Network Discovery and Recommendation via Joint Network and Topic Modeling

Ayan Acharya
Dept. of ECE, UT Austin
aacharya@utexas.edu

Dean Teffer
ARL, UT Austin
dean.teffer@arlut.utexas.edu

Mingyuan Zhou
Dept. of IROM, UT Austin
mzhou@utexas.edu

Joydeep Ghosh
Dept. of ECE, UT Austin
ghosh@ece.utexas.edu

ABSTRACT

Many complex social and biological interactions can be represented as graphs. Often such interaction networks also come with count-valued side-information that influence these interactions. Similarly, in applications like recommender systems, the count-valued rating matrix, capturing the interaction between users and items, may also be associated with auxiliary information in the form of a social network of users, which can conveniently be represented as a binary matrix. Jointly modeling both the primary as well as side information can result in better predictive performance. This paper introduces one such model, Joint Gamma Process Poisson Factorization (J-GPPF), which jointly models a binary symmetric interaction matrix $\mathbf{B}$ and a count matrix $\mathbf{Y}$. The model discovers the ideal number of latent factors for modeling both $\mathbf{B}$ and $\mathbf{Y}$ from the data itself and outperforms strong baselines that do not use the side-information at all. We derive closed form updates for Gibbs sampling and predict the missing values in both $\mathbf{B}$ and $\mathbf{Y}$ in a mathematically consistent way.

Keywords

Network modeling, Poisson factorization, Gamma Process

1. INTRODUCTION

Social networks and other relational datasets can conveniently be represented using a binary symmetric adjacency matrix $\mathbf{B} \in \{0,1\}^{N\times N}$. Often, the nodes in such datasets are also associated with "side information", such as documents read or written, movies rated, or messages sent by these nodes, all of which can be represented as count-valued side information matrix $\mathbf{Y} \in \mathbb{Z}^{D\times V}$, where $\mathbb{Z} = \{0, 1, \ldots\}$. For example, $\mathbf{B}$ may represent a coauthor network and $\mathbf{Y}$ may correspond to a document-by-word count matrix representing the documents written by all these authors. In another example, $\mathbf{B}$ may represent a user-by-user social network and $\mathbf{Y}$ may represent a user-by-item rating matrix. In this particular setting, $\mathbf{B}$ can be considered as "side information" and used for modeling $\mathbf{Y}$ better. Incorporating such side information can result in better modeling of both $\mathbf{B}$ and $\mathbf{Y}$ when either of these matrices lacks sufficient information.

This paper proposes **J**oint **G**amma **P**rocess **P**oisson **F**actorization (J-GPPF), to jointly factorize $\mathbf{B}$ and $\mathbf{Y}$ in a nonparametric Bayesian manner. The paper makes the following contributions:

- We present a fast and effective model that uses both $\mathbf{B}$ and $\mathbf{Y}$ to help discover better network structures and cater better recommendation.
- We perform nonparametric Bayesian modeling for discovering latent structures in both $\mathbf{B}$ and $\mathbf{Y}$ and predict on missing entries in both $\mathbf{B}$ and $\mathbf{Y}$.
- Our model scales with the number of non-zero entries $S_{\mathbf{B}}$ in the network and the number of non-zero entries $S_{\mathbf{Y}}$ in the count matrix as $\mathrm{O}\left(S_{\mathbf{B}}K_{\mathbf{B}} + S_{\mathbf{Y}}K_{\mathbf{Y}}\right)$, where $K_{\mathbf{B}}$ is the number of network groups and $K_{\mathbf{Y}}$ is the number of latent factors in the count matrix, preset at a high value while performing nonparametric modeling.

The remainder of the paper is organized as follows. We present background material in Section 2. J-GPPF and its inference algorithm are explained in Section 3. The related works are presented in Section 4. Experimental results are reported in Section 5, followed by conclusions and scope of future work in Section 6.

2. BACKGROUND

2.1 Negative Binomial Distribution

The definition of the negative binomial (NB) distribution is given in [1]. Here we introduce a few relevant lemmas which we use to derive the Gibbs sampling updates in Section 3.

Lemma 2.1 ([31]). *If $m \sim \mathrm{NB}(r,p)$ is represented under its compound Poisson representation, then the conditional posterior of l given m and r has PMF:*

$$\Pr(l=j|m,r) = \frac{\Gamma(r)}{\Gamma(m+r)}|s(m,j)|r^j,\ j=0,1,\cdots,m, \tag{1}$$

where $|s(m,j)|$ are unsigned Stirling numbers of the first kind. We denote this conditional posterior as $(l|m,r) \sim \mathrm{CRT}(m,r)$, a Chinese restaurant table (CRT) count random variable, which can be generated via $l = \sum_{n=1}^{m} z_n, z_n \sim \mathrm{Bernoulli}(r/(n-1+r))$.

Lemma 2.2. *Let $X = \sum_{k=1}^{K} x_k$, $x_k \sim \mathrm{Pois}(\zeta_k)\ \forall k$, and $\zeta = \sum_{k=1}^{K}\zeta_k$. If $(y_1,\cdots,y_K|X) \sim \mathrm{Mult}(X,\zeta_1/\zeta,\cdots,\zeta_K/\zeta)$ and $X \sim \mathrm{Pois}(\zeta)$, then the following holds:*

$$P(X,x_1,\cdots,x_K) = P(X,y_1,\cdots,y_K). \tag{2}$$

Lemma 2.3. *If* $x_i \sim \text{Pois}(m_i\lambda),\ \lambda \sim \text{Gam}(r, 1/c)$*, then* $x = \sum_i x_i \sim \text{NB}(r, p)$*, where* $p = (\sum_i m_i)/(c + \sum_i m_i)$*.*

Lemma 2.4. *If* $x_i \sim \text{Pois}(m_i\lambda),\ \lambda \sim \text{Gam}(r, 1/c)$*, then*

$$(\lambda|\{x_i\}, r, c) \sim \text{Gam}\left(r + \sum_i x_i, \frac{1}{c + \sum_i m_i}\right). \quad (3)$$

Lemma 2.5. *If* $r_i \sim \text{Gam}(a_i, 1/b)\ \forall i,\ b \sim \text{Gam}(c, 1/d)$*, then we have:*

$$(b|\{r_i, a_i\}, c, d) \sim \text{Gam}\left(\sum_i a_i + c, \frac{1}{\sum_i r_i + d}\right). \quad (4)$$

The proofs of Lemmas 2.3, 2.4 and 2.5 follow from the definitions of Gamma, Poisson and Negative Binomial distributions.

Lemma 2.6. *If* $x_i \sim \text{Pois}(m_i r_2),\ r_2 \sim \text{Gam}(r_1, 1/d),\ r_1 \sim \text{Gam}(a, 1/b),$ *, then* $(r_1|-) \sim \text{Gam}(a + \ell, 1/(b - log(1-p)))$ *where* $(\ell|x, r_1) \sim \text{CRT}(\sum_i x_i, r_1), p = \sum_i m_i/(d + \sum_i m_i)$*. The proof and illustration can be found in Section 3.3 of [1].*

2.2 Gamma Process

For non-parametric modeling, we employ Gamma processes, a detailed description of which is provided in [1; 11; 28].

