On the Theory and Practice of Privacy-Preserving Bayesian Data Analysis

James Foulds
Calit2 & CSE Department
UC San Diego
jfoulds@ucsd.edu

Joseph Geumlek
CSE Department
UC San Diego
jgeumlek@cs.ucsd.edu

Max Welling
Informatics Institute & QUVA Lab
University of Amsterdam
m.welling@uva.nl

Kamalika Chaudhuri
CSE Department
UC San Diego
kamalika@cs.ucsd.edu

Abstract

Bayesian inference has great promise for the privacy-preserving analysis of sensitive data, as posterior sampling automatically preserves differential privacy, an algorithmic notion of data privacy, under certain conditions (Dimitrakakis et al., 2014; Wang et al., 2015b). While this *one posterior sample* (OPS) approach elegantly provides privacy "for free," it is data inefficient in the sense of asymptotic relative efficiency (ARE). We show that a simple alternative based on the Laplace mechanism, the workhorse of differential privacy, is as asymptotically efficient as non-private posterior inference, under general assumptions. This technique also has practical advantages including efficient use of the privacy budget for MCMC. We demonstrate the practicality of our approach on a time-series analysis of sensitive military records from the Afghanistan and Iraq wars disclosed by the Wikileaks organization.

1 INTRODUCTION

Probabilistic models trained via Bayesian inference are widely and successfully used in application domains where privacy is invaluable, from text analysis (Blei et al., 2003; Goldwater and Griffiths, 2007), to personalization (Salakhutdinov and Mnih, 2008), to medical informatics (Husmeier et al., 2006), to MOOCs (Piech et al., 2013). In these applications, data scientists must carefully balance the benefits and potential insights from data analysis against the privacy concerns of the individuals whose data are being studied (Daries et al., 2014).

Dwork et al. (2006) placed the notion of privacy-preserving data analysis on a solid foundation by introducing *differential privacy* (Dwork and Roth, 2013), an algorithmic formulation of privacy which is a gold standard for privacy-preserving data-driven algorithms. Differential privacy measures the privacy "cost" of an algorithm. When designing privacy-preserving methods, the goal is to achieve a good trade-off between privacy and utility, which ideally improves with the amount of available data.

As observed by Dimitrakakis et al. (2014) and Wang et al. (2015b), Bayesian posterior sampling behaves synergistically with differential privacy because it automatically provides a degree of differential privacy under certain conditions. However, there are substantial gaps between this elegant theory and the practical reality of Bayesian data analysis. Privacy-preserving posterior sampling is hampered by data inefficiency, as measured by asymptotic relative efficiency (ARE). In practice, it generally requires artificially selected constraints on the spaces of parameters as well as data points. Its privacy properties are also not typically guaranteed for approximate inference.

This paper identifies these gaps between theory and practice, and begins to mend them via an extremely simple alternative technique based on the workhorse of differential privacy, the Laplace mechanism (Dwork et al., 2006). Our approach is equivalent to a generalization of Zhang et al. (2016)'s recently and independently proposed algorithm for beta-Bernoulli systems. We provide a theoretical analysis and empirical validation of the advantages of the proposed method. We extend both our method and Dimitrakakis et al. (2014); Wang et al. (2015b)'s *one posterior sample (OPS)* method to the case of approximate inference with privacy-preserving MCMC. Finally, we demonstrate the practical applicability of this technique by showing how to use a privacy-preserving HMM model to analyze sensitive military records from the Iraq and Afghanistan wars leaked by the Wikileaks organization. Our primary contributions are as follows:

- We analyze the privacy cost of posterior sampling for exponential family posteriors via OPS.
- We explore a simple Laplace mechanism alternative to OPS for exponential families.

- Under weak conditions we establish the consistency of the Laplace mechanism approach and its data efficiency advantages over OPS.
- We extend the OPS and Laplace mechanism methods to approximate inference via MCMC.
- We demonstrate the practical implications with a case study on sensitive military records.

2 BACKGROUND

We begin by discussing preliminaries on differential privacy and its application to Bayesian inference. Our novel contributions will begin in Section 3.1.

2.1 DIFFERENTIAL PRIVACY

Differential privacy is a formal notion of the privacy of data-driven algorithms. For an algorithm to be differentially private the probabilities of the outputs of the algorithms may not change much when one individual's data point is modified, thereby revealing little information about any one individual's data. More precisely, a randomized algorithm $\mathcal{M}(\mathbf{X})$ is said to be (ϵ, δ)-differentially private if

$$Pr(\mathcal{M}(\mathbf{X}) \in \mathcal{S}) \leq \exp(\epsilon) Pr(\mathcal{M}(\mathbf{X}') \in \mathcal{S}) + \delta \quad (1)$$

for all measurable subsets $\mathcal{S}$ of the range of $\mathcal{M}$ and for all datasets $\mathbf{X}, \mathbf{X}'$ differing by a single entry (Dwork and Roth, 2013). If $\delta = 0$, the algorithm is said to be ϵ-differentially private.

2.1.1 The Laplace Mechanism

One straightforward method for obtaining ϵ-differential privacy, known as the *Laplace mechanism* (Dwork et al., 2006), adds Laplace noise to the revealed information, where the amount of noise depends on ϵ, and a quantifiable notion of the sensitivity to changes in the database. Specifically, the $L1$ sensitivity $\triangle h$ for function h is defined as

$$\triangle h = \max_{\mathbf{X}, \mathbf{X}'} \|h(\mathbf{X}) - h(\mathbf{X}')\|_1 \quad (2)$$

for all datasets $\mathbf{X}, \mathbf{X}'$ differing in at most one element. The Laplace mechanism adds noise via

$$\mathcal{M}_L(\mathbf{X}, h, \epsilon) = h(\mathbf{X}) + (Y_1, Y_2, \ldots, Y_d)\,, \quad (3)$$
$$Y_j \sim \text{Laplace}(\triangle h/\epsilon), \forall j \in \{1, 2, \ldots, d\}\,,$$

where d is the dimensionality of the range of h. The $\mathcal{M}_L(\mathbf{X}, h, \epsilon)$ mechanism is ϵ-differentially private.

