theta}_j^{(t)} \,\big|\, \{\boldsymbol{\Phi}^{(\ell)}, c_j^{(\ell)}\}_{t+1,T}, \boldsymbol{r}\big] = \left[\prod_{\ell=t+1}^{T} \boldsymbol{\Phi}^{(\ell)}\right] \frac{\boldsymbol{r}}{\prod_{\ell=t+1}^{T+1} c_j^{(\ell)}}. \tag{19}$$

Thus for visualization, it is straightforward to project the $K_t$ topics/hidden units/factor loadings/nodes of layer $t \in \{1, \ldots, T\}$ to the bottom data layer as the columns of the $V \times K_t$ matrix

$$\prod_{\ell=1}^{t} \boldsymbol{\Phi}^{(\ell)}, \tag{20}$$

and rank their popularities using the $K_t$ dimensional nonnegative weight vector

$$\boldsymbol{r}^{(t)} := \left[\prod_{\ell=t+1}^{T} \boldsymbol{\Phi}^{(\ell)}\right] \boldsymbol{r}\,. \tag{21}$$

To measure the connection strength between node $k$ of layer $t$ and node $k'$ of layer $t-1$, we use the value of

$$\boldsymbol{\Phi}^{(t)}(k', k),$$

which is also expressed as $\boldsymbol{\phi}_k^{(t)}(k')$ or $\phi_{k'k}^{(t)}$.

Our intuition is that examining the nodes of the hidden layers, via their projections to the bottom data layer, from the top to bottom layers will gradually reveal less general and more specific aspects of the data. To verify this intuition and further understand the relationships between the general and specific aspects of the data, we consider extracting a tree for each node of layer $t$, where $t \geq 2$, to help visualize the inferred multilayer deep structure. To be more specific, to construct a tree rooted at a node of layer $t$, we grow the tree downward by linking the root node (if at layer $t$) or each leaf node of the tree (if at a layer below layer $t$) to all the nodes at the layer below that are connected to the root/leaf node with non-negligible weights. Note that a tree in our definition permits a node to have more than one parent, which means that different branches of the tree can overlap with each other. In addition, we also consider extracting subnetworks, each of which consists of multiple related trees from the full deep network. For example, shown in the left of Figure 2 is the tree extracted from the network in Figure 1 using node 5_1 as the root, shown in the middle is the tree using node 3_3 as the root, and shown in the right is a subnetwork consisting of two related trees that are rooted at nodes 3_1 and 3_3, respectively.

### 3.2.1 Visualizing nodes of different layers

Before presenting the technical details, we first provide some example results obtained with the PGBN on extracting multilayer representations from the 11,269 training documents of the 20newsgroups[1] dataset. Given a fixed budget of $K_{1\,\max} = 800$ on the width of the first layer, with $\eta^{(t)} = 0.1$ for all $t$, a five-layer deep network inferred by the PGBN has a network structure as $[K_1, K_2, K_3, K_4, K_5] = [386, 63, 58, 54, 51]$, meaning that there are 386, 63, 58, 54, and 51 nodes at layers one to five, respectively.

For visualization, we first relabel the nodes at each layer based on their weights $\{r_k^{(t)}\}_{1,K_t}$, calculated as in (21), with a more popular (larger weight) node assigned with a smaller label. We visualize node $k$ of layer $t$ by displaying its top 12 words ranked according to their probabilities in $\big(\prod_{\ell=1}^{t-1} \boldsymbol{\Phi}^{(\ell)}\big)\boldsymbol{\phi}_k^{(t)}$, the $k$th column of the projected representation calculated as in (20). We set the font size of node $k$ of layer $t$ proportional to $\big(r_k^{(t)}/r_1^{(t)}\big)^{\frac{1}{10}}$ in each subplot, and color the outside border of a text box as red, green, orange, blue, or black for a node of layer five, four, three, two, or one, respectively. For better interpretation, we also exclude from the vocabulary the top 30 words of node 1 of layer one: "don just like people think know time good make way does writes edu ve want say really article use right did things point going better thing need sure used little," and the top 20 words of node 2 of layer one: "edu writes article com apr cs ca just know don like think news cc david university john org wrote world." These 50 words are not in the standard list of stopwords but can be considered as stopwords specific to the 20newsgroups corpus discovered by the PGBN.

[1] http://qwone.com/∼jason/20Newsgroups/

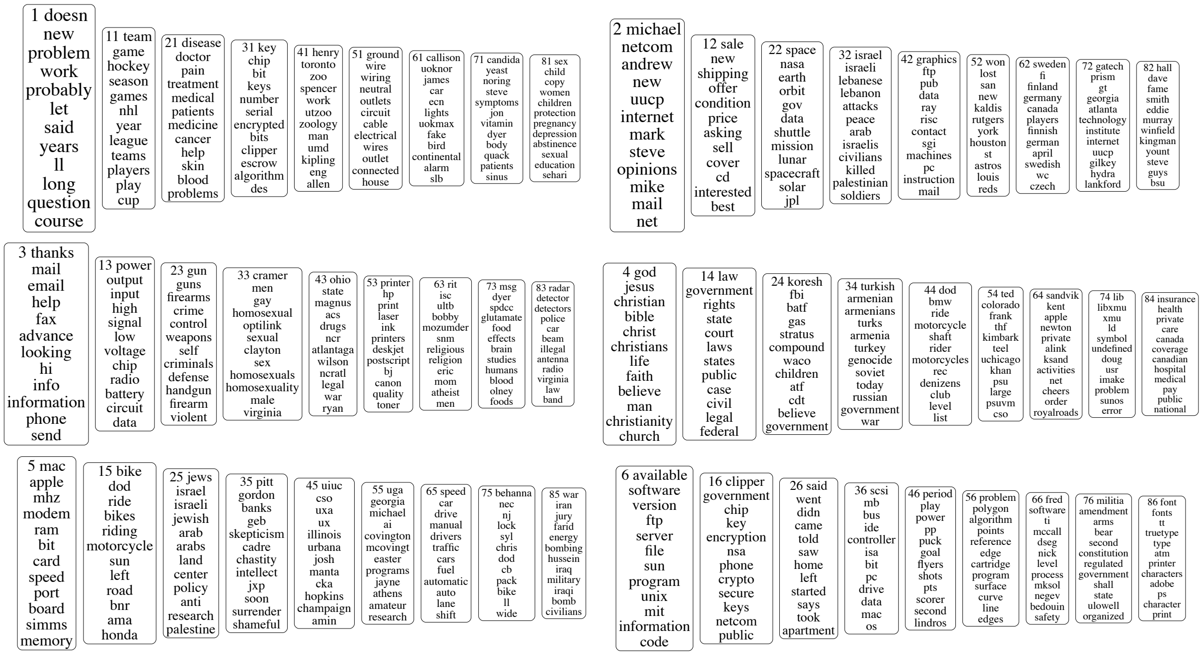

1 doesn new problem work probably let said years ll long question course
11 team game hockey season games nhl year league teams players play cup
21 disease doctor pain treatment medical patients medicine cancer help skin blood problems
31 key chip bit keys number serial encrypted bits clipper escrow algorithm des
41 henry toronto zoo spencer work utzoo zoology man umd kipling eng allen
51 ground wire wiring neutral outlets circuit cable electrical wires outlet connected house
61 callison uoknor james car ecn lights uokmax fake bird continental alarm slb
71 candida yeast noring steve symptoms jon vitamin dyer body quack patients sinus
81 sex child copy women children protection pregnancy depression abstinence sexual education sehari
2 michael netcom andrew new uucp internet mark steve opinions mike mail net
12 sale new shipping offer condition price asking sell cover cd interested best
22 space nasa earth orbit gov data shuttle mission lunar spacecraft solar jpl
32 israel israeli lebanese lebanon attacks peace arab israelis civilians killed palestinian soldiers
42 graphics ftp pub data ray risc contact sgi machines pc instruction mail
52 won lost san new kaldis rutgers york houston st astros louis reds
62 sweden fi finland germany canada players finnish german april swedish wc czech
72 gatech prism gt georgia atlanta technology institute internet uucp gilkey hydra lankford
82 hall dave fame smith eddie murray winfield kingman yount steve guys bsu
3 thanks mail email help fax advance looking hi info information phone send
13 power output input high signal low voltage chip radio battery circuit data
23 gun guns firearms crime control weapons self criminals defense handgun firearm violent
33 cramer men gay homosexual optilink sexual clayton sex homosexuals homosexuality male virginia
43 ohio state magnus acs drugs ncr atlantaga wilson ncratl legal war ryan
53 printer hp print laser ink printers deskjet postscript bj canon quality toner
63 rit isc ultb bobby mozumder snm religious religion eric mom atheist men
73 msg dyer spdcc glutamate food effects brain studies humans blood olney foods
83 radar detector detectors police car beam illegal antenna radio virginia law band
4 god jesus christian bible christ christians life faith believe man christianity church
14 law government rights state court laws states public case civil legal federal
24 koresh fbi batf gas stratus compound waco children atf cdt believe government
34 turkish armenian armenians turks armenia turkey genocide soviet today russian government war
44 dod bmw ride motorcycle shaft rider motorcycles rec denizens club level list
54 ted colorado frank thf kimbark teel uchicago khan psu large psuvm cso
64 sandvik kent apple newton private alink ksand activities net cheers order royalroads
74 lib libxmu xmu ld symbol undefined doug usr imake problem sunos error
84 insurance health private care canada coverage canadian hospital medical pay public national
5 mac apple mhz modem ram bit card speed port board simms memory
15 bike dod ride bikes riding motorcycle sun left road bnr ama honda
25 jews israel israeli jewish arab arabs land center policy anti research palestine
35 pitt gordon banks geb skepticism cadre chastity intellect jxp soon surrender shameful
45 uiuc cso uxa ux illinois urbana josh manta cka hopkins champaign amin
55 uga georgia michael ai covington mcovingt easter programs jayne athens amateur research
65 speed car drive manual drivers traffic cars fuel automatic auto lane shift
75 behanna nec nj lock syl chris dod cb pack bike ll wide
85 war iran jury farid energy bombing hussein iraq military iraqi bomb civilians
6 available software version ftp server file sun program unix mit information code
16 clipper government chip key encryption nsa phone crypto secure keys netcom public
26 said went didn came told saw home left started says took apartment
36 scsi mb bus ide controller isa bit pc drive data mac os
46 period play power pp puck goal flyers shots pts scorer second lindros
56 problem polygon algorithm points reference edge cartridge program surface curve line edges
66 fred software ti mccall dseg nick level process mksol negev bedouin safety
76 militia amendment arms bear second constitution regulated government shall state ulowell organized
86 font fonts tt truetype type atm printer characters adobe ps character print

Figure 3: Example topics of layer one of the PGBN learned on the 20newsgroups corpus.

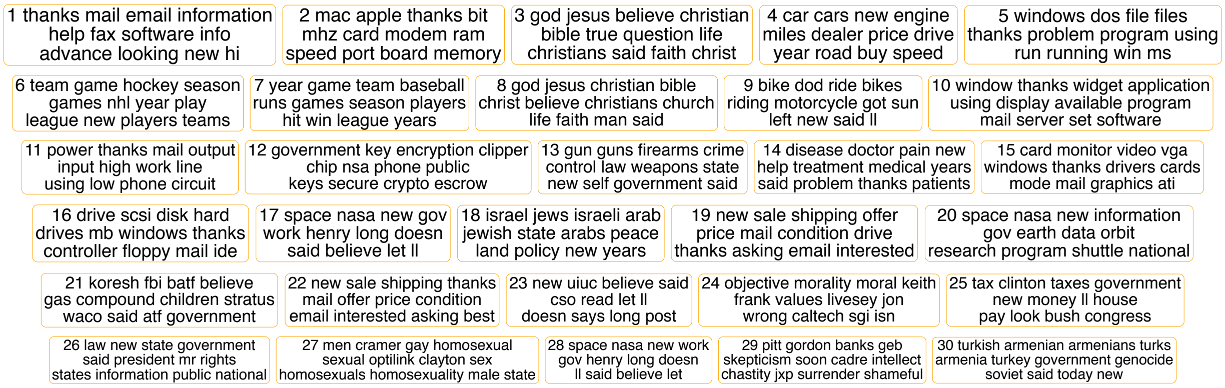

1 thanks mail email information help fax software info advance looking new hi
2 mac apple thanks bit mhz card modem ram speed port board memory
3 god jesus believe christian bible true question life christians said faith christ
4 car cars new engine miles dealer price drive year road buy speed
5 windows dos file files thanks problem program using run running win ms
6 team game hockey season games nhl year play league new players teams
7 year game team baseball runs games season players hit win league years
8 god jesus christian bible christ believe christians church life faith man said
9 bike dod ride bikes riding motorcycle got sun left new said ll
10 window thanks widget application using display available program mail server set software
11 power thanks mail output input high work line using low phone circuit
12 government key encryption clipper chip nsa phone public keys secure crypto escrow
13 gun guns firearms crime control law weapons state new self government said
14 disease doctor pain new help treatment medical years said problem thanks patients
15 card monitor video vga windows thanks drivers cards mode mail graphics ati
16 drive scsi disk hard drives mb windows thanks controller floppy mail ide
17 space nasa new gov work henry long doesn said believe let ll
18 israel jews israeli arab jewish state arabs peace land policy new years
19 new sale shipping offer price mail condition drive thanks asking email interested
20 space nasa new information gov earth data orbit research program shuttle national
21 koresh fbi batf believe gas compound children stratus waco said atf government
22 new sale shipping thanks mail offer price condition email interested asking best
23 new uiuc believe said cso read let ll doesn says long post
24 objective morality moral keith frank values livesey jon wrong caltech sgi isn
25 tax clinton taxes government new money ll house pay look bush congress
26 law new state government said president mr rights states information public national
27 men cramer gay homosexual sexual optilink clayton sex homosexuals homosexuality male state
28 space nasa new work gov henry long doesn ll said believe let
29 pitt gordon banks geb skepticism soon cadre intellect chastity jxp surrender shameful
30 turkish armenian armenians turks armenia turkey government genocide soviet said today new

Figure 4: The top 30 topics of layer three of the PGBN learned on the 20newsgroups corpus.

For the $[386, 63, 58, 54, 51]$ PGBN learned on the 20newsgroups corpus, we plot 60 example topics of layer one, the top 30 topics of layer three, and the top 30 topics of layer five in Figures 3, 4, and 5, respectively. Figure 3 clearly shows that the topics of layer one, except for topics 1-3 of layer one that mainly consist of common functional words of the corpus, are all very specific. For example, topics 71 and 81 shown in the first row are about "candida yeast symptoms" and "sex," respectively, topics 53, 73, 83, and 84 shown in the second row are about "printer," "msg," "police radar detector," and "Canadian health care system," respectively, and topics 46 and 76 shown in third row are about "ice hockey" and "second amendment," respectively. By contrast, the topics of layer three, shown in Figure 4, and those of layer five, shown in Figure 5, are much less specific and

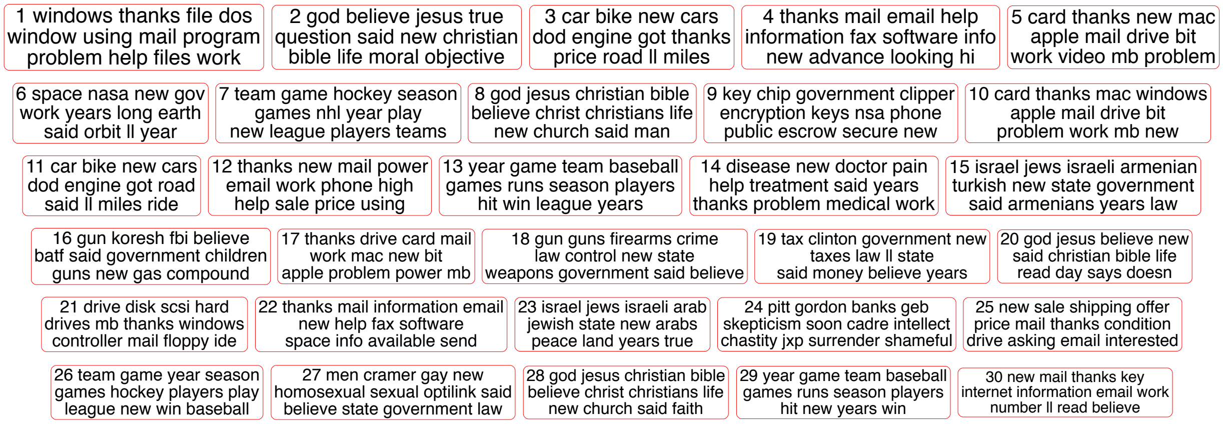

1 windows thanks file dos window using mail program problem help files work
2 god believe jesus true question said new christian bible life moral objective
3 car bike new cars dod engine got thanks price road ll miles
4 thanks mail email help information fax software info new advance looking hi
5 card thanks new mac apple mail drive bit work video mb problem
6 space nasa new gov work years long earth said orbit ll year
7 team game hockey season games nhl year play new league players teams
8 god jesus christian bible believe christ christians life new church said man
9 key chip government clipper encryption keys nsa phone public escrow secure new
10 card thanks mac windows apple mail drive bit problem work mb new
11 car bike new cars dod engine got road said ll miles ride
12 thanks new mail power email work phone high help sale price using
13 year game team baseball games runs season players hit win league years
14 disease new doctor pain help treatment said years thanks problem medical work
15 israel jews israeli armenian turkish new state government said armenians years law
16 gun koresh fbi believe batf said government children guns new gas compound
17 thanks drive card mail work mac new bit apple problem power mb
18 gun guns firearms crime law control new state weapons government said believe
19 tax clinton government new taxes law ll state said money believe years
20 god jesus believe new said christian bible life read day says doesn
21 drive disk scsi hard drives mb thanks windows controller mail floppy ide
22 thanks mail information email new help fax software space info available send
23 israel jews israeli arab jewish state new arabs peace land years true
24 pitt gordon banks geb skepticism soon cadre intellect chastity jxp surrender shameful
25 new sale shipping offer price mail thanks condition drive asking email interested
26 team game year season games hockey players play league new win baseball
27 men cramer gay new homosexual sexual optilink said believe state government law
28 god jesus christian bible believe christ christians life new church said faith
29 year game team baseball games runs season players hit new years win
30 new mail thanks key internet information email work number ll read believe

Figure 5: The top 30 topics of layer five of the PGBN learned on the 20newsgroups corpus.

can in general be matched to one or two news groups out of the 20 news groups, including comp.{graphics, os.ms-windows.misc, sys.ibm.pc.hardware, sys.mac.hardware, windows.x}, rec.{autos, motorcycles}, rec.sport.{baseball, hockey}, sci.{crypt, electronics, med, space,} misc.forsale, talk.politics.{misc, guns, mideast}, and {talk.religion.misc, alt.atheism, soc.religion.christian}.

### 3.2.2 Visualizing trees rooted at the top-layer hidden units

While it is interesting to examine the topics of different layers to understand the general and specific aspects of the corpus used to train the PGBN, it would be more informative to further illustrate how the topics of different layers are related to each other. Thus we consider constructing trees to visualize the PGBN. We first pick a node as the root of a tree and grow the tree downward by drawing a line from node $k$ at layer $t$, the root or a leaf node of the tree, to node $k'$ at layer $t-1$ for all $k'$ in the set $\{k' : \mathbf{\Phi}^{(t)}(k', k) > \tau_t/K_{t-1}\}$, where we set the width of the line connecting node $k$ of layer $t$ to node $k'$ of layer $t-1$ be proportional to $\sqrt{\mathbf{\Phi}^{(t)}(k', k)}$ and use $\tau_t$ to adjust the complexity of a tree. In general, increasing $\tau_t$ would discard more weak connections and hence make the tree simpler and easier to visualize.