3. JOINT GAMMA PROCESS POISSON FACTORIZATION (J-GPPF)

Let there be a network of N users encoded in an $N \times N$ binary matrix $\boldsymbol{B}$. The users in the network participate in writing D documents summarized in a $D \times V$ count matrix $\boldsymbol{Y}$, where V is the size of the vocabulary. Additionally, a binary matrix $\mathbf{Z}$ of dimension $D \times N$ can also be maintained, where the unity entries in each column indicate the set of documents in which the corresponding user contributes. In applications where $\boldsymbol{B}$ represents a user-by-user social network and $\boldsymbol{Y}$ represents a user-by-item rating matrix, $\mathbf{Z}$ turns out to be an N-dimensional identity matrix and hence can be considered as a special case of the document-author framework, which we pursue to describe the model. Also, to make the notations more explicit, the variables associated with the side information have $\boldsymbol{Y}$ as a subscript (e.g., $G_{\boldsymbol{Y}}$) and those associated with the network make similar use of the subscript $\boldsymbol{B}$ (e.g., $G_{\boldsymbol{B}}$).

We employ two separate Gamma Processes. The first one models the latent factors in the network. A draw from this Gamma Process $G_{\boldsymbol{B}} \sim \Gamma\text{P}(c_{\boldsymbol{B}}, H_{\boldsymbol{B}})$ is expressed as $G_{\boldsymbol{B}} = \sum_{k_{\boldsymbol{B}}=1}^{\infty} \rho_{k_{\boldsymbol{B}}} \delta_{\phi_{k_{\boldsymbol{B}}}}$, where $\boldsymbol{\phi}_{k_{\boldsymbol{B}}} \in \Omega_{\boldsymbol{B}}$ is an atom drawn from an N-dimensional base distribution as $\boldsymbol{\phi}_{k_{\boldsymbol{B}}} \sim \prod_{n=1}^{N} \text{Gam}(a_{\boldsymbol{B}}, 1/\sigma_n)$ and $\rho_{k_{\boldsymbol{B}}} = G_{\boldsymbol{B}}(\boldsymbol{\phi}_{k_{\boldsymbol{B}}})$ is the associated weight. The second Gamma Process models the latent groups of side information. A draw from this gamma process $G_{\boldsymbol{Y}} \sim \Gamma\text{P}(c_{\boldsymbol{Y}}, H_{\boldsymbol{Y}})$ is expressed as $G_{\boldsymbol{Y}} = \sum_{k_{\boldsymbol{Y}}=1}^{\infty} r_{k_{\boldsymbol{Y}}} \delta_{\boldsymbol{\beta}_{k_{\boldsymbol{Y}}}}$, where $\boldsymbol{\beta}_{k_{\boldsymbol{Y}}} \in \Omega_{\boldsymbol{Y}}$ is an atom drawn from a V-dimensional base distribution as $\boldsymbol{\beta}_{k_{\boldsymbol{Y}}} \sim \prod_{w=1}^{V} \text{Gam}(\xi_{\boldsymbol{Y}}, 1/\zeta_w)$ and $r_{k_{\boldsymbol{Y}}} = G_{\boldsymbol{Y}}(\boldsymbol{\beta}_{k_{\boldsymbol{Y}}})$ is the associated weight. Also, $\gamma_{\boldsymbol{B}} = H_{\boldsymbol{B}}(\Omega_{\boldsymbol{B}})$ is defined as the mass parameter corresponding to the base measure $H_{\boldsymbol{B}}$ and $\gamma_{\boldsymbol{Y}} = H_{\boldsymbol{Y}}(\Omega_{\boldsymbol{Y}})$ is defined as the mass parameter corresponding to the base measure $H_{\boldsymbol{Y}}$. The $(n, m)^{\text{th}}$ entry in the matrix $\boldsymbol{B}$ is assumed to be derived from a latent count as:

$$b_{nm} = \mathbb{I}_{\{x_{nm} \geq 1\}},\ x_{nm} \sim \text{Pois}\left(\lambda_{nm}\right), \lambda_{nm} = \sum_{k_{\boldsymbol{B}}} \lambda_{nmk_{\boldsymbol{B}}},$$

where $\lambda_{nmk_{\boldsymbol{B}}} = \rho_{k_{\boldsymbol{B}}} \phi_{nk_{\boldsymbol{B}}} \phi_{mk_{\boldsymbol{B}}}$. This is called as the Poisson-Bernoulli (PoBe) link in [1; 30]. The distribution of b_{nm} given λ_{nm} is named as the Poisson-Bernoulli distribution, with the PMF:

$$f(b_{nm}|\lambda_{nm}) = e^{-\lambda_{nm}(1-b_{nm})}(1 - e^{-\lambda_{nm}})^{b_{nm}}.$$

One may consider $\lambda_{nmk_{\boldsymbol{B}}}$ as the strength of mutual latent community membership between nodes n and m in the network for latent community $k_{\boldsymbol{B}}$, and λ_{nm} as the interaction strength aggregating all possible community membership. Using Lemma 2.2, one may augment the above representation as $x_{nm} = \sum_{k_{\boldsymbol{B}}} x_{nmk_{\boldsymbol{B}}}$, $x_{nmk_{\boldsymbol{B}}} \sim \text{Pois}\left(\lambda_{nmk_{\boldsymbol{B}}}\right)$. Thus each interaction pattern contributes a count and the total latent count aggregates the countably infinite interaction patters. Unlike the usual approach that links the binary observations to latent Gaussian random variables with a logistic or probit function, the above approach links the binary observations to Poisson random variables, providing several potential advantages. First, it is more interpretable because $\rho_{k_{\boldsymbol{B}}}$ and $\boldsymbol{\phi}_{k_{\boldsymbol{B}}}$ are non-negative and the aggregation of different interaction patterns increases the probability of establishing a link between two nodes. Second, the computational benefit is significant since the computational complexity is approximately linear in the number of non-zeros $S_{\boldsymbol{B}}$ in the observed binary adjacency matrix $\boldsymbol{B}$.

