Gamma Process Poisson Factorization for Joint Modeling of Network and Documents

Ayan Acharya[1(✉)], Dean Teffer[2], Jette Henderson[2], Marcus Tyler[2], Mingyuan Zhou[3], and Joydeep Ghosh[1]

[1] Department of ECE, University of Texas at Austin, Austin, USA
{aacharya,jghosh}@utexas.edu
[2] Applied Research Laboratories, University of Texas at Austin, Austin, USA
{dean.teffer,jhende,mtyler}@arlut.utexas.edu
[3] Department of IROM, University of Texas at Austin, Austin, USA
mzhou@utexas.edu

Abstract. Developing models to discover, analyze, and predict clusters within networked entities is an area of active and diverse research. However, many of the existing approaches do not take into consideration pertinent auxiliary information. This paper introduces Joint Gamma Process Poisson Factorization (J-GPPF) to jointly model network and side-information. J-GPPF naturally fits sparse networks, accommodates separately-clustered side information in a principled way, and effectively addresses the computational challenges of analyzing large networks. Evaluated with hold-out link prediction performance on sparse networks (both synthetic and real-world) with side information, J-GPPF is shown to clearly outperform algorithms that only model the network adjacency matrix.

Keywords: Network modeling · Poisson factorization · Gamma process

1 Introduction

Social networks and other relational datasets often involve a large number of nodes N with sparse connections between them. If the relationship is symmetric, it can be represented compactly using a binary symmetric adjacency matrix $\mathbf{B} \in \{0,1\}^{N\times N}$, where $b_{ij} = b_{ji} = 1$ if and only if nodes i and j are linked. 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. Integer-valued side information are commonly observed and can be naturally represented by a count 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 that adds nuance and support to the network data. Incorporating such side information can result in better community identification

© Springer International Publishing Switzerland 2015
A. Appice et al. (Eds.): ECML PKDD 2015, Part I, LNAI 9284, pp. 283–299, 2015.
DOI: 10.1007/978-3-319-23528-8_18

and superior link prediction performance as compared to modeling only the network adjacency matrix $\mathbf{B}$, especially for sparse networks.

Many of the popular network models [2,13,18,25,28] are demonstrated to work well for small size networks. However, small networks are often dense, while larger real-world networks tend to be much sparser and hence challenge existing modeling approaches. Incorporating auxiliary information associated with the nodes has the potential to address such challenges, as it may help better identify latent communities and predict missing links. A model that takes advantage of such side information has the potential to outperform network-only models. However, the side information may not necessarily suggest the same community structure as the existing links. Thus a network latent factor model that allows separate factors for side information and network interactions, but at the same time is equipped with a mechanism to capture dependencies between the two types of factors, is desirable.

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: 1) we present a fast and effective model that uses side information to help discover latent network structures, 2) we perform nonparametric Bayesian modeling for discovering latent structures in both $\mathbf{B}$ and $\mathbf{Y}$, and 3) our model scales with the number of non-zero entries in the network $S_{\mathbf{B}}$ as $\mathrm{O}\left(S_{\mathbf{B}} K_{\mathbf{B}}\right)$, where $K_{\mathbf{B}}$ is the number of network groups inferred from the data.

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

2 Background and Related Work

This section presents the related literature and the background materials that are useful for understanding the framework described in Section 3.

2.1 Negative Binomial Distribution

The negative binomial (NB) distribution $m \sim \mathrm{NB}(r, p)$, with probability mass function (PMF) $\Pr(M = m) = \frac{\Gamma(m+r)}{m!\Gamma(r)} p^m (1-p)^r$ for $m \in \mathbb{Z}$, can be augmented into a gamma-Poisson construction as $m \sim \mathrm{Pois}(\lambda),\ \lambda \sim \mathrm{Gamma}(r, p/(1-p))$, where the gamma distribution is parameterized by its shape r and scale $p/(1-p)$. It can also be augmented under a compound Poisson representation as $m = \sum_{t=1}^{l} u_t, u_t \overset{iid}{\sim} \mathrm{Log}(p), l \sim \mathrm{Pois}(-r\ln(1-p))$, where $u \sim \mathrm{Log}(p)$ is the logarithmic distribution [17]. Consequently, we have the following Lemma.

Lemma 1 ([41])**.** *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. *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 3. *If* $x_i \sim \mathrm{Pois}(m_i\lambda)$, $\lambda \sim \mathrm{Gamma}(r, 1/c)$, *then* $x = \sum_i x_i \sim \mathrm{NB}(r,p)$, *where* $p = (\sum_i m_i)/(c + \sum_i m_i)$.

Lemma 4. *If* $x_i \sim \mathrm{Pois}(m_i\lambda)$, $\lambda \sim \mathrm{Gamma}(r, 1/c)$, *then*

$$(\lambda|\{x_i\}, r, c) \sim \mathrm{Gamma}\left(r + \sum_i x_i, 1/(c + \sum_i m_i)\right). \tag{3}$$

Lemma 5. *If* $r_i \sim \mathrm{Gamma}(a_i, 1/b)\ \forall i$, $b \sim \mathrm{Gamma}(c, 1/d)$, *then we have:*

$$(b|\{r_i, a_i\}, c, d) \sim \mathrm{Gamma}\left(\sum_i a_i + c, 1/(\sum_i r_i + d)\right). \tag{4}$$

The proofs of Lemmas 3, 4 and 5 follow from the definitions of Gamma, Poisson and Negative Binomial distributions.

Lemma 6. *If* $x_i \sim \mathrm{Pois}(m_i r_2)$, $r_2 \sim \mathrm{Gamma}(r_1, 1/d)$, $r_1 \sim \mathrm{Gamma}(a, 1/b)$, *then* $(r_1|-) \sim \mathrm{Gamma}(a+\ell, 1/(b - log(1-p)))$ *where* $(\ell|x, r_1) \sim \mathrm{CRT}(\sum_i x_i, r_1)$ *and* $p = \sum_i m_i/(d + \sum_i m_i)$. *The proof and illustration can be found in Section 3.3 of [1].*

2.2 Gamma Process

The Gamma Process [12,36] $G \sim \Gamma\mathrm{P}(c, H)$ is a completely random measure defined on the product space $\mathbb{R}_+ \times \Omega$, with concentration parameter c and a finite and continuous base measure H over a complete separable metric space Ω, such that $G(A_i) \sim \mathrm{Gamma}(H(A_i), 1/c)$ are independent gamma random variables for disjoint partition $\{A_i\}_i$ of Ω. The Lévy measure of the Gamma Process can be expressed as $\nu(drd\omega) = r^{-1}e^{-cr}drH(d\omega)$. Since the Poisson intensity $\nu^+ = \nu(\mathbb{R}_+ \times \Omega) = \infty$ and the value of $\int_{\mathbb{R}_+\times\Omega} r\nu(drd\omega)$ is finite, a draw from the Gamma Process consists of countably infinite atoms, which can be expressed as follows:

$$G = \sum_{k=1}^{\infty} r_k \delta_{\omega_k},\ (r_k, \omega_k) \overset{iid}{\sim} \pi(drd\omega),\ \pi(drd\omega)\nu^+ \equiv \nu(drd\omega). \tag{5}$$

A gamma process based model has an inherent shrinkage mechanism, as in the prior the number of atoms with weights greater than $\tau \in \mathbb{R}_+$ follows a Poisson distribution with parameter $H(\Omega)\int_\tau^\infty r^{-1}\exp(-cr)dr$, the value of which decreases as τ increases.