We set $\tau_t = 3$ for all $t$ to visualize both a five-layer tree rooted at the top ranked node of the top hidden layer, as shown in Figure 6, and a five-layer tree rooted at the second ranked node of the top hidden layer, as shown in Figure 7. For the tree in Figure 6, while it is somewhat vague to determine the actual meanings of both node 1 of layer five and node 1 of layer four based on their top words, examining the more specific topics of layers three and two within the tree clearly indicate that this tree is primarily about "windows," "window system," "graphics," "information," and "software," which are relatively specific concepts that are all closely related to each other. The similarities and differences between the five

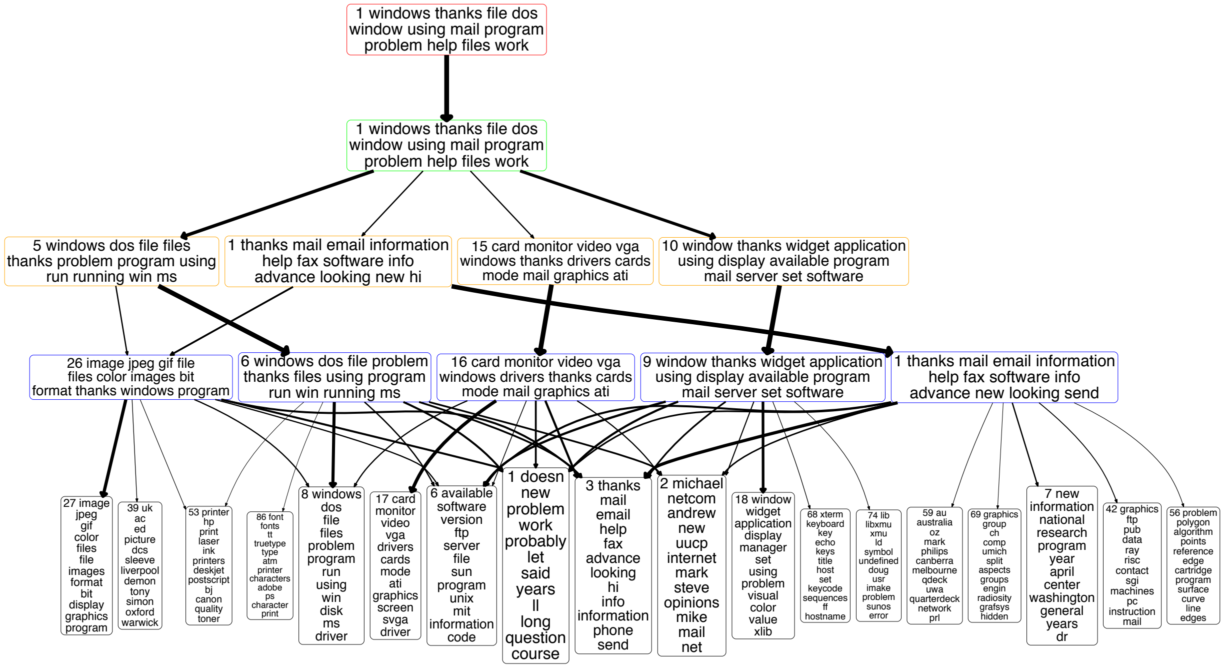

1 windows thanks file dos window using mail program problem help files work
1 windows thanks file dos window using mail program problem help files work
5 windows dos file files thanks problem program using run running win ms
1 thanks mail email information help fax software info advance looking new hi
15 card monitor video vga windows thanks drivers cards mode mail graphics ati
10 window thanks widget application using display available program mail server set software
26 image jpeg gif file files color images bit format thanks windows program
6 windows dos file problem thanks files using program run win running ms
16 card monitor video vga windows drivers thanks cards mode mail graphics ati
9 window thanks widget application using display available program mail server set software
1 thanks mail email information help fax software info advance new looking send
27 image jpeg gif color files file images format bit display graphics program
39 uk ac ed picture dcs sleeve liverpool demon tony simon oxford warwick
53 printer hp print laser ink printers deskjet postscript bj canon quality toner
86 font fonts tt truetype type atm printer characters adobe ps character print
8 windows dos file files problem program run using win disk ms driver
17 card monitor video vga drivers cards mode ati graphics screen svga driver
6 available software version ftp server file sun program unix mit information code
1 doesn new problem work probably let said years ll long question course
3 thanks mail email help fax advance looking hi info information phone send
2 michael netcom andrew new uucp internet mark steve opinions mike mail net
18 window widget application display manager set using problem visual color value xlib
68 xterm keyboard key echo keys title host set keycode sequences ff hostname
74 lib libxmu xmu ld symbol undefined doug usr imake problem sunos error
59 au australia oz mark philips canberra melbourne qdeck uwa quarterdeck network prl
69 graphics group ch comp umich split aspects groups engin radiosity grafsys hidden
7 new information national research program year april center washington general years dr
42 graphics ftp pub data ray risc contact sgi machines pc instruction mail
56 problem polygon algorithm points reference edge cartridge program surface curve line edges

Figure 6: A $[18, 5, 4, 1, 1]$ tree that includes all the lower-layer nodes (directly or indirectly) linked with non-negligible weights to the top ranked node of the top layer, taken from the full $[386, 63, 58, 54, 51]$ network inferred by the PGBN on the 11,269 training documents of the 20newsgroups corpus, with $\eta^{(t)} = 0.1$ for all $t$. A line from node $k$ at layer $t$ to node $k'$ at layer $t-1$ indicates that $\mathbf{\Phi}^{(t)}(k', k) > 3/K_{t-1}$, with the width of the line proportional to $\sqrt{\mathbf{\Phi}^{(t)}(k', k)}$. For each node, the rank (in terms of popularity) at the corresponding layer and the top 12 words of the corresponding topic are displayed inside the text box, where the text font size monotonically decreases as the popularity of the node decreases, and the outside border of the text box is colored as red, green, orange, blue, or black if the node is at layer five, four, three, two, or one, respectively.

nodes of layer two can be further understood by examining the nodes of layer one that are connected to them. For example, while nodes 26 and 16 of layer two share their connections to multiple nodes of layer one, node 27 of layer one on “image” is strongly connected to node 26 of layer two but not to node 16 of layer two, and node 17 of layer one on “video” is strongly connected to node 16 of layer two but not to node 26 of layer two.

Following the branches of each tree shown in both figures, it is clear that the topics become more and more specific when moving along the tree from the top to bottom. Taking the tree on “religion” shown in Figure 7 for example, the root node splits into two nodes when moving from layers five to four: while the left node is still mainly about “religion,” the right node is on “objective morality.” When moving from layers four to three, node 5 of layer four splits into a node about “Christian” and another node about “Islamic.” When moving from layers three to two, node 3 of layer three splits into a node about “God, Jesus, & Christian,” and another node about “science, atheism, & question of the existence of God.” When moving from layers two to one, all four nodes of layer two split into multiple

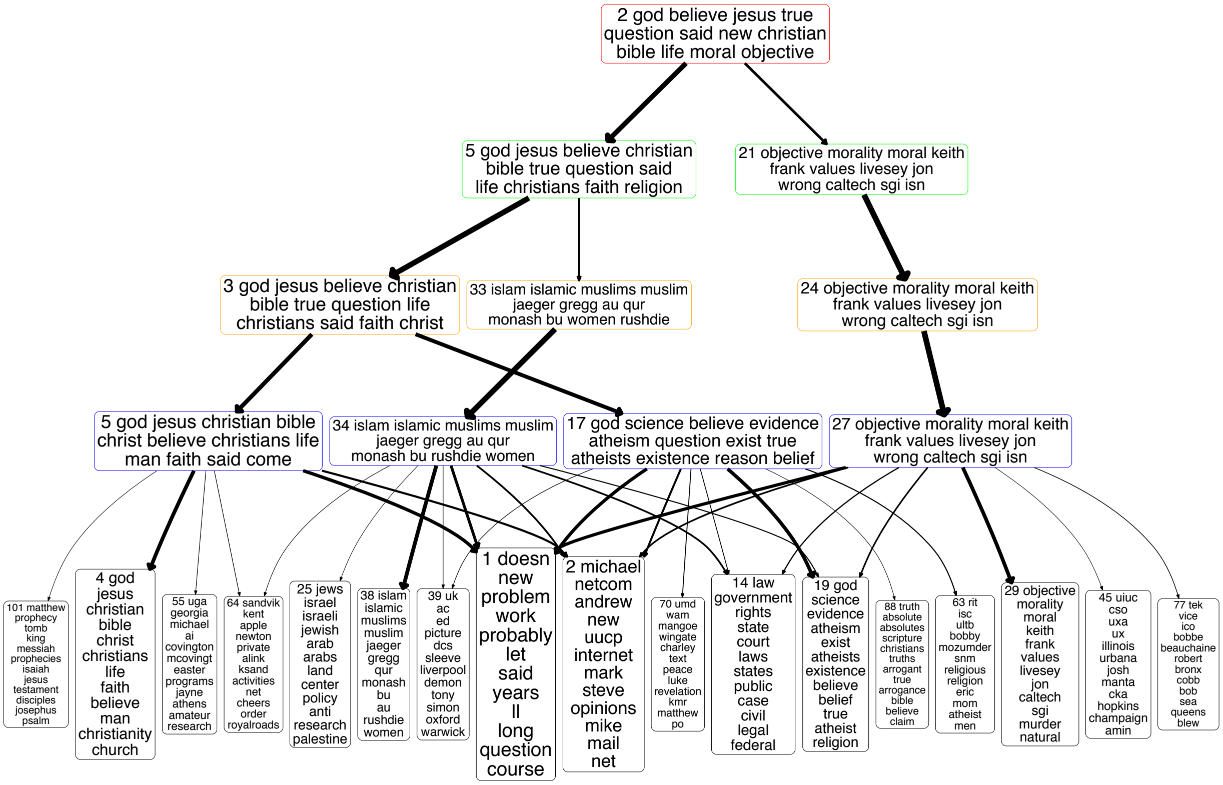

2 god believe jesus true question said new christian bible life moral objective
5 god jesus believe christian bible true question said life christians faith religion
21 objective morality moral keith frank values livesey jon wrong caltech sgi isn
3 god jesus believe christian bible true question life christians said faith christ
33 islam islamic muslims muslim jaeger gregg au qur monash bu women rushdie
24 objective morality moral keith frank values livesey jon wrong caltech sgi isn
5 god jesus christian bible christ believe christians life man faith said come
34 islam islamic muslims muslim jaeger gregg au qur monash bu rushdie women
17 god science believe evidence atheism question exist true atheists existence reason belief
27 objective morality moral keith frank values livesey jon wrong caltech sgi isn
101 matthew prophecy tomb king messiah prophecies isaiah jesus testament disciples josephus psalm
4 god jesus christian bible christ christians life faith believe man christianity church
55 uga georgia michael ai covington mcovingt easter programs jayne athens amateur research
64 sandvik kent apple newton private alink ksand activities net cheers order royalroads
25 jews israel israeli jewish arab arabs land center policy anti research palestine
38 islam islamic muslims muslim jaeger gregg qur monash bu au rushdie women
39 uk ac ed picture dcs sleeve liverpool demon tony simon oxford warwick
1 doesn new problem work probably let said years ll long question course
2 michael netcom andrew new uucp internet mark steve opinions mike mail net
70 umd wam mangoe wingate charley text peace luke revelation kmr matthew po
14 law government rights state court laws states public case civil legal federal
19 god science evidence atheism exist atheists existence believe belief true atheist religion
88 truth absolute absolutes scripture christians truths arrogant true arrogance bible believe claim
63 rit isc ultb bobby mozumder snm religious religion eric mom atheist men
29 objective morality moral keith frank values livesey jon caltech sgi murder natural
45 uiuc cso uxa ux illinois urbana josh manta cka hopkins champaign amin
77 tek vice ico bobbe beauchaine robert bronx cobb bob sea queens blew

Figure 7: Analogous plot to Figure 6 for a tree on "religion," rooted at node 2 of the top-layer.

topics, and they are all strongly connected to both topics 1 and 2 of layer one, whose top words are those that appear frequently in the 20newsgroups corpus.

### 3.2.3 Visualizing subnetworks consisting of related trees

Examining the top-layer topics shown in Figure 5, one may find that some of the nodes seem to be closely related to each other. For example, topics 3 and 11 share eleven words out of the top twelve ones; topics 15 and 23 both have "Israel" and "Jews" as their top two words; topics 16 and 18 are both related to "gun;" and topics 7, 13, and 26 all share "team(s)," "game(s)," "player(s)," "season," and "league."

To understand the relationships and distinctions between these related nodes, we construct subnetworks that include the trees rooted at them, as shown in Figures 8-11. It is clear from Figure 8 that the top-layer topic 3 differs from topic 11 in that it is not only strongly connected to topic 2 of layer four on"car & bike," but also has a non-negligible connection to topic 27 of layer four on "sales." It is clear from Figure 9 that topic 15 differs from topic 23 in that it is not only about "Israel & Arabs," but also about "Israel, Armenia, & Turkey." It is clear from Figure 10 in that topic 16 differs from topic 18 in that it is mainly about Waco siege happened in 1993 involving David Koresh, the Federal Bureau of

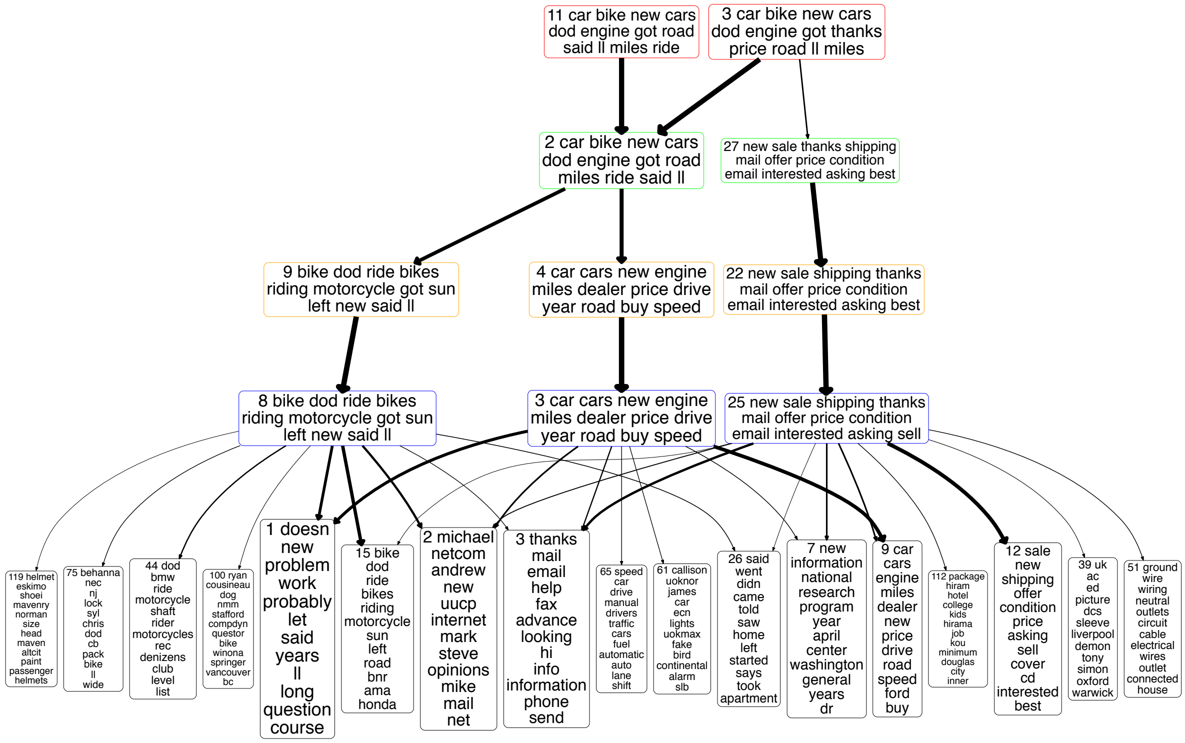

11 car bike new cars dod engine got road said ll miles ride
3 car bike new cars dod engine got thanks price road ll miles
2 car bike new cars dod engine got road miles ride said ll
27 new sale thanks shipping mail offer price condition email interested asking best
9 bike dod ride bikes riding motorcycle got sun left new said ll
4 car cars new engine miles dealer price drive year road buy speed
22 new sale shipping thanks mail offer price condition email interested asking best
8 bike dod ride bikes riding motorcycle got sun left new said ll
3 car cars new engine miles dealer price drive year road buy speed
25 new sale shipping thanks mail offer price condition email interested asking sell
119 helmet eskimo shoei mavenry norman size head maven altcit paint passenger helmets
75 behanna nec nj lock syl chris dod cb pack bike ll wide
44 dod bmw ride motorcycle shaft rider motorcycles rec denizens club level list
100 ryan cousineau dog nmm stafford compdyn questor bike winona springer vancouver bc
1 doesn new problem work probably let said years ll long question course
15 bike dod ride bikes riding motorcycle sun left road bnr ama honda
2 michael netcom andrew new uucp internet mark steve opinions mike mail net
3 thanks mail email help fax advance looking hi info information phone send
65 speed car drive manual drivers traffic cars fuel automatic auto lane shift
61 callison uoknor james car ecn lights uokmax fake bird continental alarm slb
26 said went didn came told saw home left started says took apartment
7 new information national research program year april center washington general years dr
9 car cars engine miles dealer new price drive road speed ford buy
112 package hiram hotel college kids hirama job kou minimum douglas city inner
12 sale new shipping offer condition price asking sell cover cd interested best
39 uk ac ed picture dcs sleeve liverpool demon tony simon oxford warwick
51 ground wire wiring neutral outlets circuit cable electrical wires outlet connected house

Figure 8: Analogous plots to Figure 6 for a subnetwork on "car & bike", consisting of the tree rooted at node 11 of layer one and the tree rooted at node 3 of layer one.