To model the matrix $\boldsymbol{Y}$, its $(d, w)^{\text{th}}$ entry y_{dw} is generated as:

$$y_{dw} \sim \text{Pois}(\zeta_{dw}), \zeta_{dw} = \left(\sum_{k_{\boldsymbol{Y}}} \zeta_{\boldsymbol{Y}dwk_{\boldsymbol{Y}}} + \sum_{k_{\boldsymbol{B}}} \zeta_{\boldsymbol{B}dwk_{\boldsymbol{B}}}\right),$$

where $\zeta_{\boldsymbol{Y}dwk_{\boldsymbol{Y}}} = r_{k_{\boldsymbol{Y}}} \theta_{dk_{\boldsymbol{Y}}} \beta_{wk_{\boldsymbol{Y}}}$, $Z_{nd} \in \{0, 1\}$ and $Z_{nd} = 1$ if and only if author n is one of the authors of paper d and $\zeta_{\boldsymbol{B}dwk_{\boldsymbol{B}}} = \epsilon\rho_{k_{\boldsymbol{B}}} \left(\sum_n Z_{nd}\phi_{nk_{\boldsymbol{B}}}\right) \psi_{wk_{\boldsymbol{B}}}$. One can consider ζ_{dw} as the affinity of document d for word w, This affinity is influenced by two different components, one of which comes from the network modeling. Without the contribution from network modeling, the joint model reduces to a gamma process Poisson matrix factorization model, in which the matrix $\boldsymbol{Y}$ is factorized in such a way that $y_{dw} \sim \text{Pois}\left(\sum_{k_{\boldsymbol{Y}}} r_{k_{\boldsymbol{Y}}} \theta_{dk_{\boldsymbol{Y}}} \beta_{wk_{\boldsymbol{Y}}}\right)$. Here, $\boldsymbol{\Theta} \in \mathbb{R}_+^{D \times \infty}$ is the factor score matrix, $\boldsymbol{\beta} \in \mathbb{R}_+^{V \times \infty}$ is the factor loading matrix (or dictionary) and $r_{k_{\boldsymbol{Y}}}$ signifies the weight of the $k_{\boldsymbol{Y}}^{\text{th}}$ factor. The number of latent factors, possibly smaller than both D and V, would be inferred from the data.

In the proposed joint model, $\boldsymbol{Y}$ is also determined by the users participating in writing the d^{th} document. We assume that the distribution over word counts for a document is a function of *both* its topic distribution *as well as* the characteristics of the users associated with it. In the author-document framework, the authors employ different writing styles and have expertise in different domains. In the user-rating framework, the entries in $\boldsymbol{Y}$ are also believed to be influenced by the interaction network of the users. Such influence of the authors is modeled by the interaction of the authors in the latent communities *via* the latent factors $\boldsymbol{\phi} \in \mathbb{R}_+^{N \times \infty}$ and $\boldsymbol{\psi} \in \mathbb{R}_+^{V \times \infty}$, which encodes the writing style of the authors belonging to different latent communities. Since an infinite number of network communities is maintained, each entry y_{dw} is assumed to come from an infinite dimensional interaction. $\rho_{k_{\boldsymbol{B}}}$ signifies the interaction strength corresponding to the $k_{\boldsymbol{B}}^{\text{th}}$ network community. The contributions of the interaction from all the authors participating in a given document are accumulated to produce the total contribution from the networks in generating y_{dw}. Since $\mathbf{B}$ and $\mathbf{Y}$ might have different levels of sparsity and the range of integers in $\mathbf{Y}$ can be quite large, a parameter ϵ is required to balance the contribution of the network communities in dictating the structure of $\mathbf{Y}$. A low value of ϵ forces disjoint modeling of $\mathbf{B}$ and $\mathbf{Y}$, while a higher value implies joint modeling of $\mathbf{B}$ and $\mathbf{Y}$ where information can flow both ways, from network discovery to topic discovery and vice-versa. To complete the generative process, we put Gamma

$$(\phi_{nk_{\boldsymbol{B}}}|-) \sim \text{Gam}\left(a_{\boldsymbol{B}} + \sum_{m=1}^{(n-1)} x_{mnk_{\boldsymbol{B}}} + \sum_{m=(n+1)}^{N} x_{nmk_{\boldsymbol{B}}} + y_{\cdot n\cdot k_{\boldsymbol{B}}}, \left(\sigma_n + \rho_{k_{\boldsymbol{B}}}\left(\sum_{\substack{m=1\\ m\neq n}}^{N} \phi_{mk_{\boldsymbol{B}}} + \epsilon \sum_{d,w} Z_{nd}\psi_{wk_{\boldsymbol{B}}}\right)\right)^{-1}\right),$$

$$(\psi_{wk_{\boldsymbol{B}}}|-) \sim \text{Gam}\left(\xi_{\boldsymbol{B}} + y_{\cdot\cdot wk_{\boldsymbol{B}}}, \frac{1}{\zeta_w + \epsilon\rho_{k_{\boldsymbol{B}}}\sum_{d,n} Z_{nd}\phi_{nk_{\boldsymbol{B}}}}\right), (\theta_{dk_{\boldsymbol{Y}}}|-) \sim \text{Gam}\left(a_{\boldsymbol{Y}} + y_{d\cdot k_{\boldsymbol{Y}}}, \frac{1}{\varsigma_d + r_{k_{\boldsymbol{Y}}}\beta_{\cdot k_{\boldsymbol{Y}}}}\right), (\varsigma_d|-) \sim \text{Gam}\left(\alpha_{\boldsymbol{Y}} + K_{\boldsymbol{Y}}a_{\boldsymbol{Y}}, \frac{1}{\varepsilon_{\boldsymbol{Y}} + \theta_{d\cdot}}\right),$$

$$(\rho_{k_{\boldsymbol{B}}}|-) \sim \text{Gam}\left(\frac{\gamma_{\boldsymbol{B}}}{K_{\boldsymbol{B}}} + \sum_{\substack{(n,m)\\ n<m}} x_{nmk_{\boldsymbol{B}}} + y_{\cdots k_{\boldsymbol{B}}}, \left(c_{\boldsymbol{B}} + \sum_{\substack{(n,m)\\ n<m}} \phi_{nk_{\boldsymbol{B}}}\phi_{mk_{\boldsymbol{B}}} + \epsilon \sum_{n,d,w} Z_{nd}\phi_{nk_{\boldsymbol{B}}}\psi_{wk_{\boldsymbol{B}}}\right)^{-1}\right),$$

$$(r_{k_{\boldsymbol{Y}}}|-) \sim \text{Gam}\left(\frac{\gamma_{\boldsymbol{Y}}}{K_{\boldsymbol{Y}}} + y_{\cdot\cdot k_{\boldsymbol{Y}}}, \frac{1}{c_{\boldsymbol{Y}} + \theta_{\cdot k_{\boldsymbol{Y}}}\beta_{\cdot k_{\boldsymbol{Y}}}}\right), (\beta_{wk_{\boldsymbol{Y}}}|-) \sim \text{Gam}\left(\xi_{\boldsymbol{Y}} + y_{\cdot wk_{\boldsymbol{Y}}}, \frac{1}{\eta_w + r_{k_{\boldsymbol{Y}}}\theta_{\cdot k_{\boldsymbol{Y}}}}\right),$$