2.1.2 The Exponential Mechanism

The exponential mechanism (McSherry and Talwar, 2007) aims to output responses of high utility while maintaining privacy. Given a utility function $u(\mathbf{X}, \mathbf{r})$ that maps database $\mathbf{X}$/output $\mathbf{r}$ pairs to a real-valued score, the exponential mechanism $\mathcal{M}_E(\mathbf{X}, u, \epsilon)$ produces random outputs via

$$Pr(\mathcal{M}_E(\mathbf{X}, u, \epsilon) = \mathbf{r}) \propto \exp\left(\frac{\epsilon u(\mathbf{X}, \mathbf{r})}{2\triangle u}\right), \quad (4)$$

where the sensitivity of the utility function is

$$\triangle u \triangleq \max_{r, (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})} \|u(\mathbf{X}^{(1)}, r) - u(\mathbf{X}^{(2)}, r)\|_1\,, \quad (5)$$

in which $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ are pairs of databases that differ in only one element.

2.1.3 Composition Theorems

A key property of differential privacy is that it holds under composition, via an additive accumulation.

Theorem 1. *If $\mathcal{M}_1$ is (ϵ_1, δ_1)-differentially private, and $\mathcal{M}_2$ is (ϵ_2, δ_2)-differentially private, then $\mathcal{M}_{1,2}(\mathbf{X}) = (\mathcal{M}_1(\mathbf{X}), \mathcal{M}_2(\mathbf{X}))$ is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-differentially private.*

This allows us to view the total ϵ and δ of our procedure as a privacy "budget" that we spend across the operations of our analysis. There also exists an "advanced composition" theorem which provides privacy guarantees in an adversarial adaptive scenario called k-fold composition, and also allows an analyst to trade an increased δ for a smaller ϵ in this scenario (Dwork et al., 2010). Differential privacy is also immune to data-independent post-processing.

2.2 PRIVACY AND BAYESIAN INFERENCE

Suppose we would like a differentially private draw of parameters and latent variables of interest θ from the posterior $Pr(\theta|\mathbf{X})$, where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is the private dataset. We can accomplish this by interpreting posterior sampling as an instance of the exponential mechanism with utility function $u(\mathbf{X}, \theta) = \log Pr(\theta, \mathbf{X})$, i.e. the log joint probability of the chosen θ assignment and the dataset $\mathbf{X}$ (Wang et al., 2015b). We then draw θ via

$$f(\theta; \mathbf{X}, \epsilon) \propto \exp\left(\frac{\epsilon \log Pr(\theta, \mathbf{X})}{2\triangle \log Pr(\theta, \mathbf{X})}\right) = Pr(\theta, \mathbf{X})^{\frac{\epsilon}{2\triangle \log Pr(\theta, \mathbf{X})}} \quad (6)$$

where the sensitivity is $\triangle \log Pr(\theta, \mathbf{X}) \triangleq$

$$\max_{\theta, (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})} \|\log Pr(\theta, \mathbf{X}^{(1)}) - \log Pr(\theta, \mathbf{X}^{(2)})\|_1 \quad (7)$$

in which $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ differ in one element. If the data points are conditionally independent given θ,

$$\log Pr(\theta, \mathbf{X}) = \log Pr(\theta) + \sum_{i=1}^{N} \log Pr(\mathbf{x}_i|\theta)\,, \quad (8)$$

where $Pr(\theta)$ is the prior and $Pr(\mathbf{x}_i|\theta)$ is the likelihood term for data point $\mathbf{x}_i$. Since the prior does not depend

on the data, and each data point is associated with a single log-likelihood term $\log Pr(\mathbf{x}_i|\theta)$ in $\log Pr(\theta, \mathbf{X})$, from the above two equations we have

$$\triangle \log Pr(\theta, \mathbf{X}) = \max_{\mathbf{x},\mathbf{x}',\theta} |\log Pr(\mathbf{x}'|\theta) - \log Pr(\mathbf{x}|\theta)| . \tag{9}$$

This gives us the privacy cost of posterior sampling:

Theorem 2. *If* $\max_{\mathbf{x},\mathbf{x}'\in\chi,\theta\in\Theta} |\log Pr(\mathbf{x}'|\theta) - \log Pr(\mathbf{x}|\theta)| \leq C$*, releasing one sample from the posterior distribution* $Pr(\theta|\mathbf{X})$ *with any prior is* $2C$*-differentially private.*

Wang et al. (2015b) derived this form of the result from first principles, while noting that the exponential mechanism can be used, as we do here. Although they do not explicitly state the theorem, they implicitly use it to show two noteworthy special cases, referred to as the *One Posterior Sample (OPS)* procedure. We state the first of these cases:

Theorem 3. *If* $\max_{\mathbf{x}\in\chi,\theta\in\Theta} |\log Pr(\mathbf{x}|\theta)| \leq B$*, releasing one sample from the posterior distribution* $Pr(\theta|\mathbf{X})$ *with any prior is* $4B$*-differentially private.*

This follows directly from Theorem 2, since if $|\log Pr(\mathbf{x}|\theta)| \leq B$, $C = \triangle \log Pr(\theta, \mathbf{X}) = 2B$.

Under the exponential mechanism, ϵ provides an adjustable knob trading between privacy and fidelity. When $\epsilon = 0$, the procedure samples from a uniform distribution, giving away no information about $\mathbf{X}$. When $\epsilon = 2\triangle \log Pr(\theta, \mathbf{X})$, the procedure reduces to sampling θ from the posterior $Pr(\theta|\mathbf{X}) \propto Pr(\theta, \mathbf{X})$. As ϵ approaches infinity the procedure becomes increasingly likely to sample the θ assignment with the highest posterior probability. Assuming that our goal is to sample rather than to find a mode, we would cap ϵ at $2\triangle \log Pr(\theta, \mathbf{X})$ in the above procedure in order to correctly sample from the true posterior. More generally, if our privacy budget is ϵ', and $\epsilon' \geq 2q\triangle \log Pr(\theta, \mathbf{X})$, for integer q, we can draw q posterior samples within our budget.