2.3 Network Modeling, Topic Modeling and Count Matrix Factorization

The Infinite Relational Model (IRM [18]) 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. Though the MMSB has been applied successfully to discover complex network structure in a variety of applications, the computational complexity of the underlying inference mechanism is in the order of N^2, which limits its use to small networks. Computation complexity is also a problem with many other existing latent variable network models, such as the latent feature relational model [25] and its max margin version [44], and the infinite latent attribute model [28]. The Assortative Mixed-Membership Stochastic Blockmodel (a-MMSB [13]) bypasses the quadratic complexity of the MMSB by making certain assumptions about the network structure that might not be true in general. The hierarchical Dirichlet process relational model [19] allows mixed membership with an unbounded number of latent communities; however, it is built on the a-MMSB whose assumptions could be restrictive.

Some of the existing approaches handle sparsity in real-world networks by using some auxiliary information [21,24,39]. For example, in a protein-protein interaction network, the features describing the biological properties of each protein can be used [24]. In an extremely sparse social network, information about each user's profile can be used to better recommend friends [21]. Recommender system and text mining researchers, in contrast, tend to take an orthogonal approach. In recommender systems [10,22], $\mathbf{Y}$ 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 [22]. Similarly, in the text mining community, many existing models [3,23,26,35] 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 [11] discovers links between documents based on their topic distributions, obtained through unsupervised exploration. The Author-Topic framework (AT [30]) and the Author-Recipient-Topic model (ART [23]) jointly model documents along with the authors of the documents. Block-LDA [3], on the other hand, provides a generative model for the links between authors and recipients in addition to documents. The Group-Topic model [34] addresses the task of modeling events pertaining to pairs of entities with textual attributes that annotate the event. 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) [43], which factorizes a count matrix $\mathbf{Y} \in \mathbb{Z}^{D \times V}$ under the Poisson likelihood as $\mathbf{Y} \sim \text{Pois}(\boldsymbol{\Phi}\boldsymbol{\Theta})$, where $\boldsymbol{\Phi} \in \mathbb{R}_+^{D \times K}$ is the factor loading matrix or dictionary, $\boldsymbol{\Theta} \in \mathbb{R}_+^{K \times V}$ is the factor score matrix. A wide variety of algorithms, although constructed with different motivations and for distinct problems, can all be viewed as PFA with different prior distributions imposed on $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$. For example, non-negative matrix factorization [9,20], with the objective to minimize the Kullback-Leibler divergence between $\mathbf{N}$ and its factorization $\boldsymbol{\Phi}\boldsymbol{\Theta}$, is essentially PFA solved with maximum likelihood estimation. LDA [6] is equivalent to PFA, in terms of both block Gibbs sampling and variational inference, if Dirichlet distribution priors are imposed on both $\boldsymbol{\phi}_k \in \mathbb{R}_+^D$, the columns of $\boldsymbol{\Phi}$, and $\boldsymbol{\theta}_k \in \mathbb{R}_+^V$, the columns of $\boldsymbol{\Theta}$. The gamma-Poisson model [8,32] is PFA with gamma priors on $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$. A family of negative binomial (NB) processes, such as the beta-NB [7,43] and gamma-NB processes [41,42], impose different gamma priors on $\{\theta_{vk}\}$, the marginalization of which leads to differently parameterized NB distributions to explain the latent counts. Both the beta-NB and gamma-NB process PFAs are nonparametric Bayesian models that allow K to grow without limits [16].

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 [22]). 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. Additionally, the model allows mixed membership across an unbounded number of latent communities using the gamma Process as a prior. The authors in [4] also use Poisson factorization to model a binary interaction matrix. However, this is a parametric model and a KL-divergence based objective is optimized w.r.t. the latent factors without using any prior information. To model the binary observations of the network matrix $\mathbf{B}$, J-GPPF additionally uses a novel Poisson-Bernoulli (PoBe) link, discussed in detail in Section 3, that transforms the count values from the Poisson factorization to binary values. Similar transformation has also been used in the BigCLAM model [37] which builds on the works of [4]. This model was later extended to include non-network information in the form of binary attributes [38]. Neither BigCLAM nor its extension allows non-parametric modeling or imposing prior structure on the latent factors, thereby limiting the flexibility of the models and making the obtained solutions more sensitive to initialization. The collaborative topic Poisson factorization (CTPF) framework proposed in [15] 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.

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. However, in the following model description we consider the more general document-author framework. 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., $G_{\boldsymbol{B}}$). Also, if $\boldsymbol{Y}$ represents a matrix of dimension $D \times V$, then $y_{.w}$ represents the sum over all the rows for the entries in the wth column, and $y_{d.}$ represents the sum over all the columns for the entries in the dth row.

Before providing an explicit description of the model, we introduce two separate Gamma Processes. The first one models the latent factors in the network and also contributes to generate the count matrix. A draw from this Gamma Process $G_{\boldsymbol{B}} \sim \Gamma\text{P}(c_{\boldsymbol{B}}, H_{\boldsymbol{B}})$ is expressed as: $G_{\boldsymbol{B}} = \sum_{k_{\boldsymbol{B}}=1}^{\infty} \rho_{k_{\boldsymbol{B}}} \delta_{\boldsymbol{\phi}_{k_{\boldsymbol{B}}}}$, where $\boldsymbol{\phi}_{k_{\boldsymbol{B}}} \in \Omega_{\boldsymbol{B}}$ is an atom drawn from an N-dimensional base distribution as $\boldsymbol{\phi}_{k_{\boldsymbol{B}}} \sim \prod_{n=1}^{N} \text{Gamma}(a_{\boldsymbol{B}}, 1/\sigma_n)$, $\rho_{k_{\boldsymbol{B}}} = G_{\boldsymbol{B}}(\boldsymbol{\phi}_{k_{\boldsymbol{B}}})$ is the associated weight, and $H_{\boldsymbol{B}}$ is the corresponding base measure. 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 $\boldsymbol{\beta}_{k_{\boldsymbol{Y}}} \sim \text{Dir}(\boldsymbol{\xi}_{\boldsymbol{Y}})$, $r_{k_{\boldsymbol{Y}}} = G_{\boldsymbol{Y}}(\boldsymbol{\beta}_{k_{\boldsymbol{Y}}})$ is the associated weight, and $H_{\boldsymbol{Y}}$ is the corresponding base measure. 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 $H_{\boldsymbol{Y}}$. In the following paragraphs, we explain how these Gamma processes, with the atoms and their associated weights, are used for modeling both $\boldsymbol{B}$ and $\boldsymbol{Y}$.