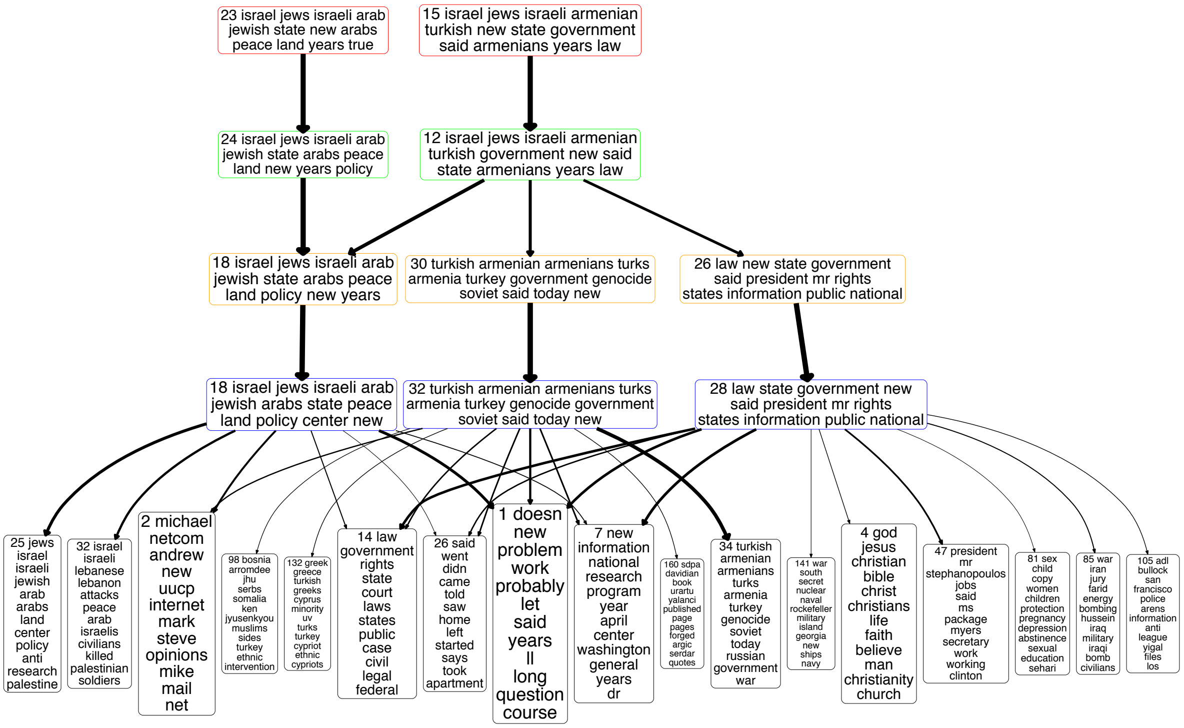

23 israel jews israeli arab jewish state new arabs peace land years true
15 israel jews israeli armenian turkish new state government said armenians years law
24 israel jews israeli arab jewish state arabs peace land new years policy
12 israel jews israeli armenian turkish government new said state armenians years law
18 israel jews israeli arab jewish state arabs peace land policy new years
30 turkish armenian armenians turks armenia turkey government genocide soviet said today new
26 law new state government said president mr rights states information public national
18 israel jews israeli arab jewish arabs state peace land policy center new
32 turkish armenian armenians turks armenia turkey genocide government soviet said today new
28 law state government new said president mr rights states information public national
25 jews israel israeli jewish arab arabs land center policy anti research palestine
32 israel israeli lebanese lebanon attacks peace arab israelis civilians killed palestinian soldiers
2 michael netcom andrew new uucp internet mark steve opinions mike mail net
98 bosnia arromdee jhu serbs somalia ken jyusenkyou muslims sides turkey ethnic intervention
132 greek greece turkish greeks cyprus minority uv turks turkey cypriot ethnic cypriots
14 law government rights state court laws states public case civil legal federal
26 said went didn came told saw home left started says took apartment
1 doesn new problem work probably let said years ll long question course
7 new information national research program year april center washington general years dr
160 sdpa davidian book urartu yalanci published page pages forged argic serdar quotes
34 turkish armenian armenians turks armenia turkey genocide soviet today russian government war
141 war south secret nuclear naval rockefeller military island georgia new ships navy
4 god jesus christian bible christ christians life faith believe man christianity church
47 president mr stephanopoulos jobs said ms package myers secretary work working clinton
81 sex child copy women children protection pregnancy depression abstinence sexual education sehari
85 war iran jury farid energy bombing hussein iraq military iraqi bomb civilians
105 adl bullock san francisco police arens information anti league yigal files los

Figure 9: Analogous plot to Figure 6 for a subnetwork on "Middle East."

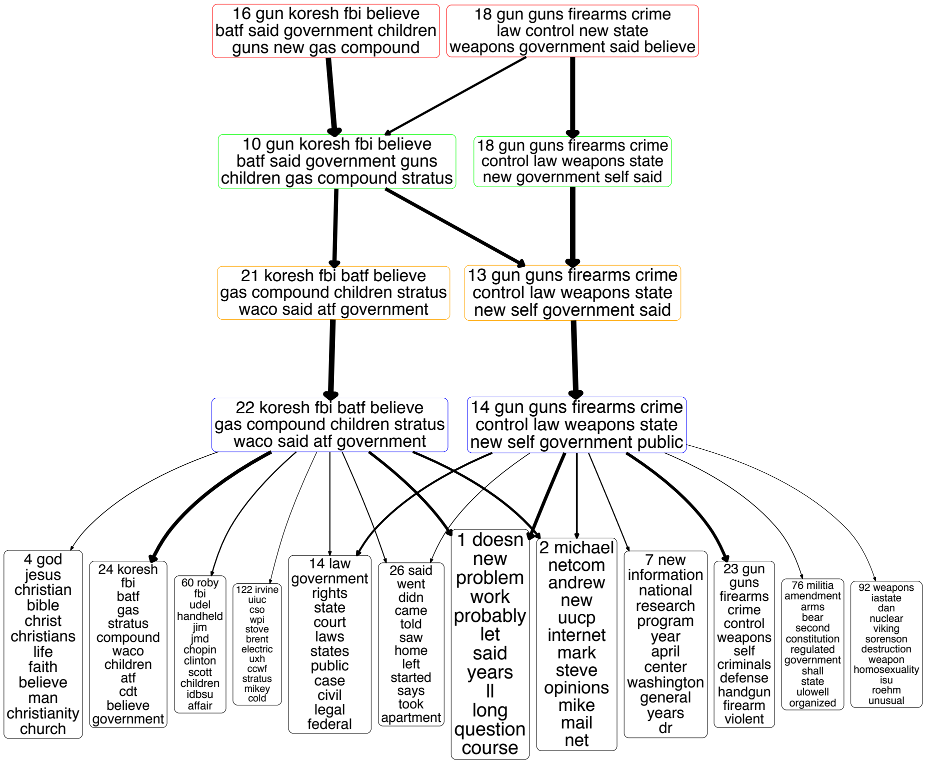

16 gun koresh fbi believe batf said government children guns new gas compound
18 gun guns firearms crime law control new state weapons government said believe
10 gun koresh fbi believe batf said government guns children gas compound stratus
18 gun guns firearms crime control law weapons state new government self said
21 koresh fbi batf believe gas compound children stratus waco said atf government
13 gun guns firearms crime control law weapons state new self government said
22 koresh fbi batf believe gas compound children stratus waco said atf government
14 gun guns firearms crime control law weapons state new self government public
4 god jesus christian bible christ christians life faith believe man christianity church
24 koresh fbi batf gas stratus compound waco children atf cdt believe government
60 roby fbi udel handheld jim jmd chopin clinton scott children idbsu affair
122 irvine uiuc cso wpi stove brent electric uxh ccwf stratus mikey cold
14 law government rights state court laws states public case civil legal federal
26 said went didn came told saw home left started says took apartment
1 doesn new problem work probably let said years ll long question course
2 michael netcom andrew new uucp internet mark steve opinions mike mail net
7 new information national research program year april center washington general years dr
23 gun guns firearms crime control weapons self criminals defense handgun firearm violent
76 militia amendment arms bear second constitution regulated government shall state ulowell organized
92 weapons iastate dan nuclear viking sorenson destruction weapon homosexuality isu roehm unusual

Figure 10: Analogous plot to Figure 6 for a subnetwork related to “gun.”

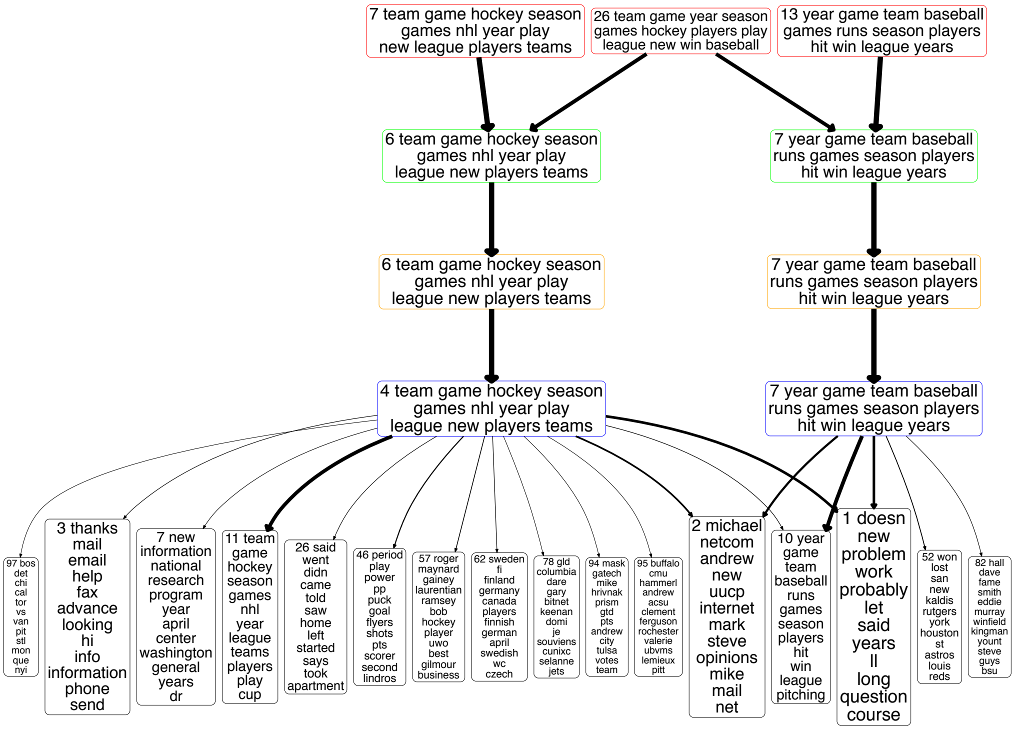

7 team game hockey season games nhl year play new league players teams
26 team game year season games hockey players play league new win baseball
13 year game team baseball games runs season players hit win league years
6 team game hockey season games nhl year play league new players teams
7 year game team baseball runs games season players hit win league years
6 team game hockey season games nhl year play league new players teams
7 year game team baseball runs games season players hit win league years
4 team game hockey season games nhl year play league new players teams
7 year game team baseball runs games season players hit win league years
97 bos det chi cal tor vs van pit stl mon que nyi
3 thanks mail email help fax advance looking hi info information phone send
7 new information national research program year april center washington general years dr
11 team game hockey season games nhl year league teams players play cup
26 said went didn came told saw home left started says took apartment
46 period play power pp puck goal flyers shots pts scorer second lindros
57 roger maynard gainey laurentian ramsey bob hockey player uwo best gilmour business
62 sweden fi finland germany canada players finnish german april swedish wc czech
78 gld columbia dare gary bitnet keenan domi je souviens cunixc selanne jets
94 mask gatech mike hrivnak prism gtd pts andrew city tulsa votes team
95 buffalo cmu hammerl andrew acsu clement ferguson rochester valerie ubvms lemieux pitt
2 michael netcom andrew new uucp internet mark steve opinions mike mail net
10 year game team baseball runs games season players hit win league pitching
1 doesn new problem work probably let said years ll long question course
52 won lost san new kaldis rutgers york houston st astros louis reds
82 hall dave fame smith eddie murray winfield kingman yount steve guys bsu

Figure 11: Analogous plot to Figure 6 for a subnetwork on “ice hockey” and “baseball”.

Investigation (FBI), and the Bureau of Alcohol, Tobacco, Firearms and Explosives (BATF). It is clear from Figure 11 that topics 7 and 13 are mainly about "ice hockey" and "baseball," respectively, and topic 26 is a mixture of both.

### 3.2.4 Capturing correlations between nodes

For the GBN, as in (25), given the weight vector $\boldsymbol{\theta}_j^{(1)}$, we have

$$\mathbb{E}\big[\boldsymbol{x}_j^{(1)}\,\big|\,\boldsymbol{\Phi}^{(1)},\boldsymbol{\theta}_j^{(1)}\big] = \boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}. \tag{22}$$

A distinction between a shallow GBN with $T=1$ hidden layers and a deep GBN with $T\geq 2$ hidden layers is that the prior for $\boldsymbol{\theta}_j^{(1)}$ changes from $\boldsymbol{\theta}_j^{(1)}\sim \mathrm{Gam}(\boldsymbol{r},1/c_j^{(2)})$ for $T=1$ to $\boldsymbol{\theta}_j^{(1)}\sim \mathrm{Gam}(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)},1/c_j^{(2)})$ for $T\geq 2$. For the GBN with $T=1$, given the shared weight vector $\boldsymbol{r}$, we have

$$\mathbb{E}\big[\boldsymbol{x}_j^{(1)}\,\big|\,\boldsymbol{\Phi}^{(1)},\boldsymbol{r}\big] = \boldsymbol{\Phi}^{(1)}\boldsymbol{r}/c_j^{(2)}; \tag{23}$$

for the GBN with $T=2$, given the shared weight vector $\boldsymbol{r}$, we have

$$\mathbb{E}\big[\boldsymbol{x}_j^{(1)}\,\big|\,\boldsymbol{\Phi}^{(1)},\boldsymbol{\Phi}^{(2)},\boldsymbol{r}\big] = \boldsymbol{\Phi}^{(1)}\boldsymbol{\Phi}^{(2)}\boldsymbol{r}\Big/\Big(c_j^{(2)}c_j^{(3)}\Big); \tag{24}$$

and for the GBN with $T\geq 2$, given the weight vector $\boldsymbol{\theta}_j^{(2)}$, we have

$$\mathbb{E}\big[\boldsymbol{x}_j^{(1)}\,\big|\,\boldsymbol{\Phi}^{(1)},\boldsymbol{\Phi}^{(2)},\boldsymbol{\theta}_j^{(2)}\big] = \boldsymbol{\Phi}^{(1)}\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)}/c_j^{(2)}. \tag{25}$$

Thus in the prior, the co-occurrence patterns of the columns of $\boldsymbol{\Phi}^{(1)}$ are modeled by only a single vector $\boldsymbol{r}$ when $T=1$, but are captured in the columns of $\boldsymbol{\Phi}^{(2)}$ when $T\geq 2$. Similarly, in the prior, if $T\geq t+1$, the co-occurrence patterns of the $K_t$ columns of the projected topics $\prod_{\ell=1}^{t}\boldsymbol{\Phi}^{(\ell)}$ will be captured in the columns of the $K_t\times K_{t+1}$ matrix $\boldsymbol{\Phi}^{(t+1)}$.

To be more specific, we show in Figure 12 three example trees rooted at three different nodes of layer three, where we lower the threshold to $\tau_t=1$ to reveal more weak links between the nodes of adjacent layers. The top subplot reveals that, in addition to strongly co-occurring with the top two topics of layer one, topic 21 of layer one on "medicine" tends to co-occur not only with topics 7, 21, and 26, which are all common topics that frequently appear, but also with some much less common topics that are related to very specific diseases or symptoms, such as topic 67 on "msg" and "Chinese restaurant syndrome," topic 73 on "candida yeast symptoms," and topic 180 on "acidophilous" and "astemizole (hismanal)."

The middle subplot reveals that topic 31 of layer two on "encryption & cryptography" tends to co-occur with topic 13 of layer two on "government & encryption," and it also

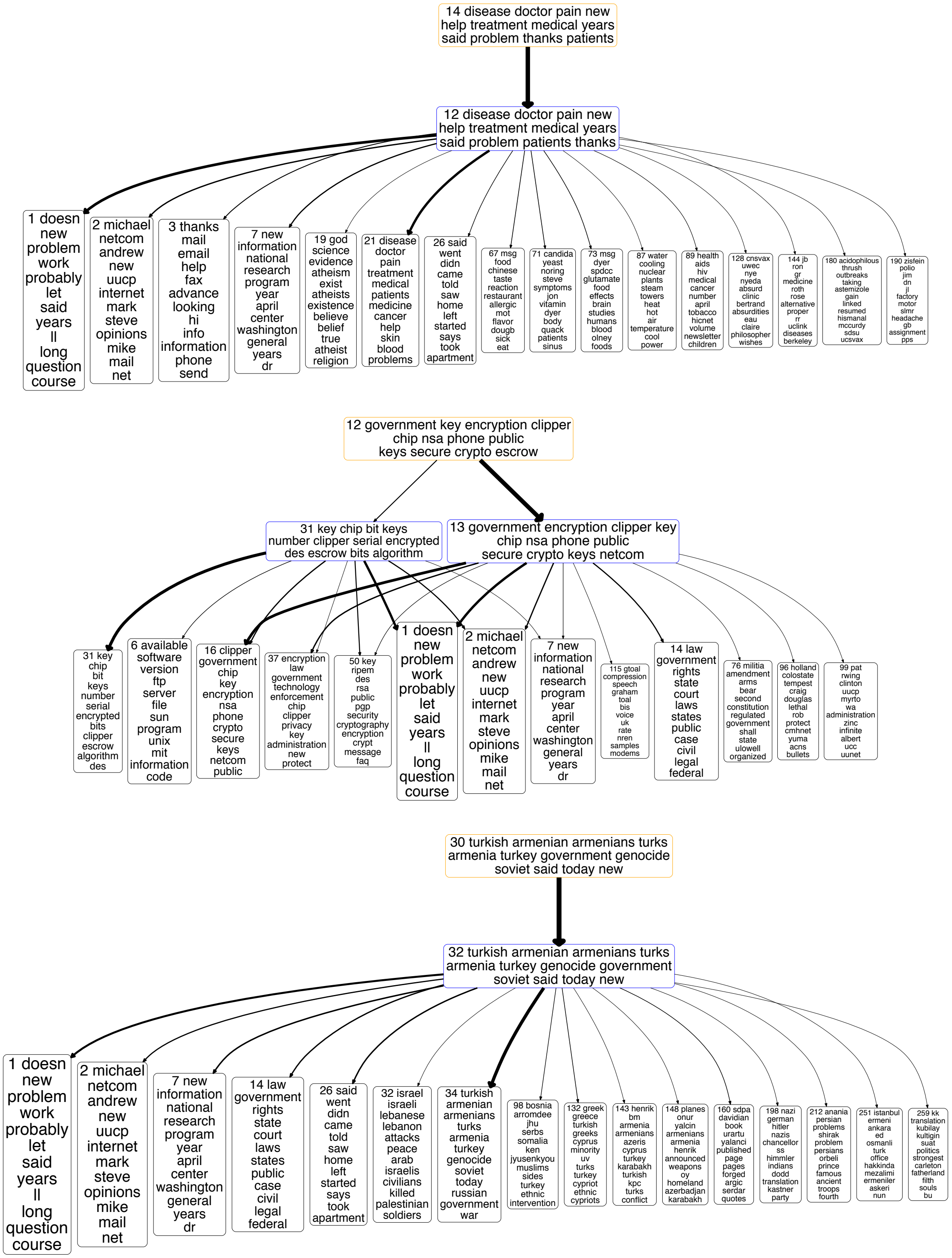

14 disease doctor pain new help treatment medical years said problem thanks patients
12 disease doctor pain new help treatment medical years said problem patients thanks
1 doesn new problem work probably let said years ll long question course
2 michael netcom andrew new uucp internet mark steve opinions mike mail net
3 thanks mail email help fax advance looking hi info information phone send
7 new information national research program year april center washington general years dr
19 god science evidence atheism exist atheists existence believe belief true atheist religion
21 disease doctor pain treatment medical patients medicine cancer help skin blood problems
26 said went didn came told saw home left started says took apartment
67 msg food chinese taste reaction restaurant allergic mot flavor dougb sick eat
71 candida yeast noring steve symptoms jon vitamin dyer body quack patients sinus
73 msg dyer spdcc glutamate food effects brain studies humans blood olney foods
87 water cooling nuclear plants steam towers heat hot air temperature cool power
89 health aids hiv medical cancer number april tobacco hicnet volume newsletter children
128 cnsvax uwec nye nyeda absurd clinic bertrand absurdities eau claire philosopher wishes
144 jb ron gr medicine roth rose alternative proper rr uclink diseases berkeley
180 acidophilous thrush outbreaks taking astemizole gain linked resumed hismanal mccurdy sdsu ucsvax
190 zisfein polio jim dn jl factory motor slmr headache gb assignment pps
12 government key encryption clipper chip nsa phone public keys secure crypto escrow
31 key chip bit keys number clipper serial encrypted des escrow bits algorithm
13 government encryption clipper key chip nsa phone public secure crypto keys netcom
31 key chip bit keys number serial encrypted bits clipper escrow algorithm des
6 available software version ftp server file sun program unix mit information code
16 clipper government chip key encryption nsa phone crypto secure keys netcom public
37 encryption law government technology enforcement chip clipper privacy key administration new protect
50 key ripem des rsa public pgp security cryptography encryption crypt message faq
1 doesn new problem work probably let said years ll long question course
2 michael netcom andrew new uucp internet mark steve opinions mike mail net
7 new information national research program year april center washington general years dr
115 gtoal compression speech graham toal bis voice uk rate nren samples modems
14 law government rights state court laws states public case civil legal federal
76 militia amendment arms bear second constitution regulated government shall state ulowell organized
96 holland colostate tempest craig douglas lethal rob protect cmhnet yuma acns bullets
99 pat rwing clinton uucp myrto wa administration zinc infinite albert ucc uunet
30 turkish armenian armenians turks armenia turkey government genocide soviet said today new
32 turkish armenian armenians turks armenia turkey genocide government soviet said today new
1 doesn new problem work probably let said years ll long question course
2 michael netcom andrew new uucp internet mark steve opinions mike mail net
7 new information national research program year april center washington general years dr
14 law government rights state court laws states public case civil legal federal
26 said went didn came told saw home left started says took apartment
32 israel israeli lebanese lebanon attacks peace arab israelis civilians killed palestinian soldiers
34 turkish armenian armenians turks armenia turkey genocide soviet today russian government war
98 bosnia arromdee jhu serbs somalia ken jyusenkyou muslims sides turkey ethnic intervention
132 greek greece turkish greeks cyprus minority uv turks turkey cypriot ethnic cypriots
143 henrik bm armenia armenians azeris cyprus turkey karabakh turkish kpc turks conflict
148 planes onur yalcin armenians armenia henrik announced weapons oy homeland azerbadjan karabakh
160 sdpa davidian book urartu yalanci published page pages forged argic serdar quotes
198 nazi german hitler nazis chancellor ss himmler indians dodd translation kastner party
212 anania persian problems shirak problem persians orbeli prince famous ancient troops fourth
251 istanbul ermeni ankara ed osmanli turk office hakkinda mezalimi ermeniler askeri nun
259 kk translation kubilay kultigin suat politics strongest carleton fatherland filth souls bu

Figure 12: Analogous plots to Figure 6, with $\tau_t = 1$ to reveal more weak links. Top: the tree rooted at node 14 of layer three on "medicine." Middle: the tree rooted at node 12 of layer three on "encryption." Bottom: the tree rooted at node 30 of layer three on "Turkey & Armenia."

indicates that topic 31 of layer one is more purely about "encryption" and more isolated from "government" in comparison to the other topics of layer one.