As observed by Huang and Kannan (2012), the exponential mechanism can be understood via statistical mechanics. We can write it as a Boltzmann distribution (a.k.a. a Gibbs measure)

$$f(\theta; \mathbf{x}, \epsilon) \propto \exp\left(\frac{-E(\theta)}{T}\right), T = \frac{2\triangle u(\mathbf{X}, \theta)}{\epsilon} , \tag{10}$$

where $E(\theta) = -u(\mathbf{X}, \theta) = -\log Pr(\theta, \mathbf{X})$ is the energy of state θ in a physical system, and T is the temperature of the system (in units such that Boltzmann's constant is one). Reducing ϵ corresponds to increasing the temperature, which can be understood as altering the distribution such that a Markov chain moves through the state space more rapidly.

3 PRIVACY FOR EXPONENTIAL FAMILIES: EXPONENTIAL VS LAPLACE

By analyzing the privacy cost of sampling from exponential family posteriors in the general case we can recover the privacy properties of many standard distributions. These results can be applied to full posterior sampling, when feasible, or to Gibbs sampling updates, as we discuss in Section 4. In this section we analyze the privacy cost of sampling from exponential family posterior distributions exactly (or at an appropriate temperature) via the exponential mechanism, following Dimitrakakis et al. (2014) and Wang et al. (2015b), and via a method based on the Laplace mechanism, which is a generalization of Zhang et al. (2016). The properties of the two methods are compared in Table 1.

3.1 THE EXPONENTIAL MECHANISM

Consider exponential family models with likelihood

$$Pr(\mathbf{x}|\theta) = h(\mathbf{x})g(\theta)\exp\left(\theta^\intercal S(\mathbf{x})\right),$$

where $S(\mathbf{x})$ is a vector of sufficient statistics for data point $\mathbf{x}$, and θ is a vector of natural parameters. For N i.i.d. data points, we have

$$Pr(\mathbf{X}|\theta) = \Big(\prod_{i=1}^{N} h(\mathbf{x}^{(i)})\Big) g(\theta)^N \exp\left(\theta^\intercal \sum_{i=1}^{N} S(\mathbf{x}^{(i)})\right).$$

Further suppose that we have a conjugate prior which is also an exponential family distribution,

$$Pr(\theta|\chi, \alpha) = f(\chi, \alpha)g(\theta)^\alpha \exp\left(\alpha\theta^\intercal \chi\right),$$

where α is a scalar, the number of prior "pseudo-counts," and χ is a parameter vector. The posterior is proportional to the prior times the likelihood,

$$Pr(\theta|\mathbf{X}, \chi, \alpha) \propto g(\theta)^{N+\alpha} \exp\Big(\theta^\intercal \big(\sum_{i=1}^{N} S(\mathbf{x}^{(i)}) + \alpha\chi\big)\Big). \tag{11}$$

To compute the sensitivity of the posterior, we have

$$\begin{aligned} &|\log Pr(\mathbf{x}'|\theta) - \log Pr(\mathbf{x}|\theta)| \qquad (12)\\ &\quad = |\theta^\intercal \Big(S(\mathbf{x}') - S(\mathbf{x})\Big) + \log h(\mathbf{x}') - \log h(\mathbf{x})| . \end{aligned}$$

From Equation 9, we obtain $\triangle \log Pr(\theta, \mathbf{X}) =$

$$\sup_{\mathbf{x},\mathbf{x}'\in\chi,\theta\in\Theta} |\theta^\intercal \Big(S(\mathbf{x}') - S(\mathbf{x})\Big) + \log h(\mathbf{x}') - \log h(\mathbf{x})| . \tag{13}$$

A posterior sample at temperature T,

$$\begin{aligned} Pr_T(\theta|\mathbf{X}, \chi, \alpha) \propto g(\theta)^{\frac{N+\alpha}{T}} \exp\left(\theta^\intercal \frac{\sum_{i=1}^{N} S(\mathbf{x}^{(i)}) + \alpha\chi}{T}\right),\\ T = \frac{2\triangle \log p(\theta, X)}{\epsilon} , \end{aligned} \tag{14}$$

Mechanism	Sensitivity	$S(\mathbf{X})$ is	Release	ARE	Pay Gibbs cost
Laplace	$\sup_{\mathbf{X},\mathbf{X}'} \| \sum_{i=1}^{N} S(\mathbf{x}'^{(i)}) - \sum_{i=1}^{N} S(\mathbf{x}^{(i)}) \|_1$	Noised	Statistics	1	Once
Exponential (OPS)	$\sup_{\mathbf{x},\mathbf{x}' \in \chi, \theta \in \Theta} \|\theta^\intercal \Big(S(\mathbf{x}') - S(\mathbf{x}) \Big) + \log h(\mathbf{x}') - \log h(\mathbf{x})\|$	Rescaled	One Sample	$1 + T$	Per update (unless converged)

Table 1: Comparison of the properties of the two methods for private Bayesian inference.

has privacy cost ϵ, by the exponential mechanism. As an example, consider a beta-Bernoulli model,

$$\begin{aligned} Pr(p|\alpha,\beta) &= \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\ &= \frac{1}{B(\alpha,\beta)} \exp\big((\alpha-1)\log p + (\beta-1)\log(1-p)\big) \\ Pr(x|p) = p^x(1-p)^{1-x} &= \exp(x\log p + (1-x)\log(1-p)) \end{aligned}$$

where $B(\alpha,\beta)$ is the beta function. Given N binary-valued data points $\mathbf{X} = x^{(1)}, \ldots, x^{(N)}$ from the Bernoulli distribution, the posterior is

$$\begin{aligned} &Pr(p|\mathbf{X},\alpha,\beta) \propto \\ &\quad \exp\Big(\big(n_+ + \alpha - 1\big)\log p + \big(n_- + \beta - 1\big)\log(1-p)\Big) \\ &\quad n_+ = \sum_{i=1}^{N} x^{(i)}, \qquad n_- = \sum_{i=1}^{N} (1 - x^{(i)}) \,. \end{aligned}$$

The sufficient statistics for each data point are $S(x) = [x, 1-x]^\intercal$. The natural parameters for the posterior are $\theta = [\log p, \log(1-p)]^\intercal$, and $h(x) = 0$. The exponential mechanism sensitivity for a *truncated* version of this model, where $a_0 \le p \le 1 - a_0$, can be computed from Equation 13, $\triangle \log Pr(\theta, \mathbf{X}) =$

$$\begin{aligned} \sup_{x,x' \in \{0,1\}, p \in [a_0, 1-a_0]} & |x \log p + (1-x)\log(1-p) \\ & - \big(x' \log p + (1-x')\log(1-p)\big)| \\ & = -\log a_0 + \log(1-a_0) \,. \end{aligned} \tag{15}$$

Note that if $a_0 = 0$, corresponding to a standard untruncated beta distribution, the sensitivity is unbounded. This makes intuitive sense because some datasets are impossible if $p = 0$ or $p = 1$, which violates differential privacy.