The $(n, m)^{\text{th}}$ entry in the matrix $\boldsymbol{B}$ is assumed to be derived from a latent count as:

$$b_{nm} = \mathbb{I}_{\{x_{nm} \geq 1\}},\ x_{nm} \sim \text{Pois}\left(\lambda_{nm}\right), \lambda_{nm} = \sum_{k_{\boldsymbol{B}}} \lambda_{nmk_{\boldsymbol{B}}}, \tag{6}$$

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 [1,40]. 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, 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. Thus, this approach transforms the problem of modeling binary network interaction into a count modeling problem, 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 $\boldsymbol{B}$. This benefit is especially pertinent in many real-word datasets where $S_{\boldsymbol{B}}$ is significantly smaller than N^2.

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

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

$$\zeta_{\boldsymbol{Y}dwk_{\boldsymbol{Y}}} = r_{k_{\boldsymbol{Y}}}\theta_{dk_{\boldsymbol{Y}}}\beta_{wk_{\boldsymbol{Y}}}, \zeta_{\boldsymbol{B}dwk_{\boldsymbol{B}}} = \epsilon\rho_{k_{\boldsymbol{B}}}\left(\sum_{n} Z_{nd}\phi_{nk_{\boldsymbol{B}}}\right)\psi_{wk_{\boldsymbol{B}}},$$

where $Z_{nd} \in \{0, 1\}$ and $Z_{nd} = 1$ if and only if author n is one of the authors of paper d. 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 dicticonary) 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 dth 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. For example, an author from machine learning and statistics would use words like "probability", "classifiers", "patterns", "prediction" more often than an author with an economics background. Frameworks such as author-topic model [23,30] were motivated by a related concept. 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. We present a thorough discussion of the effect of ϵ in Section 4.1. To complete the generative process, we put Gamma priors over σ_n, ς_d, $c_{\boldsymbol{B}}$, $c_{\boldsymbol{Y}}$ and ϵ as:

$$c_{\boldsymbol{B}} \sim \text{Gamma}(g_{\boldsymbol{B}}, 1/h_{\boldsymbol{B}}), c_{\boldsymbol{Y}} \sim \text{Gamma}(g_{\boldsymbol{Y}}, 1/h_{\boldsymbol{Y}}), \epsilon \sim \text{Gamma}(g_0, 1/f_0). \quad (7)$$

$$\sigma_n \sim \text{Gamma}(\alpha_{\boldsymbol{B}}, 1/\varepsilon_{\boldsymbol{B}}), \varsigma_d \sim \text{Gamma}(\alpha_{\boldsymbol{Y}}, 1/\varepsilon_{\boldsymbol{Y}}). \quad (8)$$

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. Instead of marginalizing out the underlying stochastic process [5,27] or using slice sampling [33] for non-parametric modeling, for simplicity, 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 $\rho_{k_{\boldsymbol{B}}} \sim \text{Gamma}(\gamma_{\boldsymbol{B}}/K_{\boldsymbol{B}}, 1/c_{\boldsymbol{B}})$ and $r_{k_{\boldsymbol{Y}}} \sim \text{Gamma}(\gamma_{\boldsymbol{Y}}/K_{\boldsymbol{Y}}, 1/c_{\boldsymbol{Y}})$. Such approximation approaches the original infinite model as both $K_{\boldsymbol{B}}$ and $K_{\boldsymbol{Y}}$ approach infinity. With such finite approximation, the generative process of J-GPPF is summarized in Table 1. For notational convenience, we represent the set of documents the nth author contributes to as $\mathcal{Z}_n$ and the set of authors contributing to the dth document as $\mathcal{Z}_d$.

Sampling of $(x_{nmk_{\boldsymbol{B}}})_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}}$: We first sample the network links according to the following:

$$(x_{nm}|-) \sim b_{nm}\text{Pois}_+\left(\sum_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}} \lambda_{nmk_{\boldsymbol{B}}}\right). \quad (9)$$

Sampling from a truncated Poisson distribution is described in detail in [40]. Since, one can augment $x_{nm} \sim \text{Pois}\left(\sum_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}} \lambda_{nmk_{\boldsymbol{B}}}\right)$ as $x_{nm} = \sum_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}} x_{nmk_{\boldsymbol{B}}}$, where $x_{nmk_{\boldsymbol{B}}} \sim \text{Pois}\left(\lambda_{nmk_{\boldsymbol{B}}}\right)$, equivalently, one obtains the following:

$$\left((x_{nmk_{\boldsymbol{B}}})_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}}|-\right) \sim \text{Mult}\left(x_{nm}, \left(\lambda_{nmk_{\boldsymbol{B}}}/\sum_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}} \lambda_{nmk_{\boldsymbol{B}}}\right)_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}}\right). \quad (10)$$