The bottom subplot reveals that in layer one, topic 14 on "law & government," topic 32 on "Israel & Lebanon," topic 34 on "Turkey, Armenia, Soviet Union, & Russian," topic 132 on "Greece, Turkey, & Cyprus," topic 98 on "Bosnia, Serbs, & Muslims," topic 143 on "Armenia, Azeris, Cyprus, Turkey, & Karabakh," and several other very specific topics related to Turkey and/or Armenia all tend to co-occur with each other.

We note that capturing the co-occurrence patterns between the topics not only helps exploratory data analysis, but also helps extract better features for classification in an unsupervised manner and improves prediction for held-out data, as will be demonstrated in detail in Section 5.

## 3.3 Related models

The structure of the GBN resembles the sigmoid belief network and the recently proposed deep exponential family model (Ranganath et al., 2015). Such kind of gamma distribution based network and its inference procedure were vaguely hinted in Corollary 2 of Zhou and Carin (2015), and had been exploited by Acharya et al. (2015) to develop a gamma Markov chain to model the temporal evolution of the factor scores of a dynamic count matrix, but have not yet been investigated for extracting multilayer data representations.

### 3.3.1 Sigmoid and deep belief networks

Under the hierarchical model in (12), given the connection weight matrices, the joint distribution of the observed/latent counts and gamma hidden units of the GBN can be expressed, similar to those of the sigmoid and deep belief networks (Bengio et al., 2015), as

$$P\left(\boldsymbol{x}_j^{(1)}, \{\boldsymbol{\theta}_j^{(t)}\}_t \,\middle|\, \{\boldsymbol{\Phi}^{(t)}\}_t\right) = P\left(\boldsymbol{x}_j^{(1)} \,\middle|\, \boldsymbol{\Phi}^{(1)}, \boldsymbol{\theta}_j^{(1)}\right) \left[\prod_{t=1}^{T-1} P\left(\boldsymbol{\theta}_j^{(t)} \,\middle|\, \boldsymbol{\Phi}^{(t+1)}, \boldsymbol{\theta}_j^{(t+1)}\right)\right] P\left(\boldsymbol{\theta}_j^{(T)}\right).$$

With $\boldsymbol{\phi}_{v:}$ representing the $v$th row $\boldsymbol{\Phi}$, for the gamma hidden units $\theta_{vj}^{(t)}$ we have

$$P\left(\theta_{vj}^{(t)} \,\middle|\, \boldsymbol{\phi}_{v:}^{(t+1)}, \boldsymbol{\theta}_j^{(t+1)}, c_{j+1}^{(t+1)}\right) = \frac{\left(c_{j+1}^{(t+1)}\right)^{\boldsymbol{\phi}_{v:}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}}}{\Gamma\left(\boldsymbol{\phi}_{v:}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}\right)} \left(\theta_{vj}^{(t)}\right)^{\boldsymbol{\phi}_{v:}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}-1} e^{-c_{j+1}^{(t+1)}\theta_{vj}^{(t)}}, \tag{26}$$

which are highly nonlinear functions that are strongly desired in deep learning. By contrast, with the sigmoid function $\sigma(x) = 1/(1 + e^{-x})$ and bias terms $b_v^{(t+1)}$, a sigmoid/deep belief network would connect the binary hidden units $\theta_{vj}^{(t)} \in \{0,1\}$ of layer $t$ (for deep belief networks, $t < T-1$ ) to the product of the connection weights and binary hidden units of the next layer with

$$P\left(\theta_{vj}^{(t)} = 1 \middle| \boldsymbol{\phi}_{v:}^{(t+1)}, \boldsymbol{\theta}_j^{(t+1)}, b_v^{(t+1)}\right) = \sigma\left(b_v^{(t+1)} + \boldsymbol{\phi}_{v:}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}\right). \tag{27}$$

Comparing (26) with (27) clearly shows the distinctions between the gamma distributed nonnegative hidden units and the sigmoid link function based binary hidden units. As a new alternative to binary hidden units, the gamma distributed nonnegative real hidden units have the potential to carry richer information and model more complex nonlinearities given the same network structure. Note that the rectified linear units have emerged as powerful alternatives of sigmoid units to introduce nonlinearity (Nair and Hinton, 2010). It would be interesting to investigate whether the gamma units can be used to introduce nonlinearity into the positive region of the rectified linear units.

### 3.3.2 Deep Poisson factor analysis

With $T = 1$, the PGBN specified by (12)-(14) and (15) reduces to Poisson factor analysis (PFA) using the (truncated) gamma-negative binomial process (Zhou and Carin, 2015), with a truncation level of $K_1$. As discussed in (Zhou and Carin, 2015, Zhou et al., 2012), with priors imposed on neither $\boldsymbol{\phi}_k^{(1)}$ nor $\boldsymbol{\theta}_j^{(1)}$, PFA is related to nonnegative matrix factorization (Lee and Seung, 2001), and with the Dirichlet priors imposed on both $\boldsymbol{\phi}_k^{(1)}$ and $\boldsymbol{\theta}_j^{(1)}$, PFA is related to latent Dirichlet allocation (Blei et al., 2003).

Related to the PGBN and the dynamic model in (Acharya et al., 2015), the deep exponential family model of Ranganath et al. (2015) also considers a gamma chain under Poisson observations, but it is the gamma scale parameters that are chained and factorized, which allows learning the network parameters using black box variational inference (Ranganath et al., 2014). In the proposed PGBN, we chain the gamma random variables via the gamma shape parameters. Both strategies worth through investigation. We prefer chaining the shape parameters in this paper, which leads to efficient upward-downward Gibbs sampling via data augmentation and makes it clear how the latent counts are propagated across layers, as discussed in detail in the following sections. The sigmoid belief network has also been recently incorporated into PFA for deep factorization of count data (Gan et al., 2015), however, that deep structure captures only the correlations between binary factor usage patterns but not the full connection weights. In addition, neither Ranganath et al. (2015) nor Gan et al. (2015) provide a principled way to learn the network structure, whereas the proposed GBN uses the gamma negative binomial process together with a greedy layer-wise training strategy to automatically infer the widths of the hidden layers, which will be described in Section 4.4.

### 3.3.3 Correlated and tree-structured topic models

The PGBN with $T = 2$ can also be related to correlated topic models (Blei and Lafferty, 2006, Chen et al., 2013, Linderman et al., 2015, Paisley et al., 2012, Ranganath and Blei, 2015), which typically use the logistic normal distributions to replace the topic-proportion Dirichlet distributions used in latent Dirichlet allocation (Blei et al., 2003), capturing the co-occurrence patterns between the topics in the latent Gaussian space using a covariance matrix. By contrast, the PGBN factorizes the topic usage weights (not proportions) under the gamma likelihood, capturing the co-occurrence patterns between the topics of the first layer (i.e., the columns of $\mathbf{\Phi}^{(1)}$) in the columns of $\mathbf{\Phi}^{(2)}$, the latent weight matrix connecting the hidden units of layers two and one. For the PGBN, the computation does not involve matrix inversion and scales linearly with the number of topics, hence it is suitable to be used to capture the correlations between hundreds of or thousands of topics.

As in Figures 7-12, trees and subnetworks can be extracted from the inferred deep network to visualize the data. Tree-structured topic models have also been proposed before, such as those in Blei et al. (2010), Adams et al. (2010), and Paisley et al. (2015), but they usually artificially impose the tree structures to be learned, whereas the PGBN learns a directed network, from which trees and subnetworks can be extracted for visualization, without the need to specify the number of nodes per layer, restrict the number of branches per node, and forbid a node to have multiple parents.

# 4 Model Properties and Inference

In this section, we break the inference of the GBN of $T$ hidden layers into $T$ related subproblems, each of which is solved with the same subroutine. Thus for implementation, it is straightforward for the GBN to adjust its depth $T$.

## 4.1 The upward propagation of latent counts

**Lemma 1** (Augment-and-conquer the gamma belief network)**.** *With* $p_j^{(1)} := 1 - e^{-1}$ *and*

$$p_j^{(t+1)} := -\ln(1 - p_j^{(t)}) \Big/ \left[c_j^{(t+1)} - \ln(1 - p_j^{(t)})\right] \tag{28}$$

*for $t = 1, \ldots, T$, one may connect the observed or latent counts $\boldsymbol{x}_j^{(t)} \in \mathbb{Z}^{K_{t-1}}$ to the product $\boldsymbol{\Phi}^{(t)}\boldsymbol{\theta}_j^{(t)}$ at layer $t$ under the Poisson likelihood as*

$$\boldsymbol{x}_j^{(t)} \sim \text{Pois}\left[-\boldsymbol{\Phi}^{(t)}\boldsymbol{\theta}_j^{(t)} \ln\left(1 - p_j^{(t)}\right)\right]. \tag{29}$$

*Proof.* By definition (29) is true for layer $t = 1$. Suppose that (29) is also true for layer $t > 1$, then we can augment each count $x_{vj}^{(t)}$ into the summation of $K_t$ latent counts, which are smaller than or equal to $x_{vj}^{(t)}$, as

$$x_{vj}^{(t)} = \sum_{k=1}^{K_t} x_{vjk}^{(t)}, \quad x_{vjk}^{(t)} \sim \text{Pois}\left[-\phi_{vk}^{(t)}\theta_{kj}^{(t)} \ln\left(1 - p_j^{(t)}\right)\right], \tag{30}$$

where $v \in \{1, \ldots, K_{t-1}\}$. With

$$m_{kj}^{(t)(t+1)} := x_{\cdot jk}^{(t)} := \sum_{v=1}^{K_{t-1}} x_{vjk}^{(t)}$$

representing the number of times that factor $k \in \{1, \ldots, K_t\}$ of layer $t$ appears in observation $j$ and $\boldsymbol{m}_j^{(t)(t+1)} := \left(x_{\cdot j1}^{(t)}, \ldots, x_{\cdot jK_t}^{(t)}\right)'$, since $\sum_{v=1}^{K_{t-1}} \phi_{vk}^{(t)} = 1$, we can marginalize out $\boldsymbol{\Phi}^{(t)}$ as in (Zhou et al., 2012), leading to

$$\boldsymbol{m}_j^{(t)(t+1)} \sim \text{Pois}\left[-\boldsymbol{\theta}_j^{(t)} \ln\left(1 - p_j^{(t)}\right)\right].$$

Further marginalizing out the gamma distributed $\boldsymbol{\theta}_j^{(t)}$ from the above Poisson likelihood leads to

$$\boldsymbol{m}_j^{(t)(t+1)} \sim \text{NB}\left(\boldsymbol{\Phi}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}, p_j^{(t+1)}\right). \tag{31}$$

The $k$th element of $\boldsymbol{m}_j^{(t)(t+1)}$ can be augmented under its compound Poisson representation as

$$m_{kj}^{(t)(t+1)} = \sum_{\ell=1}^{x_{kj}^{(t+1)}} u_\ell, \quad u_\ell \sim \text{Log}(p_j^{(t+1)}), \quad x_{kj}^{(t+1)} \sim \text{Pois}\left[-\boldsymbol{\phi}_{k:}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)} \ln\left(1 - p_j^{(t+1)}\right)\right].$$

Thus if (29) is true for layer $t$, then it is also true for layer $t + 1$. □

**Corollary 2** (Propagate the latent counts upward)**.** *Using Lemma 4.1 of (Zhou et al., 2012) on (30) and Theorem 1 of (Zhou and Carin, 2015) on (31), we can propagate the latent counts*

$x_{vj}^{(t)}$ *of layer $t$ upward to layer $t+1$ as*

$$\left\{\left(x_{vj1}^{(t)},\ldots,x_{vjK_t}^{(t)}\right) \,\Big|\, x_{vj}^{(t)}, \boldsymbol{\phi}_{v:}^{(t)}, \boldsymbol{\theta}_j^{(t)}\right\} \sim \text{Mult}\left(x_{vj}^{(t)}, \frac{\phi_{v1}^{(t)}\theta_{1j}^{(t)}}{\sum_{k=1}^{K_t}\phi_{vk}^{(t)}\theta_{kj}^{(t)}},\ldots,\frac{\phi_{vK_t}^{(t)}\theta_{K_tj}^{(t)}}{\sum_{k=1}^{K_t}\phi_{vk}^{(t)}\theta_{kj}^{(t)}}\right), \quad (32)$$

$$\left(x_{kj}^{(t+1)} \,\Big|\, m_{kj}^{(t)(t+1)}, \boldsymbol{\phi}_{k:}^{(t+1)}, \boldsymbol{\theta}_j^{(t+1)}\right) \sim \text{CRT}\left(m_{kj}^{(t)(t+1)}, \boldsymbol{\phi}_{k:}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}\right). \quad (33)$$

Note that $x_{\cdot j}^{(t)} = m_{\cdot j}^{(t)(t+1)}$ and as the number of tables occupied by the customers is in the same order as the logarithm of the customer number in a Chinese restaurant process, $x_{kj}^{(t+1)}$ is in the same order as $\ln\left(m_{kj}^{(t)(t+1)}\right)$. Thus the total count of layer $t+1$ as $\sum_j x_{\cdot j}^{(t+1)}$ would often be much smaller than that of layer $t$ as $\sum_j x_{\cdot j}^{(t)}$, and hence one may use the total count $\sum_j x_{\cdot j}^{(T)}$ as a simple criterion to decide whether it is necessary to add more layers to the GBN. In addition, if the latent count $x_{k'\cdot k}^{(t)} := \sum_j x_{k'jk}^{(t)}$ becomes close or equal to zero, then the posterior mean of $\boldsymbol{\Phi}^{(t)}(k',k)$ could become so small that node $k'$ of layer $t-1$ can be considered to be disconnected from node $k$ of layer $t$.

## 4.2 Modeling data variability with distributed representation

In comparison to a single-layer model with $T=1$ that assumes the hidden units of layer one are independent in the prior, the multilayer model with $T \geq 2$ captures the correlations between them. Note that for the extreme case that $\boldsymbol{\Phi}^{(t)} = \mathbf{I}_{K_t}$ for $t \geq 2$ are all identity matrices, which indicates that there are no correlations between the features of $\boldsymbol{\theta}_j^{(t-1)}$ left to be captured, the deep structure could still provide benefits as it helps model latent counts $\boldsymbol{m}_j^{(1)(2)}$ that may be highly overdispersed. For example, let us assume $\boldsymbol{\Phi}^{(t)} = \mathbf{I}_{K_2}$ for all $t \geq 2$, then from (12) and (31) we have

$$m_{kj}^{(1)(2)} \sim \text{NB}(\theta_{kj}^{(2)}, p_j^{(2)}),\ \ldots,\ \theta_{kj}^{(t)} \sim \text{Gam}(\theta_{kj}^{(t+1)}, 1/c_j^{(t+1)}),\ \ldots,\ \theta_{kj}^{(T)} \sim \text{Gam}(r_k, 1/c_j^{(T+1)}).$$

Using the laws of total expectation and total variance, we have

$$\mathbb{E}\big[\theta_{kj}^{(2)} \,|\, r_k\big] = \frac{r_k}{\prod_{t=3}^{T+1} c_j^{(t)}}, \qquad \text{Var}\big[\theta_{kj}^{(2)} \,|\, r_k\big] = r_k \sum_{t=3}^{T+1}\left[\prod_{\ell=3}^{t}\left(c_j^{(\ell)}\right)^{-2}\right]\left[\prod_{\ell=t+1}^{T+1}\left(c_j^{(\ell)}\right)^{-1}\right].$$

Further applying the same laws, we have

$$\mathbb{E}\big[m_{kj}^{(1)(2)} \,|\, r_k\big] = \frac{r_k p_j^{(2)}}{\left(1-p_j^{(2)}\right)\prod_{t=3}^{T+1} c_j^{(t)}}$$

and

$$\mathrm{Var}\big[m_{kj}^{(1)(2)}\,|\,r_k\big] = \frac{r_k p_j^{(2)}}{\big(1-p_j^{(2)}\big)^2 \prod_{t=3}^{T+1} c_j^{(t)}} \left\{1 + p_j^{(2)} \sum_{t=3}^{T+1} \left[\prod_{\ell=3}^{t} \left(c_j^{(\ell)}\right)^{-1}\right]\right\}.$$

Thus the variance to mean ratio (VMR) of the latent count $m_{kj}^{(1)(2)}$ given $r_k$ can be expressed as

$$\mathrm{VMR}\big[m_{kj}^{(1)(2)}\,|\,r_k\big] = \frac{1}{\big(1-p_j^{(2)}\big)} \left\{1 + p_j^{(2)} \sum_{t=3}^{T+1} \left[\prod_{\ell=3}^{t} \left(c_j^{(\ell)}\right)^{-1}\right]\right\}. \tag{34}$$

In comparison to PFA with $m_{kj}^{(1)(2)} \sim \mathrm{NB}(r_k, p_j^{(2)})$ given $r_k$, with a VMR of $1/(1-p_j^{(2)})$, the GBN with $T$ hidden layers, which mixes the shape of $m_{kj}^{(1)(2)} \sim \mathrm{NB}(\theta_{kj}^{(2)}, p_j^{(2)})$ with a chain of gamma random variables, increases $\mathrm{VMR}\big[m_{kj}^{(1)(2)}\,|\,r_k\big]$ by a factor of

$$1 + p_j^{(2)} \sum_{t=3}^{T+1} \left[\prod_{\ell=3}^{t} \left(c_j^{(\ell)}\right)^{-1}\right],$$

which is equal to

$$1 + (T-1)p_j^{(2)}$$

if we further assume $c_j^{(t)} = 1$ for all $t \geq 3$. Therefore, by increasing the depth of the network to distribute the variability into more layers, the multilayer structure could increase its capability to model data variability.

## 4.3 Upward-downward Gibbs sampling

### 4.3.1 Inference for the PGBN

With Lemma 1 and Corollary 2 and the width of the first layer being bounded by $K_{1\max}$, we first consider multivariate count observations and develop an upward-downward Gibbs sampler for the PGBN, each iteration of which proceeds as follows.