3.2 THE LAPLACE MECHANISM

One limitation of the exponential mechanism / OPS approach to private Bayesian inference is that the temperature T of the approximate posterior is fixed for any ϵ that we are willing to pay, regardless of the number of data points N (Equation 10). While the posterior becomes more accurate as N increases, and the OPS approximation becomes more accurate by proxy, the OPS approximation remains a factor of T flatter than the posterior at N data points. This is not simply a limitation of the analysis. An adversary can choose data such that the dataset-specific privacy cost of posterior sampling approaches the worst case given by the exponential mechanism as N increases, by causing the posterior to concentrate on the worst-case θ (see the supplement for an example).

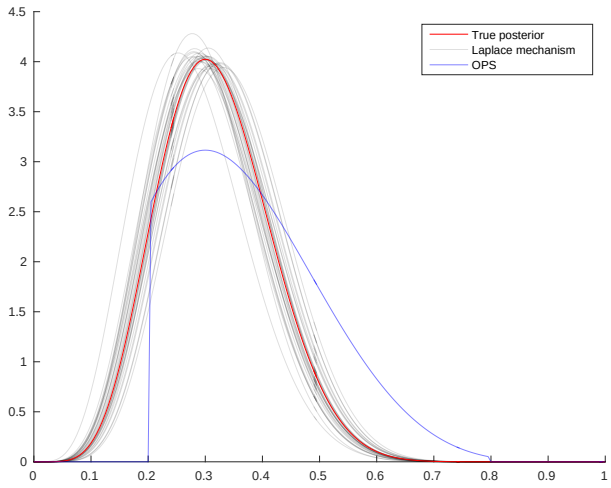

Figure 1: Privacy-preserving approximate posteriors for a beta-Bernoulli model ($\epsilon = 1$, the true parameter $p = 0.3$, OPS truncation point $a_0 = 0.2$, and number of observations $N = 20$). For the Laplace mechanism, 30 privatizing draws are rendered.

Here, we provide a simple Laplace mechanism alternative for exponential family posteriors, which becomes increasingly faithful to the true posterior with N data points, as N increases, for any fixed privacy cost ϵ, under general assumptions. The approach is based on the observation that for exponential family posteriors, as in Equation 11, the data interacts with the distribution only through the aggregate sufficient statistics, $S(\mathbf{X}) = \sum_{i=1}^{N} S(\mathbf{x}^{(i)})$. If we release privatized versions of these statistics we can use them to perform any further operations that we'd like, including drawing samples, computing moments and quantiles, and so on. This can straightforwardly be accomplished via the Laplace mechanism:

$$\begin{aligned} \hat{S}(\mathbf{X}) &= \text{proj}(S(\mathbf{X}) + (Y_1, Y_2, \ldots, Y_d)) \,, \\ Y_j &\sim \text{Laplace}(\triangle S(\mathbf{X})/\epsilon), \forall j \in \{1, 2, \ldots, d\} \,, \end{aligned} \tag{16}$$

where $\text{proj}(\cdot)$ is a projection onto the space of sufficient statistics, if the Laplace noise takes it out of this region.

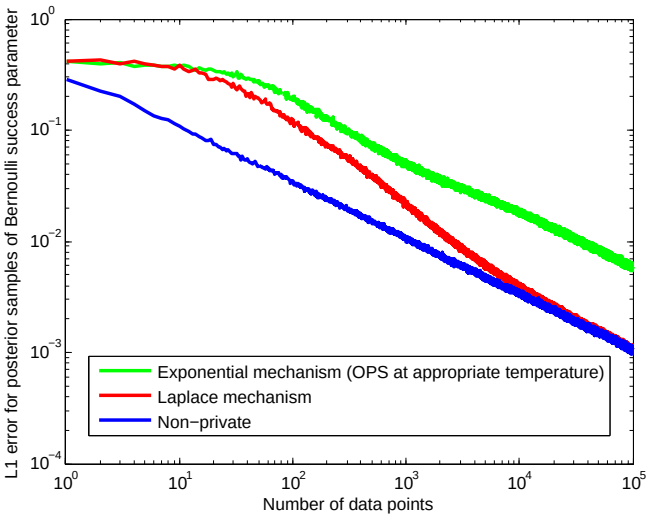

Figure 2: L1 error for private approximate samples from a beta posterior over a Bernoulli success parameter p, as a function of the number of Bernoulli(p) observations, averaged over 1000 repeats. The true parameter was $p = 0.1$, the exponential mechanism posterior was truncated at $a_0 = 0.05$, and $\epsilon = 0.1$.

For example, if the statistics are counts, the projection ensures that they are non-negative. The L_1 sensitivity of the aggregate statistics is

$$\begin{aligned} \triangle S(\mathbf{X}) &= \sup_{\mathbf{X},\mathbf{X}'} \| \sum_{i=1}^{N} S(\mathbf{x}'^{(i)}) - \sum_{i=1}^{N} S(\mathbf{x}^{(i)}) \|_1 \qquad (17) \\ &= \sup_{\mathbf{x},\mathbf{x}'} \|S(\mathbf{x}') - S(\mathbf{x})\|_1 \,, \end{aligned}$$

where $\mathbf{X}$, $\mathbf{X}'$ differ in at most one element. Note that perturbing the sufficient statistics is equivalent to perturbing the parameters, which was recently and independently proposed by Zhang et al. (2016) for beta-Bernoulli models such as Bernoulli naive Bayes.