$$(c_{\boldsymbol{B}}|-) \sim \text{Gam}\left(g_{\boldsymbol{B}} + \gamma_{\boldsymbol{B}}, \frac{1}{h_{\boldsymbol{B}} + \sum_{k_{\boldsymbol{B}}}\rho_{k_{\boldsymbol{B}}}}\right), (c_{\boldsymbol{Y}}|-) \sim \text{Gam}\left(g_{\boldsymbol{Y}} + \gamma_{\boldsymbol{Y}}, \frac{1}{h_{\boldsymbol{Y}} + \sum_{k_{\boldsymbol{Y}}} r_{k_{\boldsymbol{Y}}}}\right), (\epsilon|-) \sim \text{Gam}\left(f_0 + \sum_{k=1}^{K_{\boldsymbol{B}}} y_{\cdots k}, \left(g_0 + \sum_{k=1}^{K_{\boldsymbol{B}}}\rho_{k_{\boldsymbol{B}}}\sum_{n=1}^{N}|\mathcal{Z}_n|\phi_{nk_{\boldsymbol{B}}}\right)^{-1}\right),$$

$$(\zeta_w|-) \sim \text{Gam}\left(a_0 + K_{\boldsymbol{B}}\xi_{\boldsymbol{B}}, \frac{1}{b_0 + \psi_{w\cdot}}\right), (\eta_w|-) \sim \text{Gam}\left(c_0 + K_{\boldsymbol{Y}}\xi_{\boldsymbol{Y}}, \frac{1}{d_0 + \beta_{w\cdot}}\right),$$

Table 1: Gibbs sampling updates in J-GPPF

$$(\phi_{nk_{\boldsymbol{B}}}|-) \sim \text{Gam}\left(a_{\boldsymbol{B}} + \sum_{\substack{m=1\\ (n,m)\notin\mathcal{M}_{\boldsymbol{B}}}}^{(n-1)} x_{mnk_{\boldsymbol{B}}} + \sum_{\substack{m=(n+1)\\ (n,m)\notin\mathcal{M}_{\boldsymbol{B}}}}^{N} x_{nmk_{\boldsymbol{B}}} + \sum_{\substack{d,w\\ (d,w)\notin\mathcal{M}_{\boldsymbol{Y}}}} y_{dnwk_{\boldsymbol{B}}}, \left(\sigma_n + \rho_{k_{\boldsymbol{B}}}\left(\sum_{\substack{m=1\\ (n,m)\notin\mathcal{M}_{\boldsymbol{B}}}}^{N} \phi_{mk_{\boldsymbol{B}}} + \epsilon \sum_{\substack{d,w\\ (d,w)\notin\mathcal{M}_{\boldsymbol{Y}}}} Z_{nd}\psi_{wk_{\boldsymbol{B}}}\right)\right)^{-1}\right),$$

$$(\psi_{wk_{\boldsymbol{B}}}|-) \sim \text{Gam}\left(\xi_{\boldsymbol{B}} + \sum_{\substack{d\\ (d,w)\notin\mathcal{M}_{\boldsymbol{Y}}}} y_{dwk_{\boldsymbol{B}}}, \left(\eta_w + \epsilon\rho_{k_{\boldsymbol{B}}} \sum_{\substack{n,d\\ (d,w)\notin\mathcal{M}_{\boldsymbol{Y}}}} Z_{nd}\phi_{nk_{\boldsymbol{B}}}\right)^{-1}\right),$$

$$(\rho_{k_{\boldsymbol{B}}}|-) \sim \text{Gam}\left(\frac{\gamma_{\boldsymbol{B}}}{K_{\boldsymbol{B}}} + \sum_{\substack{n=1,n<m\\ (n,m)\notin\mathcal{M}_{\boldsymbol{B}}}}^{(N-1)} x_{nmk_{\boldsymbol{B}}} + \sum_{\substack{n,d,w\\ (d,w)\notin\mathcal{M}_{\boldsymbol{Y}}}} y_{dnwk_{\boldsymbol{B}}}, \left(c_{\boldsymbol{B}} + \sum_{\substack{n=1,n<m\\ (n,m)\notin\mathcal{M}_{\boldsymbol{B}}}}^{(N-1)} \phi_{nk_{\boldsymbol{B}}}\phi_{mk_{\boldsymbol{B}}} + \epsilon \sum_{\substack{n,d,w\\ (d,w)\notin\mathcal{M}_{\boldsymbol{Y}}}} Z_{nd}\phi_{nk_{\boldsymbol{B}}}\psi_{wk_{\boldsymbol{B}}}\right)^{-1}\right).$$

Table 2: Sampling of $\phi_{nk_{\boldsymbol{B}}}$, $\psi_{wk_{\boldsymbol{B}}}$, $\rho_{k_{\boldsymbol{B}}}$ in J-GPPF with missing entries

priors over $c_{\boldsymbol{B}}$, $c_{\boldsymbol{Y}}$, σ_n, ς_d and ϵ as:

$$c_{\boldsymbol{B}} \sim \text{Gam}(g_{\boldsymbol{B}}, 1/h_{\boldsymbol{B}}), c_{\boldsymbol{Y}} \sim \text{Gam}(g_{\boldsymbol{Y}}, 1/h_{\boldsymbol{Y}}),$$
$$\epsilon \sim \text{Gam}(g_0, 1/f_0),$$
$$\sigma_n \sim \text{Gam}(\alpha_{\boldsymbol{B}}, 1/\varepsilon_{\boldsymbol{B}}), \varsigma_d \sim \text{Gam}(\alpha_{\boldsymbol{Y}}, 1/\varepsilon_{\boldsymbol{Y}}).$$

3.1 Inference *via* Gibbs Sampling

Though J-GPPF supports countably infinite number of latent communities for network modeling and infinite number of latent factors for topic modeling, in practice it is impossible to instantiate all of them. We consider a finite approximation of the infinite model by truncating the number of graph communities and the latent topics to $K_{\boldsymbol{B}}$ and $K_{\boldsymbol{Y}}$ respectively, by letting $\rho_{k_{\boldsymbol{B}}} \sim \text{Gam}(\gamma_{\boldsymbol{B}}/K_{\boldsymbol{B}}, 1/c_{\boldsymbol{B}})$ and $r_{k_{\boldsymbol{Y}}} \sim \text{Gam}(\gamma_{\boldsymbol{Y}}/K_{\boldsymbol{Y}}, 1/c_{\boldsymbol{Y}})$. Such approximation approaches the original infinite model as both $K_{\boldsymbol{B}}$ and $K_{\boldsymbol{Y}}$ approach infinity.