***Sample*** $x_{vjk}^{(t)}$. We can sample $x_{vjk}^{(t)}$ for all layers using (32). But for the first hidden layer, we may treat each observed count $x_{vj}^{(1)}$ as a sequence of word tokens at the $v$th term (in a vocabulary of size $V := K_0$) in the $j$th document, and assign the $x_{\cdot j}^{(1)}$ words $\{v_{ji}\}_{i=1,x_{\cdot j}^{(1)}}$ one after another to the latent factors (topics), with both the topics $\boldsymbol{\Phi}^{(1)}$ and topic weights $\boldsymbol{\theta}_j^{(1)}$ marginalized out, as

$$P(z_{ji} = k\,|\,-) \propto \frac{\eta^{(1)} + x_{v_{ji}\cdot k}^{(1)^{-ji}}}{V\eta^{(1)} + x_{\cdot\cdot k}^{(1)^{-ji}}} \left(x_{\cdot jk}^{(1)^{-ji}} + \boldsymbol{\phi}_{k:}^{(2)} \boldsymbol{\theta}_j^{(2)}\right), \quad k \in \{1, \ldots, K_{1\max}\}, \tag{35}$$

where $z_{ji}$ is the topic index for $v_{ji}$ and $x_{vjk}^{(1)} := \sum_i \delta(v_{ji} = v, z_{ji} = k)$ counts the number of times that term $v$ appears in document $j$; we use the $\cdot$ symbol to represent summing over the corresponding index, *e.g.*, $x_{\cdot jk}^{(t)} := \sum_v x_{vjk}^{(t)}$, and use $x^{-ji}$ to denote the count $x$ calculated without considering word $i$ in document $j$. The collapsed Gibbs sampling update equation shown above is related to the one developed in (Griffiths and Steyvers, 2004) for latent Dirichlet allocation, and the one developed in (Zhou, 2014) for PFA using the beta-negative binomial process. When $T = 1$, we would replace the terms $\boldsymbol{\phi}_{k:}^{(2)}\boldsymbol{\theta}_j^{(2)}$ with $r_k$ for PFA built on the gamma-negative binomial process (Zhou and Carin, 2015) (or with $\alpha\pi_k$ for hierarchical Dirichlet process latent Dirichlet allocation, see (Teh et al., 2006) and (Zhou, 2014) for details), and add an additional term to account for the possibility of creating an additional factor (Zhou, 2014). For simplicity, in this paper, we truncate the nonparametric Bayesian model with $K_{1\max}$ factors and let $r_k \sim \mathrm{Gam}(\gamma_0/K_{1\max}, 1/c_0)$ if $T = 1$. Note that although we use collapsed Gibbs sampling inference in this paper, if one desires embarrassingly parallel inference and possibly lower computation, then one may consider explicitly sampling $\{\boldsymbol{\phi}_k^{(1)}\}_k$ and $\{\boldsymbol{\theta}_j^{(1)}\}_j$ and sampling $x_{vjk}^{(1)}$ with (32).

***Sample*** $\boldsymbol{\phi}_k^{(t)}$. Given these latent counts, we sample the factors/topics $\boldsymbol{\phi}_k^{(t)}$ as

$$(\boldsymbol{\phi}_k^{(t)} \,|\, -) \sim \mathrm{Dir}\left(\eta_1^{(t)} + x_{1\cdot k}^{(t)}, \ldots, \eta_{K_{t-1}}^{(t)} + x_{K_{t-1}\cdot k}^{(t)}\right). \tag{36}$$

***Sample*** $x_{vj}^{(t+1)}$. We sample $\boldsymbol{x}_j^{(t+1)}$ using (33), where we replace the term $\boldsymbol{\phi}_{v:}^{(T+1)}\boldsymbol{\theta}_j^{(T+1)}$ with $r_v$.

***Sample*** $\boldsymbol{r}$. Both $\gamma_0$ and $c_0$ are sampled using related equations in (Zhou and Carin, 2015), omitted here for brevity. We sample $\boldsymbol{r}$ as

$$(r_v \,|\, -) \sim \mathrm{Gam}\left(\gamma_0/K_T + x_{v\cdot}^{(T+1)}, \left[c_0 - \textstyle\sum_j \ln\left(1 - p_j^{(T+1)}\right)\right]^{-1}\right). \tag{37}$$

***Sample*** $\boldsymbol{\theta}_j^{(t)}$. Using (29) and the gamma-Poisson conjugacy, we sample $\boldsymbol{\theta}_j$ as

$$\begin{aligned}
(\boldsymbol{\theta}_j^{(T)} \,|\, -) &\sim \mathrm{Gam}\left(\boldsymbol{r} + \boldsymbol{m}_j^{(T)(T+1)}, \left[c_j^{(T+1)} - \ln\left(1 - p_j^{(T)}\right)\right]^{-1}\right), \\
&\vdots \\
(\boldsymbol{\theta}_j^{(t)} \,|\, -) &\sim \mathrm{Gam}\left(\boldsymbol{\Phi}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)} + \boldsymbol{m}_j^{(t)(t+1)}, \left[c_j^{(t+1)} - \ln\left(1 - p_j^{(t)}\right)\right]^{-1}\right), \\
&\vdots \\
(\boldsymbol{\theta}_j^{(1)} \,|\, -) &\sim \mathrm{Gam}\left(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)} + \boldsymbol{m}_j^{(1)(2)}, \left[c_j^{(2)} - \ln\left(1 - p_j^{(1)}\right)\right]^{-1}\right),
\end{aligned} \tag{38}$$

***Sample*** $c_j^{(t)}$. With $\theta_{\cdot j}^{(t)} := \sum_{k=1}^{K_t} \theta_{kj}^{(t)}$ for $t \leq T$ and $\theta_{\cdot j}^{(T+1)} := r_{\cdot}$, we sample $p_j^{(2)}$ and $\{c_j^{(t)}\}_{t\geq 3}$ as

$$(p_j^{(2)}\,|\,-) \sim \text{Beta}\left(a_0 + m_{\cdot j}^{(1)(2)}, b_0 + \theta_{\cdot j}^{(2)}\right), \ \ (c_j^{(t)}\,|\,-) \sim \text{Gam}\left(e_0 + \theta_{\cdot j}^{(t)}, \left[f_0 + \theta_{\cdot j}^{(t-1)}\right]^{-1}\right), \tag{39}$$

and calculate $c_j^{(2)}$ and $\{p_j^{(t)}\}_{t\geq 3}$ with (28).

### 4.3.2 Handling binary and nonnegative real observations

For binary observations that are linked to the latent counts at layer one as $b_{vj}^{(1)} = \mathbf{1}(x_{vj}^{(1)} \geq 1)$, we first sample the latent counts at layer one from the truncated Poisson distribution as

$$\left(x_{vj}^{(1)}\,|\,-\right) \sim b_{vj}^{(1)} \cdot \text{Pois}_+\left(\sum_{k=1}^{K_1} \phi_{vk}^{(1)} \theta_{kj}^{(1)}\right) \tag{40}$$

and then sample $x_{vjk}^{(t)}$ for all layers using (32).

For nonnegative real observations $y_{vj}^{(1)}$ that are linked to the latent counts at layer one as

$$y_{vj}^{(1)} \sim \text{Gam}(x_{vj}^{(1)}, 1/a_j),$$

we let $x_{vj}^{(1)} = 0$ if $y_{vj}^{(1)} = 0$ and sample $x_{vj}^{(1)}$ from the truncated Bessel distribution as

$$\left(x_{vj}^{(1)}\,|\,-\right) \sim \text{Bessel}_{-1}\left(2\sqrt{a_j y_{vj}^{(1)} \sum_{k=1}^{K_1} \phi_{vk}^{(1)} \theta_{kj}^{(1)}}\right) \tag{41}$$

if $y_{vj}^{(1)} > 0$. We let $a_j \sim \text{Gam}(e_0, 1/f_0)$ in the prior and sample $a_j$ as

$$(a_j\,|\,-) \sim \text{Gam}\left(e_0 + \sum_v x_{vj}^{(1)}, \frac{1}{f_0 + \sum_v y_{vj}^{(1)}}\right). \tag{42}$$

We then sample $x_{vjk}^{(t)}$ for all layers using (32).

## 4.4 Learning the network structure with layer-wise training

As jointly training all layers together is often difficult, existing deep networks are typically trained using a greedy layer-wise unsupervised training algorithm, such as the one proposed in (Hinton et al., 2006) to train the deep belief networks. The effectiveness of this training

**Algorithm 1** The PGBN upward-downward Gibbs sampler that uses a layer-wise training strategy to train a set of networks, each of which adds an additional hidden layer on top of the previously inferred network, retrains all its layers jointly, and prunes inactive factors from the last layer.
**Inputs:** observed counts $\{x_{vj}\}_{v,j}$, upper bound of the width of the first layer $K_{1\max}$, upper bound of the number of layers $T_{\max}$, number of iterations $\{B_T, S_T\}_{1,T_{\max}}$, and hyper-parameters.
**Outputs:** A total of $T_{\max}$ jointly trained PGBNs with depths $T=1$, $T=2$, …, and $T=T_{\max}$.

```
 1: for $T = 1, 2, \ldots, T_{\max}$ do Jointly train all the $T$ layers of the network
 2:     Set $K_{T-1}$, the inferred width of layer $T-1$, as $K_{T\max}$, the upper bound of layer $T$'s width.
 3:     for $iter = 1 : B_T + C_T$ do Upward-downward Gibbs sampling
 4:         Sample $\{z_{ji}\}_{j,i}$ using collapsed inference; Calculate $\{x_{vjk}^{(1)}\}_{v,k,j}$; Sample $\{x_{vj}^{(2)}\}_{v,j}$ ;
 5:         for $t = 2, 3, \ldots, T$ do
 6:             Sample $\{x_{vjk}^{(t)}\}_{v,j,k}$ ; Sample $\{\boldsymbol{\phi}_k^{(t)}\}_k$ ; Sample $\{x_{vj}^{(t+1)}\}_{v,j}$ ;
 7:         end for
 8:         Sample $p_j^{(2)}$ and Calculate $c_j^{(2)}$; Sample $\{c_j^{(t)}\}_{j,t}$ and Calculate $\{p_j^{(t)}\}_{j,t}$ for $t = 3, \ldots, T+1$;
 9:         for $t = T, T-1, \ldots, 2$ do
10:             Sample $\boldsymbol{r}$ if $t = T$; Sample $\{\boldsymbol{\theta}_j^{(t)}\}_j$ ;
11:         end for
12:         if $iter = B_T$ then
13:             Prune layer $T$'s inactive factors $\{\boldsymbol{\phi}_k^{(T)}\}_{k:x_{\cdot\cdot k}^{(T)}=0}$;
14:             let $K_T = \sum_k \delta(x_{\cdot\cdot k}^{(T)} > 0)$ and update $\boldsymbol{r}$;
15:         end if
16:     end for
17:     Output the posterior means (according to the last MCMC sample) of all remaining factors
        $\{\boldsymbol{\phi}_k^{(t)}\}_{k,t}$ as the inferred network of $T$ layers, and $\{r_k\}_{k=1}^{K_T}$ as the gamma shape parameters of
        layer $T$'s hidden units.
18: end for
```

strategy is further analyzed in (Bengio et al., 2007). By contrast, the GBN has a simple Gibbs sampler to jointly train all its hidden layers, as described in Section 4.3, and hence does not require greedy layer-wise training, but the same as these commonly used deep learning algorithms, it still needs to specify the number of layers and the width of each layer.

**Algorithm 2** The upward-downward Gibbs samplers for the Ber-GBN and PRG-GBN are constructed by using Lines 1-8 shown below to substitute Lines 4-11 of the PGBN Gibbs sampler shown in Algorithm 1.

```
1: Sample $\{x_{vj}^{(1)}\}_{v,j}$ using (40) for binary observations; Sample $\{x_{vj}^{(1)}\}_{v,j}$ using (41) and sample $a_j$
   using (42) for nonnegative real observations;
2: for $t = 1, 2, \ldots, T$ do
3:     Sample $\{x_{vjk}^{(t)}\}_{v,j,k}$ ; Sample $\{\boldsymbol{\phi}_k^{(t)}\}_k$ ; Sample $\{x_{vj}^{(t+1)}\}_{v,j}$ ;
4: end for
5: Sample $p_j^{(2)}$ and Calculate $c_j^{(2)}$; Sample $\{c_j^{(t)}\}_{j,t}$ and Calculate $\{p_j^{(t)}\}_{j,t}$ for $t = 3, \ldots, T+1$;
6: for $t = T, T-1, \ldots, 1$ do
7:     Sample $\boldsymbol{r}$ if $t = T$; Sample $\{\boldsymbol{\theta}_j^{(t)}\}_j$ ;
8: end for
```

In this paper, we adopt the idea of layer-wise training for the GBN, not because of the lack of an effective joint-training algorithm, but for the purpose of learning the width of each hidden layer in a greedy layer-wise manner, given a fixed budget on the width of the first layer. The proposed layer-wise training strategies are summarized in Algorithm 1 for multivariate count data, and in Algorithm 2 for multivariate binary and nonnegative real data. With a GBN of $T-1$ layer that has already been trained, the key idea is to use a truncated gamma-negative binomial process (Zhou and Carin, 2015, Zhou et al., 2015b) to model the latent count matrix for the newly added top layer as $m_{kj}^{(T)(T+1)} \sim \mathrm{NB}(r_k, p_j^{(T+1)})$, $r_k \sim \mathrm{Gam}(\gamma_0/K_{T\max}, 1/c_0)$, and rely on that stochastic process's shrinkage mechanism to prune inactive factors (connection weight vectors) of layer $T$, and hence the inferred $K_T$ would be smaller than $K_{T\max}$ if $K_{T\max}$ is set to be sufficiently large. The newly added layer and all the layers below it would be jointly trained, but with the structure below the newly added layer kept unchanged. Note that when $T=1$, the GBN would infer the number of active factors if $K_{1\max}$ is set large enough, otherwise, it would still assign the factors with different weights $r_k$, but may not be able to prune any of them.

# 5 Experimental Results

## 5.1 Deep topic modeling

We first analyze multivariate count data with the Poisson gamma belief network (PGBN). We apply the PGBNs for topic modeling of text corpora, each document of which is represented as a term-frequency count vector. Note that the PGBN with a single hidden layer is identical to the (truncated) gamma-negative binomial process PFA of Zhou and Carin (2015), which is a nonparametric Bayesian algorithm that performs similarly to the hierarchical Dirichlet process latent Dirichlet allocation of Teh et al. (2006) for text analysis, and is considered as a strong baseline. Thus we will focus on making comparison to the PGBN with a single layer, with its layer width set to be large to approximate the performance of the gamma-negative binomial process PFA. We evaluate the PGBNs' performance by examining both how well they unsupervisedly extract low-dimensional features for document classification, and how well they predict heldout word tokens. Matlab code will be available in http://mingyuanzhou.github.io/.

We use Algorithm 1 to learn, in a layer-wise manner, from the training data the connection weight matrices $\mathbf{\Phi}^{(1)}, \ldots, \mathbf{\Phi}^{(T_{\max})}$ and the top-layer hidden units' gamma shape parameters $\boldsymbol{r}$: to add layer $T$ to a previously trained network with $T-1$ layers, we use $B_T$ iterations to jointly train $\mathbf{\Phi}^{(T)}$ and $\boldsymbol{r}$ together with $\{\mathbf{\Phi}^{(t)}\}_{1,T-1}$, prune the inactive factors of layer $T$, and continue the joint training with another $C_T$ iterations. We set the hyper-parameters as $a_0 = b_0 = 0.01$ and $e_0 = f_0 = 1$. Given the trained network, we apply the upward-downward Gibbs sampler to collect 500 MCMC samples after 500 burnins to estimate the posterior mean of the feature usage proportion vector $\boldsymbol{\theta}_j^{(1)}/\theta_{\boldsymbol{\cdot}j}^{(1)}$ at the first hidden layer, for every document in both the training and testing sets.

### 5.1.1 Feature learning for binary classification

We consider the 20newsgroups dataset (http://qwone.com/∼jason/20Newsgroups/) that consists of 18,774 documents from 20 different news groups, with a vocabulary of size $K_0 = 61{,}188$. It is partitioned into a training set of 11,269 documents and a testing set of 7,505 ones. We first consider two binary classification tasks that distinguish between the *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware*, and between the *sci.electronics* and *sci.med* news groups. For each binary classification task, we remove a standard list of stop words and only consider the terms that appear at least five times, and report the classification accuracies based on 12 independent random trials. With the upper bound of the first layer's width set as $K_{1\max} \in \{25, 50, 100, 200, 400, 600, 800\}$, and $B_t = C_t = 1000$ and $\eta^{(t)} = 0.01$ for all $t$, we use Algorithm 1 to train a network with $T \in \{1, 2, \ldots, 8\}$ layers. Denote $\bar{\boldsymbol{\theta}}_j$ as the estimated $K_1$ dimensional feature vector for document $j$, where $K_1 \leq K_{1\max}$ is the inferred number of active factors of the first layer that is bounded by the pre-specified truncation level $K_{1\max}$. We use the $L_2$ regularized logistic regression provided by the LIBLINEAR package (Fan et al., 2008) to train a linear classifier on $\bar{\boldsymbol{\theta}}_j$ in the training set and use it to classify $\bar{\boldsymbol{\theta}}_j$ in the test set, where the regularization parameter is five-folder cross-validated on the training set from $(2^{-10}, 2^{-9}, \ldots, 2^{15})$.

As shown in Figure 13, modifying the PGBN from a single-layer shallow network to a multilayer deep one clearly improves the qualities of the unsupervisedly extracted feature vectors. In a random trial, with $K_{1\max} = 800$, we infer a network structure of $[K_1, \ldots, K_8] = [512, 154, 75, 54, 47, 37, 34, 29]$ for the first binary classification task, and $[K_1, \ldots, K_8] = [491, 143, 74, 49, 36, 32, 28, 26]$ for the second one. Figures 13(c)-(d) also show that increasing the network depth in general improves the performance, but the first-layer width clearly plays a critical role in controlling the ultimate network capacity. This insight is further illustrated below.

### 5.1.2 Feature learning for multi-class classification

We test the PGBNs for multi-class classification on 20newsgroups. After removing a standard list of stopwords and the terms that appear less than five times, we obtain a vocabulary

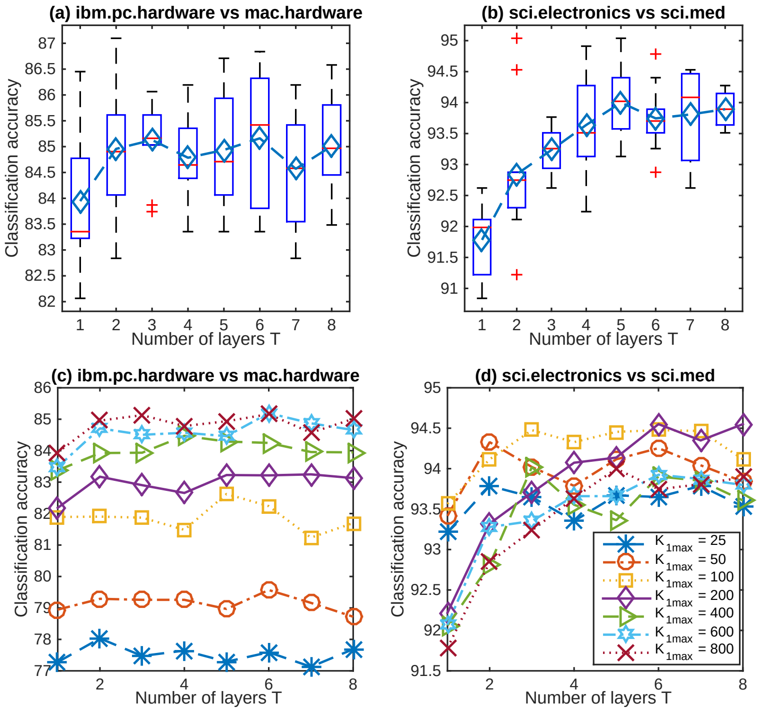


Figure 13: Classification accuracy (%) as a function of the network depth $T$ for two 20newsgroups binary classification tasks, with $\eta^{(t)} = 0.01$ for all layers. (a)-(b): the boxplots of the accuracies of 12 independent runs with $K_{1\max} = 800$. (c)-(d): the average accuracies of these 12 runs for various $K_{1\max}$ and $T$. Note that $K_{1\max} = 800$ is large enough to cover all active first-layer topics (inferred to be around 500 for both binary classification tasks), whereas all the first-layer topics would be used if $K_{1\max} = 25$, 50, 100, or 200.

with $V = 33,420$. We set $C_t = 500$ and $\eta^{(t)} = 0.05$ for all $t$; we set $B_t = 1000$ for all $t$ if $K_{1\max} \leq 400$, and set $B_1 = 1000$ and $B_t = 500$ for $t \geq 2$ if $K_{1\max} > 400$. We use all 11,269 training documents to infer a set of networks with $T_{\max} \in \{1, \dots, 5\}$ and $K_{1\max} \in \{50, 100, 200, 400, 600, 800\}$, and mimic the same testing procedure used for binary classification to extract low-dimensional feature vectors, with which each testing document is classified to one of the 20 news groups using the $L_2$ regularized logistic regression.

