Locally Private Bayesian Inference for Count Models

Aaron Schein [1] Zhiwei Steven Wu [2] Mingyuan Zhou [3] Hanna Wallach [2]

Abstract

As more aspects of social interaction are digitally recorded, there is a growing need to develop privacy-preserving data analysis methods. Social scientists will be more likely to adopt these methods if doing so entails minimal change to their current methodology. Toward that end, we present a general and modular method for privatizing Bayesian inference for Poisson factorization, a broad class of models that contains some of the most widely used models in the social sciences. Our method satisfies local differential privacy, which ensures that no single centralized server need ever store the non-privatized data. To formulate our local-privacy guarantees, we introduce and focus on limited-precision local privacy—the local privacy analog of limited-precision differential privacy (Flood et al., 2013). We present two case studies, one involving social networks and one involving text corpora, that test our method's ability to form the posterior distribution over latent variables under different levels of noise, and demonstrate our method's utility over a naïve approach, wherein inference proceeds as usual, treating the privatized data as if it were not privatized.

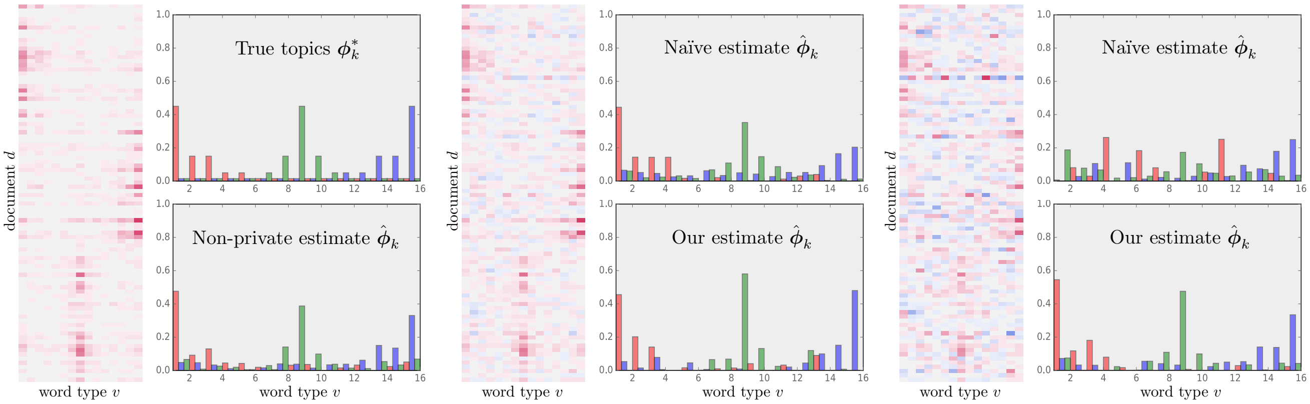
1. Introduction

Data from social processes often take the form of discrete observations (e.g., edges in a social network, word tokens in an email) and these observations often contain sensitive information about the people involved. As more aspects of social interaction are digitally recorded, the opportunities for social scientific insights grow; however, so too does the risk of unacceptable privacy violations. As a result, there is a growing need to develop privacy-preserving data analysis methods.

In practice, social scientists will be more likely to adopt these methods if doing so entails minimal change to their current methodology. Toward that end, under the framework of differential privacy (Dwork et al., 2006), we present a method for privatizing Bayesian inference for Poisson factorization (Titsias, 2008; Cemgil, 2009; Zhou & Carin, 2012; Gopalan & Blei, 2013; Paisley et al., 2014), a broad class of models for learning latent structure from discrete data. This class contains some of the most widely used models in the social sciences, including topic models for text corpora (Blei et al., 2003; Buntine & Jakulin, 2004; Canny, 2004), genetic population models (Pritchard et al., 2000), stochastic block models for social networks (Ball et al., 2011; Gopalan & Blei, 2013; Zhou, 2015), and tensor factorization for dyadic data (Welling & Weber, 2001; Chi & Kolda, 2012; Schmidt & Morup, 2013; Schein et al., 2015; 2016b); it further includes deep hierarchical models (Ranganath et al., 2015; Zhou et al., 2015), dynamic models (Charlin et al., 2015; Acharya et al., 2015; Schein et al., 2016a), and many others. Our method is general and modular, allowing social scientists to build on (instead of replace) their existing derivations and implementations of non-private Poisson factorization. To derive our method, we rely on a novel reinterpretation of the geometric mechanism (Ghosh et al., 2012), as well as a previously unknown general relationship between the Skellam (Skellam, 1946), Bessel (Yuan & Kalbfleisch, 2000), and Poisson distributions; we note that these new results may be of independent interest in other contexts.

Our method satisfies a strong variant of differential privacy—i.e., local privacy—under which the sensitive data is privatized (or noised) via a randomized response method before inference. This ensures that no single centralized server need ever store the non-privatized data—a condition that is non-negotiable in many real-world settings. The key challenge introduced by local privacy is how to infer the latent variables (including model parameters) given the privatized data. One option is a naïve approach, wherein inference proceeds as usual, treating the privatized data as if it were not privatized. In the context of maximum likelihood estimation, the naïve approach has been shown to exhibit pathologies when observations are discrete or count-valued; researchers have therefore advocated for treating the non-privatized observations as latent variables to be inferred (Yang et al., 2012; Karwa et al., 2014; Bernstein et al., 2017). We embrace this approach and extend it to Bayesian inference, where our aim is to form the posterior distribution over the

[1] University of Massachusetts Amherst [2] Microsoft Research, New York [3] University of Texas at Austin. Correspondence to: Aaron Schein <aschein@cs.umass.edu>, Zhiwei Steven Wu <steven7woo@gmail.com>, Mingyuan Zhou <mingyuan.zhou@mccombs.utexas.edu>, Hanna Wallach <hanna@dirichlet.net>.



(a) No noise.

(b) Low noise.

(c) High noise.

Figure 1. Topic recovery: our method vs. the naïve approach. (a) We generated the non-privatized data synthetically so that the true topics were known. We then privatized the data using (b) a low noise level and (c) a high noise level. The heatmap in each subfigure visualizes the data, using red to denote positive counts and blue to denote negative counts. With a high noise level, the naïve approach overfits the noise and therefore fails to recover the true topics. We describe this experiment in more detail in section 5.2.

latent variables conditioned on the privatized data and the randomized response method; our method is asymptotically guaranteed to draw samples from this posterior distribution.

We present two case studies applying our method to 1) overlapping community detection in social networks and 2) topic modeling for text corpora. In order to formulate our local-privacy guarantees, we introduce and focus on limited-precision local privacy—the local privacy analog of limited-precision differential privacy, originally proposed by Flood et al. (2013). For each case study, we report a suite of experiments that test our method's ability to form the posterior distribution over latent variables under different levels of noise. These experiments also demonstrate the utility of our method over the naïve approach for both case studies; we provide an illustrative example in figure 1.

2. Background and problem formulation

Differential privacy. Differential privacy (Dwork et al., 2006) is a rigorous privacy criterion that guarantees that no single observation in a data set will have a significant influence on the information obtained by analyzing that data set.

Definition 2.1. *A randomized algorithm $\mathcal{A}(\cdot)$ satisfies ϵ-differential privacy if for all pairs of neighboring data sets Y and Y' that differ in only a single observation*

$$P\left(\mathcal{A}(Y) \in \mathcal{S}\right) \leq e^{\epsilon} P\left(\mathcal{A}(Y') \in \mathcal{S}\right) \qquad (1)$$

for all subsets $\mathcal{S}$ in the range of $\mathcal{A}(\cdot)$.

Local differential privacy. We focus on local differential privacy, which we refer to as local privacy. In this setting, the observations remain private from even the data analysis algorithm. The algorithm only sees privatized versions of the observations, often constructed by adding noise from specific distributions. The process of adding noise is known as randomized response—a reference to survey-sampling methods originally developed in the social sciences prior to the development of differential privacy (Warner, 1965).

Definition 2.2. *A randomized response method $\mathcal{R}(\cdot)$ is ϵ-private if for all pairs of observations $y, y' \in \mathcal{Y}$*

$$P\left(\mathcal{R}(y) \in \mathcal{S}\right) \leq e^{\epsilon} P\left(\mathcal{R}(y') \in \mathcal{S}\right) \qquad (2)$$

for all subsets $\mathcal{S}$ in the range of $\mathcal{R}(\cdot)$. If a data analysis algorithm sees only the observations' ϵ-private responses, then the data analysis itself satisfies ϵ-local privacy.

Limited-precision local privacy. Definition 2.2 requires that condition 2 hold for all pairs of observations $y, y' \in \mathcal{Y}$. In practice, this is notoriously difficult to achieve when $\mathcal{Y}$ is extremely large, meaning that any pair of observations may be arbitrarily different, as is often the case with data from social processes. We therefore introduce and focus on limited-precision local privacy—the local privacy analog of limited-precision differential privacy, originally proposed by Flood et al. (2013) and subsequently used to privatize analyses of geographic location data (Andrés et al., 2013) and financial network data (Papadimitriou et al., 2017). Although limited-precision local privacy is weaker than local privacy, it can still provide reasonably strong guarantees.

Definition 2.3. *If N is a positive integer, then a randomized response method $\mathcal{R}(\cdot)$ is (N, ϵ)-private if for all pairs of observations $y, y' \in \mathcal{Y}$ such that $\|y - y'\|_1 \leq N$*

$$P\left(\mathcal{R}(y) \in \mathcal{S}\right) \leq e^{\epsilon} P\left(\mathcal{R}(y') \in \mathcal{S}\right) \qquad (3)$$

for all subsets $\mathcal{S}$ in the range of $\mathcal{R}(\cdot)$. If a data analysis algorithm sees only the observations' (N, ϵ)-private responses, then the data analysis itself satisfies (N, ϵ)-limited-precision local privacy. If $\|y\|_1 \leq N$ for all $y \in \mathcal{Y}$, then (N, ϵ)-limited-precision local privacy implies ϵ-local privacy.