Table 1. Generative Process of J-GPPF

$$b_{nm} = I_{\{x_{nm}\geq 1\}}, x_{nm} \sim \text{Pois}\left(\sum_{k_{\boldsymbol{B}}} \rho_{k_{\boldsymbol{B}}} \phi_{nk_{\boldsymbol{B}}} \phi_{mk_{\boldsymbol{B}}}\right),$$
$$y_{dw} \sim \text{Pois}\left(\sum_{k_{\boldsymbol{Y}}} r_{k_{\boldsymbol{Y}}} \theta_{dk_{\boldsymbol{Y}}} \beta_{wk_{\boldsymbol{Y}}} + \epsilon \sum_{k_{\boldsymbol{B}}} \rho_{k_{\boldsymbol{B}}} \left(\sum_{n} Z_{nd} \phi_{nk_{\boldsymbol{B}}}\right) \psi_{wk_{\boldsymbol{B}}}\right),$$
$$\boldsymbol{\phi}_{k_{\boldsymbol{B}}} \sim \prod_{n=1}^{N} \text{Gamma}(a_{\boldsymbol{B}}, 1/\sigma_n),\ \boldsymbol{\psi}_{k_{\boldsymbol{B}}} \sim \text{Dir}(\boldsymbol{\xi}_{\boldsymbol{B}}),$$
$$\boldsymbol{\theta}_{k_{\boldsymbol{Y}}} \sim \prod_{d=1}^{D} \text{Gamma}(a_{\boldsymbol{Y}}, 1/\varsigma_d),\ \boldsymbol{\beta}_{k_{\boldsymbol{Y}}} \sim \text{Dir}(\boldsymbol{\xi}_{\boldsymbol{Y}}),\ \epsilon \sim \text{Gamma}(f_0, 1/g_0),$$
$$\sigma_n \sim \text{Gamma}(\alpha_{\boldsymbol{B}}, 1/\varepsilon_{\boldsymbol{B}}),\ \rho_{k_{\boldsymbol{B}}} \sim \text{Gamma}(\gamma_{\boldsymbol{B}}/K_{\boldsymbol{B}}, 1/c_{\boldsymbol{B}}),$$
$$\gamma_{\boldsymbol{B}} \sim \text{Gamma}(e_{\boldsymbol{B}}, 1/f_{\boldsymbol{B}}),\ c_{\boldsymbol{B}} \sim \text{Gamma}(g_{\boldsymbol{B}}, 1/h_{\boldsymbol{B}}),$$
$$\varsigma_d \sim \text{Gamma}(\alpha_{\boldsymbol{Y}}, 1/\varepsilon_{\boldsymbol{Y}}),\ r_{k_{\boldsymbol{Y}}} \sim \text{Gamma}(\gamma_{\boldsymbol{Y}}/K_{\boldsymbol{Y}}, 1/c_{\boldsymbol{Y}}),$$
$$\gamma_{\boldsymbol{Y}} \sim \text{Gamma}(e_{\boldsymbol{Y}}, 1/f_{\boldsymbol{Y}}),\ c_{\boldsymbol{Y}} \sim \text{Gamma}(g_{\boldsymbol{Y}}, 1/h_{\boldsymbol{Y}}).$$

Sampling of $(y_{dwk})_k$: Since, one can augment $y_{dw} \sim \text{Pois}(\zeta_{dw})$ as $y_{dw} = \sum_{k_{\boldsymbol{Y}}=1}^{K_{\boldsymbol{Y}}} y_{dwk_{\boldsymbol{Y}}} + \sum_{n\in\mathcal{Z}_d} \sum_{k_{\boldsymbol{B}}=1}^{K_{\boldsymbol{B}}} y_{dnwk_{\boldsymbol{B}}}$, $y_{dwk_{\boldsymbol{Y}}} \sim \text{Pois}(\zeta_{dwk_{\boldsymbol{Y}}})$, $y_{dnwk_{\boldsymbol{B}}} \sim \text{Pois}(\zeta_{dnwk_{\boldsymbol{B}}})$, again following Lemma 2, we have:

$$\left((y_{dwk_{\boldsymbol{Y}}})_{k_{\boldsymbol{Y}}=1}^{K_{\boldsymbol{Y}}}, (y_{dnwk_{\boldsymbol{B}}})_{k_{\boldsymbol{B}}=1, n\in\mathcal{Z}_d}^{K_{\boldsymbol{B}}} | -\right) \sim \text{Mult}\left(y_{dw}, \frac{\{\zeta_{dwk_{\boldsymbol{Y}}}\}_{k_{\boldsymbol{Y}}}, \{\zeta_{dnwk_{\boldsymbol{B}}}\}_{n\in\mathcal{Z}_d, k_{\boldsymbol{B}}}}{\sum_{k_{\boldsymbol{Y}}} \zeta_{dwk_{\boldsymbol{Y}}} + \sum_{n\in\mathcal{Z}_d} \sum_{k_{\boldsymbol{B}}} \zeta_{dnwk_{\boldsymbol{B}}}}\right). \quad (11)$$

Sampling of $\phi_{nk_{\boldsymbol{B}}}$**,** $\rho_{k_{\boldsymbol{B}}}$**,** $\theta_{dk_{\boldsymbol{Y}}}$**,** $r_{k_{\boldsymbol{Y}}}$ **and** ϵ : Sampling of these parameters follow from Lemma 4 and are given as follows:

$$(\phi_{nk_{\boldsymbol{B}}}|-) \sim \text{Gamma}\left(a_{\boldsymbol{B}} + x_{n\cdot k_{\boldsymbol{B}}} + y_{.n.k_{\boldsymbol{B}}}, \frac{1}{\sigma_n + \rho_{k_{\boldsymbol{B}}}(\phi_{k_{\boldsymbol{B}}}^{-n} + \epsilon|\mathcal{Z}_n|)}\right), \quad (12)$$

$$(\rho_{k_{\boldsymbol{B}}}|-) \sim \text{Gamma}\left(\frac{\gamma_{\boldsymbol{B}}}{K_{\boldsymbol{B}}} + x_{\cdot\cdot k_{\boldsymbol{B}}} + y_{\cdot\cdot\cdot k_{\boldsymbol{B}}}, \frac{1}{c_{\boldsymbol{B}} + \sum_n \phi_{nk_{\boldsymbol{B}}} \phi_{k_{\boldsymbol{B}}}^{-n} + \epsilon \sum_n |\mathcal{Z}_n| \phi_{nk_{\boldsymbol{B}}}}\right), \quad (13)$$

$$(\theta_{dk_{\boldsymbol{Y}}}|-) \sim \text{Gamma}\left(a_{\boldsymbol{Y}} + y_{d\cdot k_{\boldsymbol{Y}}}, \frac{1}{\varsigma_d + r_{k_{\boldsymbol{Y}}}}\right), (r_{k_{\boldsymbol{Y}}}|-) \sim \text{Gamma}\left(\frac{\gamma_{\boldsymbol{Y}}}{K_{\boldsymbol{Y}}} + y_{\cdot\cdot k_{\boldsymbol{Y}}}, \frac{1}{c_{\boldsymbol{Y}} + \theta_{\cdot k_{\boldsymbol{Y}}}}\right), \quad (14)$$

$$(\epsilon|-) \sim \text{Gamma}\left(f_0 + \sum_{k=1}^{K_{\boldsymbol{B}}} y_{\cdot\cdot\cdot k}, \frac{1}{g_0 + q_0}\right), q_0 = \sum_{k=1}^{K_{\boldsymbol{B}}} \rho_{k_{\boldsymbol{B}}} \sum_{n=1}^{N} |\mathcal{Z}_n| \phi_{nk_{\boldsymbol{B}}}. \quad (15)$$