Figure 14 shows a clear trend of improvement in classification accuracy by increasing the network depth with a limited first-layer width, or by increasing the upper bound of the width of the first layer with the depth fixed. For example, a single-layer PGBN with $K_{1\max} = 100$ could add one or more layers to slightly outperform a single-layer PGBN with $K_{1\max} = 200$, and a single-layer PGBN with $K_{1\max} = 200$ could add layers to clearly outperform a single-layer PGBN with $K_{1\max}$ as large as 800. We also note that each iteration of jointly training multiple layers costs moderately more than that of training a single layer, e.g., with $K_{1\max} = 400$, a training iteration on a single core of an Intel Xeon 2.7 GHz CPU takes about 5.59, 6.68, 7.09 seconds for the PGBN with 1, 3, and 5 layers, respectively.

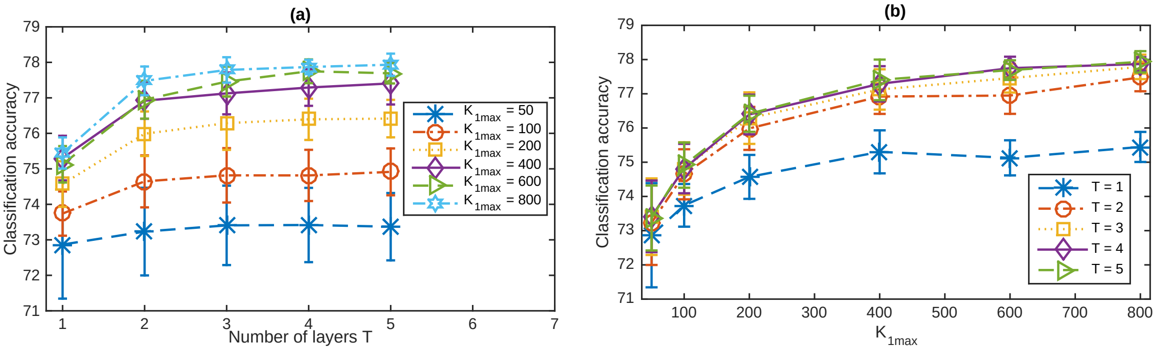


Figure 14: Classification accuracy (%) of the PGBNs with Algorithm 1 for 20newsgroups multi-class classification (a) as a function of the depth $T$ with various $K_{1\max}$ and (b) as a function of $K_{1\max}$ with various depths, with $\eta^{(t)} = 0.05$ for all layers. The widths of the hidden layers are automatically inferred. In a random trial, the inferred network widths $[K_1, \ldots, K_5]$ for $K_{1\max} = 50, 100, 200, 400, 600$, and 800 are $[50, 50, 50, 50, 50]$, $[100, 99, 99, 94, 87]$, $[200, 161, 130, 94, 63]$, $[396, 109, 99, 82, 68]$, $[528, 129, 109, 98, 91]$, and $[608, 100, 99, 96, 89]$, respectively.

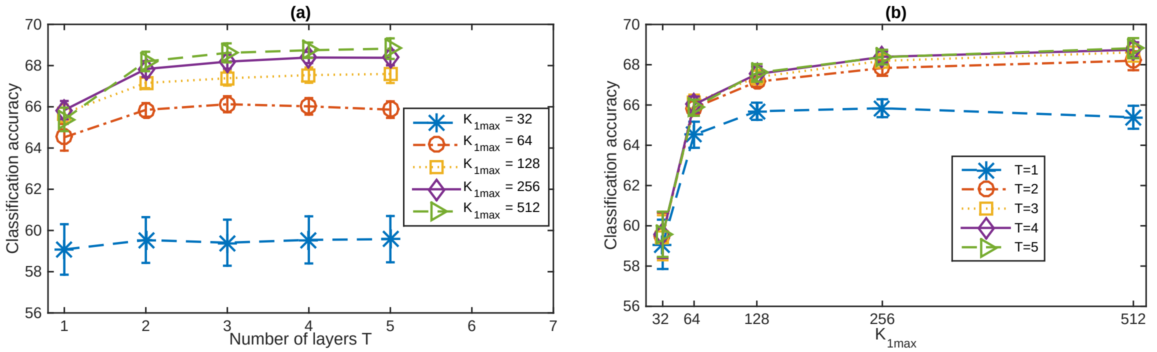


Figure 15: Analogous plots to Figure 14 with the vocabulary size restricted to be 2000, including the most frequent 2000 terms after removing a standard list of stopwords. The widths of the hidden layers are automatically inferred. In a random trial, the inferred network widths $[K_1, \ldots, K_5]$ for $K_{1\max} = 32, 64, 128, 256$, and 512 are $[32, 32, 32, 32, 32]$, $[64, 64, 64, 59, 59]$, $[128, 125, 118, 106, 87]$, $[256, 224, 124, 83, 65]$, and $[512, 187, 89, 78, 62]$, respectively.

Examining the inferred network structure also reveals interesting details. For example, in a random trial with Algorithm 1, with $\eta^{(t)} = 0.05$ for all $t$, the inferred network widths $[K_1, \ldots, K_5]$ for $K_{1\max} = 50, 100, 200, 400, 600$, and 800 are $[50, 50, 50, 50, 50]$, $[100, 99, 99, 94, 87]$, $[200, 161, 130, 94, 63]$, $[396, 109, 99, 82, 68]$, $[528, 129, 109, 98, 91]$, and [608, 100, 99, 96, 89], respectively. This indicates that for a network with an insufficient budget on its first-layer width, as the network depth increases, its inferred layer widths decay more slowly than a network with a sufficient or surplus budget on its first-layer width; and a network with a surplus budget on its first-layer width may only need relatively small widths for its higher hidden layers.

In order to make comparison to related algorithms, we also consider restricting the vocabulary to the 2000 most frequent terms of the vocabulary after moving a standard list of stopwords. We repeat the same experiments with the same settings except that we set $K_{1\max} \in \{32, 64, 128, 256, 512\}$, $B_1 = 1000$, $C_1 = 500$, and $B_t = C_t = 500$ for all $t \geq 2$. We show the results in Figure 15. Again, we observe a clear trend of improvement by increasing the network depth with a limited first-layer width, or by increasing the upper bound of the width of the first layer with the depth fixed. In a random trial with Algorithm 1, the inferred network widths $[K_1, \ldots, K_5]$ for $K_{1\max} = 32, 64, 128, 256$, and 512 are $[32, 32, 32, 32, 32]$, $[64, 64, 64, 59, 59]$, $[128, 125, 118, 106, 87]$, $[256, 224, 124, 83, 65]$, and $[512, 187, 89, 78, 62]$, respectively.

For comparison, we first consider the same $L_2$ regularized logistic regression multi-class classifier, trained either on the raw word counts or normalized term-frequencies of the 20newsgroups training documents using five-folder cross-validation. As summarized in Table 1, when using the raw term-frequency word counts as covariates, the same classifier achieves 69.8% (68.2%) accuracy on the 20newsgroups test documents if using the top 2000 terms that exclude (include) a standard list of stopwords, achieves 75.8% if using all the $61,188$ terms in the vocabulary, and achieves 78.0% if using the $33,420$ terms remained after removing a standard list of stopwords and the terms that appear less than five times; and when using the normalized term-frequencies as covariates, the corresponding accuracies are 70.8% (67.9%) if using the top 2000 terms excluding (including) stopwords, 77.6% with all the $61,188$ terms, and 79.4% with the $33,420$ selected terms.

Table 1: Multi-class classification accuracy of $L_2$ regularized logistic regression.

| $V = 61,188$<br>with stopwords<br>with rare words<br>raw word counts | $V = 61,188$<br>with stopwords<br>with rare words<br>term frequencies | $V = 33,420$<br>remove stopwords<br>remove rare words<br>raw word counts | $V = 33,420$<br>remove stopwords<br>remove rare words<br>term frequencies |
|---|---|---|---|
| 75.8% | 77.6% | 78.0% | 79.4% |

| $V = 2000$<br>with stopwords<br>raw counts | $V = 2000$<br>with stopwords<br>term frequencies | $V = 2000$<br>remove stopwords<br>raw counts | $V = 2000$<br>remove stopwords<br>term frequencies |
|---|---|---|---|
| 68.2% | 67.9% | 69.8% | 70.8% |

As summarized in Table 2, for multi-class classification on the same dataset, with a vocabulary size of 2000 that consisits of the 2000 most frequent terms after removing stopwords and stemming, the DocNADE (Larochelle and Lauly, 2012) and the over-replicated softmax (Srivastava et al., 2013) provide the accuracies of 67.0% and 66.8%, respectively, for a feature dimension of $K = 128$, and provide the accuracies of 68.4% and 69.1%, respectively, for a feature dimension of $K = 512$.

Table 2: Multi-class classification accuracy of the DocNADE (Larochelle and Lauly, 2012) and over-replicated softmax (Srivastava et al., 2013).

| | $V = 2000$, $K = 128$<br>remove stopwords, stemming | $V = 2000$, $K = 512$<br>remove stopwords, stemming |
|---|---|---|
| DocNADE | 67.0% | 68.4% |
| Over-replicated softmax | 66.8% | 69.1% |

As shown in Figure 15 and summarized in Table 3, with the same vocabulary size of 2000 (but different terms due to different preprocessing), the proposed PGBN provides 65.9% (67.5%) with $T = 1$ ($T = 5$) for $K_{1\,\max} = 128$, and 65.9% (69.2%) with $T = 1$ ($T = 5$) for $K_{1\,\max} = 512$, which may be further improved if we also consider the stemming step, as done in these two algorithms, for word preprocessing, or if we set the values of $\eta^{(t)}$ to be smaller than 0.05 to encourage a more complex network structure. We also summarize in Table 3 the classification accuracies shown in Figure 14 for the PGBNs with $V = 33,420$. Note that the accuracies in Tables 2 and 3 are provided to show that the PGBNs are in the same ballpark as both the DocNADE (Larochelle and Lauly, 2012) and over-replicated softmax (Srivastava et al., 2013). Note these results are not intended to provide a head-to-head comparison, which is possible if the same data preprocessing and classifier were used and the error bars were shown in Srivastava et al. (2013), or we could obtain the code to replicate the experiments using the same preprocessed data and classifier.

Table 3: Classification accuracy of the PGBN trained with $\eta^t = 0.05$ for all $t$.

| | $V = 2000$, $K_{1\,\max} = 128$<br>remove stopwords | $V = 2000$, $K_{1\,\max} = 256$<br>remove stopwords | $V = 2000$, $K_{1\,\max} = 512$<br>remove stopwords |
|---|---|---|---|
| PGBN ($T = 1$) | $65.9\% \pm 0.4\%$ | $66.3\% \pm 0.4\%$ | $65.9\% \pm 0.4\%$ |
| PGBN ($T = 2$) | $67.1\% \pm 0.5\%$ | $67.9\% \pm 0.4\%$ | $68.3\% \pm 0.3\%$ |
| PGBN ($T = 3$) | $67.3\% \pm 0.3\%$ | $68.6\% \pm 0.5\%$ | $69.0\% \pm 0.4\%$ |
| PGBN ($T = 5$) | $67.5\% \pm 0.4\%$ | $68.8\% \pm 0.3\%$ | $69.2\% \pm 0.4\%$ |

| | $V = 33,420$, $K_{1\,\max} = 200$<br>remove stopwords<br>remove rare words | $V = 33,420$, $K_{1\,\max} = 400$<br>remove stopwords<br>remove rare words | $V = 33,420$, $K_{1\,\max} = 800$<br>remove stopwords<br>remove rare words |
|---|---|---|---|
| PGBN ($T = 1$) | $74.6\% \pm 0.6\%$ | $75.3\% \pm 0.6\%$ | $75.4\% \pm 0.4\%$ |
| PGBN ($T = 2$) | $76.0\% \pm 0.6\%$ | $76.9\% \pm 0.5\%$ | $77.5\% \pm 0.4\%$ |
| PGBN ($T = 3$) | $76.3\% \pm 0.8\%$ | $77.1\% \pm 0.6\%$ | $77.8\% \pm 0.4\%$ |
| PGBN ($T = 5$) | $76.4\% \pm 0.5\%$ | $77.4\% \pm 0.6\%$ | $77.9\% \pm 0.3\%$ |

### 5.1.3 Perplexities for heldout words

In addition to examining the performance of the PGBN for unsupervised feature learning, we also consider a more direct approach that we randomly choose 30% of the word tokens in each document as training, and use the remaining ones to calculate per-heldout-word perplexity. We consider both all the 18,774 documents of the 20newsgroups corpus, limiting the vocabulary to the 2000 most frequent terms after removing a standard list of stopwords, and the NIPS12 (http://www.cs.nyu.edu/∼roweis/data.html) corpus whose stopwords have already been removed, limiting the vocabulary to the 2000 most frequent terms. We set $\eta^{(t)} = 0.05$ and $C_t = 500$ for all $t$, set $B_1 = 1000$ and $B_t = 500$ for $t \geq 2$, and consider five random trials. Among the $B_t + C_t$ Gibbs sampling iterations used to train layer $t$, we collect one sample per five iterations during the last 500 iterations, for each of which we draw the topics $\{\boldsymbol{\phi}_k^{(1)}\}_k$ and topics weights $\boldsymbol{\theta}_j^{(1)}$, to compute the per-heldout-word perplexity using Equation (34) of Zhou and Carin (2015). This evaluation method is similar to those used in Newman et al. (2009), Wallach et al. (2009), and Paisley et al. (2012).

As shown in both Figures 16 and 17, we observe a clear trend of improvement by increasing both $K_{1\max}$ and $T$. We have also examined the topics and network structure learned on the NIPS12 corpus. Similar to the exploratory data analysis performed on the 20newsgroups corpus, as described in detail in Section 3.2, the inferred deep networks also allow us to extract trees and subnetworks to visualize various aspects of the NIPS12 corpus from general to specific and reveal how they are related to each other. We omit these details for brevity and instead provide a brief description: with $K_{1\max} = 200$ and $T = 5$, the PGBN infers a network with $[K_1, \ldots, K_5] = [200, 164, 106, 60, 42]$ in one of the five random trials. The ranks, according to the weights $r_k^{(t)}$ calculated in (21), and the top five words of three example topics for layer $T = 5$ are "6 network units input learning training," "15 data model learning set image," and "34 network learning model input neural;" while these of five example topics of layer $T = 1$ are "19 likelihood em mixture parameters data," "37 bayesian posterior prior log evidence," "62 variables belief networks conditional inference," "126 boltzmann binary machine energy hinton," and "127 speech speaker acoustic vowel phonetic." It is clear that the topics of the bottom hidden layers are very specific whereas these of the top hidden layer are quite general.

### 5.1.4 Generating synthetic documents

We have also tried drawing $\boldsymbol{\theta}_{j'}^{(T)} \sim \text{Gam}\big(\boldsymbol{r}, 1/c_{j'}^{(T+1)}\big)$ and downward passing it through a $T$-layer network trained on a text corpus to generate synthetic documents, which are found to be quite interpretable and reflect various general aspects of the corpus used to train the

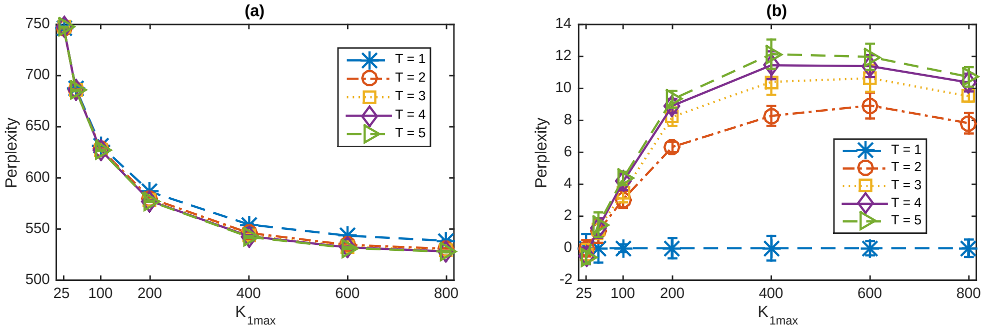


Figure 16: (a) per-heldout-word perplexity (the lower the better) for the NIPS12 corpus (using the 2000 most frequent terms) as a function of the upper bound of the first layer width $K_{1\,\max}$ and network depth $T$, with 30% of the word tokens in each document used for training and $\eta^{(t)} = 0.05$ for all $t$. (b) for visualization, each curve in (a) is reproduced by subtracting its values from the average perplexity of the single-layer network. In a random trial, the inferred network widths $[K_1, \ldots, K_5]$ for $K_{1\,\max} = 25, 50, 100, 200, 400, 600,$ and 800 are $[25, 25, 25, 25, 25]$, $[50, 50, 50, 49, 42]$, $[100, 99, 93, 78, 54]$, $[200, 164, 106, 60, 42]$, $[400, 130, 83, 52, 39]$, $[596, 71, 68, 58, 37]$, and $[755, 57, 53, 46, 42]$, respectively.

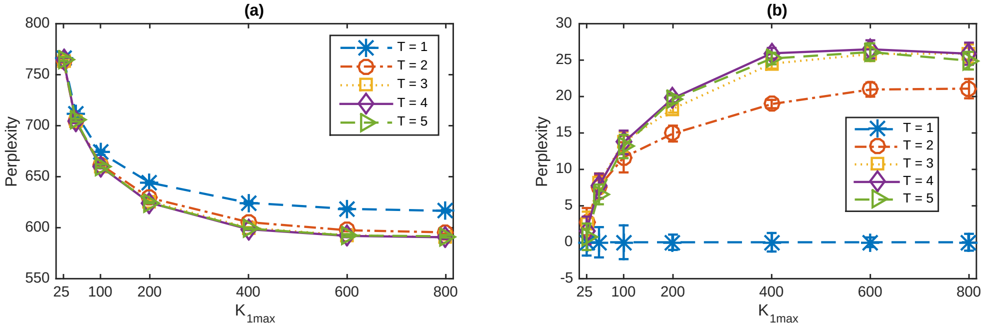


Figure 17: Analogous plots to Figure 16 for the 20newsgroups corpus (using the 2000 most frequent terms after removing a standard list of stopwords). In a random trial, the inferred network widths $[K_1, \ldots, K_5]$ for $K_{1\,\max} = 25, 50, 100, 200, 400, 600,$ and 800 are $[25, 25, 25, 25, 25]$, $[50, 50, 50, 50, 50]$, $[100, 99, 99, 97, 97]$, $[200, 194, 177, 152, 123]$, $[398, 199, 140, 116, 105]$, $[557, 156, 133, 118, 103]$, and $[701, 119, 116, 112, 103]$, respectively.

network. We consider the PGBN with $[K_1, \ldots, K_5] = [608, 100, 99, 96, 89]$, which is trained on the training set of the 20newsgroups corpus with $K_{1\,\max} = 800$ and $\eta^{(t)} = 0.05$ for all $t$. We set $c_{j'}^{(t)}$ as the median of the inferred $\{c_j^t\}_j$ of the training documents for all $t$. Given $\{\boldsymbol{\Phi}^{(t)}\}_{1,T}$ and $\boldsymbol{r}$, we first generate $\boldsymbol{\theta}_{j'}^{(T)} \sim \text{Gam}\left(\boldsymbol{r}, 1/c_{j'}^{(T+1)}\right)$ and then downward pass it through the network by drawing nonnegative real random variables, one layer after another, from the gamma distributions as in (12). With the simulated $\boldsymbol{\theta}_{j'}^{(1)}$, we calculate the Poisson rates for all the $V$ words using $\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_{j'}^{(1)}$ and display the top 100 words ranked by their Poisson rates. As shown below and in the Appendix, the synthetic documents generated in this manner are all easy to interpret and reflect various general aspects of the 20newsgroups corpus on which the PGBN is trained.