Geometric mechanism. There are several standard randomized response methods in the differential privacy toolbox, many of which involve adding independently generated noise to each element of each observation. Unfortunately, the most commonly used noise mechanisms—the Gaussian and Laplace mechanisms—are poor choices for count data because they involve real-valued distributions. We therefore focus on the geometric mechanism (Ghosh et al., 2012), which can be viewed as the discrete analog of the Laplace mechanism. The geometric mechanism adds noise drawn from a two-sided geometric distribution to each element of each observation. A two-sided geometric random variable $\tau \sim 2\text{Geo}(\alpha)$ is an integer $\tau \in \mathbb{Z}$. The PMF for the two-sided geometric distribution is as follows:

$$2\text{Geo}(\tau; \alpha) = \frac{1-\alpha}{1+\alpha}\, \alpha^{|\tau|}. \tag{4}$$

Theorem 2.4. (Proof in appendix.) *If N is a positive integer and randomized response method $\mathcal{R}(\cdot)$ is the geometric mechanism with parameter α, then for any pair of observations $y, y' \in \mathcal{Y}$ such that $\|y - y'\|_1 \leq N$, $\mathcal{R}(\cdot)$ satisfies*

$$P\left(\mathcal{R}(y) \in \mathcal{S}\right) \leq e^{\epsilon} P\left(\mathcal{R}(y') \in \mathcal{S}\right) \tag{5}$$

for all subsets $\mathcal{S}$ in the range of $\mathcal{R}(\cdot)$, where

$$\epsilon = N \ln\left(\frac{1}{\alpha}\right). \tag{6}$$

Therefore, the geometric mechanism with parameter α is an (N, ϵ)-private randomized response method with $\epsilon = N \ln\left(\frac{1}{\alpha}\right)$. If a data analysis algorithm sees only the observations' (N, ϵ)-private responses, then the data analysis itself satisfies (N, ϵ)-limited precision local privacy.

Differentially Private Bayesian inference. In Bayesian statistics, we begin with a probabilistic model $\mathcal{M}$ that relates observable variables Y to latent variables Z via a joint distribution $P_{\mathcal{M}}(Y, Z)$. The goal of inference is then to compute the posterior distribution $P_{\mathcal{M}}(Z \,|\, Y)$ over the latent variables conditioned on observed values of Y. The posterior is almost always analytically intractable and thus inference involves approximating it. The two most common methods of approximate Bayesian inference are variational inference, wherein we fit the parameters of an approximating distribution $Q(Z \,|\, Y)$, and Markov chain Monte Carlo (MCMC), wherein we approximate the posterior with a finite set of samples $\{Z^{(s)}\}_{s=1}^{S}$ generated via a Markov chain whose stationary distribution is the exact posterior. We can conceptualize each of these methods as a randomized algorithm $\mathcal{A}(\cdot)$ that returns an approximation to the posterior distribution $P_{\mathcal{M}}(Z \,|\, Y)$; in general $\mathcal{A}(\cdot)$ does not satisfy ϵ-differential privacy. However, if $\mathcal{A}(\cdot)$ is an MCMC algorithm that returns a single sample from the posterior, it guarantees privacy (Dimitrakakis et al., 2014; Wang et al., 2015; Foulds et al., 2016). Adding noise to posterior samples can also guarantee privacy (Zhang et al., 2016), though this set of noised samples $\{\tilde{Z}^{(s)}\}_{s=1}^{S}$ collectively approximate some distribution $\tilde{P}_{\mathcal{M}}(Z \,|\, Y)$ that depends on ϵ and is different than the exact posterior (but close, in some sense, and equal when $\epsilon \to 0$). For specific models, we can also noise the transition kernel of the MCMC algorithm to construct a Markov chain whose stationary distribution is again not the exact posterior, but something close that guarantees privacy (Foulds et al., 2016). We can also take an analogous approach to privatize variational inference, wherein we add noise to the sufficient statistics computed in each iteration (Park et al., 2016).

Locally private Bayesian inference. We first formalize the general objective of Bayesian inference under local privacy. Given a generative model $\mathcal{M}$ for non-privatized data Y and latent variables Z with joint distribution $P_{\mathcal{M}}(Y, Z)$, we further assume a randomized response method $\mathcal{R}(\cdot)$ that generates privatized data sets: $\tilde{Y} \sim P_{\mathcal{R}}(\tilde{Y} \,|\, Y)$. The aim of Bayesian inference is then to form the following posterior:

$$\begin{aligned} P_{\mathcal{M},\mathcal{R}}(Z \,|\, \tilde{Y}) &= \mathbb{E}_{P_{\mathcal{R}}(Y \,|\, \tilde{Y})}\left[P_{\mathcal{M}}(Z \,|\, Y)\right] \\ &= \int P_{\mathcal{M}}(Z \,|\, Y)\, P_{\mathcal{R}}(Y \,|\, \tilde{Y})\, dY. \end{aligned} \tag{7}$$

This distribution correctly characterizes our uncertainty about the latent variables Z, conditioned on all of our observations and assumptions—i.e., the privatized data $\tilde{Y}$, the model $\mathcal{M}$, and the randomized response method $\mathcal{R}$. The expansion in equation 7 shows that this posterior implicitly treats the non-privatized data Y as a latent variable and marginalizes over it using the mixing distribution $P_{\mathcal{R}}(Y \,|\, \tilde{Y})$ which is itself a posterior that characterizes our uncertainty about Y given $\tilde{Y}$ and the randomized response method. The key observation here is that if we can generate samples from $P_{\mathcal{R}}(Y \,|\, \tilde{Y})$, we can use them to approximate the expectation in equation 7, assuming that we already have a method for approximating the non-private posterior $P_{\mathcal{M}}(Z \,|\, Y)$. In the context of MCMC, alternating between sampling values of the non-privatized data from its complete conditional—i.e., $Y^{(s)} \sim P_{\mathcal{M},\mathcal{R}}(Y \,|\, Z^{(s-1)}, \tilde{Y})$—and sampling values of the latent variables—i.e., $Z^{(s)} \sim P_{\mathcal{M}}(Z \,|\, Y^{(s)})$—constitutes a Markov chain whose stationary distribution is $P_{\mathcal{M},\mathcal{R}}(Z, Y \,|\, \tilde{Y})$. In scenarios where we already have derivations and implementations for sampling from $P_{\mathcal{M}}(Z \,|\, Y)$, we need only be able to sample efficiently from $P_{\mathcal{M},\mathcal{R}}(Y \,|\, Z, \tilde{Y})$ in order to obtain a locally private Bayesian inference algorithm; whether we can do this depends heavily on our assumptions about $\mathcal{M}$ and $\mathcal{R}$.

We note that the objective of Bayesian inference under local privacy, as defined in equation 7, is similar to that of Williams & McSherry (2010), who identify their key barrier to inference as being unable to analytically form the marginal likelihood that links the privatized data to Z:

$$P_{\mathcal{M},\mathcal{R}}(\tilde{Y} \mid Z) = \int P_{\mathcal{R}}(\tilde{Y} \mid Y)\, P_{\mathcal{M}}(Y \mid Z)\, dY. \quad (8)$$

In the next sections, we show that if $\mathcal{M}$ is a Poisson factorization model and $\mathcal{R}$ is the geometric mechanism, then we can analytically form this marginal likelihood and derive an efficient MCMC algorithm that is asymptotically guaranteed to generate samples from the posterior in equation 7.

3. Locally private Poisson factorization

Poisson factorization. We assume that Y is a count-valued data set. We further assume that each count $y_n \in \mathbb{Z}_+$ in this data set is an independent Poisson random variable $y_n \sim \text{Pois}(\mu_n)$, where the count's latent rate parameter μ_n is a function of the latent variables Z. This class of models is known as Poisson factorization and, as described in section 1, includes many widely used models in social science. For example, the mixed-membership stochastic block model for social networks (Ball et al., 2011; Gopalan & Blei, 2013; Zhou, 2015) corresponds to the case where Y is a $V \times V$ count matrix; $n = (i, j)$, where $i, j \in [V]$; $Z = \{\Theta, \Pi\}$; Θ and Π are $V \times C$ and $C \times C$ non-negative, real-valued matrices, respectively; and $\mu_{ij} = \sum_{c=1}^{C} \sum_{d=1}^{C} \theta_{ic}\, \theta_{jd}\, \pi_{cd}$. Similarly, latent Dirichlet allocation (Blei et al., 2003)—a well-known topic model for text corpora—corresponds to the case where Y is a $D \times V$ count matrix; $n = (d, v)$, where $d \in [D]$ and $v \in [V]$; $Z = \{\Theta, \Phi\}$, where Θ and Φ are $D \times K$ and $K \times V$ non-negative, real-valued matrices, respectively; and $\mu_{dv} = \sum_{k=1}^{K} \theta_{dk}\, \phi_{kv}$. In both cases, it is standard to assume independent gamma priors over the elements of the latent matrices that comprise Z; doing so facilitates efficient Bayesian inference of these matrices via gamma–Poisson conjugacy (when conditioned on Y).

Geometric mechanism. We focus on the geometric mechanism (Ghosh et al., 2012) because it is a natural choice for count data. By reinterpreting the geometric mechanism as involving Skellam noise and deriving a general relationship between the Skellam, Bessel, and Poisson distributions, we are able to obtain analytic tractability and efficient Bayesian inference while also maintaining local privacy guarantees. In particular, we show that augmenting our model with auxiliary variables $\boldsymbol{\lambda}_n = (\lambda_{n1}, \lambda_{n2})$ allows us to analytically form the marginal likelihood $P_{\mathcal{M},\mathcal{R}}(\tilde{y}_n \mid \mu_n, \boldsymbol{\lambda}_n)$ and sample efficiently from $P_{\mathcal{M},\mathcal{R}}(y_n \mid \tilde{y}_n, \mu_n, \boldsymbol{\lambda}_n)$, as desired.

Each non-privatized count y_n is generated by our model $\mathcal{M}$—i.e., $y_n \sim \text{Pois}(\mu_n)$—and then privatized as follows:

$$\tau_n \sim 2\text{Geo}(\alpha), \;\; \tilde{y}_n^{(\pm)} := y_n + \tau_n. \quad (9)$$

We use $(\pm)$ to emphasize that unlike y_n (which must be non-negative) $\tilde{y}_n^{(\pm)} \in \mathbb{Z}$ may be non-negative or negative.

Theorem 3.1. (Proof in appendix.) *A two-sided geometric random variable $\tau \sim 2Geo(\alpha)$ can be generated as follows:*

$$\lambda_1, \lambda_2 \sim \textit{Exp}(\tfrac{\alpha}{1-\alpha}), \;\; \tau \sim \textit{Skel}(\lambda_1, \lambda_2), \quad (10)$$

where the Skellam distribution is the marginal distribution over the difference $\tau := g_1 - g_2$ of two independent Poisson random variables $g_1 \sim \textit{Pois}(\lambda_1)$ and $g_2 \sim \textit{Pois}(\lambda_2)$.

Via theorem C.1, we can express the generative process for $\tilde{y}_n^{(\pm)}$ in three equivalent ways, shown in figure 2, each of which provides a unique and necessary insight. The first way (process 1) is useful for showing that our MCMC algorithm guarantees privacy, since two-sided geometric noise is an existing privacy mechanism. The second way (process 2) represents the two-sided geometric noise in terms of a pair of Poisson random variables with exponentially distributed rates; in so doing, it reveals the auxiliary variables that facilitate inference. The third way (process 3) marginalizes out all three Poisson random variables (including y_n), so that $\tilde{y}_n^{(\pm)}$ is directly drawn from a Skellam distribution, which also happens to be the desired marginal likelihood $P_{\mathcal{M},\mathcal{R}}(\tilde{y}_n \mid \mu_n, \boldsymbol{\lambda}_n)$ under the geometric mechanism. To derive the second and third ways, we use theorem C.1, the definition of the Skellam distribution, and the additive property of two or more Poisson random variables.