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 $\boldsymbol{\psi}_{k_{\boldsymbol{B}}}$: Since $y_{dnwk_{\boldsymbol{B}}} \sim \text{Pois}(\epsilon \rho_{k_{\boldsymbol{B}}} \phi_{nk_{\boldsymbol{B}}} \psi_{wk_{\boldsymbol{B}}})$, using Lemma 2 we have: $(y_{\cdot\cdot wk_{\boldsymbol{B}}})_{w=1}^{V} \sim \text{Mult}(y_{\cdot\cdot\cdot k_{\boldsymbol{B}}}, (\psi_{wk_{\boldsymbol{B}}})_{w=1}^{V})$. Since the Dirichlet distribution is conjugate to the multinomial, the posterior of $\boldsymbol{\psi}_{k_{\boldsymbol{B}}}$ also becomes a Dirichlet distribution and can be sampled as:

$$(\boldsymbol{\psi}_{k_{\boldsymbol{B}}}|-) \sim \text{Dir}\left(\boldsymbol{\xi}_{\boldsymbol{B}1} + y_{\cdot\cdot 1k_{\boldsymbol{B}}}, \cdots, \boldsymbol{\xi}_{\boldsymbol{B}V} + y_{\cdot\cdot Vk_{\boldsymbol{B}}}\right). \quad (16)$$

Sampling of $\boldsymbol{\beta}_{k_{\boldsymbol{Y}}}$: Since $y_{dwk_{\boldsymbol{Y}}} \sim \text{Pois}(r_{k_{\boldsymbol{Y}}}\theta_{dk_{\boldsymbol{Y}}}\beta_{wk_{\boldsymbol{Y}}})$, again using Lemma 2, we have:

$$(y_{.wk_{\boldsymbol{Y}}})_{w=1}^{V} \sim \text{Mult}\left(y_{..k_{\boldsymbol{Y}}}, (\beta_{wk_{\boldsymbol{Y}}})_{w=1}^{V}\right).$$

Using conjugacy, the posterior of $\boldsymbol{\beta}_{k_{\boldsymbol{Y}}}$ can be sampled as:

$$(\boldsymbol{\beta}_{k_{\boldsymbol{Y}}}|-) \sim \text{Dir}\left(\boldsymbol{\xi}_{\boldsymbol{Y}1} + y_{.1k_{\boldsymbol{Y}}}, \cdots, \boldsymbol{\xi}_{\boldsymbol{Y}V} + y_{.V+k_{\boldsymbol{Y}}}\right). \tag{17}$$

Sampling of σ_n, ς_d, $c_{\boldsymbol{B}}$ and $c_{\boldsymbol{Y}}$: Sampling of these parameters follow from Lemma 5 and are given as:

$$(\sigma_n|-) \sim \text{Gamma}\left(\alpha_{\boldsymbol{B}} + K_{\boldsymbol{B}}a_{\boldsymbol{B}}, \frac{1}{\varepsilon_{\boldsymbol{B}}+\phi_{n.}}\right), (\varsigma_d|-) \sim \text{Gamma}\left(\alpha_{\boldsymbol{Y}} + K_{\boldsymbol{Y}}a_{\boldsymbol{Y}}, \frac{1}{\varepsilon_{\boldsymbol{Y}}+\theta_{d.}}\right), \tag{18}$$

$$(c_{\boldsymbol{B}}|-) \sim \text{Gamma}\left(g_{\boldsymbol{B}} + \gamma_{\boldsymbol{B}}, \frac{1}{h_{\boldsymbol{B}}+\sum_{k_{\boldsymbol{B}}}\rho_{k_{\boldsymbol{B}}}}\right), (c_{\boldsymbol{Y}}|-) \sim \text{Gamma}\left(g_{\boldsymbol{Y}} + \gamma_{\boldsymbol{Y}}, \frac{1}{h_{\boldsymbol{Y}}+\sum_{k_{\boldsymbol{Y}}}r_{k_{\boldsymbol{Y}}}}\right). \tag{19}$$

Sampling of $\gamma_{\boldsymbol{B}}$: Using Lemma 2, one can show that $x_{..k_{\boldsymbol{B}}} \sim \text{Pois}(\rho_{k_{\boldsymbol{B}}})$. Integrating $\rho_{k_{\boldsymbol{B}}}$ and using Lemma 4, one can have $x_{..k_{\boldsymbol{B}}} \sim \text{NB}(\gamma_{\boldsymbol{B}}, p_{\boldsymbol{B}})$, where $p_{\boldsymbol{B}} = 1/(c_{\boldsymbol{B}}+1)$. Similarly, $y_{..k_{\boldsymbol{B}}} \sim \text{Pois}(\rho_{k_{\boldsymbol{B}}})$ and after integrating $\rho_{k_{\boldsymbol{B}}}$ and using Lemma 4, we have $y_{..k_{\boldsymbol{B}}} \sim \text{NB}(\gamma_{\boldsymbol{B}}, p_{\boldsymbol{B}})$. We now augment $l_{k_{\boldsymbol{B}}} \sim \text{CRT}(x_{..k_{\boldsymbol{B}}} + y_{..k_{\boldsymbol{B}}}, \gamma_{\boldsymbol{B}})$ and then following Lemma 6 sample

$$(\gamma_{\boldsymbol{B}}|-) \sim \text{Gamma}\left(e_{\boldsymbol{B}} + \sum_{k_{\boldsymbol{B}}} l_{k_{\boldsymbol{B}}}, \frac{1}{f_{\boldsymbol{B}}-q_{\boldsymbol{B}}}\right), q_{\boldsymbol{B}} = \sum_{k_{\boldsymbol{B}}} \frac{q_{k_{\boldsymbol{B}}}}{K_{\boldsymbol{B}}}, q_{k_{\boldsymbol{B}}} = \log\left(\frac{c_{\boldsymbol{B}}}{c_{\boldsymbol{B}}+\sum_n \phi_{nk_{\boldsymbol{B}}}\phi_{k_{\boldsymbol{B}}}^{-n}}\right). \tag{20}$$