A comparison of Equations 17 and 13 reveals that the L1 sensitivity and exponential mechanism sensitivities are closely related. The L1 sensitivity is generally easier to control as it does not involve θ or $h(\mathbf{x})$ but otherwise involves similar terms to the exponential mechanism sensitivity. For example, in the beta posterior case, where $S(\mathbf{x}) = [x, 1-x]$ is a binary indicator vector, the L1 sensitivity is 2. This should be contrasted to the exponential mechanism sensitivity of Equation 15, which depends heavily on the truncation point, and is unbounded for a standard untruncated beta distribution. The L1 sensitivity is fixed regardless of the number of data points N, and so the amount of Laplace noise to add becomes smaller relative to the total $S(\mathbf{X})$ as N increases.

Figure 1 illustrates the differences in behavior between the two privacy-preserving Bayesian inference algorithms for a beta distribution posterior with Bernoulli observations. The OPS estimator requires the distribution be truncated, here at $a_0 = 0.2$. This controls the exponential mechanism sensitivity, which determines the temperature T of the distribution, i.e. the extent to which the distribution is flattened, for a given ϵ. Here, $T = 2.7$. In contrast, the Laplace mechanism achieves privacy by adding noise to the sufficient statistics, which in this case are the pseudo-counts of successes and failures for the posterior distribution. In Figure 2 we illustrate the fidelity benefits of posterior sampling based on the Laplace mechanism instead of the exponential mechanism as the amount of data increases. In this case the exponential mechanism performs better than the Laplace mechanism only when the number of data points is very small (approximately $N = 10$), and is quickly overtaken by the Laplace mechanism sampling procedure. As N increases the accuracy of sampling from the Laplace mechanism's approximate posterior converges to the performance of samples from the true posterior at the current number of observations N, while the exponential mechanism behaves similarly to the posterior with fewer than N observations. We show this formally in the next subsection.

3.3 THEORETICAL RESULTS

First, we show that the Laplace mechanism approximation of exponential family posteriors approaches the true posterior distribution *evaluated at N data points*. Proofs are given in the supplementary.

Lemma 1. *For a minimal exponential family given a conjugate prior, where the posterior takes the form* $Pr(\theta|\mathbf{X}, \chi, \alpha) \propto g(\theta)^{n+\alpha} \exp\left(\theta^\intercal \left(\sum_{i=1}^{n} S(\mathbf{x}^{(i)}) + \alpha\chi\right)\right)$*, where $p(\theta|\eta)$ denotes this posterior with a natural parameter vector η, if there exists a $\delta > 0$ such that these assumptions are met:*

1. *The data $\mathbf{X}$ comes i.i.d. from a minimal exponential family distribution with natural parameter $\theta_0 \in \Theta$*
2. *θ_0 is in the interior of Θ*
3. *The function $A(\theta)$ has all derivatives for θ in the interior of Θ*
4. *$cov_{Pr(\mathbf{x}|\theta)}(S(\mathbf{x})))$ is finite for $\theta \in \mathcal{B}(\theta_0, \delta)$*
5. *$\exists w > 0$ s.t. $\det(cov_{Pr(\mathbf{x}|\theta)}(S(\mathbf{x})))) > w$ for $\theta \in \mathcal{B}(\theta_0, \delta)$*
6. *The prior $Pr(\theta|\chi, \alpha)$ is integrable and has support on a neighborhood of θ^**

then for any mechanism generating a perturbed posterior $\tilde{p}_N = p(\theta|\eta_N + \gamma)$ against a noiseless posterior $p_N = p(\theta|\eta_N)$ where γ comes from a distribution that does not

depend on the number of data observations N and has finite covariance, this limit holds:

$\lim_{N\to\infty} E[KL(\tilde{p}_N||p_N)] = 0\,.$

Corollary 2. *The Laplace mechanism on an exponential family satisfies the noise distribution requirements of Lemma 1 when the sensitivity of the sufficient statistics is finite and either the exponential family is minimal, or if the exponential family parameters θ are identifiable.*

These assumptions correspond to the data coming from a distribution where the Laplace regularity assumptions hold and the posterior satisfies the asymptotic normality given by the Bernstein-von Mises theorem. For example, in the beta-Bernoulli setting, these assumptions hold as long as the success parameter p is in the open interval $(0, 1)$. For $p = 0$ or 1, the relevant parameter is not in the interior of Θ, and the result does not apply. In the setting of learning a normal distribution's mean μ where the variance $\sigma^2 > 0$ is known, the assumptions of Lemma 1 always hold, as the natural parameter space is an open set. However, Corollary 2 does not apply in this setting because the sensitivity is infinite (unless bounds are placed on the data). Our efficiency result, in Theorem 4, follows from Lemma 1 and the Bernstein-von Mises theorem.

Theorem 4. *Under the assumptions of Lemma 1, the Laplace mechanism has an asymptotic posterior of $\mathcal{N}(\theta_0, 2\mathbb{I}^{-1}/N)$ from which drawing a single sample has an asymptotic relative efficiency of 2 in estimating θ_0, where $\mathbb{I}$ is the Fisher information at θ_0.*

Above, the asymptotic posterior refers to the normal distribution, whose variance depends on N, that the posterior distribution approaches as N increases. This ARE result should be contrasted to that of the exponential mechanism (Wang et al., 2015b).

Theorem 5. *The exponential mechanism applied to the exponential family with temperature parameter $T \geq 1$ has an asymptotic posterior of $\mathcal{N}(\theta^*, (1+T)\mathbb{I}^{-1}/N)$ and a single sample has an asymptotic relative efficiency of $(1 + T)$ in estimating θ^*, where $\mathbb{I}$ is the Fisher information at θ^*.*

Here, the ARE represents the ratio between the variance of the estimator and the optimal variance $\mathbb{I}^{-1}/N$ achieved by the posterior mean in the limit. Sampling from the posterior itself has an ARE of 2, due to the stochasticity of sampling, which the Laplace mechanism approach matches. These theoretical results provide an explanation for the difference in the behavior of these two methods as N increases seen in Figure 2. The Laplace mechanism will eventually approach the true posterior and the impact of privacy on accuracy will diminish when the data size increases. However, for the exponential mechanism with $T > 1$, the ratio of variances between the sampled posterior and the true posterior given N data points approaches $(1 + T)/2$, making the sampled posterior more spread out than the true posterior even as N grows large.

So far we have compared the ARE values for *sampling*, as an apples-to-apples comparison. In reality, the Laplace mechanism has a further advantage as it releases a full posterior with privatized parameters, while the exponential mechanism can only release a finite number of samples with a finite ϵ, which we discuss in Remark 1.