4. MCMC algorithm

We now rely on a previously unknown general relationship between the Skellam, Bessel, and Poisson distributions to derive an efficient way to draw samples from $P_{\mathcal{M},\mathcal{R}}(y_n \mid \tilde{y}_n, \mu_n, \boldsymbol{\lambda}_n)$. As explained in section 2, this is all we need to obtain an locally private MCMC algorithm for drawing samples of the latent variables given the privatized data, provided we already have a way to draw samples of the latent variables given the non-private data. The input to this MCMC algorithm is the privatized data set $\tilde{Y}^{(\pm)}$.

Theorem 4.1. (Proof in appendix.) *Consider two Poisson random variables $y_1 \sim \textit{Pois}(\lambda_1)$ and $y_2 \sim \textit{Pois}(\lambda_2)$. Their minimum $m := \textit{min}\{y_1, y_2\}$ and their difference $\delta := y_1 - y_2$ are deterministic functions of y_1 and y_2. However, if not conditioned on y_1 and y_2, the random variables m and δ can be marginally generated as follows:*

$$\delta \sim \textit{Skel}(\lambda_1, \lambda_2), \;\; m \sim \textit{Bes}\left(|\delta|, 2\sqrt{2\lambda_1\lambda_2}\right). \quad (11)$$

Yuan & Kalbfleisch (2000) give details of the Bessel distribution, which can be sampled efficiently (Devroye, 2002).[1]

Lemma D.1 means that we can generate two independent Poisson random variables by first generating their difference

[1] We have released our implementation of Bessel sampling. It is the only open-source version of which we are aware.

Process 1	**Process 2**	**Process 3**
—	$\lambda_{n1}, \lambda_{n2} \sim \text{Exp}(\frac{\alpha}{1-\alpha})$	$\lambda_{n1}, \lambda_{n2} \sim \text{Exp}(\frac{\alpha}{1-\alpha})$
$\tau_n \sim \text{2Geo}(\alpha)$	$g_{nl} \sim \text{Pois}(\lambda_{nl})$ for $l \in \{1, 2\}$	—
$y_n \sim \text{Pois}(\mu_n)$	$y_n \sim \text{Pois}(\mu_n)$	—
$\tilde{y}_n^{(\pm)} := y_n + \tau_n$	$\tilde{y}_n^{(\pm)} := y_n + g_{n1} - g_{n2}$	$\tilde{y}_n^{(\pm)} \sim \text{Skel}(\lambda_{n1}{+}\mu_n,\, \lambda_{n2})$

Figure 2. Three equivalent ways to generate $\tilde{y}_n^{(\pm)}$.

δ and then their minimum m. Because $\delta = y_1 - y_2$, if δ is positive, then y_2 must be the minimum and thus $y_1 = \delta - m$. In practice, this means that if we only get to observe the difference of two Poisson-distributed counts, we can still "recover" the counts by drawing a Bessel random variable.

Assuming that $\tilde{y}_n^{(\pm)} \sim \text{Skel}(\lambda_{n1} + \mu_n, \lambda_{n2})$ via theorem C.1, we can represent $\tilde{y}_n^{(\pm)}$ explicitly as the difference between two latent non-negative counts: $\tilde{y}_n^{(\pm)} = \tilde{y}_n^{(+)} - g_{n2}$. We can then define the minimum of these latent counts to be $m_n = \min\{\tilde{y}_n^{(+)}, g_{n2}\}$. Given randomly initialized latent variables, we can then sample a value of m_n from its conditional posterior, which is a Bessel distribution:

$$\left(m_n \mid - \right) \sim \text{Bes}\left(|\tilde{y}_n^{(\pm)}|,\, 2\sqrt{(\lambda_{n1}{+}\mu_n)\lambda_{n2}}\right). \quad (12)$$

Using this value, we can then compute $\tilde{y}_n^{(+)}$ and g_{n2}:

$$\tilde{y}_n^{(+)} := m_n, \;\; g_{n2} := \tilde{y}_n^{(+)} - \tilde{y}_n^{(\pm)} \;\text{ if } \tilde{y}_n^{(\pm)} \leq 0 \quad (13)$$

$$g_{n2} := m_n, \;\; \tilde{y}_n^{(+)} := g_{n2} + \tilde{y}_n^{(\pm)} \;\text{ otherwise.} \quad (14)$$

Because $\tilde{y}_n^{(+)}$ is the sum of y_n and g_{n1}—two independent Poisson random variables—we can then sample y_n from its conditional posterior, which is a binomial distribution:

$$\left(y_n \mid - \right) \sim \text{Binom}\left(\tilde{y}_n^{(+)}, \tfrac{\mu_n}{\mu_n + \lambda_{n1}}\right) \quad (15)$$

Equations 12 through 15 constitute a way to draw samples from $P_{\mathcal{M},\mathcal{R}}(y_n \mid \tilde{y}_n, \mu_n, \boldsymbol{\lambda}_n)$. Given a sampled Y, we can then draw samples of the latent variables from their conditional posteriors, which are the same as in non-private Poisson factorization. Finally, we can also sample λ_{n1} and λ_{n2}:

$$\left(\lambda_{nl} \mid - \right) \sim \Gamma\left(1 + g_{nl}, \tfrac{\alpha}{1-\alpha} + 1\right) \text{ for } l \in \{1, 2\}. \quad (16)$$

Equation 16 follows from gamma–Poisson conjugacy and the fact that the exponential prior over λ_{nl} can be expressed as a gamma prior with shape parameter equal to one—i.e., $\lambda_{nl} \sim \Gamma(1, \frac{\alpha}{1-\alpha})$. Equations 12–16, along with the conditional posteriors for the latent variables, define an MCMC algorithm that is asymptotically guaranteed to generate samples from $P_{\mathcal{M},\mathcal{R}}(Z \mid \tilde{Y}^{(\pm)})$ as desired.

5. Case studies

We now present two case studies applying our method to 1) overlapping community detection in social networks and 2) topic modeling for text corpora. For each case study, we formulate local-privacy guarantees and ground them in illustrative examples. We then report a suite of experiments that test our method's ability to form the posterior distribution over latent variables for different types of data under different levels of noise. We focus on synthetic and semi-synthetic data to control for the effects of model mismatch (i.e., non-Poisson observations); although model mismatch is an important problem, it is outside the scope of this paper. Using synthetic and semi-synthetic data also allows us to vary high-level properties of the data (e.g., scale or sparsity).

Reference methods. We compare the performance of our method to two references methods: 1) non-private Poisson factorization on the non-privatized data and 2) non-private Poisson factorization on the privatized data—i.e., the naïve approach, wherein inference proceeds as usual, treating the privatized data as if it were not privatized.[2] Throughout our experiments, we use MCMC for both reference methods.

Performance measure. Ideally, we would directly compare our method's posterior distribution and the naïve approach's posterior distribution to that of non-private Poisson factorization on the non-privatized data. Unfortunately, all three posteriors are analytically intractable. However, because we use MCMC to approximate each posterior with a finite set of samples of the latent variables, we can instead form the expected value of μ_n with respect to each one—e.g.,

$$\hat{\mu}_n = \frac{1}{S}\sum_{s=1}^{S} \mu_n^{(s)} \approx \mathbb{E}_{P_{\mathcal{M},\mathcal{R}}(Z \mid \tilde{Y}^{(\pm)})}\left[\mu_n\right]. \quad (17)$$

Furthermore, because we focus on synthetic and semi-synthetic data, we can use an aggregate loss function to compare the expected values to the values used to generate the data: $\frac{1}{N}\sum_{n=1}^{N} \ell(\hat{\mu}_n, \mu_n^*)$, where μ_n^* is the "true" value. We define $\ell(\hat{\mu}_n, \mu_n^\star)$ to be the KL divergence of the Poisson distribution implied by $\hat{\mu}_n$ from the Poisson distribution implied by $\mu_n^\star$. Comparing the value of this aggregate loss

[2] The naïve approach first truncates negative counts to zero and thus uses the *truncated* geometric mechanism (Ghosh et al., 2012).

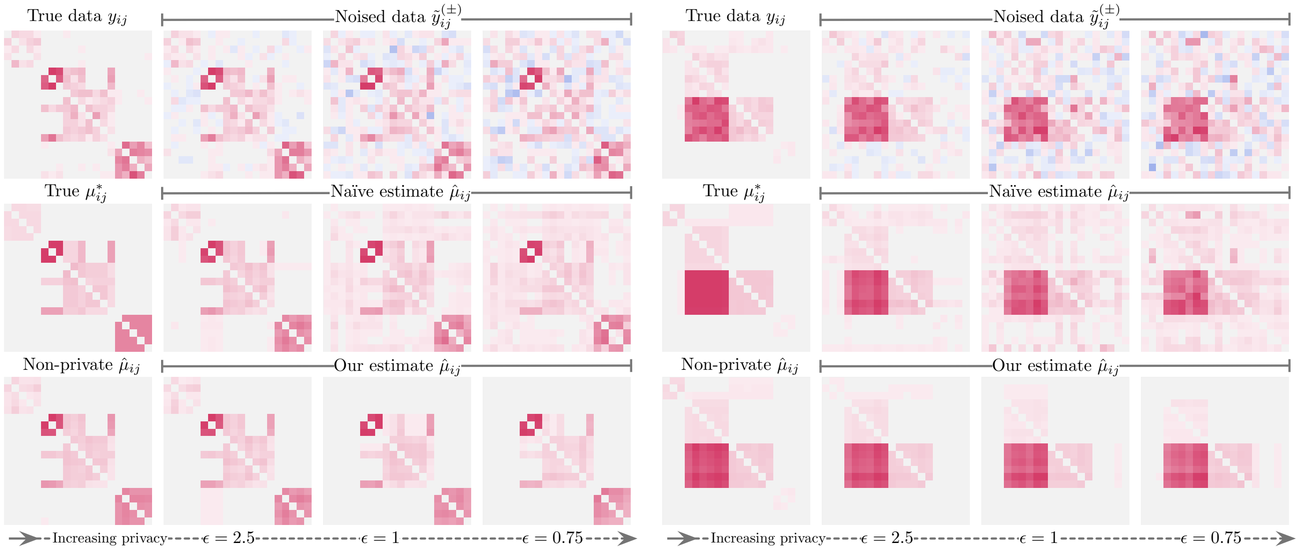


Figure 3. Block structure recovery: our method vs. the naïve approach. We generated the non-privatized data synthetically. We then privatized the data using three different levels of noise. The top row depicts the data, using red to denote positive counts and blue to denote negative counts. As the noise level increases, the naïve approach overfits the noise and fails to recover the true $\mu^\star_{ij}$ values, predicting high values even for sparse parts of the matrix. In contrast, our method recovers the latent structure, even for high noise levels.

function for our method and the value for naïve approach to the value for non-private Poisson factorization provides a proxy for measuring the divergence of their posterior distributions from that of non-private Poisson factorization.