Sampling of $\gamma_{\boldsymbol{Y}}$: Using Lemma 2, one can show that $y_{..(K_{\boldsymbol{B}}+k_{\boldsymbol{Y}})} \sim \text{Pois}(r_{k_{\boldsymbol{Y}}})$ and after integrating $r_{k_{\boldsymbol{Y}}}$ and using Lemma 4, we have $y_{..(K_{\boldsymbol{B}}+k_{\boldsymbol{Y}})} \sim \text{NB}(\gamma_{\boldsymbol{Y}}, p_{\boldsymbol{Y}})$, where $p_{\boldsymbol{Y}} = 1/(c_{\boldsymbol{Y}}+1)$. We now augment $m_{k_{\boldsymbol{Y}}} \sim \text{CRT}(y_{..(K_{\boldsymbol{B}}+k_{\boldsymbol{Y}})}, \gamma_{\boldsymbol{Y}})$ and then following Lemma 6 sample

$$(\gamma_{\boldsymbol{Y}}|-) \sim \text{Gamma}\left(e_{\boldsymbol{Y}} + \sum_{k_{\boldsymbol{Y}}} m_{k_{\boldsymbol{Y}}}, \frac{1}{f_{\boldsymbol{Y}}-q_{\boldsymbol{Y}}}\right), q_{\boldsymbol{Y}} = \sum_{k_{\boldsymbol{Y}}} \frac{q_{k_{\boldsymbol{Y}}}}{K_{\boldsymbol{Y}}}, q_{k_{\boldsymbol{Y}}} = \log\left(\frac{c_{\boldsymbol{Y}}}{c_{\boldsymbol{Y}}+\theta_{.k_{\boldsymbol{Y}}}}\right). \tag{21}$$

Table 2. Generative Process of N-GPPF

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

3.2 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 generative model of N-GPPF is given in Table 2. The update equations of variables corresponding to N-GPPF can be obtained with $\mathbf{Z} = \mathbf{0}$ in the corresponding equations. As mentioned in

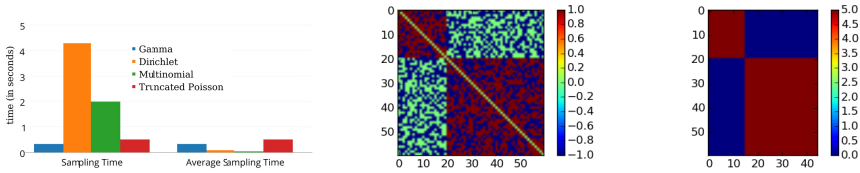
Table 3. Generative Process of C-GPPF

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

Section 2.3, N-GPPF can be considered as the gamma process infinite edge partition model (EPM) proposed in [40], which is shown to well model assortative networks but not necessarily disassortative ones. Using the techniques developed in [40] to capture community-community interactions, it is relatively straightforward to extend J-GPPF to model disassortative networks 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 C-GPPF is given in Table 3.

3.3 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 [31] that work with the matrix $\boldsymbol{Y}$ and incur $O(DVK_{\boldsymbol{Y}})$ computation cost. Interestingly, the inference in [14] incurs cost $O(K_{\boldsymbol{Y}}^2D + K_{\boldsymbol{Y}}V + K_{\boldsymbol{Y}}S_{\boldsymbol{Y}})$ 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 [14]. However, one needs to keep in mind that [14] use variational approximation for which the updates are available in closed form solution. Our method does not use any approximation to joint distribution but uses Gibbs sampling, the computation cost of which should also be taken into account. In



(a) (b) (c)

Fig. 1. (a) Time to generate a million of samples, (b) $\mathbf{B}$ with held-out data, (c) $\mathbf{Y}$

Fig. 1(a), we show the computation time for generating one million samples from Gamma, Dirichlet (of dimension 50), multinomial (of dimension 50) and truncated Poisson distributions using the samplers available from GNU Scientific Library (GSL) on an Intel 2127U machine with 2 GB of RAM and 1.90 GHz of processor base frequency. To highlight the average complexity of sampling from Dirichlet and multinomial distributions, we further display another plot where the computation time is divided by 50 for these samplers only. One can see that to draw one million samples, our implementation of the sampler for truncated Poisson distribution takes the longest, though the difference from the Gamma sampler in GSL is not that significant.

4 Experimental Results

4.1 Experiments with Synthetic Data

We generate a synthetic network of size 60×60 ($\mathbf{B}$) and a count data matrix of size 60×45 ($\mathbf{Y}$). Each user in the network writes exactly one document and a user and the corresponding document are indexed by the same row-index in $\mathbf{B}$ and $\mathbf{Y}$ respectively. To evaluate the quality of reconstruction in presence of side-information and less of network structure, we hold-out 50% of links and equal number of non-links from $\mathbf{B}$. This is shown in Fig. 1(b) where the links are presented by brown, the non-links by green and the held-out data by deep blue. Clearly, the network consists of two groups. $\mathbf{Y} \in \{0, 5\}^{60\times 45}$, shown in Fig. 1(c), is also assumed to consist of the same structure as $\mathbf{B}$ where the zeros are presented by deep blue and the non-zeros are represented by brown. The performance of N-GPPF is displayed in Fig. 2(a). Evidently, there is not much structure visible in the discovered partition of $\mathbf{B}$ from N-GPPF and that is reflected in the poor value of AUC in Fig. 3(a). The parameter ϵ, when fixed at a given value, plays an important role in determining the quality of reconstruction for J-GPPF. As $\epsilon \to 0$, J-GPPF approaches the performance of N-GPPF on $\mathbf{B}$ and we observe as poor a quality of reconstruction as in Fig. 2(a). When ϵ is increased and set to 1.0, J-GPPF departs from N-GPPF and performs much better in terms of both structure recovery and prediction on held-out data as shown in Fig. 2(e) and Fig. 3(b). With $\epsilon = 10.0$, perfect reconstruction and prediction are recorded as shown in Fig. 2(i) and Fig. 3(c) respectively. In this synthetic example, $\mathbf{Y}$ is purposefully designed to reinforce the structure of $\mathbf{B}$ when most of its links and non-links are held-out. However, in real applications, $\mathbf{Y}$ might not contain as much of information and the Gibbs sampler needs to find a suitable value of ϵ that can carefully glean information from $\mathbf{Y}$.

There are few more interesting observations from the experiment with synthetic data that characterize the behavior of the model and match our intuitions. In our experiment with synthetic data $K_{\mathbf{B}} = K_{\mathbf{Y}} = 20$ is used. Fig. 2(b) demonstrates the assignment of the users in the network communities and Fig. 2(d) illustrates the assignment of the documents to the combined space of network communities and the topics (with the network communities appearing before the topics in the plot). For $\epsilon = 0.001$, we observe disjoint modeling of $\mathbf{B}$ and $\mathbf{Y}$, with