Sampling of $(x_{nmk_{\boldsymbol{B}}})_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}}$: First, the total latent count corresponding to the non-zero entries can be derived as:

$$(x_{nm}|-) \sim b_{nm}\text{Pois}_+\left(\sum_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}} \lambda_{nmk_{\boldsymbol{B}}}\right). \quad (5)$$

After which, following Lemma 2.2 one can derive:

$$\left((x_{nmk_{\boldsymbol{B}}})_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}}|-\right) \sim \text{Mult}\left(x_{nm}, \left(\frac{\lambda_{nmk_{\boldsymbol{B}}}}{\sum_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}}\lambda_{nmk_{\boldsymbol{B}}}}\right)_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}}\right). \quad (6)$$

Sampling of $(y_{dwk})_k$: Again, following Lemma 2.2, we have:

$$\left((y_{dwk_{\boldsymbol{Y}}})_{k_{\boldsymbol{Y}}=1}^{K_{\boldsymbol{Y}}}, (y_{dnwk_{\boldsymbol{B}}})_{k_{\boldsymbol{B}}=1,n\in\mathcal{Z}_d}^{K_{\boldsymbol{B}}}|-\right) \sim \quad (7)$$
$$\text{Mult}\left(y_{dw}, \frac{\{\zeta_{dwk_{\boldsymbol{Y}}}\}_{k_{\boldsymbol{Y}}}, \{\zeta_{dnwk_{\boldsymbol{B}}}\}_{n\in\mathcal{Z}_d,k_{\boldsymbol{B}}}}{\sum_{k_{\boldsymbol{Y}}}\zeta_{dwk_{\boldsymbol{Y}}} + \sum_{n\in\mathcal{Z}_d}\sum_{k_{\boldsymbol{B}}}\zeta_{dnwk_{\boldsymbol{B}}}}\right).$$

Sampling of $\phi_{nk_{\boldsymbol{B}}}$, $\psi_{wk_{\boldsymbol{B}}}$, $\rho_{k_{\boldsymbol{B}}}$, $\theta_{dk_{\boldsymbol{Y}}}$, $\beta_{wk_{\boldsymbol{Y}}}$, $r_{k_{\boldsymbol{Y}}}$ **and** ϵ : Sampling of these parameters follow from Lemma 2.4 and are given in Table 1. The sampling of parameters $\phi_{nk_{\boldsymbol{B}}}$ and $\rho_{k_{\boldsymbol{B}}}$ exhibits how information from the count matrix $\mathbf{Y}$ influences the discovery of the latent network structure. The latent counts from $\mathbf{Y}$ impact the shape parameters for both the posterior gamma distribution of $\phi_{nk_{\boldsymbol{B}}}$ and $\rho_{k_{\boldsymbol{B}}}$, while $\mathbf{Z}$ influences the corresponding scale parameters.

Sampling of σ_n, ς_d, ϵ, ζ_w, η_w, $c_{\boldsymbol{B}}$ **and** $c_{\boldsymbol{Y}}$: Sampling of these parameters follow from Lemma 2.5 and are given in Table 1.

Sampling of $\gamma_{\boldsymbol{B}}$: Using Lemma 2.2, one can show that $x_{\cdot\cdot k_{\boldsymbol{B}}} \sim \text{Pois}(\rho_{k_{\boldsymbol{B}}})$. Integrating $\rho_{k_{\boldsymbol{B}}}$ and using Lemma 2.4, one can have $x_{\cdot\cdot k_{\boldsymbol{B}}} \sim \text{NB}(\gamma_{\boldsymbol{B}}, p_{\boldsymbol{B}})$, where $p_{\boldsymbol{B}} = 1/(c_{\boldsymbol{B}} + 1)$. Similarly, $y_{\cdot\cdot k_{\boldsymbol{B}}} \sim \text{Pois}(\rho_{k_{\boldsymbol{B}}})$ and after integrating $\rho_{k_{\boldsymbol{B}}}$ and using Lemma 2.4, we have $y_{\cdot\cdot k_{\boldsymbol{B}}} \sim \text{NB}(\gamma_{\boldsymbol{B}}, p_{\boldsymbol{B}})$. We now augment $l_{k_{\boldsymbol{B}}} \sim \text{CRT}(x_{\cdot\cdot k_{\boldsymbol{B}}} + y_{\cdot\cdot k_{\boldsymbol{B}}}, \gamma_{\boldsymbol{B}})$ and then following Lemma 2.6 sample:

$$(\gamma_{\boldsymbol{B}}|-) \sim \text{Gam}\left(e_{\boldsymbol{B}} + \sum_{k_{\boldsymbol{B}}} l_{k_{\boldsymbol{B}}}, (f_{\boldsymbol{B}} - q_{\boldsymbol{B}})^{-1}\right), \quad (8)$$

where $q_{\boldsymbol{B}} = \sum_{k_{\boldsymbol{B}}} \log\left(c_{\boldsymbol{B}}/(c_{\boldsymbol{B}} + \sum_n \phi_{nk_{\boldsymbol{B}}}\phi_{k_{\boldsymbol{B}}}^{-n})\right)/K_{\boldsymbol{B}}$.

Sampling of $\gamma_{\boldsymbol{Y}}$: Using Lemma 2.2, one can show that $y_{\cdot\cdot(K_{\boldsymbol{B}}+k_{\boldsymbol{Y}})} \sim$

$$(\theta_{dk_{\boldsymbol{Y}}}|-) \sim \text{Gam}\left(a_{\boldsymbol{Y}} + \sum_{w:(d,w)\notin\mathcal{M}_{\boldsymbol{Y}}} y_{dwk_{\boldsymbol{Y}}}, \left(\varsigma_d + r_{k_{\boldsymbol{Y}}} \sum_{w:(d,w)\notin\mathcal{M}_{\boldsymbol{Y}}} \beta_{wk_{\boldsymbol{Y}}}\right)^{-1}\right),$$

$$(\beta_{wk_{\boldsymbol{Y}}}|-) \sim \text{Gam}\left(\xi_{\boldsymbol{Y}} + \sum_{d:(d,w)\notin\mathcal{M}_{\boldsymbol{Y}}} y_{dwk_{\boldsymbol{Y}}}, \left(\eta_w + r_{k_{\boldsymbol{Y}}} \sum_{d,n:(d,w)\notin\mathcal{M}_{\boldsymbol{Y}}} Z_{dn}\phi_{nk_{\boldsymbol{Y}}}\right)^{-1}\right),$$

$$(r_{k_{\boldsymbol{Y}}}|-) \sim \text{Gam}\left(\frac{\gamma_{\boldsymbol{Y}}}{K_{\boldsymbol{Y}}} + \sum_{d,w:(d,w)\notin\mathcal{M}_{\boldsymbol{Y}}} y_{dwk_{\boldsymbol{Y}}}, \left(c_{\boldsymbol{Y}} + \sum_{d,w:(d,w)\notin\mathcal{M}_{\boldsymbol{Y}}} \theta_{dk_{\boldsymbol{Y}}}\beta_{wk_{\boldsymbol{Y}}}\right)^{-1}\right).$$