5.1. Case study 1: Overlapping community detection

Organizations often want to know whether their employees are interacting as efficiently and productively as possible. For example, are there missing connections between employees that, if present, would significantly reduce duplication of effort? Do the natural “communities” that emerge from digitally recorded employee interactions match up with the formal organizational structure? To answer these and other questions, many organizations want to partner with social scientists in order to gain actionable insights based on their employees’ interactions. However, sharing such interaction data increases the risk of privacy violations. Moreover, standard anonymization procedures can be reverse-engineered adversarially and thus do not provide privacy guarantees (Narayanan & Shmatikov, 2009). In contrast, the formal privacy guarantees provided by differential privacy may be sufficient for employees to consent to sharing their data.

Limited-precision local privacy. In this scenario, data set Y is a $V \times V$ count matrix, where each element $y_{ij} \in \mathbb{Z}_+$ in this matrix is the number of interactions from actor $i \in V$ to actor $j \in [V]$. A single observation in this data set is a single element. Via theorem 2.4, $\tilde{y}_{ij}^{(\pm)} := y_{ij} + \tau_{ij}$, where $\tau_{ij} \sim 2\text{Geo}(\alpha)$, is (N, ϵ)-private, where N is the precision level and $\epsilon = N \ln\left(\frac{1}{\alpha}\right)$. Informally, this means that if the difference between two observations is N or less, then their privatized versions will be indistinguishable, provided ϵ is sufficiently small. Furthermore, if $y_{ij} \leq N$, then its privatized version will be indistinguishable from the privatized version of $y_{ij} = 0$. For example, if i interacted with j three times (i.e., $y_{ij} = 3$) and $N = 3$, then an adversary would be unable to tell from $\tilde{y}_{ij}^{(\pm)}$ whether i had interacted with j at all, provided ϵ is sufficiently small. We note that if $y_{ij} \gg N$, then an adversary would be able to tell that i had interacted with j, though not the exact number of times.

Poisson factorization. As explained in section 3, the mixed-membership stochastic block model for learning latent overlapping community structure in social networks (Ball et al., 2011; Gopalan & Blei, 2013; Zhou, 2015) is a special case of Poisson factorization where Y is a $V \times V$ count matrix; $n = (i, j)$, where $i, j \in [V]$; $Z = \{\Theta, \Pi\}$; Θ and Π are $V \times C$ and $C \times C$ non-negative, real-valued matrices, respectively; and $\mu_{ij} = \sum_{c=1}^{C} \sum_{d=1}^{C} \theta_{ic}\, \theta_{jd}\, \pi_{cd}$. The factors θ_{ic} and θ_{jd} represent how much actors i and j participate in communities c and d, respectively, while the factor π_{cd} represents how much actors in community c interact with actors in community d. It is standard to assume independent gamma priors over the factors—i.e., $\theta_{ic}, \pi_{cd} \sim \text{Gamma}(a_0, b_0)$, where a_0 and b_0 are shape and rate hyperparameters, respectively.

Synthetic data. We generated social networks of $V = 20$ actors with $C = 5$ communities. We randomly generated the true parameters $\theta^*_{ic}, \pi^*_{cd} \sim \Gamma(a_0, b_0)$ by setting $a_0 = 0.01$ and $b_0 = 0.5$ to encourage sparsity; doing so exaggerates the block structure in the network. We then sampled a data set $y_{ij} \sim \text{Pois}(\mu^*_{ij})$ and noised it $\tau_{ij} \sim 2\text{Geo}(\alpha)$ for three increasing values of α. Since the magnitude of the counts y_{ij}

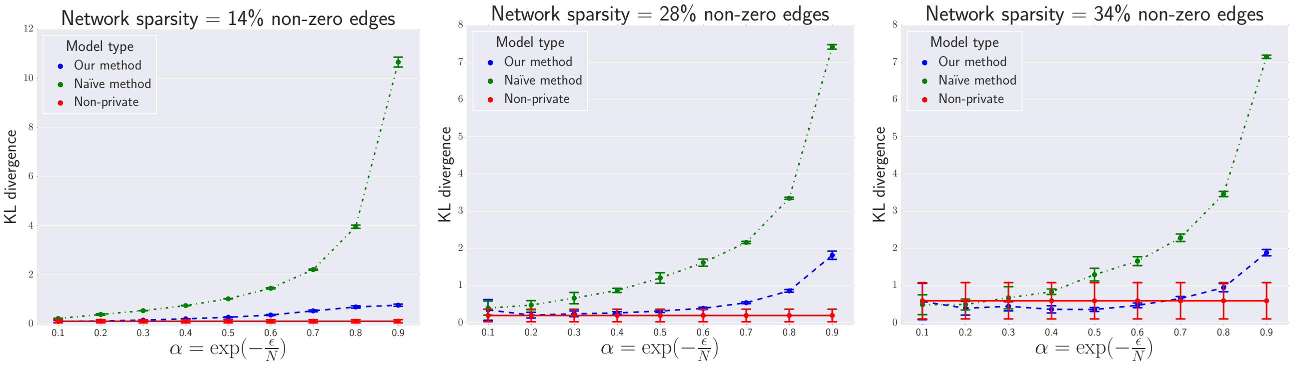


Figure 4. Mean KL divergence of the Poisson distribution implied by $\hat{\mu}_{ij}$ from the Poisson distribution implied by μ^*_{ij}; lower is better. Error bars denote standard deviation across five replications. Each subfigure reports the results across nine values of α (higher values mean more noise) for a given setting of e_0 which controls the sparsity and magnitude of the count observations. As the counts grow larger and denser, the performance of both private methods approaches the performance of non-private Poisson factorization for small values of α. Our method almost always outperforms the naïve approach, with the difference being especially stark at higher noise levels.

varied across trials, we allowed the three values of α to vary by setting the precision to the empirical mean of the data $N := \hat{\mathbb{E}}[y_{ij}]$ and setting $\alpha := \exp(-\epsilon/N)$ for three values of $\epsilon \in \{2.5, 1, 0.75\}$. For each model, we ran 8,500 sampling iterations, saving every 25th sample after the first 1,000 and using these samples to compute $\hat{\mu_{ij}}$, as given in equation 17. In figure 3, we visually compare the estimates of $\hat{\mu}_{ij}$ by our method and the naïve approach, under the three different noise levels, to the estimate by the non-private method and to the true values μ^*_{ij}. We see that the naïve approach overfits the noise, predicting high rates in sparse parts of the matrix. In contrast, our method maintains a precise representation of the data even under high noise levels.

Enron. We processed the Enron corpus (Klimt & Yang, 2004) to obtain a $V \times V$ adjacency matrix Y where y_{ij} is the number of emails sent from actor i to actor j. We included an actor if they sent at least one email and sent or received at least one hundred emails, yielding $V = 161$ actors. When an email included multiple recipients, we incremented the corresponding counts by one. We fit non-private Poisson factorization to the data Y and obtained point estimates of the factors θ^*_{ic} and π^*_{cd}. We then generated semi-synthetic data sets $y_{ij} \sim \text{Pois}(\mu^*_{ij})$ where we fixed π^*_{cd} but re-scaled $\theta^*_{ic} = \gamma_i\, \theta^*_{ic}$ allowing γ_i to vary. Rescaling changes the overall interaction rate of actor i without changing their relative frequency of participation across communities; this allows us to vary the scale and sparsity of the network while maintaining realistic community block structure. We randomly generated $\gamma_i \sim \Gamma(e_0 f_0, f_0)$ for $f_0 = 0.1$ and varying e_0. Under this parameterization of the gamma distribution the mean is $\mathbb{E}[\gamma_i] = e_0$; this allows us to increase the scale and density of the data. We generated five semi-synthetic data sets for three values of $e_0 \in \{0.75, 1, 1.5\}$ and nine levels of noise $\alpha \in \{0.1, \ldots, 0.9\}$. We applied our method and the naïve approach to each noised data set by running 15,000 sampling iterations and saving every 25th sample after the first 3,000 to compute $\hat{\mu}_{ij}$. In figure 4, we report the mean KL divergence of the Poisson distribution implied by $\hat{\mu}_{ij}$ from the Poisson distribution implied by μ^*_{ij} for both models.

5.2. Case study 2: Topic modeling

Topic models are in widely used in the social sciences for learning latent topics (i.e., probability distributions over some vocabulary) from text corpora, often to characterize high-level thematic structure (e.g., Ramage et al., 2009; Grimmer & Stewart, 2013; Mohr & Bogdanov, 2013; Roberts et al., 2013). In many settings, these corpora contain sensitive information about the people involved (e.g., emails, survey responses). As a result, people may be unwilling to consent to sharing their data without formal privacy guarantees, such as those provided by differential privacy.

Limited-precision local privacy. In this scenario, data set Y is a $D \times V$ count matrix, where each element $y_{dv} \in \mathbb{Z}_+$ in this matrix is the number of times word type $v \in [V]$ occurred in document $d \in [D]$. Similar to the community-detection scenario, a single observation in this data set might correspond to a single element. In this case, if $y_{dv} \leq N$, an adversary would be unable to tell from $\tilde{y}^{(\pm)}_{dv}$ whether v occurred in d, provided ϵ is sufficiently small. However, a more natural interpretation would be to assume that a single observation is an entire document—i.e., $\boldsymbol{y}_d = (y_{d1}, \ldots, y_{dV})$. In this case, if the ℓ_1 norm of the difference between two documents is N or less, then their privatized versions will be indistinguishable, provided ϵ is sufficiently small. For example, if $N = 4$, then the privatized version of email that includes the sentence "I hate my boss" will be indistinguishable from that of an email without the sentence. We note that it is also natural to consider heterogeneous document-specific precision levels—i.e., N_d leading to $\epsilon_d = N_d \ln\left(\frac{1}{\alpha_d}\right)$—to enable the author of document d to choose how much to noise this document before sharing it. For example, if an author wanted to make sure that an adversary would be unable to tell that she wrote "surprise party" five times in an email, she would first set $N_d := 5 \cdot 2 = 10$. Then, to achieve $\epsilon_d = 1$, she would set $\tilde{y}_{dv}^{(\pm)} := y_{dv} + \tau_{dv}$, where $\tau_{dv} \sim$ 2Geo(α_d) and $\alpha_d = \exp\left(-\frac{\epsilon_d}{N_d}\right) \approx 0.9$.

Poisson factorization. As explained in section 3, latent Dirichlet allocation (Blei et al., 2003)—a well-known topic model for text corpora—is a special case of Poisson factorization where Y is a $D \times V$ count matrix; $n = (d, v)$, where $d \in [D]$ and $v \in [V]$; $Z = \{\Theta, \Phi\}$, where Θ and Φ are $D \times K$ and $K \times V$ non-negative, real-valued matrices, respectively; and $\mu_{dv} = \sum_{k=1}^{K} \theta_{dk}\, \phi_{kv}$. The factor θ_{dk} represents how much topic k is used in document d, while the factor ϕ_{kv} represents how much word type v is used in topic k. Again, it is standard to assume independent Gamma priors over the factors—i.e., $\theta_{dk}, \phi_{kv} \sim$ Gamma(a_0, b_0), where a_0 and b_0 are shape and rate hyperparameters, respectively.