Remark 1. *Under the the assumptions of Lemma 1, by using the full privatized posterior instead of just a sample from it, the Laplace mechanism can release the privatized posterior's mean, which has an asymptotic relative efficiency of 1 in estimating θ^*.*

4 PRIVATE GIBBS SAMPLING

We now shift our discussion to the case of approximate Bayesian inference. While the analysis of Dimitrakakis et al. (2014) and Wang et al. (2015b) shows that posterior sampling is differentially private under certain conditions, exact sampling is not in general tractable. It does not directly follow that approximate sampling algorithms such as MCMC are also differentially private, or private at the same privacy level. Wang et al. (2015b) give two results towards understanding the privacy properties of approximate sampling algorithms. First, they show that if the approximate sampler is "close" to the true distribution in a certain sense, then the privacy cost will be close to that of a true posterior sample:

Proposition 3. *If procedure $\mathcal{A}$ which produces samples from distribution $P_{\mathbf{X}}$ is ϵ-differentially private, then any approximate sampling procedures $\mathcal{A}'$ that produces a sample from $P'_{\mathbf{X}}$ such that $\|P_{\mathbf{X}} - P'_{\mathbf{X}}\|_1 \leq \delta$ for any $\mathbf{X}$ is $(\epsilon, (1 + \exp(\epsilon)\delta)$-differentially private.*

Unfortunately, it is not in general feasible to verify the convergence of an MCMC algorithm, and so this criterion is not generally verifiable in practice. In their second result, Wang et al. study the privacy properties of stochastic gradient MCMC algorithms, including stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011) and its extensions. SGLD is a stochastic gradient method with noise injected in the gradient updates which converges in distribution to the target posterior.

In this section we study the privacy cost of MCMC, allowing us to quantify the privacy of many real-world MCMC-based Bayesian analyses. We focus on the case of Gibbs sampling, under exponential mechanism and Laplace mechanism approaches. By reinterpreting Gibbs sampling as an instance of the exponential mechanism, we obtain the "privacy for free" cost of Gibbs sampling. Metropolis-Hastings and annealed importance sampling also have privacy guarantees, which we show in the supplementary materials.

4.1 EXPONENTIAL MECHANISM

We consider the privacy cost of a Gibbs sampler, where data $\mathbf{X}$ are behind the privacy wall, current sampled values of parameters and latent variables $\theta = [\theta_1, \ldots, \theta_D]$ are publicly known, and a Gibbs update is a randomized algorithm which queries our private data in order to randomly select a new value θ'_l for the current variable θ_l. The transition kernel for a Gibbs update of θ_l is

$$T^{(Gibbs,l)}(\theta, \theta') = Pr(\theta'_l | \theta_{\neg l}, \mathbf{X}) \,, \tag{18}$$

where $\theta_{\neg l}$ refers to all entries of θ except l, which are held fixed, i.e. $\theta'_{\neg l} = \theta_{\neg l}$. This update can be understood via the exponential mechanism:

$$T^{(Gibbs,l,\epsilon)}(\theta, \theta') \propto Pr(\theta'_l, \theta_{\neg l}, \mathbf{X})^{\frac{\epsilon}{2\triangle \log Pr(\theta'_l, \theta_{\neg l}, \mathbf{X})}} \,, \tag{19}$$

with utility function $u(\mathbf{X}, \theta'_l; \theta_{\neg l}) = \log Pr(\theta'_l, \theta_{\neg l}, \mathbf{X})$, over the space of possible assignments to θ_l, holding $\theta_{\neg l}$ fixed. A Gibbs update is therefore ϵ-differentially private, with $\epsilon = 2\triangle \log Pr(\theta'_l, \theta_{\neg l}, \mathbf{X})$. This update corresponds to Equation 6 except that the set of responses for the exponential mechanism is restricted to those where $\theta'_{\neg l} = \theta_{\neg l}$. Note that

$$\triangle \log Pr(\theta'_l, \theta_{\neg l}, \mathbf{X}) \leq \triangle \log Pr(\theta, \mathbf{X}) \tag{20}$$

as the worst case is computed over a strictly smaller set of outcomes. In many cases each parameter and latent variable θ_l is associated with only the lth data point $\mathbf{x}_l$, in which case the privacy cost of a Gibbs scan can be improved over simple additive composition. In this case a random sequence scan Gibbs pass, which updates all N θ_l's exactly once, is $2\triangle \log Pr(\theta, \mathbf{X})$-differentially private by parallel composition (Song et al., 2013). Alternatively, a random scan Gibbs sampler, which updates a random Q out of N θ_l's, is $4\triangle \log Pr(\theta, \mathbf{X})\frac{Q}{N}$-differentially private from the *privacy amplification* benefit of subsampling data (Li et al., 2012).

4.2 LAPLACE MECHANISM

Suppose that the conditional posterior distribution for a Gibbs update is in the exponential family. Having privatized the sufficient statistics arising from the data for the likelihoods involved in each update, via Equation 16, and publicly released them with privacy cost ϵ, we may now perform the update by drawing a sample from the approximate conditional posterior, i.e. Equation 11 but with $S(\mathbf{X}) = \sum_{i=1}^{N}(\mathbf{x}^{(i)})$ replaced by $\hat{S}(\mathbf{X})$. Since the privatized statistics can be made public, we can also subsequently draw from an approximate posterior based on $\hat{S}(\mathbf{X})$ with any other prior (selected based on public information only), without paying any further privacy cost. This is especially valuable in a Gibbs sampling context, where the "prior" for a Gibbs update often consists of factors from other variables and parameters to be sampled, which are updated during the course of the algorithm.

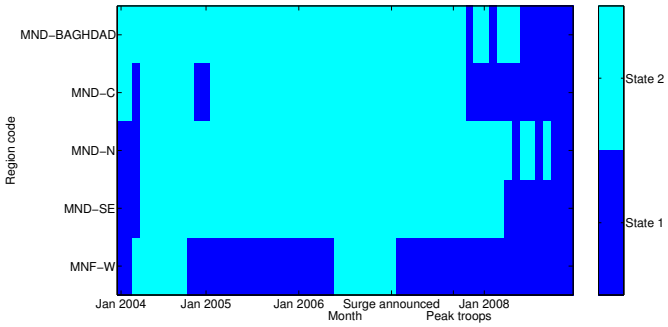

Figure 3: State assignments of privacy-preserving HMM on Iraq (Laplace mechanism, $\epsilon = 5$).