Table 3: Sampling of $\theta_{dk_{\boldsymbol{Y}}}$, $\beta_{wk_{\boldsymbol{Y}}}$, $r_{k_{\boldsymbol{Y}}}$ in J-GPPF with missing entries

$$b_{nm} = I_{\{x_{nm}\geq 1\}}, x_{nm} \sim \text{Pois}\left(\sum_{k_{\boldsymbol{B}}=1}^{\infty} \lambda_{nmk_{\boldsymbol{B}}}\right),$$
$$r_{k_{\boldsymbol{B}}} \sim \text{Gam}(\gamma_{\boldsymbol{B}}/K_{\boldsymbol{B}}, 1/c_{\boldsymbol{B}}), \boldsymbol{\phi}_{k_{\boldsymbol{B}}} \sim \textstyle\prod_{n=1}^{N} \text{Gam}(a_{\boldsymbol{B}}, 1/\sigma_n),$$
$$\sigma_n \sim \text{Gam}(\alpha_{\boldsymbol{B}}, 1/\varepsilon_{\boldsymbol{B}}),$$
$$\gamma_{\boldsymbol{B}} \sim \text{Gam}(e_{\boldsymbol{B}}, 1/f_{\boldsymbol{B}}), c_{\boldsymbol{B}} \sim \text{Gam}(g_{\boldsymbol{B}}, 1/h_{\boldsymbol{B}}),$$

Table 4: Generative Process of N-GPPF

$$y_{dw} \sim \text{Pois}\left(\sum_{k_{\boldsymbol{Y}}=1}^{\infty} r_{k_{\boldsymbol{Y}}}\theta_{dk_{\boldsymbol{Y}}}\beta_{wk_{\boldsymbol{Y}}}\right),$$
$$\boldsymbol{\theta}_{k_{\boldsymbol{Y}}} \sim \textstyle\prod_{d=1}^{D} \text{Gam}(a_{\boldsymbol{Y}}, 1/\varsigma_d), \boldsymbol{\beta}_{k_{\boldsymbol{Y}}} \sim \textstyle\prod_{w=1}^{V} \text{Gam}(\xi_{\boldsymbol{Y}}, 1/\zeta_w),$$
$$\varsigma_d \sim \text{Gam}(\alpha, 1/\varepsilon), r_{k_{\boldsymbol{Y}}} \sim \text{Gam}(\gamma_{\boldsymbol{Y}}/K_{\boldsymbol{Y}}, 1/c_{\boldsymbol{Y}}),$$
$$\gamma_{\boldsymbol{Y}_{\boldsymbol{Y}}} \sim \text{Gam}(e_{\boldsymbol{Y}}, 1/f_{\boldsymbol{Y}}), c_{\boldsymbol{Y}} \sim \text{Gam}(g_{\boldsymbol{Y}}, 1/h_{\boldsymbol{Y}}).$$

Table 5: Generative Process of C-GPPF

Pois($r_{k_{\boldsymbol{Y}}}$) and after integrating $r_{k_{\boldsymbol{Y}}}$ and using Lemma 2.4, we have $y_{\cdot\cdot(K_{\boldsymbol{B}}+k_{\boldsymbol{Y}})} \sim \text{NB}(\gamma_{\boldsymbol{Y}}, p_{\boldsymbol{Y}})$, where $p_{\boldsymbol{Y}} = 1/(c_{\boldsymbol{Y}}+1)$. We now augment $m_{k_{\boldsymbol{Y}}} \sim \text{CRT}(y_{\cdot\cdot(K_{\boldsymbol{B}}+k_{\boldsymbol{Y}})}, \gamma_{\boldsymbol{Y}})$ and then following Lemma 2.6 sample:

$$(\gamma_{\boldsymbol{Y}}|-) \sim \text{Gam}\left(e_{\boldsymbol{Y}} + \sum_{k_{\boldsymbol{Y}}} m_{k_{\boldsymbol{Y}}}, (f_{\boldsymbol{Y}} - q_{\boldsymbol{Y}})^{-1}\right), \quad (9)$$

where $q_{\boldsymbol{Y}} = \sum_{k_{\boldsymbol{Y}}} \log\left(c_{\boldsymbol{Y}}/(c_{\boldsymbol{Y}} + \theta_{\cdot k_{\boldsymbol{Y}}})\right)/K_{\boldsymbol{Y}}$.

3.2 Gibbs Sampling for J-GPPF with Missing Entries

Parameters whose update get affected in presence of missing entires are $\rho_{k_{\boldsymbol{B}}}$, $\phi_{nk_{\boldsymbol{B}}}$, $\psi_{wk_{\boldsymbol{B}}}$, $r_{k_{\boldsymbol{Y}}}$, $\theta_{dk_{\boldsymbol{Y}}}$, $\beta_{wk_{\boldsymbol{Y}}}$. Sampling of these parameters follow from Lemma 2.4 and are given in Table 2 and 3. Here $\mathcal{M}_{\boldsymbol{B}}$ and $\mathcal{M}_{\boldsymbol{Y}}$ denote the set of missing entries in $\boldsymbol{B}$ and $\boldsymbol{Y}$ respectively.

3.3 Special cases: Network Only GPPF (N-GPPF) and Corpus Only GPPF (C-GPPF)

A special case of J-GPPF appears when only the binary matrix $\boldsymbol{B}$ is modeled without the auxiliary matrix $\boldsymbol{Y}$. The update equations of variables corresponding to N-GPPF can be obtained with $\mathbf{Z} = \mathbf{0}$. Another special case of J-GPPF appears when only the count matrix $\boldsymbol{Y}$ is modeled without using the contribution from the network matrix $\boldsymbol{B}$. The generative model of N-GPPF and C-GPPF are given in Table 4 and 5 respectively.

3.4 Computation Complexity

The Gibbs sampling updates of J-GPPF can be calculated in $O(K_{\boldsymbol{B}}S_{\boldsymbol{B}} + (K_{\boldsymbol{B}}+K_{\boldsymbol{Y}})S_{\boldsymbol{Y}} + NK_{\boldsymbol{B}} + DK_{\boldsymbol{Y}} + V(K_{\boldsymbol{B}}+K_{\boldsymbol{Y}}))$ time, where $S_{\boldsymbol{B}}$ is the number of non-zero entries in $\boldsymbol{B}$ and $S_{\boldsymbol{Y}}$ is the number of non-zero entries in $\boldsymbol{Y}$. It is obvious that for large matrices the computation is primarily of the order of $K_{\boldsymbol{B}}S_{\boldsymbol{B}} + (K_{\boldsymbol{B}} + K_{\boldsymbol{Y}})S_{\boldsymbol{Y}}$. Such complexity is a huge saving when compared to other methods like MMSB [2], that only models $\boldsymbol{B}$ and incurs computation cost of $O(N^2K_{\boldsymbol{B}})$; and standard matrix factorization approaches [25] that work with the matrix $\boldsymbol{Y}$ and incur $O(DVK_{\boldsymbol{Y}})$ computation cost. Interestingly, the inference in [13] incurs cost $O(K_{\boldsymbol{Y}}^2 D + K_{\boldsymbol{Y}}V + K_{\boldsymbol{Y}}S_{\boldsymbol{Y}})$ with $K_{\boldsymbol{Y}}$ signifying the termination point of stick breaking construction in their model. C-GPPF incurs computation cost $O(DK_{\boldsymbol{Y}} + K_{\boldsymbol{Y}}S_{\boldsymbol{Y}} + VK_{\boldsymbol{Y}})$, an apparent improvement over that of [13].