Synthetic data. We generated a synthetic data set of $D = 90$ documents, with $K = 3$ topics and $V = 15$ word types. We set Φ^* so that the topics were well separated, with each putting the majority of its mass on five different word types. We also ensured that the documents were well separated into three equal groups of thirty, with each putting the majority of its mass on a different topic. We then sampled a data set $y^*_{dv} \sim$ Pois(μ^*_{dv}) where $\mu^*_{dv} = \sum_{k=1}^{K} \theta^*_{dk} \phi^*_{kv}$. We then generated a heterogeneously-noised data set by sampling the d^{th} document's noise level $\alpha_d \sim$ Beta$\big(c\,\alpha_0,\, c\,(1-\alpha_0)\big)$ from a Beta distribution with mean α_0 and concentration parameter $c = 10$ and then sampling $\tau_{dv} \sim$ 2Geo(α_d) for each word type v. We repeated this for a small and large value of α_0. For each model, we ran 6,000 sampling iterations, saving every 25th sample after the first 1,000. We selected $\hat{\Phi}$ to be from the posterior sample with the highest joint probability. Note that, due to label-switching, we cannot average the samples of Φ. Following Newman et al. (2009), we then aligned the topic indices of $\hat{\Phi}$ to Φ^* using the Hungarian bipartite matching algorithm. We visualize the results in figure 1 where we see that the naïve approach performs poorly at recovering the topics in the high noise case.

Enron. For these experiments we created a corpus by treating each sent email in the Enron corpus as a single document. After removing stopwords we ran latent Dirichlet allocation (LDA)—a special case of Poisson factorization—using MALLET (McCallum, 2002) with default settings to obtain a point estimate of the parameters θ^*_{dk} and ϕ^*_{kv}. For each of the most frequently used $K = 25$ topics, we sub-selected the top 50% of word types v used by the topic and top 5% documents d that use the topic. The resultant data set included $D = 1,000$ documents and $V = 1,400$ word types. As in the previous study, we used a semi-synthetic experimental design to allow us to vary the length and density of documents and control for model mismatch. In each trial, we generated a data set $y_{dv} \sim$ Pois(μ^*_{dv}) where we rescaled $\phi^*_{kv} = \gamma_v\, \phi^*_{kv}$ and $\theta^*_{dk} = \gamma_d\, \theta^*_{dk}$, allowing γ_v and γ_d to vary across trials. We randomly generated $\gamma_d, \gamma_v \sim \Gamma(e_0 f_0, f_0)$ for $f_0 = 0.001$ and varying e_0. We generated five semi-synthetic data sets for three values of $e_0 \in \{5, 10, 50\}$ and nine levels of noise $\alpha \in \{0.1, \ldots, 0.9\}$. We applied our method and the naïve approach to each data set by running 15,000 sampling iterations and saving every 25th sample after the first 3,000. Using these samples we approximated the posterior mean of $\hat{\mu}_{dv}$ and calculated the mean KL divergence of the Poisson distribution implied by $\hat{\mu}_{dv}$ from the Poisson distribution implied by μ^*_{dv}. We include a plot of these results in the appendix; they tell a similar story to those shown in figure 4.

6. Conclusion and future directions

We presented a general and modular method for privatizing Bayesian inference for Poisson factorization, a broad class of models that contains some of the most widely used models in the social sciences. Our method satisfies local differential privacy. To formulate our local-privacy guarantees, we introduced limited-precision local privacy—the local privacy analog of limited-precision differential privacy. Finally, via two case studies, we demonstrated our method's utility over a naïve approach, wherein inference proceeds as usual, treating the privatized data as if it were not privatized.

The key to our method is being able to efficiently sample values of the non-privatized data. We accomplish this by introducing auxiliary variables and exploiting special relationships between the Bessel, Skellam, and Poisson distributions to obtain a sequence of closed-form conditional distributions for every variable. A straightforward alternative to our approach is to sample from the unnormalized density $P_{\mathcal{M},\mathcal{R}}(y_n\,, \tilde{y}_n^{(\pm)} \mid \mu_n) \propto P_{\mathcal{M},\mathcal{R}}(y_n \,| \tilde{y}_n^{(\pm)},\, \mu_n)$ using black-box techniques (e.g., rejection sampling). Although this is conceptually simpler, there are many benefits of closed-formedness. Most importantly, our algorithm can be easily built on in the future to develop stochastic inference algorithms for massive data sets. When all complete conditionals are closed-form exponential families, there are simple recipes for translating MCMC algorithms into coordinate-ascent variational inference (CAVI) algorithms (Hoffman et al., 2013). For our method, all complete conditionals are well known to be exponential families, except for the Bessel distribution; however, we show that it is in theorem E below. This allows us to derive a full CAVI algorithm—we include the derivation of this algorithm in the appendix and leave for future work the development of a stochastic version that will scale to massive data sets.

Theorem 6.1. (Proof in appendix.) *The Bessel distribution $m \sim$ Bes(ν, a) for fixed ν is an exponential family with sufficient statistic $T_\nu(m) = 2m + \nu$, natural parameter $\eta_\nu(m) = \log\left(\frac{a}{2}\right)$, and base measure $h_\nu(m) = \frac{1}{m!\,\Gamma(m+\nu+1)}$.*

References

Acharya, Ayan, Ghosh, Joydeep, and Zhou, Mingyuan. Nonparametric Bayesian factor analysis for dynamic count matrices. arXiv:1512.08996, 2015.

Andrés, Miguel E., Bordenabe, Nicolás E., Chatzikokolakis, Konstantinos, and Palamidessi, Catuscia. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer &#38; Communications Security*, CCS '13, pp. 901–914, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2477-9. doi: 10.1145/2508859.2516735. URL http://doi.acm.org/10.1145/2508859.2516735.

Ball, Brian, Karrer, Brian, and Newman, Mark E. J. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3):036103, 2011.

Bernstein, Garrett, McKenna, Ryan, Sun, Tao, Sheldon, Daniel, Hay, Michael, and Miklau, Gerome. Differentially private learning of undirected graphical models using collective graphical models. *arXiv preprint arXiv:1706.04646*, 2017.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Braun, Michael and McAuliffe, Jon. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

Buntine, Wray and Jakulin, Aleks. Applying discrete PCA in data analysis. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 59–66, 2004.

Canny, John. GaP: a factor model for discrete data. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 122–129, 2004.

Cemgil, Ali Taylan. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009.

Charlin, Laurent, Ranganath, Rajesh, McInerney, James, and Blei, David M. Dynamic Poisson factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pp. 155–162, 2015.

Chi, Eric C. and Kolda, Tamara G. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299, 2012.

Devroye, Luc. Simulating Bessel random variables. *Statistics & Probability Letters*, 57(3):249–257, 2002.

Dimitrakakis, Christos, Nelson, Blaine, Mitrokotsa, Aikaterini, and Rubinstein, Benjamin I. P. Robust and private Bayesian inference. In *International Conference on Algorithmic Learning Theory*, pp. 291–305, 2014.

Dwork, Cynthia, McSherry, Frank, Nissim, Kobbi, and Smith, Adam. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*, volume 3876, pp. 265–284, 2006.

Flood, Mark, Katz, Jonathan, Ong, Stephen, and Smith, Adam. Cryptography and the economics of supervisory information: Balancing transparency and confidentiality. 2013. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2354038#.

Foulds, James, Geumlek, Joseph, Welling, Max, and Chaudhuri, Kamalika. On the theory and practice of privacy-preserving Bayesian data analysis. 2016.

Ghosh, Arpita, Roughgarden, Tim, and Sundararajan, Mukund. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012.

Gopalan, Prem K. and Blei, David M. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539, 2013.

Grimmer, Justin and Stewart, Brandon M. Text as data: The promise and pitfalls fo automatic content analysis methods for political texts. *Political Analysis*, pp. 1–31, 2013.

Hoffman, Matthew D, Blei, David M, Wang, Chong, and Paisley, John. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Johnson, Norman L., Kemp, Adrienne W., and Kotz, Samuel. *Univariate discrete distributions*. 2005.

Karwa, Vishesh, Slavković, Aleksandra B, and Krivitsky, Pavel. Differentially private exponential random graphs. In *International Conference on Privacy in Statistical Databases*, pp. 143–155. Springer, 2014.

Klimt, Bryan and Yang, Yiming. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pp. 217–226. Springer, 2004.

McCallum, Andrew Kachites. Mallet: A machine learning for language toolkit. 2002.

Mohr, John and Bogdanov, Petko (eds.). *Poetics: Topic Models and the Cultural Sciences*, volume 41. 2013.

Narayanan, Arvind and Shmatikov, Vitaly. De-anonymizing social networks. In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, pp. 173–187, 2009.

Newman, David, Asuncion, Arthur, Smyth, Padhraic, and Welling, Max. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10(Aug):1801–1828, 2009.

Paisley, John, Blei, David M., and Jordan, Michael I. Bayesian nonnegative matrix factorization with stochastic variational inference. In Airoldi, Edoardo M., Blei, David M., Erosheva, Elena A., and Fienberg, Stephen E. (eds.), *Handbook of Mixed Membership Models and Their Applications*, pp. 203–222. 2014.

Papadimitriou, Antonis, Narayan, Arjun, and Haeberlen, Andreas. Dstress: Efficient differentially private computations on distributed data. In *Proceedings of the Twelfth European Conference on Computer Systems*, EuroSys '17, pp. 560–574, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4938-3. doi: 10.1145/3064176.3064218. URL http://doi.acm.org/10.1145/3064176.3064218.

Park, Mijung, Foulds, James, Chaudhuri, Kamalika, and Welling, Max. Private topic modeling. *arXiv:1609.04120*, 2016.

Pritchard, Jonathan K., Stephens, Matthew, and Donnelly, Peter. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Ramage, Daniel, Rosen, Evan, Chuang, Jason, Manning, Christopher D., and McFarland, Daniel A. Topic modeling for the social sciences. In *NIPS Workshop on Applications for Topic Models*, 2009.

Ranganath, Rajesh, Tang, Linpeng, Charlin, Laurent, and Blei, David. Deep exponential families. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 762–771, 2015.

Roberts, Margaret E., Stewart, Brandon M., Tingley, Dustin, and Airoldi, Edoardo M. The structural topic model and applied social science. In *NIPS Workshop on Topic Models: Computation, Application, and Evaluation*, 2013.

Schein, Aaron, Paisley, John, Blei, David M., and Wallach, Hanna. Bayesian Poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1045–1054, 2015.