In particular, consider a Bayesian model where a Gibbs sampler interacts with data only via conditional posteriors and their corresponding likelihoods that are exponential family distributions. We can privatize the sufficient statistics of the likelihood just once at the beginning of the MCMC algorithm via the Laplace mechanism with privacy cost ϵ, and then approximately sample from the posterior by running the entire MCMC algorithm based on these privatized statistics without paying any further privacy cost. This is typically much cheaper in the privacy budget than exponential mechanism MCMC which pays a privacy cost for every Gibbs update, as we shall see in our case study in Section 5. The MCMC algorithm does not need to converge to obtain privacy guarantees, unlike the OPS method. This approach applies to a very broad class of models, including Bayesian parameter learning for fully-observed MRF and Bayesian network models. Of course, for this technique to be useful in practice, the aggregate sufficient statistics for each Gibbs update must be large relative to the Laplace noise. For latent variable models, this typically corresponds to a setting with many data points per latent variable, such as the HMM model with multiple emissions per timestep which we study in the next section.

5 CASE STUDY: WIKILEAKS IRAQ & AFGHANISTAN WAR LOGS

A primary goal of this work is to establish the practical feasibility of privacy-preserving Bayesian data analysis using complex models on real-world datasets. In this section we investigate the performance of the methods studied in this paper for the analysis of sensitive military data. In July and October 2010, the Wikileaks organization disclosed collections of internal U.S. military field reports from the wars in Afghanistan and Iraq, respectively. Both disclosures contained data from between January 2004 to December 2009, with ~75,000 entries from the war in Afghanistan, and ~390,000 entries from Iraq. Hillary Clinton, at that time

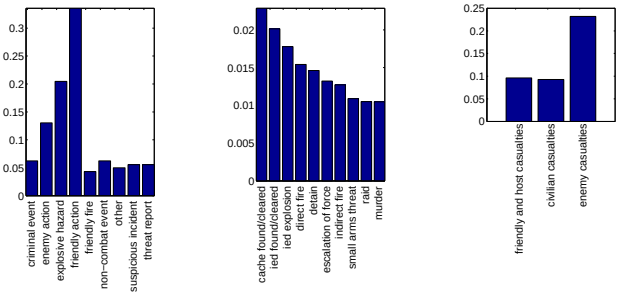

Figure 4: State 1 for Iraq (*type*, *category*, *casualties*).

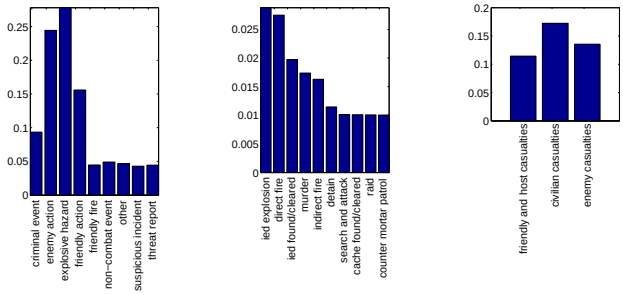

Figure 5: State 2 for Iraq (*type*, *category*, *casualties*).

the U.S. Secretary of State, criticized the disclosure, stating that it "puts the lives of United States and its partners' service members and civilians at risk."[1] These risks, and the motivations for the leak, could potentially have been mitigated by releasing a differentially private analysis of the data, which protects the contents of each individual log entry while revealing high-level trends. Note that since the data are publicly available, although our *models* were differentially private, other aspects of this manuscript such as the evaluation may reveal certain information, as in other works such as Wang et al. (2015a,b).

The disclosed war logs each correspond to an individual event, and contain textual reports, as well as fields such as coarse-grained *types* (*friendly action*, *explosive hazard*, ...), fine-grained *categories* (*mine found/cleared*, *show of force*, ...), and casualty counts (*wounded/killed/detained*) for the different factions (*Friendly*, *HostNation* (i.e. Iraqi and Afghani forces), *Civilian*, and *Enemy*, where the names are relative to the U.S. military's perspective). We use the techniques discussed in this paper to privately infer a hidden Markov model on the log entries. The HMM was fit to the non-textual fields listed above, with one timestep per month, and one HMM chain per region code. A naive Bayes conditional independence assumption was used in the emission probabilities for simplicity and parameter-count parsimony. Each field was modeled via a discrete distribution per latent state, with casualty counts binarized (0 versus > 0), and with *wounded/killed/detained* and *Friendly/HostNation* features combined, respectively, via disjunction of the binary values. This decreased the number of features to privatize, while slightly increasing the size of the counts per field to protect and simplifying the model for visualization purposes. After preprocessing to remove empty timesteps and near-empty region codes (see the supplementary), the median number of log entries per region/timestep pair was 972 for Iraq, and 58 for Afghanistan. The number of log entries per timestep was highly skewed for Afghanistan, due to an increase in density over time.

The models were trained via Gibbs sampling, with the transition probabilities collapsed out, following Goldwater and Griffiths (2007). We did not collapse out the naive Bayes parameters in order to keep the conditional likelihood in the exponential family. The details of the model and inference algorithm are given in the supplementary material. We trained the models for 200 Gibbs iterations, with the first 100 used for burn-in. Both privatization methods have the same overall computational complexity as the non-private sampler. The Laplace mechanism's computational overhead is paid once up-front, and did not greatly affect the runtime, while OPS roughly doubled the runtime. For visualization purposes we recovered parameter estimates via the posterior mean based on the latent variable assignments of the final iteration, and we reported the most frequent latent variable assignments over the non-burn-in iterations. We trained a 2-state model on the Iraq data, and a 3-state model for the Afghanistan data, using the Laplace approach with total $\epsilon = 5$ ($\epsilon = 1$ for each of 5 features).

Interestingly, when given 10 states, the privacy-preserving model only assigned substantial numbers of data points to these 2-3 states, while a non-private HMM happily fit a 10-state model to the data. The Laplace noise therefore appears to play the role of a regularizer, consistent with the noise being interpreted as a "random prior," and along the lines of noise-based regularization techniques such as (Srivastava et al., 2014; van der Maaten et al., 2013), although of course it may correspond to more regularization than we would typically like. This phenomenon potentially merits further study, beyond the scope of this paper.