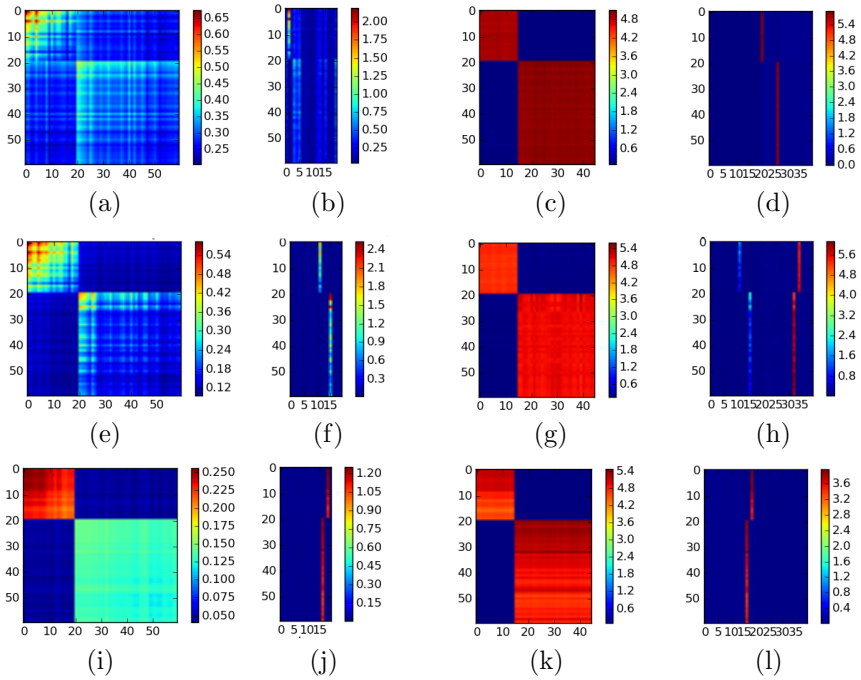


Fig. 2. Performance of J-GPPF: $\epsilon = 10^{-3}$ (top row), $\epsilon = 1$ (middle row), $\epsilon = 10$ (bottom row)

two latent factors modeling $\mathbf{Y}$ and multiple latent factors modeling $\mathbf{B}$ without any clear assignment. As we increase ϵ, we start observing joint modeling of $\mathbf{B}$ and $\mathbf{Y}$. For $\epsilon = 1.0$, as Fig. 2(h) reveals, two of the network latent factors and two of the factors for count data together model $\mathbf{Y}$, the contribution from the network factors being expectedly small. Fig. 2(f) shows how two of the exact same latent factors model $\mathbf{B}$ as well. Fig. 2(j) and Fig. 2(l) show how two of the latent factors corresponding to $\mathbf{B}$ dictate the modeling of both $\mathbf{B}$ and $\mathbf{Y}$ when $\epsilon = 10.0$. This implies that the discovery of latent groups in $\mathbf{B}$ is dictated mostly by the information contained in $\mathbf{Y}$. In all these cases, however, we observe perfect reconstruction of $\mathbf{Y}$ as shown in Fig. 2(c), Fig. 2(g) and Fig. 2(k).

4.2 Experiments with Real World Data

To evaluate the performance of J-GPPF, we consider N-GPPF, the infinite relational model (IRM) of [18] and the Mixed Membership Stochastic Block Model (MMSB) [2] as the baseline algorithms.

NIPS Authorship Network. This dataset contains the papers and authors from NIPS 1988 to 2003. We took the 234 authors who published with the most

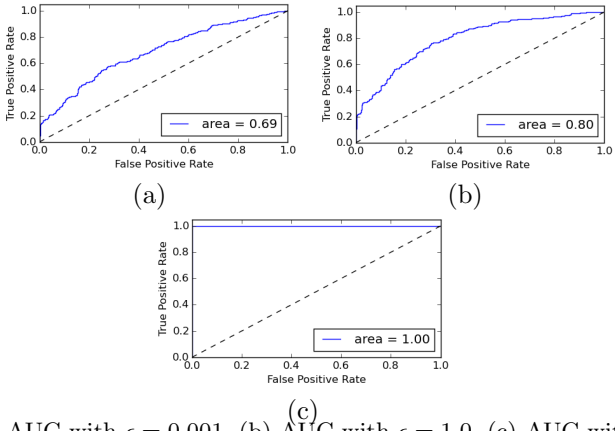


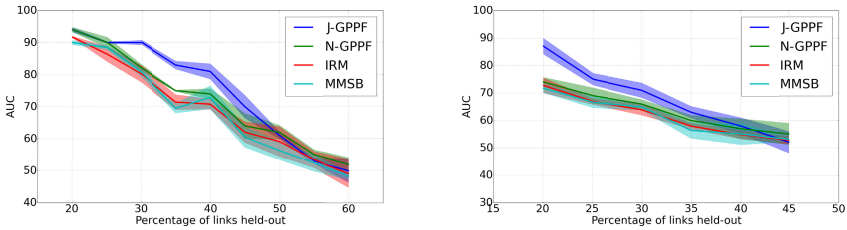
Fig. 3. (a) AUC with $\epsilon = 0.001$, (b) AUC with $\epsilon = 1.0$, (c) AUC with $\epsilon = 10.0$

other people and looked at their co-authors. After pre-processing and removing words that appear less than 50 times, the number of users in the graph is 225 and the total number of unique words is 1354. The total number of documents is 1165.

GoodReads Data. Using the Goodreads API, we collected a base set of users with recent activity on the website. The friends and friends of friends of these users were collected. Up to 200 reviews were saved per user, each consisting of a book ID and a rating from 0 to 5. A similar dataset was used in [10]. After pre-processing and removing words that appear less than 10 times, the number of users in the graph is 84 and the total number of unique words is 189.

Twitter Data. The Twitter dataset is a set of geo-tagged tweets collected by the authors in [29]. We extracted a subset of users located in San Francisco for our analysis. We created a graph within the subset by collecting follower information from the Twitter API. The side information consists of tweets aggregated by user, with one document per user. After pre-processing and removing words that appear less than 25 times, the number of users in the graph is 670 and the total number of unique words is 538.

Experimental Setup and Results. In all the experiments, we initialized ϵ to 2 and let the sampler decide what value works best for joint modeling. We used $K_{\mathbf{B}} = K_{\mathbf{Y}} = 50$ and initialized all the hyper-parameters to 1. For each dataset, we ran 20 different experiments and display the mean AUC and one standard error. Fig. 4 and 5 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



(a) (b)

Fig. 4. (a) NIPS Data, (b) GoodReads Data

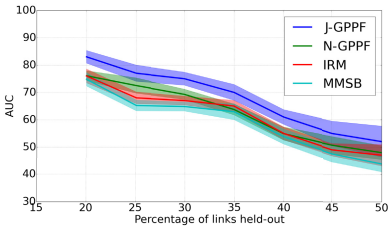


Fig. 5. Twitter Data

of network-only models in highly sparse networks, as the sampler faces additional difficulty in extracting information from both the network and the count matrix.