- mac apple bit mhz ram simms mb like memory just don cpu people chip chips think color board ibm speed does know se video time machines motherboard hardware lc cache meg ns simm need upgrade built vram good quadra want centris price dx run way processor card clock slots make fpu internal did macs cards ve pin power really machine say faster said software intel macintosh right week writes slot going sx performance things edu years nubus possible thing monitor work point expansion rom iisi ll add dram better little slow let sure pc ii didn ethernet lciii case kind

- image jpeg gif file color files images format bit display convert quality formats colors programs program tiff picture viewer graphics bmp bits xv screen pixel read compression conversion zip shareware scale view jpg original save quicktime jfif free version best pcx viewing bitmap gifs simtel viewers don mac usenet resolution animation menu scanner pixels sites gray quantization displays better try msdos tga want current black faq converting white setting mirror xloadimage section ppm fractal amiga write algorithm mpeg pict targa arithmetic export scodal archive converted grasp lossless let space human grey directory pictures rgb demo scanned old choice grayscale compress

- medical health disease doctor pain patients treatment medicine cancer edu hiv blood use years patient writes cause skin don like just aids symptoms number article help diseases drug com effects information doctors infection physician normal chronic think taking care volume condition drugs page says cure people tobacco hicnet know newsletter effective therapy problem common time women prevent surgery children center immune research called april control effect weeks low syndrome hospital physicians states clinical diagnosed day med age good make caused severe reported public safety child said cdc usually diet national studies tissue months way cases causing migraine smokeless infections does

- key use chip like encryption don used time just keys bit new people number think des make clipper know does data way good information work law using security algorithm need encrypted government escrow enforcement bits want say serial point ve edu available really random public right writes secure different case message better going secret probably second things technology block order nsa let problem long high com cryptography possible real example little note thing did game chips change access end ll standard sure problems doesn called large provide actually crypt non years agencies given idea privacy unit far able faq best

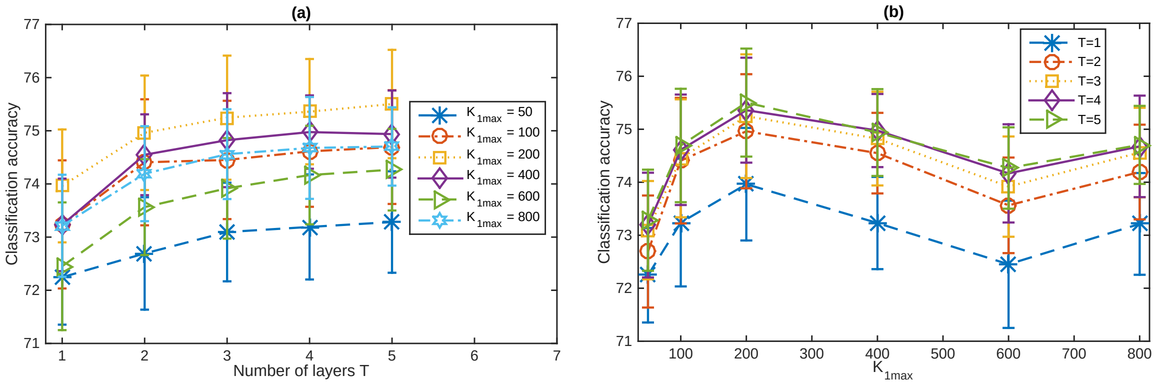


Figure 18: Analogous plots to Figure 14 for the BerPo-GBNs on the binarized 20newsgroups term-document count matrix. The widths of the hidden layers are automatically inferred. In a random trial with Algorithm 2, the inferred network widths $[K_1, \ldots, K_5]$ for $K_{1\,\max} = 50, 100, 200, 400, 600,$ and 800 are $[50, 50, 50, 50, 50]$, $[100, 97, 95, 90, 82]$, $[178, 145, 122, 97, 72]$, $[184, 139, 119, 101, 75]$, $[172, 165, 158, 138, 110]$, and $[156, 151, 147, 134, 117]$, respectively.

## 5.2 Multilayer representation for binary data

We apply the BerPo-GBN to extract multilayer representations for high-dimensional sparse binary vectors. The BerPo link is proposed in Zhou (2015) to construct edge partition models for network analysis, whose computation is mainly spent on pairs of linked nodes and hence is scalable to big sparse relational networks. That link function and its inference procedure have also been recently adopted by Hu et al. (2015) to analyze big sparse binary tensors.

We consider the same problem of feature learning for multi-class classification studied in detail in Section 5.1.2. We consider the same setting except that the original term-document word count matrix is now binarized into a term-document indicator matrix, the $(v, j)$ element of which is set as one if and only if $n_{vj} \geq 1$ and set as zero otherwise. We test the BerPo-GBNs on the 20newsgroups corpus, with $\eta^{(t)} = 0.05$ for all layers. As shown in Figure 18, given the same upper-bound on the width of the first layer, increasing the depth of the network clearly improves the performance. Whereas given the same number of hidden layers, the performance initially improves and then fluctuates as the upper-bound of the first layer increases. Such kind of fluctuations when $K_{1\,\max}$ reaches over 200 are expected, since the width of the first layer is inferred to be less than 190 and hence the budget as small as $K_{1\,\max} = 200$ is already large enough to cover all active factors.

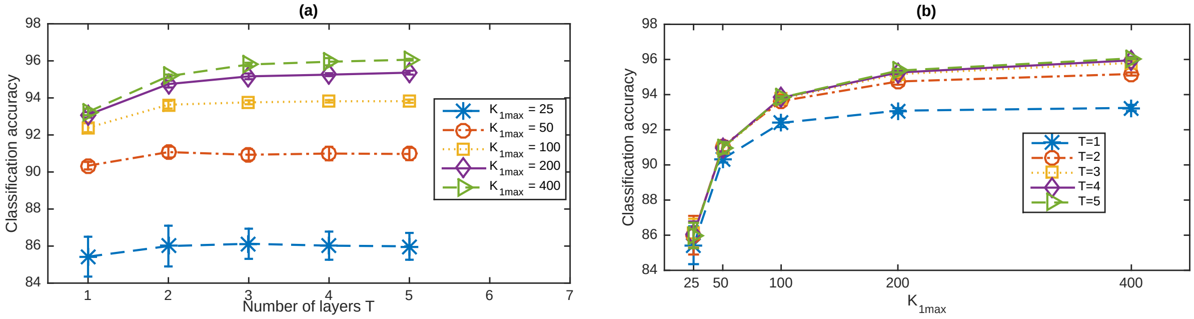
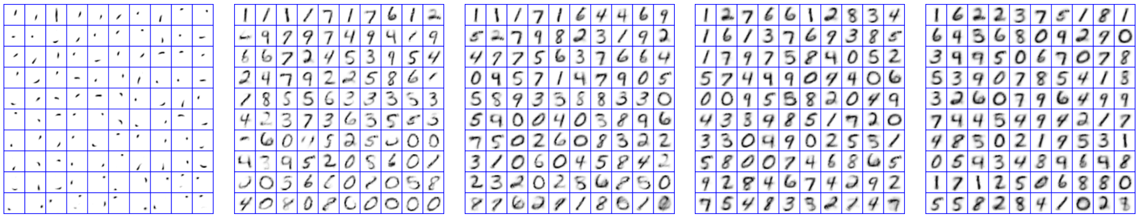


Figure 19: Analogous plots to Figure 14 for the PRG-GBNs on the MNIST dataset. In a random trial with Algorithm 2, the inferred network widths $[K_1, \ldots, K_5]$ for $K_{1\,\mathrm{max}}$=50, 100, 200, and 400 are $[50, 50, 50, 50, 50]$, $[100, 100, 100, 100, 100]$, $[200, 200, 200, 200, 200]$, and $[400, 400, 399, 385, 321]$, respectively.



(a) (b) (c) (d) (e)

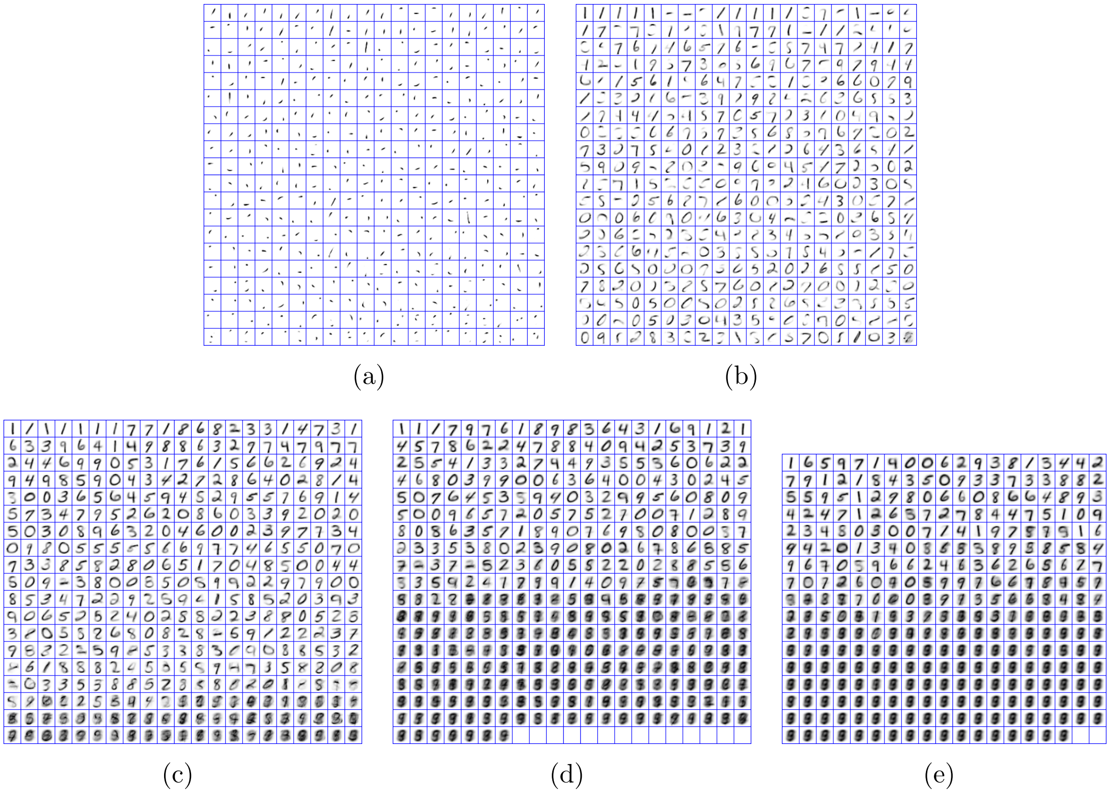
Figure 20: Visualization of the inferred $\{\Phi^{(1)}, \cdots, \Phi^{(T)}\}$ on the MNIST dataset using the PRG-GBN with $K_{1max} = 100$ and $\eta^{(t)} = 0.05$ for all $t$. The latent factors of all layers are projected to the first layer: (a) $\Phi^{(1)}$, (b) $\Phi^{(1)}\Phi^{(2)}$, (c) $\Phi^{(1)}\Phi^{(2)}\Phi^{(3)}$, (d) $\Phi^{(1)}\Phi^{(2)}\Phi^{(3)}\Phi^{(4)}$, and (e) $\Phi^{(1)}\Phi^{(2)}\Phi^{(3)}\Phi^{(4)}\Phi^{(5)}$.

## 5.3 Multilayer representation for nonnegative real data

We use the PRG-GBN to unsupervisedly extract features from nonnegative real data. We consider the MNIST dataset[2], which consists of 60000 training handwritten digits and 10000 testing ones. We divide the gray-scale pixel values of each $24 \times 24$ image by 255 and represent each image as a 784 dimensional nonnegative real vector. We set $\eta^{(1)} = 0.05$ and use all training digits to infer the PRG-GNBs with $T_{max} \in \{1, \cdots, 5\}$ and $K_{1max} \in \{50, 100, 200, 400\}$. We consider the same problem of feature extraction for multi-class classification studied in detail in Section 5.1.2, and we follow the same experimental settings over there. As shown in Figure 19, both increasing the width of the first layer and the depth of the network could clearly improve the performance in terms of unsupervisedly extracting features that are better suited for multi-class classification.

Note that the PRG distribution might not be the best distribution to fit MNIST digits, but nevertheless, displaying the inferred features at various layers as images provides a straightforward way to visualize the latent structures inferred from the data and hence provides an excellent example to understand the properties and working mechanisms of the

[2] http://yann.lecun.com/exdb/mnist/



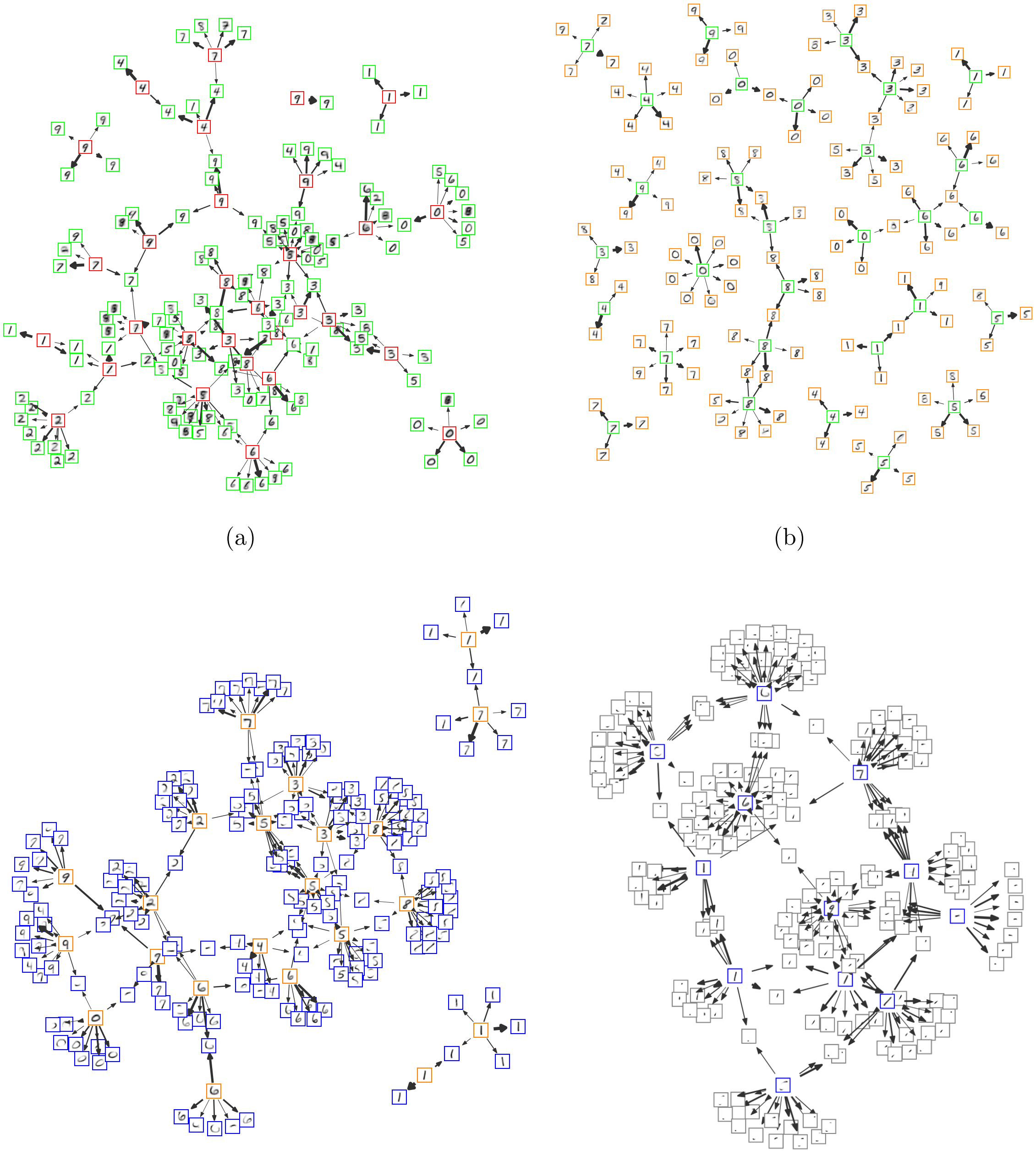
(a) (b) (c) (d) (e)

Figure 21: Visualization of the inferred $\{\Phi^{(1)}, \cdots, \Phi^{(T)}\}$ on the MNIST dataset using the PRG-GBN with $K_{1max} = 400$ and $\eta^{(t)} = 0.05$ for all $t$. The latent factors of all layers are projected to the first layer: (a) $\Phi^{(1)}$, (b) $\Phi^{(1)}\Phi^{(2)}$, (c) $\Phi^{(1)}\Phi^{(2)}\Phi^{(3)}$, (d) $\Phi^{(1)}\Phi^{(2)}\Phi^{(3)}\Phi^{(4)}$, and (e) $\Phi^{(1)}\Phi^{(2)}\Phi^{(3)}\Phi^{(4)}\Phi^{(5)}$.

GBN. We display the projections to the first layer of the factors $\boldsymbol{\Phi}^{(t)}$ at all five hidden layers as images for $K_{1\max} = 100$ and $K_{1\max} = 400$ in Figures 20 and 21, respectively, which clearly show that the inferred latent factors become increasingly more general as the layer increases. In both Figures 20 and 21, the latent factors inferred at the first hidden layer represent filters that are only active at very particular regions of the images, those inferred at the second hidden layer represent larger parts of the hidden-written digits, and those inferred at the third and deeper layers resemble the whole digits.

To visualize the relationships between the factors of different layers, we show in Figure 22 a subset of nodes of each layer and the nodes of the layer below that are connected to them with non-negligible weights.