Schein, Aaron, Wallach, Hanna, and Zhou, Mingyuan. Poisson–gamma dynamical systems. In *Advances in Neural Information Processing Systems 29*, pp. 5005–5013, 2016a.

Schein, Aaron, Zhou, Mingyuan, Blei, David M., and Wallach, Hanna. Bayesian Poisson Tucker decomposition for learning the structure of international relations. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016b.

Schmidt, Mikkel N. and Morup, Morten. Nonparametric Bayesian modeling of complex networks: an introduction. *IEEE Signal Processing Magazine*, 30(3):110–128, 2013.

Skellam, John G. The frequency distribution of the difference between two Poisson variates belonging to different populations. *Journal of the Royal Statistical Society, Series A (General)*, 109:296, 1946.

Titsias, Michalis K. The infinite gamma–Poisson feature model. In *Advances in Neural Information Processing Systems 21*, pp. 1513–1520, 2008.

Ver Hoef, Jay M. Who invented the delta method? *The American Statistician*, 66(2):124–127, 2012.

Wang, Chong and Blei, David M. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031, 2013.

Wang, Yu-Xiang, Fienberg, Stephen, and Smola, Alex. Privacy for free: posterior sampling and stochastic gradient Monte Carlo. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2493–2502, 2015.

Warner, Stanley L. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Welling, Max and Weber, Markus. Positive tensor factorization. *Pattern Recognition Letters*, 22(12):1255–1261, 2001.

Williams, Oliver and McSherry, Frank. Probabilistic inference and differential privacy. In *Advances in Neural Information Processing Systems 23*, pp. 2451–2459, 2010.

Yang, Xiaolin, Fienberg, Stephen E, and Rinaldo, Alessandro. Differential privacy for protecting multi-dimensional contingency table data: Extensions and applications. *Journal of Privacy and Confidentiality*, 4(1):5, 2012.

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

Zhang, Zuhe, Rubinstein, Benjamin I. P., and Dimitrakakis, Christos. On the differential privacy of Bayesian inference. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 2365–2371, 2016.

Zhou, Mingyuan. Infinite edge partition models for overlapping community detection and link prediction. In *Proceedings of the 18th Conference on Artificial Intelligence and Statistics*, pp. 1135–1143, 2015.

Zhou, Mingyuan and Carin, Lawrence. Augment-and-conquer negative binomial processes. In *Advances in Neural Information Processing Systems 25*, pp. 2546–2554, 2012.

Zhou, Mingyuan, Cong, Yulai, and Chen, Bo. The Poisson gamma belief network. In *Advances in Neural Information Processing Systems 28*, pp. 3043–3051, 2015.

A. Additional figures

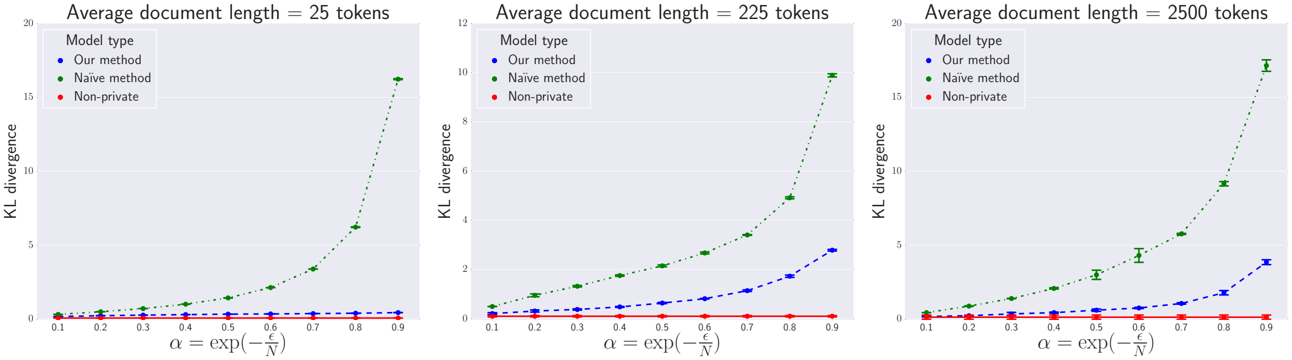


Figure 5. Mean KL divergence of the Poisson distribution implied by $\hat{\mu}_{dv}$ from the Poisson distribution implied by μ^*_{dv}; lower is better. Error bars denote standard deviation across five replications. Each subfigure reports the results across nine values of α (higher values mean more noise) for a given setting of e_0 which controls the sparsity and magnitude of the count observations. As the counts grow larger and denser, the performance of both private methods approaches the performance of non-private Poisson factorization for small values of α. Our method almost always outperforms the naïve approach, with the difference being especially stark at higher noise levels.

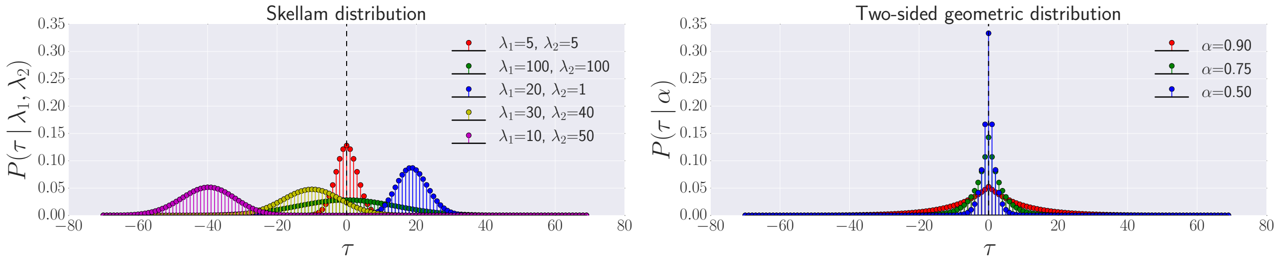


Figure 6. The two-sided geometric distribution (bottom) can be obtained by randomizing the parameters of the Skellam distribution (top). With fixed paramters, the Skellam distribution can be asymmetric and centered at a value other than zero; however, the two-sided geometric distribution is symmetric and centered at zero. It is also heavy tailed and the discrete analog of the Laplace distribution.

B. Geometric mechanism

Theorem 2.4. *If N is a positive integer and randomized response method $\mathcal{R}(\cdot)$ is the geometric mechanism with parameter α, then for any pair of observations $y, y' \in \mathcal{Y} \subseteq \mathbb{Z}^d$ such that $\|y - y'\|_1 \leq N$, $\mathcal{R}(\cdot)$ satisfies*

$$P\left(\mathcal{R}(y) \in \mathcal{S}\right) \leq e^{\epsilon} P\left(\mathcal{R}(y') \in \mathcal{S}\right) \tag{18}$$

for all subsets $\mathcal{S}$ in the range of $\mathcal{R}(\cdot)$, where

$$\epsilon = N \ln\left(\frac{1}{\alpha}\right). \tag{19}$$

Therefore, the geometric mechanism with parameter α is an (N, ϵ)-private randomized response method with $\epsilon = N \ln\left(\frac{1}{\alpha}\right)$. If a data analysis algorithm sees only the observations' (N, ϵ)-private responses, then the data analysis itself satisfies (N, ϵ)-limited precision local privacy.

Proof. It suffices to show that for any integer-valued vector $o \in \mathbb{Z}^d$, the following inequality holds for any pair of observations $y, y' \in \mathcal{Y} \subseteq \mathbb{Z}^d$ such that $\|y - y'\|_1 \leq N$:

$$\exp(-\epsilon) \leq \frac{P\left(\mathcal{R}(y) = o\right)}{P\left(\mathcal{R}(y') = o\right)} \leq \exp(\epsilon), \tag{20}$$

where $\epsilon = N \ln\left(\frac{1}{\alpha}\right)$.

Let ν denote a d-dimensional noise vector with elements drawn independently from $2\text{Geo}(\alpha)$. Then,

$$\frac{P\left(\mathcal{R}(y) = o\right)}{P\left(\mathcal{R}(y') = o\right)} = \frac{P(\nu = o - y)}{P(\nu = o - y')} \tag{21}$$

$$= \frac{\prod_{i=1}^{d} \frac{1-\alpha}{1+\alpha}\, \alpha^{|o_i - y_i|}}{\prod_{i=1}^{d} \frac{1-\alpha}{1+\alpha}\, \alpha^{|o_i - y'_i|}} \tag{22}$$

$$= \alpha^{\left(\sum_{i=1}^{d} |o_i - y_i| - |o_i - y'_i|\right)}. \tag{23}$$

By the triangle inequality, we also know that for each i,

$$-|y_i - y'_i| \leq |o_i - y_i| - |o_i - y'_i| \leq |y_i - y'_i|. \tag{24}$$

Therefore,

$$-\|y - y'\|_1 \leq \sum_{i=1}^{d} \left(|o_i - y_i| - |o_i - y'_i|\right) \leq \|y - y'\|_1. \tag{25}$$

It follows that

$$\alpha^{-N} \leq \frac{P\left(\mathcal{R}(y) = o\right)}{P\left(\mathcal{R}(y') = o\right)} \leq \alpha^{N}. \tag{26}$$

If $\epsilon = N \ln\left(\frac{1}{\alpha}\right)$, then we recover the bound in equation 20. □

C. Two-sided geometric noise as exponentially randomized Skellam noise

Theorem C.1. *A two-sided geometric random variable $\tau \sim 2Geo(\alpha)$ can be generated as follows:*

$$\lambda_1, \lambda_2 \sim \textit{Exp}\left(\tfrac{\alpha}{1-\alpha}\right), \;\; \tau \sim \textit{Skel}(\lambda_1, \lambda_2), \tag{27}$$

where the Skellam distribution is the marginal distribution over the difference $\tau := g_1 - g_2$ of two independent Poisson random variables $g_1 \sim Pois(\lambda_1)$ and $g_2 \sim Pois(\lambda_2)$.

Proof. A two-sided geometric random variable $\tau \sim 2\text{Geo}(\alpha)$ can be generated by taking the difference of two independent and identically distributed geometric random variables:[3]

$$g_1 \sim \text{Geo}(\alpha), \;\; g_2 \sim \text{Geo}(\alpha), \;\; \tau := g_1 - g_2. \tag{28}$$

The geometric distribution is a special case of the negative binomial distribution, with shape parameter equal to one (Johnson et al., 2005). Furthermore, the negative binomial distribution can be represented as a mixture of Poisson distributions with a gamma mixing distribution. We can therefore re-express equation 28 as follows:

$$\lambda_1 \sim \text{Gam}(1, \tfrac{\alpha}{1-\alpha}), \;\; \lambda_2 \sim \text{Gam}(1, \tfrac{\alpha}{1-\alpha}), \;\; g_1 \sim \text{Pois}(\lambda_1), \;\; g_2 \sim \text{Pois}(\lambda_2), \;\; \tau := g_1 - g_2. \tag{29}$$

Finally, a gamma distribution with shape parameter equal to one is an exponential distribution, while the difference of two independent Poisson random variables is marginally a Skellam random variable (Skellam, 1946). □

D. Relationship between the Bessel and Skellam distributions

Theorem D.1. *Consider two Poisson random variables $y_1 \sim Pois(\lambda_1)$ and $y_2 \sim Pois(\lambda_2)$. Their minimum $m := min\{y_1, y_2\}$ and their difference $\delta := y_1 - y_2$ are deterministic functions of y_1 and y_2. However, if not conditioned on y_1 and y_2, the random variables m and δ can be marginally generated as follows:*

$$\delta \sim \textit{Skel}(\lambda_1, \lambda_2), \;\; m \sim \textit{Bes}\left(|\delta|, 2\sqrt{2\lambda_1\lambda_2}\right). \tag{30}$$

[3] See https://www.youtube.com/watch?v=V1EyqL1cqTE.