5 Conclusion

We propose J-GPPF that jointly factorizes the network adjacency matrix and the associated side information that can represented as a count matrix. The model has the advantage of representing true sparsity in adjacency matrix, in latent group membership, and in the side information. We derived an efficient MCMC inference method, and compared our approach to several popular network algorithms that model the network adjacency matrix. Experimental results confirm the efficiency of the proposed approach in utilizing side information to improve the performance of network models.

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

References

1. Acharya, A., Ghosh, J., Zhou, M.: Nonparametric bayesian factor analysis for dynamic count matrices. In: Proc. of AISTATS (to appear, 2015)
2. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Mixed membership stochastic blockmodels. JMLR **9**, 1981–2014 (2008)

3. Balasubramanyan, R., Cohen, W.W.: Block-LDA: jointly modeling entity-annotated text and entity-entity links. In: Proc. of SDM, pp. 450–461 (2011)
4. Ball, B., Karrer, B., Newman, M.: Efficient and principled method for detecting communities in networks. Phys. Rev. E **84**, September 2011
5. Blackwell, D., MacQueen, J.: Ferguson distributions via Pólya urn schemes. The Annals of Statistics (1973)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. JMLR **3**, 993–1022 (2003)
7. Broderick, T., Mackey, L., Paisley, J., Jordan, M.I.: Combinatorial clustering and the beta negative binomial process. arXiv:1111.1802v5 (2013)
8. Canny, J.: Gap: a factor model for discrete data. In: SIGIR (2004)
9. Cemgil, A.T.: Bayesian inference for nonnegative matrix factorisation models. Intell. Neuroscience (2009)
10. Chaney, A., Gopalan, P., Blei, D.: Poisson trust factorization for incorporating social networks into personalized item recommendation. In: NIPS Workshop: What Difference Does Personalization Make? (2013)
11. Chang, J., Blei, D.: Relational topic models for document networks. In: Proc. of AISTATS (2009)
12. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Statist. (1973)
13. Gopalan, P., Mimno, D.M., Gerrish, S., Freedman, M.J., Blei, D.M.: Scalable inference of overlapping communities. In: Proc. of NIPS, pp. 2258–2266 (2012)
14. Gopalan, P., Ruiz, F., Ranganath, R., Blei, D.: Bayesian nonparametric poisson factorization for recommendation systems. In: Proc. of AISTATS (2014)
15. Gopalan, P., Charlin, L., Blei, D.: Content-based recommendations with poisson factorization. In: Proc. of NIPS, pp. 3176–3184 (2014)
16. Hjort, N.L.: Nonparametric Bayes estimators based on beta processes in models for life history data. Ann. Statist. (1990)
17. Johnson, N.L., Kemp, A.W., Kotz, S.: Univariate Discrete Distributions. John Wiley & Sons (2005)
18. Kemp, C., Tenenbaum, J., Griffiths, T., Yamada, T., Ueda, N.: Learning systems of concepts with an infinite relational model. In: Proc. of AAAI, pp. 381–388 (2006)
19. Kim, D.I., Gopalan, P., Blei, D.M., Sudderth, E.B.: Efficient online inference for bayesian nonparametric relational models. In: Proc. of NIPS, pp. 962–970 (2013)
20. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: NIPS (2001)
21. Leskovec, J., Julian, J.: Learning to discover social circles in ego networks. In: Proc. of NIPS, pp. 539–547 (2012)
22. Ma, H., Yang, H., Lyu, M.R., King, I.: Sorec: social recommendation using probabilistic matrix factorization. In: Proc. of CIKM, pp. 931–940 (2008)
23. McCallum, A., Wang, X., Corrada-Emmanuel, A.: Topic and role discovery in social networks with experiments on enron and academic email. J. Artif. Int. Res. **30**(1), 249–272 (2007)
24. Menon, A.K., Elkan, C.: Link prediction via matrix factorization. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part II. LNCS, vol. 6912, pp. 437–452. Springer, Heidelberg (2011)
25. Miller, K.T., Griffiths, T.L., Jordan, M.I.: Nonparametric latent feature models for link prediction. In: Proc. of NIPS, pp. 1276–1284 (2009)
26. Nallapati, R., Ahmed, A., Xing, E., Cohen, W.: Joint latent topic models for text and citations. In: Proc. of KDD, pp. 542–550 (2008)

27. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics (2000)
28. Palla, K., Ghahramani, Z., Knowles, D.A.: An infinite latent attribute model for network data. In: Proc. of ICML, pp. 1607–1614 (2012)
29. Roller, S., Speriosu, M., Rallapalli, S., Wing, B., Baldridge, J.: Supervised text-based geolocation using language models on an adaptive grid. In: Proc. of EMNLP-CoNLL, pp. 1500–1510 (2012)
30. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proc. of UAI, pp. 487–494 (2004)
31. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: Proc. of NIPS (2007)
32. Titsias, M.K.: The infinite gamma-poisson feature model. In: Proc. of NIPS (2008)
33. Walker, S.G.: Sampling the Dirichlet mixture model with slices. Communications in Statistics Simulation and Computation (2007)
34. Wang, X., Mohanty, N., Mccallum, A.: Group and topic discovery from relations and their attributes. In: Proc. of NIPS, pp. 1449–1456 (2006)
35. Wen, Z., Lin, C.: Towards finding valuable topics. In: Proc. of SDM, pp. 720–731 (2010)
36. Wolpert, R.L., Clyde, M.A., Tu, C.: Stochastic expansions using continuous dictionaries: Lévy Adaptive Regression Kernels. Annals of Statistics (2011)
37. Yang, J., Leskovec, J.: Overlapping community detection at scale: a nonnegative matrix factorization approach. In: Proc. of WSDM, pp. 587–596 (2013)
38. Yang, J., McAuley, J.J., Leskovec, J.: Community detection in networks with node attributes. In: Proc. of ICDM, pp. 1151–1156 (2013)
39. Yoshida, T.: Toward finding hidden communities based on user profile. J. Intell. Inf. Syst. **40**(2), 189–209 (2013)
40. Zhou, M.: Infinite edge partition models for overlapping community detection and link prediction. In: Proc. of AISTATS (to appear, 2015)
41. Zhou, M., Carin, L.: Augment-and-conquer negative binomial processes. In: Proc. of NIPS (2012)
42. Zhou, M., Carin, L.: Negative binomial process count and mixture modeling. IEEE Trans. Pattern Analysis and Machine Intelligence (2015)
43. Zhou, M., Hannah, L., Dunson, D., Carin, L.: Beta-negative binomial process and poisson factor analysis. In: Proc. of AISTATS (2012)
44. Zhu, J.: Max-margin nonparametric latent feature models for link prediction. In: Proc. of ICML (2012)