4. RELATED WORK

The Infinite Relational Model (IRM [14]) allows for multiple types of relations between entities in a network and an infinite number of clusters, but restricts these entities to belong to only one cluster. The Mixed Membership Stochastic Blockmodel (MMSB [2]) assumes that each node in the network can exhibit a mixture of communities but the computational complexity of the underlying inference mechanism is $O(N^2)$. Such quadratic computation complexity is also a problem with many other existing latent variable network models, such as the latent feature relational model [21] and its max margin version [34], and the infinite latent attribute model [23]. Some of the existing approaches handle sparsity in real-world networks by using some auxiliary information [16; 20; 29]. Recommender system and text mining researchers, in contrast, tend to take an orthogonal approach. In recommender systems [9; 17], $\mathbf{Y}$ may represent a user-by-item rating matrix and the objective in this setting is to predict the missing entries in $\mathbf{Y}$, and the social network matrix $\mathbf{B}$ plays a secondary role in providing auxiliary information to facilitate this task [17]. Similarly, in the text mining community, many existing models [4; 19; 22; 27] use the network information or other forms of side information to improve the discovery of "topics" from the document-by-word matrix $\mathbf{Y}$. The matrix $\mathbf{B}$ can represent, for example, the interaction network of authors participating in writing the documents. The Relational Topic Model [10] discovers links between documents based on their topic distributions, obtained through unsupervised exploration. The Author-Topic framework (AT [24]) and the Author-Recipient-Topic model (ART [19]) jointly model documents along with the authors of the documents. Block-LDA [4], on the other hand, provides a generative model for the links between authors and recipients in addition to documents. J-GPPF differs from these existing approaches in mathematical formulation, including more effective modeling of both sparsity and the dependence between network interactions and side information.

A large number of discrete latent variable models for count matrix factorization can be united under Poisson factor analysis (PFA)

[33], which factorizes a count matrix $\mathbf{Y} \in \mathbb{Z}^{D\times V}$ under the Poisson likelihood as $\mathbf{Y} \sim \text{Pois}(\mathbf{\Phi\Theta})$, where $\mathbf{\Phi} \in \mathbb{R}_+^{D\times K}$ is the factor loading matrix or dictionary, $\mathbf{\Theta} \in \mathbb{R}_+^{K\times V}$ is the factor score matrix. A wide variety of algorithms, such as non-negative matrix factorization [8; 15], gamma-Poisson model [7; 26], LDA [6], and gamma-NB processes [32; 33], although constructed with different motivations and for distinct problems, can all be viewed as PFA with different prior distributions imposed on $\mathbf{\Phi}$ and $\mathbf{\Theta}$. J-GPPF models both $\mathbf{Y}$ and $\mathbf{B}$ using Poisson factorization. As discussed in [1], Poisson factorization has several practical advantages over other factorization methods that use Gaussian assumptions (*e.g.* in [17]). First, zero-valued observations could be efficiently processed during inference, so the model can readily accommodate large, sparse datasets. Second, Poisson factorization is a natural representation of count data. The collaborative topic Poisson factorization (CTPF) framework proposed in [12] solves a different problem where the objective is to recommend articles to users of similar interest. CTPF is a parametric model and variational approximation is adopted to solve the inference. Although both models make use of Poisson factorization to infer low-rank matrices in order to recommend items to users, J-GPPF is a fundamentally different model. As recently summarized in [5], there are many useful approaches for employing social side information to improve recommendation. For example, in both [3; 18], the authors define a dependence for item recommendations based on each node's neighbors or community. J-GPPF is a joint model of both the social network and the item ratings, and solves for the latent space factorization of each, and missing elements of each, simultaneously. In addition to this innovation, model inference scales with the number of observed elements in each matrix and the number of latent groups, not the full dimension of each matrix.

5. EXPERIMENTAL RESULTS

5.1 NIPS Authorship Network

This dataset contains a list of all papers and authors from NIPS 1988 to 2003. We took the 234 authors who had published with the most other people and looked at their co-authorship information. After standard pre-processing and removing words that appear less than 50 times in the over-all corpus corresponding to these users, the number of users in the graph who write at least one document, is 225 and the total number of unique words is 1354. The total number of documents is 1165.

5.2 GoodReads Data

Using the Goodreads API, we collect a base set of users with recent activity on the website. For each user in the base set, the user's friends as well as friends of friends on the site are collected (two hops in the graph). This process is repeated over a $24-$hour time period, with a new base set constructed each time (*i.e.* friends are not polled recursively). By running for a full day, multiple time zones are covered and the reviews are collected for all identified users, with a maximum of 200 reviews per user. Each review consists of a book ID and a rating from 0 to 5. Similar dataset has also been used in [9]. After standard pre-processing and removing words that appear less than 10 times in the over-all corpus, the number of users in the graph is 84 and the total number of unique words is 189.

5.3 Experimental Setup and Results

In all the experiments, we initialize ϵ to 2 and let the sampler decide what value works best for joint modeling. We use $K_{\mathbf{B}} = K_{\mathbf{Y}} = 50$ and initialize all the hyper-parameters to 1. In the first set of experiments, for each dataset, we hold out data from $\mathbf{B}$ only and ran 20 different experiments and display the mean AUC and one standard error. In this setup, we consider N-GPPF, the infinite relational model (IRM) of [14] and the Mixed Membership Stochastic Block Model (MMSB) [2] as the baseline algorithms. Fig. 1 and 2 demonstrate the performances of the models in predicting the held-out data. J-GPPF clearly has advantage over other network-only models when the network is sparse enough and the auxiliary information is sufficiently strong. However, all methods fail when the sparsity increases beyond a certain point. The performance of J-GPPF also drops below the performances of network-only models in highly sparse networks, as the sampler faces additional difficulty in extracting information from both $\mathbf{B}$ and $\mathbf{Y}$.

In the second set of experiments, we hold out data from $\mathbf{Y}$ only and run 20 different experiments and display the precision@top-20 for J-GPPF. This evaluation is structured along the lines of the work in [13]. We calculate the intersection of the top 20 predicted set of words (arranged in the decreasing order of counts) and the top 20 words in a document and divide the number by 20 to get the precision for each document. We then calculate mean average precision (MAP) by taking the average of the precision over all the documents. C-GPPF and the hierarchical Poisson matrix factorization (HPMF) [13] are considered as the baselines, both of which model only $\mathbf{Y}$. Fig. 3 and 4 show that $\mathbf{B}$ helps in boosting the predictive performance in J-GPPF over a wide range of fractions of the data that is held out from $\mathbf{Y}$.