Proof.

$$
\begin{aligned}
P(y_1, y_2) &= \text{Pois}(y_1; \lambda_1)\,\text{Pois}(y_2; \lambda_2) && (31)\\
&= \frac{\lambda_1^{y_1}}{y_1!}\,e^{-\lambda_1}\,\frac{\lambda_2^{y_2}}{y_2!}\,e^{-\lambda_2} && (32)\\
&= \frac{(\sqrt{\lambda_1\lambda_2})^{y_1+y_2}}{y_1!\,y_2!}\,e^{-(\lambda_1+\lambda_2)}\left(\frac{\lambda_1}{\lambda_2}\right)^{(y_1-y_2)\,/\,2}. && (33)
\end{aligned}
$$

If $y_1 \geq y_2$, then

$$
\begin{aligned}
P(y_1, y_2) &= \frac{(\sqrt{\lambda_1\lambda_2})^{y_1+y_2}}{I_{y_1-y_2}(2\sqrt{\lambda_1\lambda_2})\,y_1!\,y_2!}\,e^{-(\lambda_1+\lambda_2)}\left(\frac{\lambda_1}{\lambda_2}\right)^{(y_1-y_2)\,/\,2} I_{y_1-y_2}(2\sqrt{\lambda_1\lambda_2}) && (34)\\
&= \text{Bes}\left(y_2;\, y_1-y_2, 2\sqrt{\lambda_1\lambda_2}\right)\text{Skel}(y_1-y_2; \lambda_1, \lambda_2); && (35)
\end{aligned}
$$

otherwise

$$
\begin{aligned}
P(y_1, y_2) &= \frac{(\sqrt{\lambda_1\lambda_2})^{y_1+y_2}}{I_{y_2-y_1}(2\sqrt{\lambda_1\lambda_2})\,y_1!\,y_2!}\,e^{-(\lambda_1+\lambda_2)}\left(\frac{\lambda_2}{\lambda_1}\right)^{(y_2-y_1)\,/\,2} I_{y_2-y_1}(2\sqrt{\lambda_1\lambda_2}) && (36)\\
&= \text{Bes}\left(y_1;\, y_2-y_1, 2\sqrt{\lambda_1\lambda_2}\right)\text{Skel}(y_2-y_1; \lambda_2, \lambda_1)\\
&= \text{Bes}\left(y_1; -(y_1-y_2), 2\sqrt{\lambda_1\lambda_2}\right)\text{Skel}(y_1-y_2; \lambda_1, \lambda_2). && (37)
\end{aligned}
$$

If

$$m := \min\{y_1, y_2\}, \;\; \delta := y_1 - y_2, \tag{38}$$

then

$$
\begin{aligned}
y_2 = m, \;\; y_1 = m + \delta \;\; &\text{if } \delta \geq 0 && (39)\\
y_1 = m, \;\; y_2 = m - \delta \;\; &\text{otherwise} && (40)
\end{aligned}
$$

and

$$\begin{vmatrix} \frac{\partial y_1}{\partial m} & \frac{\partial y_1}{\partial \delta} \\ \frac{\partial y_2}{\partial m} & \frac{\partial y_2}{\partial \delta} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & 0 \end{vmatrix}^{\delta \geq 0} \begin{vmatrix} 1 & 0 \\ 1 & -1 \end{vmatrix}^{\delta < 0} = 1, \tag{41}$$

so

$$
\begin{aligned}
P(m, \delta) &= P(y_1, y_2)\begin{vmatrix} \frac{\partial y_1}{\partial m} & \frac{\partial y_1}{\partial \delta} \\ \frac{\partial y_2}{\partial m} & \frac{\partial y_2}{\partial \delta} \end{vmatrix}\\
&= \text{Bes}\left(m; |\delta|, 2\sqrt{\lambda_1\lambda_2}\right)\text{Skel}(\delta; \lambda_1, \lambda_2). && (42)
\end{aligned}
$$

□

E. Coordinate-ascent variational inference

Theorem 6.1. *The Bessel distribution $m \sim Bes(\nu, a)$ for fixed ν is an exponential family with sufficient statistic $T_\nu(m) = 2m+\nu$, natural parameter $\eta_\nu(m) = \log\left(\frac{a}{2}\right)$, and base measure $h_\nu(m) = \frac{1}{m!\,\Gamma(m+\nu+1)}$.*

Proof. The Bessel distribution (Yuan & Kalbfleisch, 2000) is a two-parameter distribution over the non-negative integers:

$$f(n; a, \nu) = \frac{\left(\frac{a}{2}\right)^{2n+\nu}}{n!\,\Gamma(n+\nu+1) I_\nu(a)}, \tag{43}$$

where the normalizing constant $I_\nu(a)$ is a modified Bessel function of the first kind—i.e.,

$$I_\nu(a) = \sum_{n=0}^{\infty} \frac{\left(\frac{a}{2}\right)^{2n+\nu}}{n!\,\Gamma(n+\nu+1)}. \tag{44}$$

For fixed and known ν, we can rewrite the Bessel PMF as

$$f(n;a,\nu) = \frac{1}{n!\,\Gamma(n{+}\nu{+}1)} \exp\left((2n{+}\nu)\,\log\left(\frac{a}{2}\right) - \log I_\nu(a)\right). \tag{45}$$

We can then define the following functions:

$$h_\nu(n) = \frac{1}{n!\,\Gamma(n{+}\nu{+}1)} \tag{46}$$

$$T_\nu(n) = 2n + \nu \tag{47}$$

$$\eta_\nu(a) = \log\left(\frac{a}{2}\right) \tag{48}$$

$$A_\nu(a) = \log I_\nu(a). \tag{49}$$

Finally, we can rewrite the Bessel PMF in the exponential-family form:

$$f(n;a,\nu) = h_\nu(n)\exp\left(\eta_\nu(a)\cdot T_\nu(n) - A_\nu(a)\right). \tag{50}$$

□

We can derive a full coordinate-ascent variational inference (CAVI) algorithm for locally private Poisson factorization. For exposition, we focus on latent Dirichlet allocation, where $y_{dv} \sim \text{Pois}(\mu_{dv})$ and $\mu_{dv} = \sum_{k=1}^K \theta_{dk}\phi_{kv}$. It is standard to assume independent gamma priors over the factors—i.e., $\theta_{dk}, \phi_{kv} \sim \Gamma(a_0, b_0)$. We use $\mathbb{G}\left[X\right] = \exp\left(\mathbb{E}\left[\ln X\right]\right)$ to denote the geometric expected value of X.

E.1. $Q(\theta_{dk})$, $Q(\phi_{kv})$

The optimal variational distribution for the factors is same as in non-private Poisson factorization (Cemgil, 2009).

$$\begin{aligned}
Q(\theta_{dk}) &\propto \mathbb{G}_Q\left[P(y_{dk}, \theta_{dk} \,|\, -)\right] && (51)\\
&= \mathbb{G}_Q\left[\text{Pois}\left(y_{dk};\, \theta_{dk}\sum_{v=1}^V \phi_{kv}\right)\,\Gamma\left(\theta_{dk};\, a_0, b_0\right)\right] && (52)\\
&\propto \Gamma(\theta_{dk};\, \alpha_{dk},\, \beta_{dk}) && (53)\\
\alpha_{dk} &:= a_0 + \mathbb{E}_Q\left[y_{dv}\right] && (54)\\
\beta_{dk} &:= b_0 + \sum_{v=1}^V \mathbb{E}_Q\left[\phi_{kv}\right] && (55)\\
\mathbb{E}_Q\left[\theta_{dk}\right] &= \frac{\alpha_{dk}}{\beta_{dk}} && (56)\\
\mathbb{G}_Q\left[\theta_{dk}\right] &= \exp\left(\psi(\alpha_{dk}) - \ln(\beta_{dk})\right). && (57)
\end{aligned}$$

The derivation for ϕ_{kv} is analogous.

E.2. $Q\left(m_{dv}, \tilde{y}_{dv}^{(+)}, g_{dv1}, g_{dv2}, y_{dv}, (y_{dvk})_{k=1}^K\right)$

The relationships between the different count variables are as follows:

$$\underbrace{\underbrace{\underbrace{\left(\sum_{k=1}^K y_{dvk}\right)}_{=y_{dv}} + g_{dv1}}_{=\tilde{y}_{dv}^{(+)}} - g_{dv2}}_{=\tilde{y}_{dv}^{(\pm)}} \tag{58}$$

$$m_{dv} = \min\left\{\tilde{y}_{dv}^{(+)},\, g_{dv2}\right\} \tag{59}$$

We have a single variational distribution for these variables.