It is interesting to note that unlike Lee et al. (2009) and many other following works that rely on the convolutional and pooling operations, which are pioneered by LeCun et al. (1989), to extract hierarchical representation for images at different spatial scales, we show that it is not necessary to break the images into spatial patches in order to learn the factors that are active on very specific regions of the image in the bottom hidden layer and to learn these increasingly more general factors covering larger parts of the images as the number



(a)

(b)

(c)

(d)

Figure 22: Visualization of the network structures inferred by the PRG-GBN on the MNIST dataset with $K_{1max} = 400$. (a) Visualization of the factors $(\boldsymbol{\phi}_1^{(5)}, \boldsymbol{\phi}_{11}^{(5)}, \boldsymbol{\phi}_{21}^{(5)}, \ldots, \boldsymbol{\phi}_{111}^{(5)})$ of layer five and those of layer four that are strongly connected to them. (b) Visualization of the factors $(\boldsymbol{\phi}_1^{(4)}, \boldsymbol{\phi}_6^{(4)}, \boldsymbol{\phi}_{11}^{(4)}, \ldots, \boldsymbol{\phi}_{106}^{(4)})$ of layer four and those of layer three that are strongly connected to them. (c)the Visualization of the factors $(\boldsymbol{\phi}_1^{(3)}, \boldsymbol{\phi}_6^{(3)}, \boldsymbol{\phi}_{11}^{(3)}, \ldots, \boldsymbol{\phi}_{146}^{(3)})$ of layer three and those of layer two that are strongly connected to them. (d) Visualization of the factors $(\boldsymbol{\phi}_1^{(2)}, \boldsymbol{\phi}_6^{(2)}, \boldsymbol{\phi}_{11}^{(2)}, \ldots, \boldsymbol{\phi}_{146}^{(2)})$ of layer two and those of layer one that are strongly connected to them.

of layer increases. It would also be interesting to investigate whether one can introduce convolutional and pooling operations into the GBNs, which may substantially improve their performance on modeling natural images.

# 6 Conclusions

The gamma belief network (GBN) is proposed to extract a multilayer representation for high-dimensional count, binary, or nonnegative real vectors, with an efficient upward-downward Gibbs sampler to jointly train all its layers and a layer-wise training strategy to automatically infer the network structure. A GBN of $T$ layers can be broken into $T$ subproblems that are solved by repeating the same subroutine, with the computation mainly spent on training the first hidden layer. When used for deep topic modeling, the GBN extracts very specific topics at the first hidden layer and increasingly more general topics at deeper hidden layers. It provides an excellent way for exploratory data analysis through the visualization of the inferred deep network, whose hidden units of adjacent layers are sparsely connected. Its good performance is further demonstrated in unsupervisedly extracting features for document classification and predicting heldout word tokens. The extracted deep network can also be used to simulate very interpretable synthetic documents, which reflect various general aspects of the corpus that the network is trained on. When applied for image analysis, without using the convolutional and pooling operations, the GBN is already able to extract interpretable factors in the first hidden layer that are active in very specific spatial regions and interpretable factors in deeper hidden layers with increasingly more general spatial patterns covering larger spatial regions. For big data problems, in practice one may rarely has a sufficient budget to allow the first-layer width to grow without bound, thus it is natural to consider a deep network that can use a multilayer deep representation to better allocate its resource and increase its representation power with limited computational power. Our algorithm provides a natural solution to achieve a good compromise between the width of each layer and the depth of the network.

# A Additional synthetic documents

- team game games hockey year cup season playoffs edu win pittsburgh nhl toronto detroit stanley teams montreal play jets pens espn division chicago new penguins pick league players devils rangers wings boston islanders playoff ca series winnipeg gm abc tv playing quebec april time round st vancouver fans best gld bruins coach winner calgary leafs player great watch night patrick vs finals conference final just baseball coverage murray minnesota don won gary points mike like ice kings regular mario played louis caps contact washington selanne norris buffalo columbia keenan star people fan th think canadiens said canada canucks york gerald

- year game team games god baseball runs season win won hit pitching play run space lost teams think ball believe sox better league edu does fan best home phillies th mets good braves cubs wins morris offense yankees score people al hitting new hitter got say division boston winning san reds second religion christian fans series era true pitch bible early red nl truth york defense evidence toronto base bob st east staff pitchers dodgers question clemens field faith lot jays chicago players christians innings strong years west john scored giants mike pitched batting talent belief rockies hits stadium reason

- hall smith players fame career ozzie winfield nolan guys ryan dave baseball eddie murray numbers steve kingman robinson yount morris roger years bsu puckett long joe jackson hung brett garvey deserve robin evans princeton yeah frank ruth kirby rickey pitcher peak yogi hof great sick lee ha aaron johnny darrell santo time greatest stats seasons ron george reardon shortstops henderson hank mays jack liability marginal rogers average compare belong schmidt gibson willie leo ucs sgi bsuvc comment fans honestly deserves cal bell candidates wagner fielding walks ve likely history gee heck consideration mike player bonds lock rating sandberg standards apparent

- fbi koresh batf gas compound waco government atf people children tear cult davidians did bd branch agents happened assault warrant david reno tanks killed weapons clinton point country search building federal raid press started reported death proper needed illegal better house protect burned janet outside burn days media stand job arms inside right come cwru equipment followers investigation oldham believe non power kids burning fires women suicide law order cs sick blame initial alive feds agent tank religious automatic davidian deaths knock good hit said military possible died away light fault child witnesses pay instead folks daniel bureau armored going

- people government law state israel rights israeli jews right public states war fact political country arab laws article case court human federal american united support society policy civil freedom members national jewish evidence person majority force power legal citizens action crime world act countries issue arabs group police justice non control palestinian live land peace true anti center writes gaza population research constitution death edu org allowed party protection consider actions number adam apc general subject based murder igc considered life military self parties lives personal nation order cpr social question individual religious today situation free responsibility governments palestine innocent

- people government law state rights public states right country laws political case court fact federal support civil society war united american person force power freedom human evidence legal national majority action members crime act article police justice issue group control constitution anti party death protection countries actions citizens consider non murder based considered general military world allowed lives parties personal nation life live individual social responsibility governments involved innocent cases free defense order president threat history self member status position private situation minority property supreme organization population believe example policy shall local argument present number process rule today america officers

- greeks edu turks turkey turkish greek people writes com article bad person napoleon tankut called bike iastate greece software compromise talk trying uiowa salonica ti dod city goverment conflicts atan aiu things don learn just business fred mau like apr educated politics cyprus dseg today make mccall helmet comes panos think makes know history books positive requirement government french policies visa hatred relations brainwashing tamamidis konstantinople seas kind gwu miles ride ve met suffering aegean thessaloniki doing level good mr started accept removing supports baby countries millions sides treaty aggressive istanbul minorities achtung anatolia ither solun osinski sevres somebody speak

- space nasa launch gov new use shuttle satellite information nist university data research program used technology center science ncsl national year time using project high flight systems commercial available earth cost power satellites mission large work design april list international access development following institute th orbit based number washington small office marc low general dr vehicle does area long york conference radio press org clipper technical solar jsc include note csrc station including market make services california air different aerospace security end news street missions major example non orbital expn payload world state problems single rocket service board agency order

- use power used output input high voltage does using signal like data circuit chip edu time low work amp digital know thanks com good mhz new supply design electronics pin don need current speed make just build line water tv audio radio writes way control number point want analog logic different mail information ve run sound range probably help think cooling hot test little common article problems large problem frequency hp nuclear chips khz people switch ac lower example signals led circuits designed devices level parts scope john note possible available set series board hc components heat built standard ttl

- windows dos file files program problem win ms run use disk using memory running driver microsoft pc programs version swap installed os copy sys machine screen exe edu ftp access software space manager need ram change window directory problems mb ini command try just zip network install desktop config mode error com create write set tried drivers application size applications loaded cica setup novell available unix norton load fine mouse support work time bat like apps utility text device nt utilities ve group called start virtual user came card having noticed option edit help pub users shell environment free drive

- god believe does religion say people christian true truth evidence question faith christians bible belief think religious reason exist human atheists claim argument christianity atheism existence point beliefs said person way statement don fact things mean wrong agree know atheist different example based sense simply word exists accept non understand read world saying false nature answer religions view position meaning words arguments edu means certainly life matter claims context course prove gods did assume ask opinion evil conclusion strong discussion explain follow universe particular come idea knowledge history rutgers use doubt questions correct quite given response concept athos definition wrote

- god radar believe jesus does people christian trw batman bmd say bible religion true christians faith truth free evidence christ detector question lucifer omnipotent created think belief exist reason religious life human christianity existence way atheists man jbrown said world claim things argument fact know atheism point beliefs person logically don did word choice read detectors different come example mean use statement moral lord wrong agree based omniscient atheist accept pink nature saying rutgers exists answer perfect words church sense evil non understand edu love simply false means good sin knowledge meaning used form effectively gun john strong princeton omnipotence

- msg food edu chinese writes taste article com reaction eat restaurant flavor allergic different cause apr foods carl sick causes natural sodium chen mot ingredients berkeley walter reactions related vms chemical salt bee stomach syndrome sugar react extracted sol restaurants case lots people wife poisoning bellcore heard things organization glutamate suspect jason symptoms body chocolate contain caltech beef pet tongue comm candy vomiting believe hard hold brother vax add double blind headache monosodium heart bar eaten dougb just don disclaimer unless does industry sources responsible substances shock compound yellow gps anaphylactic lundby like know michael ago doug cs changes saying

- sun edu xterm server com motif key writes window article run keyboard problem file echo lib terminal mit display xdm keys running set error patch sparc line sunos apr using subject export contrib characters usr string warning following keycode widget define openwindows type hostname version ff tcp character local map just tar open bin emacs client xt home available postscript does toolkit cg don like dec lcs ip login ncd know clients libxmu root protocol xmu patches xwd end based cs vt shift fix look manager decnet information tek workstation xdefaults think source xmodmap widgets term use remote xrdb ultrix

- men homosexual sex gay sexual homosexuality male don people partners promiscuous number just bi like study homosexuals percent cramer heterosexual think did dramatically numbers straight church reform population know report pythagorean life man good accept time said considered kinsey posted general optilink irrational social gays behavior way children make published johnson survey table new activity showing million statistics american sexuality shows want right women ve article ago exclusively eating virginia masters repent really say purpose member clayton apparent kolodny writes going press society evil function engaged relationships ryan evolutionary different does compared person join figure community edu chose interesting things

- card video drivers cards driver vga mode ati graphics windows diamond vesa bus svga support gateway dx pc modes color isa board version local bit memory vlb ultra pro eisa monitor new does mb stealth hz using based speedstar orchid colors available latest ram know work chip performance resolution fast screen speed tech million trident winbench dcoleman set problems yes et ftp results winmarks plus edu bbs zeos utexas vram bios robert win higher magazine utxvms able high interlaced viper com boards site weitek tseng chipset modem turbo software non resolutions far faster accelerated supports price meg ega mhz true

- blood washed lamb bull species endangered metrics process graphite reserved religion levels choose serum animal preferably devising netland barney ntaib kfu human mg dl quack tank plasma ditch glucose haapanen ala world ucs adapter hercules indiana frog tomh peach gu iskandar taib list innocent washington kodak christian pole durant silver bbs tom real review gibson infoworld impression gave steve coded winbench sexual male non use young conversion nailed animals tend humans definitely light virgin sacrifice eric judging rabbit fasting rahul percent adams concentration harvey wooden obsession dinosaur martyrdom pharvey starowl sieferman nlm screw fingers don nih seen pull mouth reviewers

- jews nazi nazis germany german jewish hitler people history don just like anti think holocaust war know did europe good book time party make way european semitism right world want new ve really say writes race does going germans palestine homosexuality roehm came important ss alchemy chem things citizenship archives zionist edu utoronto chancellor thing berlin himmler rockwell groups point state documents ll jew sa better university sure killed let books sources aryan citizens little diplomats homosexuals homosexual evidence doesn article course need paris information fact power probably racial france government look case years official law lot early help sent

- bike dod don just like think people good know com ride edu writes ve bikes time way make really want ll bmw going say new does article motorcycle right riding rider things did thing point better sure ama let little road doesn need lot probably far look got years honda org hard problem tell motorcycles didn course believe work actually great long real said isn harley try bad doing thought hp maybe come use fact ask yes having getting question street tank mean rear case looking pretty buy trying money cb ranck won advice quite used cost best help idea

- card windows video drivers monitor com modem vga cards driver port pc mode screen ati serial graphics dos bus board irq support svga diamond vesa using memory problem dx color gateway file version ports local modes pro bit does isa colors mb know vlb mouse ultra win ram new monitors hz work eisa nec problems chip files stealth use set program speedstar orchid plus high based resolution fast software cable hardware display latest used performance ms like baud bbs tech connector run thanks speed just yes million trident winbench dcoleman available pin ibm uart connect sony window switch et disk

- points sphere plane center circle point bolson radius washington image coplanar graphics line solution normal abc cylinder algorithm lines given non fitting pm intersection steve perpendicular don just geometry define gems planes like exactly parallel books research ed edward trivial carson lie straightforward equation intersect distance think people method boy loss faq checked data immediately mercy equations embarassing farin know defined provide analysis sequence form surface edu algorithms dist hitachi passing pairs intersecting bisector good protein images time original say biology numerical solve molecular space containing unique pair shelley square parameters way make want writes centre ve watson biological genetic

- car cars engine speed drive good just like driving don miles mph time new people think ford driver better know road power com models high model fast performance turbo way parts year driven use make tires vehicle mileage brakes wheel manual buy did door toyota transmission auto writes said ve does want lot price drivers market gear say designed owner gas really right looking centerline edu test doesn speeds problem years clutch automatic taurus going rear bmw odometer work used little jim owners things looks point porsche makes mile away thing previous safety throttle motor mind autos shift small need

- bike dod ride bikes com bmw motorcycle riding rider ama road honda org new motorcycles harley hp street tank rear ll cb ranck list gs miles advice good cbr cage buy dog cc mail insurance ve seca send rec vt gear frame bikers cost seat yamaha hard thanks twin helmet parking mitre looking rd jeff faq moto gas tire fj zx suzuki moa sold kz newsgroup leather plastic cookson sale blue buying hey cycle ducati email owners heavy area ask bars randy harleys nice car state old low sr riders edu es ninja chain wrote sport recommend ra vehicle rode

- nissan electronics wagon altima delcoelect kocrsv station gm subaru sumax delco spiros hughes wax pathfinder legacy kokomo wagons smorris scott toyota seattleu don just like strong silver software luxury derek proof stanza seattle cisco morris cymbal triantafyllopoulos sportscar think people know near fool ugly proud claims flat statistics lincoln sedans bullet karl lee perth puzzled miata sentra maxima acura infiniti corolla mgb untruth verbatim good time consider way based make stand guys writes noticed want ve heavy suggestion eat steven horrible uunet studies armor fisher lust designs study definately lexus remove conversion embodied aesthetic elvis attached honey stole designing wd

- gun guns edu writes bike com article weapons dod control crime weapon apr used carry criminals police ride nra bikes self firearms use buy firearm laws concealed bmw defense home handgun criminal motorcycle anti problem car people owners ban rider riding shot just armed new don like crimes assault kill violent protect uio handguns ifi evil ama citizens state org know illegal politics texas thomas thomasp cb talk legal shooting pro road carrying abiding think att honda cs stolen defend good purchase ll law individual hp cc permit rifle issue government states parsli property ve killing federal does motorcycles time

- gun guns weapons people control government law crime state rights police laws weapon self criminals carry states public nra used defense firearms anti federal right criminal legal firearm citizens country home political case concealed handgun court fact crimes issue protect armed politics kill ban problem buy national individual support shot society violent use civil war property talk owners assault illegal handguns ifi uio united defend action allowed freedom article american amendment person member power force thomasp car human evidence threat thomas murder shooting majority killed carrying members citizen killing pro abiding group act evil texas america justice permit stolen said

# References

A. Acharya, J. Ghosh, and M. Zhou. Nonparametric Bayesian factor analysis for dynamic count matrices. In *AISTATS*, 2015.

R. P. Adams, Z. Ghahramani, and M. I. Jordan. Tree-structured stick breaking for hierarchical data. In *NIPS*, 2010.

D. Aldous. Exchangeability and related topics. *École d'ete de probabilités de Saint-Flour XIII-1983*, pages 1–198, 1985.

F. J. Anscombe. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, 1950.

C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, 1974.

Y. Bengio and Y. LeCun. Scaling learning algorithms towards AI. In Léon Bottou, Olivier Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Machines*. MIT Press, 2007.

Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *NIPS*, 2007.

Y. Bengio, I. J. Goodfellow, and A. Courville. Deep Learning. Book in preparation for MIT Press, 2015.

D. Blackwell and J. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1973.

D. Blei and J. Lafferty. Correlated topic models. *NIPS*, 2006.

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 2003.

D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of ACM*, 2010.

W. Buntine and A. Jakulin. Discrete component analysis. In *Subspace, Latent Structure and Feature Selection Techniques*. Springer-Verlag, 2006.

J. Canny. Gap: a factor model for discrete data. In *SIGIR*, 2004.

J. Chen, J. Zhu, Z. Wang, X. Zheng, and B. Zhang. Scalable inference for logistic-normal topic models. In *NIPS*, 2013.

S. C. Choi and R. Wette. Maximum likelihood estimation of the parameters of the gamma distribution and their bias. *Technometrics*, pages 683–690, 1969.

L. Devroye. Simulating Bessel random variables. *Statistics & probability letters*, 57(3):249–257, 2002.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, pages 1871–1874, 2008.

R. A. Fisher, A. Steven Corbet, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 1943.

Z. Gan, C. Chen, R. Henao, D. Carlson, and L. Carin. Scalable deep poisson factor analysis for topic modeling. In *ICML*, 2015.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 2004.

G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, pages 1771–1800, 2002.

G. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, pages 1527–1554, 2006.

C. Hu, P. Rai, and L. Carin. Zero-truncated poisson tensor factorization for massive binary tensors. In *UAI*, 2015.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete multivariate distributions*, volume 165. Wiley New York, 1997.

H. Larochelle and S. Lauly. A neural autoregressive topic model. In *NIPS*, 2012.

Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2001.

H. Lee, R. Grosse, R. Ranganath, and A. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, 2009.

S. W. Linderman, M. J. Johnson, and R. P. Adams. Dependent multinomial models made easy: Stick breaking with the Pólya-Gamma augmentation. *arXiv preprint arXiv:1506.05843*, 2015.

V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *ICML*, 2010.

R. M. Neal. Connectionist learning of belief networks. *Artificial Intelligence*, pages 71–113, 1992.

D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. *JMLR*, 2009.

J. Paisley, C. Wang, and D. M. Blei. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 2012.

J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan. Nested hierarchical dirichlet processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015.

J. Pitman. *Combinatorial stochastic processes.* Lecture Notes in Mathematics. Springer-Verlag, 2006.

R. Ranganath and D. Blei. Correlated random measures. *arXiv:1507.00720v1*, 2015.

R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *AISTATS*, 2014.

R. Ranganath, L. Tang, L. Charlin, and D. M. Blei. Deep exponential families. In *AISTATS*, 2015.

M Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *CVPR*, 2007.

R. Salakhutdinov and G. E. Hinton. Deep Boltzmann machines. In *AISTATS*, 2009.

R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba. Learning with hierarchical-deep models. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1958–1971, 2013.

L. K. Saul, T. Jaakkola, and M. I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence research*, pages 61–76, 1996.

N. Srivastava, R. Salakhutdinov, and G. Hinton. Modeling documents with a deep Boltzmann machine. In *UAI*, 2013.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 2006.

H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *ICML*, 2009.

M. Welling, M. Rosen-Zvi, and G. E. Hinton. Exponential family harmoniums with an application to information retrieval. In *NIPS*, pages 1481–1488, 2004.

S. Williamson, C. Wang, K. A. Heller, and D. M. Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*, 2010.

E. P. Xing, R. Yan, and A. G. Hauptmann. Mining associated text and images with dual-wing harmoniums. In *UAI*, 2005.

L Yuan and J. D. Kalbfleisch. On the Bessel distribution and related problems. *Annals of the Institute of Statistical Mathematics*, 52(3):438–447, 2000.

M. Zhou. Beta-negative binomial process and exchangeable random partitions for mixed-membership modeling. In *NIPS*, 2014.

M. Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *AISTATS*, 2015.

M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015.

M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, 2012.

M. Zhou, Y. Cong, and B. Chen. The Poisson gamma belief network. In *NIPS*, Dec. 2015a.

M. Zhou, O. H. M. Padilla, and J. G. Scott. Priors for random count matrices derived from a family of negative binomial processes. *to appear in J. Amer. Statist. Assoc.*, 2015b.