6. CONCLUSION

We propose J-GPPF that jointly factorizes the network adjacency matrix and the associated side information that can be represented as a count matrix. The model has the advantage of representing true sparsity in $\mathbf{B}$, $\mathbf{Y}$ and in latent group membership. We derived an efficient MCMC inference method, and compared our approach to several popular baselines that either work on $\mathbf{B}$ or $\mathbf{Y}$.

Acknowledgement

This work is supported by the United States Office of Naval Reseach, grant No. N00014-14-1-0039.

References

[1] A. Acharya, J. Ghosh, and M. Zhou. Nonparametric Bayesian Factor Analysis for Dynamic Count Matrices. In *Proc. of AISTATS*, pages 1–9. 2015.

[2] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *JMLR*, 9:1981–2014, jun 2008.

[3] J. Aranda, I. Givoni, J. Handcock, and D. Tarlow. An online social network-based recommendation system. *Toronto, Ontario, Canada*, 2007.

[4] R. Balasubramanyan and W. W. Cohen. Block-LDA: Jointly modeling entity-annotated text and entity-entity links. In *Proc. of SDM*, pages 450–461, 2011.

[5] D. Bernardes, M. Diaby, R. Fournier, F. FogelmanSoulié, and E. Viennet. A social formalism and survey for recommender systems. *SIGKDD Explor. Newsl.*, 16(2):20–37, may 2015.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *JMLR*, 3:993–1022, 2003.

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

[8] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Intell. Neuroscience*, 2009:4:1–4:17, Jan. 2009.

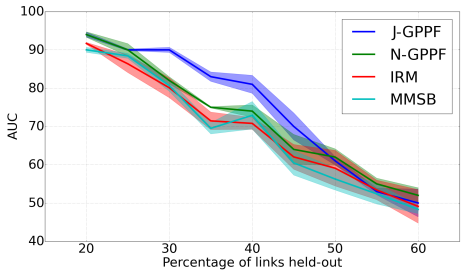

Figure 1: AUC NIPS

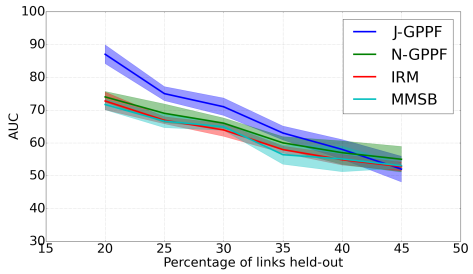

Figure 2: AUC GoodReads

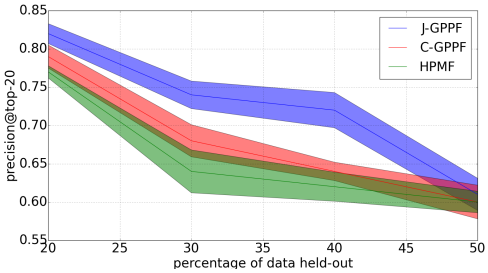

Figure 3: MAP NIPS

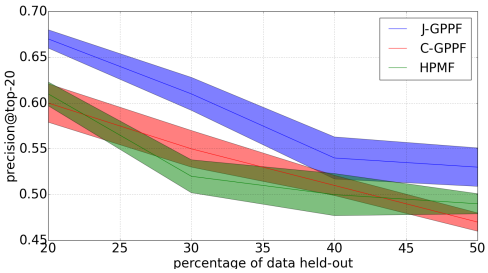

Figure 4: MAP GoodReads

[9] A. Chaney, P. Gopalan, and D. Blei. Poisson trust factorization for incorporating social networks into personalized item recommendation. In *NIPS Workshop: What Difference Does Personalization Make?*, 2013.

[10] J. Chang and D. Blei. Relational topic models for document networks. In *Proc. of AISTATS*, 2009.

[11] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1973.

[12] P. Gopalan, L. Charlin, and D. Blei. Content-based recommendations with poisson factorization. In *Proc. of NIPS*, pages 3176–3184. 2014.

[13] P. Gopalan, F. Ruiz, R. Ranganath, and D. Blei. Bayesian nonparametric poisson factorization for recommendation systems. In *Proc. of AISTATS*, 2014.

[14] C. Kemp, J. Tenenbaum, T. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proc. of AAAI*, pages 381–388, 2006.

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

[16] J. Leskovec and J. Julian. Learning to discover social circles in ego networks. In *Proc. of NIPS*, pages 539–547. 2012.

[17] H. Ma, H. Yang, M. R. Lyu, and I. King. Sorec: Social recommendation using probabilistic matrix factorization. In *Proc. of CIKM*, pages 931–940, 2008.

[18] H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King. Recommender systems with social regularization. In *Proc. of WSDM*, pages 287–296, 2011.

[19] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Int. Res.*, 30(1):249–272, Oct. 2007.

[20] A. Menon and C. Elkan. Link prediction via matrix factorization. In *Machine Learning and Knowledge Discovery in Databases*, volume 6912 of *Lecture Notes in Computer Science*, pages 437–452. Springer Berlin / Heidelberg, 2011.

[21] K. T. Miller, T. L. Griffiths, and M. I. Jordan. Nonparametric latent feature models for link prediction. In *Proc. of NIPS*, pages 1276–1284, 2009.

[22] R. Nallapati, A. Ahmed, E. Xing, and W. Cohen. Joint latent topic models for text and citations. In *Proc. of KDD*, pages 542–550, 2008.

[23] K. Palla, Z. Ghahramani, and D. A. Knowles. An infinite latent attribute model for network data. In *Proc. of ICML*, pages 1607–1614, 2012.

[24] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proc. of UAI*, pages 487–494, 2004.

[25] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Proc. of NIPS*, 2007.

[26] M. K. Titsias. The infinite gamma-Poisson feature model. In *Proc. of NIPS*, 2008.

[27] Z. Wen and C. Lin. Towards finding valuable topics. In *Proc. of SDM*, pages 720–731, 2010.

[28] R. L. Wolpert, M. A. Clyde, and C. Tu. Stochastic expansions using continuous dictionaries: Lévy Adaptive Regression Kernels. *Annals of Statistics*, 2011.

[29] T. Yoshida. Toward finding hidden communities based on user profile. *J. Intell. Inf. Syst.*, 40(2):189–209, Apr. 2013.

[30] M. Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *Proc. of AISTATS*, pages 1135–1143. 2015.

[31] M. Zhou and L. Carin. Augment-and-conquer negative binomial processes. In *Proc. of NIPS*, 2012.

[32] M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2015.

[33] M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *Proc. of AISTATS*, pages 1462–1471, 2012.

[34] J. Zhu. Max-margin nonparametric latent feature models for link prediction. In *Proc. of ICML*, 2012.