E.3. Different factorizations of the joint

To find this variational distribution, we consider the variables' joint distribution:

$$P\left(g_{dv1}, g_{dv2}, \left(y_{dvk}\right)_{k=1}^{K}, y_{dv}, \tilde{y}^{(+)}_{dv}, \tilde{y}^{(\pm)}_{dv}, m_{dv}\right). \tag{60}$$

The most straightforward factorization of this joint first generates all the Poisson random variables and then computes the remaining variables given their deterministic relationships to the underlying Poissons:

$$\begin{aligned} &P\left(g_{dv1}, g_{dv2}, \left(y_{dvk}\right)_{k=1}^{K}, y_{dv}, \tilde{y}^{(+)}_{dv}, \tilde{y}^{(\pm)}_{dv}, m_{dv}\right) \\ &\quad = \text{Pois}(g_{dv1}; \lambda_{dv1})\, \text{Pois}(g_{dv2}; \lambda_{dv2}) \left(\prod_{k=1}^{K} \text{Pois}(y_{dvk}; \theta_{dk}\phi_{kv})\right) \mathbb{1}\left(y_{dv} = \sum_{k=1}^{K} y_{dvk}\right) \\ &\quad\quad \mathbb{1}\left(\tilde{y}^{(+)}_{dv} = y_{dv} + g_{dv1}\right) \mathbb{1}\left(\tilde{y}^{(\pm)}_{dv} = \tilde{y}^{(+)}_{dv} - g_{dv2}\right) \mathbb{1}\left(m_{dv} = \min\{\tilde{y}^{(+)}_{dv}, g_{dv2}\}\right). \end{aligned} \tag{61}$$

We can equivalently first generate the sums of Poissons and then thin them using multinomial and binomial draws. In the following equation, the delta functions are implicitly present in the multinomial and binomial PMFs. Note that we write the probability parameters in the multinomial and binomial PMFs as unnormalized vectors. Also note that $\mu_{dv} = \sum_k \theta_{dk}\phi_{kv}$.

$$\begin{aligned} &P\left(g_{dv1}, g_{dv2}, \left(y_{dvk}\right)_{k=1}^{K}, y_{dv}, \tilde{y}^{(+)}_{dv}, \tilde{y}^{(\pm)}_{dv}, m_{dv}\right) \\ &\quad = \text{Pois}(g_{dv2}; \lambda_{dv2})\, \text{Pois}\left(\tilde{y}^{(+)}_{dv}; \lambda_{dv1} + \mu_{dv}\right) \\ &\quad\quad \text{Binom}\left((y_{dv}, g_{dv1});\, \tilde{y}^{(+)}_{dv}, (\mu_{dv}, \lambda_{dv1})\right) \text{Mult}\left(\left(y_{dvk}\right)_{k=1}^{K};\, y_{dv}, \left(\theta_{dk}\phi_{kv}\right)_{k=1}^{K}\right) \\ &\quad\quad \mathbb{1}\left(\tilde{y}^{(\pm)}_{dv} = \tilde{y}^{(+)}_{dv} - g_{dv2}\right) \mathbb{1}\left(m_{dv} = \min\{\tilde{y}^{(+)}_{dv}, g_{dv2}\}\right). \end{aligned} \tag{62}$$

We can equivalently first generate the difference $\tilde{y}^{(\pm)}_{dv}$ and minimum m_{dv} as Skellam and Bessel random variables. Conditioned on these variables, we can then compute $\tilde{y}^{(+)}_{dv}$ and g_{dv2} via their deterministic relationship and, finally, thin $\tilde{y}^{(+)}_{dv}$ using binomial and multinomial draws:

$$\begin{aligned} &P\left(g_{dv1}, g_{dv2}, \left(y_{dvk}\right)_{k=1}^{K}, y_{dv}, \tilde{y}^{(+)}_{dv}, \tilde{y}^{(\pm)}_{dv}, m_{dv}\right) \\ &\quad = \text{Skel}\left(\tilde{y}^{(\pm)}_{dv};\, \lambda_{dv1} + \mu_{dv},\, \lambda_{dv2}\right) \text{Bes}\left(m_{dv};\, |\tilde{y}^{(\pm)}_{dv}|,\, 2\sqrt{\lambda_{dv2}(\lambda_{dv1} + \mu_{dv})}\right) \\ &\quad\quad \mathbb{1}\left(\tilde{y}^{(+)}_{dv} = m_{dv}\right)^{\mathbb{1}(\tilde{y}^{(\pm)}_{dv} \leq 0)} \mathbb{1}\left(g_{dv2} = m_{dv}\right)^{\mathbb{1}(\tilde{y}^{(\pm)}_{dv} > 0)} \mathbb{1}\left(\tilde{y}^{(\pm)}_{dv} = \tilde{y}^{(+)}_{dv} - g_{dv2}\right) \\ &\quad\quad \text{Binom}\left((y_{dv}, g_{dv1});\, \tilde{y}^{(+)}_{dv}, (\mu_{dv}, \lambda_{dv1})\right) \text{Mult}\left(\left(y_{dvk}\right)_{k=1}^{K};\, y_{dv}, \left(\theta_{dk}\phi_{kv}\right)_{k=1}^{K}\right). \end{aligned} \tag{63}$$

It is this last factorization that enables us to derive the variational distribution.

E.4. Deriving the variational distribution

Recall that the observed data consists of the difference variables $\tilde{y}^{(\pm)}_{dv}$.

$$Q\left(m_{dv}, \tilde{y}^{(+)}_{dv}, g_{dv1}, g_{dv2}, y_{dv}, \left(y_{dvk}\right)_{k=1}^{K}\right) \propto \mathbb{G}_Q\left[P(\tilde{y}^{(\pm)}_{dv}, m_{dv}, \tilde{y}^{(+)}_{dv}, g_{dv1}, g_{dv2}, y_{dv}, \left(y_{dvk}\right)_{k=1}^{K} \mid -)\right]. \tag{64}$$

Because the likelihood (i.e., the Skellam term) in equation 63 does not depend on any of these latent variables, it disappears entirely. We can then rewrite the right-hand side of equation 64 as:

$$\begin{aligned}
&\mathbb{G}_Q\left[P(m_{dv}, \tilde{y}^{(+)}_{dv}, g_{dv1}, g_{dv2}, y_{dv}, (y_{dvk})_{k=1}^K \mid \tilde{y}^{(\pm)}_{dv} -)\right] \\
&\quad = \mathbb{G}_Q\left[\text{Bes}\left(m_{dv};\, |\tilde{y}^{(\pm)}_{dv}|,\, 2\sqrt{\lambda_{dv2}(\lambda_{dv1}+\mu_{dv})}\right)\right] \\
&\quad\quad \mathbb{1}\left(\tilde{y}^{(+)}_{dv} = m_{dv}\right)^{\mathbb{1}(\tilde{y}^{(\pm)}_{dv} \leq 0)} \mathbb{1}\left(g_{dv2} = m_{dv}\right)^{\mathbb{1}(\tilde{y}^{(\pm)}_{dv} > 0)} \mathbb{1}\left(\tilde{y}^{(\pm)}_{dv} = \tilde{y}^{(+)}_{dv} - g_{dv2}\right) \\
&\quad\quad \mathbb{G}_Q\left[\text{Binom}\left((y_{dv}, g_{dv1});\, \tilde{y}^{(+)}_{dv}, (\mu_{dv}, \lambda_{dv1})\right) \text{Mult}\left((y_{dvk})_{k=1}^K\,;\, y_{dv}, \left(\theta_{dk}\phi_{kv}\right)_{k=1}^K\right)\right].
\end{aligned} \tag{65}$$

Theorem 6.1 states that the Bessel distribution for fixed first parameter is an exponential family. We can therefore use standard results to push in the geometric expectations:

$$\begin{aligned}
&\mathbb{G}_Q\left[P(m_{dv}, \tilde{y}^{(+)}_{dv}, g_{dv1}, g_{dv2}, y_{dv}, (y_{dvk})_{k=1}^K \mid \tilde{y}^{(\pm)}_{dv} -)\right] \\
&\quad = \text{Bes}\left(m_{dv};\, |\tilde{y}^{(\pm)}_{dv}|,\, 2\sqrt{\mathbb{G}_Q\left[\lambda_{dv2}\right]\left(\mathbb{G}_Q\left[\lambda_{dv1}+\mu_{dv}\right]\right)}\right) \\
&\quad\quad \mathbb{1}\left(\tilde{y}^{(+)}_{dv} = m_{dv}\right)^{\mathbb{1}(\tilde{y}^{(\pm)}_{dv} \leq 0)} \mathbb{1}\left(g_{dv2} = m_{dv}\right)^{\mathbb{1}(\tilde{y}^{(\pm)}_{dv} > 0)} \mathbb{1}\left(\tilde{y}^{(\pm)}_{dv} = \tilde{y}^{(+)}_{dv} - g_{dv2}\right) \\
&\quad\quad \text{Binom}\left((y_{dv}, g_{dv1});\, \mathbb{E}_Q\left[\tilde{y}^{(+)}_{dv}\right], \left(\mathbb{G}_Q\left[\mu_{dv}\right], \mathbb{G}_Q\left[\lambda_{dv1}\right]\right)\right) \\
&\quad\quad \text{Mult}\left((y_{dvk})_{k=1}^K\,;\, \mathbb{E}_Q\left[y_{dv}\right], \left(\mathbb{G}_Q\left[\theta_{dk}\phi_{kv}\right]\right)_{k=1}^K\right).
\end{aligned} \tag{66}$$

There are two expectations that do not have an analytic form:

$$\mathbb{G}_Q\left[\lambda_{dv1}+\mu_{dv}\right] = \exp\left(\mathbb{E}_Q\left[\ln\left(\lambda_{dv1} + \sum_{k=1}^K \theta_{dk}\phi_{kv}\right)\right]\right) \tag{67}$$

and

$$\mathbb{G}_Q\left[\mu_{dv}\right] = \exp\left(\mathbb{E}_Q\left[\ln\left(\sum_{k=1}^K \theta_{dk}\phi_{kv}\right)\right]\right); \tag{68}$$

however, both can be very closely approximated using the delta method (Ver Hoef, 2012) which has been previously used in variational inference schemes to approximate intractable expectations (Braun & McAuliffe, 2010; Wang & Blei, 2013). In particular, for some variable $Y = f(X)$, expectation $\mathbb{E}[Y]$ is approximately:

$$\mathbb{E}[Y] = \mathbb{E}[f(X)] \approx f\left(\mathbb{E}[X]\right) + \frac{1}{2}\, f''\left(\mathbb{E}[X]\right)\mathbb{V}[X]. \tag{69}$$

On our case, we therefore have

$$\mathbb{E}_Q\left[\ln\mu_{dv}\right] = \mathbb{E}_Q\left[\ln\left(\sum_{k=1}^K \theta_{dk}\phi_{kv}\right)\right] \approx \ln\left(\mathbb{E}_Q\left[\sum_{k=1}^K \theta_{dk}\phi_{kv}\right]\right) - \frac{\mathbb{V}_Q\left[\sum_{k=1}^K \theta_{dk}\phi_{kv}\right]}{2\left(\mathbb{E}_Q\left[\sum_{k=1}^K \theta_{dk}\phi_{kv}\right]\right)^2} \tag{70}$$

$$= \ln\left(\sum_{k=1}^K \mathbb{E}_Q\left[\theta_{dk}\right]\mathbb{E}_Q\left[\phi_{kv}\right]\right) - \frac{\sum_{k=1}^K \mathbb{V}_Q\left[\theta_{dk}\phi_{kv}\right]}{2\left(\sum_{k=1}^K \mathbb{E}_Q\left[\theta_{dk}\right]\mathbb{E}_Q\left[\phi_{kv}\right]\right)^2}. \tag{71}$$

Finally, because θ_{dk} and ϕ_{kv} are independent, we have

$$\mathbb{V}_Q\left[\theta_{dk}\phi_{kv}\right] = \mathbb{V}_Q\left[\theta_{dk}\right]\mathbb{V}_Q\left[\phi_{kv}\right] + \mathbb{V}_Q\left[\theta_{dk}\right]\left(\mathbb{E}_Q\left[\phi_{kv}\right]\right)^2 + \mathbb{V}_Q\left[\phi_{kv}\right]\left(\mathbb{E}_Q\left[\theta_{dk}\right]\right)^2. \tag{72}$$