We visualized the output of the Laplace HMM for Iraq in Figures 3–5. State 1 shows the U.S. military performing well, with the most frequent outcomes for each feature being *friendly action*, *cache found/cleared*, and *enemy casualties*, while the U.S. military performed poorly in State 2 (*explosive hazard*, *IED explosion*, *civilian casualties*). State 2 was prevalent in most regions until the situation improved to State 1 after the troop surge strategy of 2007. This transition typically occurred after troops peaked in Sept.–Nov. 2007. The results for Afghanistan, in the supplementary, provide a critical lens on the US military's performance, with enemy casualty rates (including

[1]Fallon, Amy (2010). "Iraq war logs: disclosure condemned by Hillary Clinton and Nato." The Guardian. Retrieved on 2/22/2016.

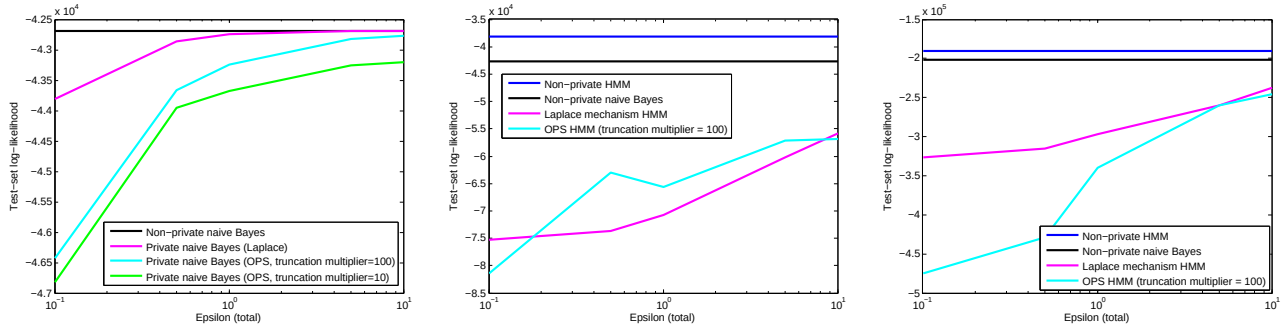

Figure 6: Log-likelihood results. **Left:** Naive Bayes (Afghanistan). **Middle**: Afghanistan. **Right**: Iraq. For OPS, Dirichlets were truncated at $a_0 = \frac{1}{MK_d}$, $M = 10$ or 100, where $K_d =$ feature d's dimensionality.

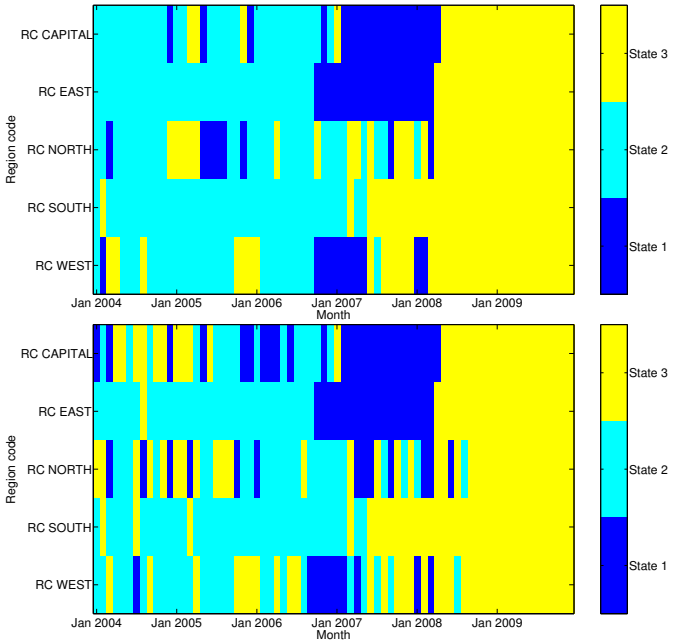

Figure 7: State assignments for OPS privacy-preserving HMM on Afghanistan. ($\epsilon = 5$, truncation point $a_0 = \frac{1}{100K_d}$). **Top**: Estimate from last 100 samples. **Bottom**: Estimate from last one sample.

detainments) lower than friendly/host casualties for all latent states, and lower than civilian casualties in 2 of 3 states.

We also evaluated the methods at prediction. A uniform random 10% of the timestep/region pairs were held out for 10 train/test splits, and we reported average test likelihoods over the splits. We estimated test log-likelihood for each split by averaging the test likelihood over the burned-in samples (Laplace mechanism), or using the final sample (OPS). All methods were given 10 latent states, and ϵ was varied between 0.1 and 10. We also considered a naive Bayes model, equivalent to a 1-state HMM. The Laplace mechanism was superior to OPS for the naive Bayes model, for which the statistics are corpus-wide counts, corresponding to a high-data regime in which our asymptotic analysis was applicable. OPS was competitive with the Laplace mechanism for the HMM on Afghanistan, where the amount of data was relatively low. For the Iraq dataset, where there was more data per timestep, the Laplace mechanism outperformed OPS, particularly in the high-privacy regime. For OPS, privacy at ϵ is only guaranteed if MCMC has converged. Otherwise, from Section 4.1, the worst case is an impractical $\epsilon^{(Gibbs)} \leq 400\epsilon$ (200 iterations of latent variable and parameter updates with worst-case cost ϵ). OPS only releases one sample, which harmed the coherency of the visualization for Afghanistan, as latent states of the final sample were noisy relative to an estimate based on all 100 post burn-in samples (Figure 7). Privatizing the Gibbs chain at a privacy cost of $\epsilon^{(Gibbs)}$ would avoid this.

6 CONCLUSION

This paper studied the practical limitations of using posterior sampling to obtain privacy "for free." We explored an alternative based on the Laplace mechanism, and analyzed it both theoretically and empirically. We illustrated the benefits of the Laplace mechanism for privacy-preserving Bayesian inference to analyze sensitive war records. The study of privacy-preserving Bayesian inference is only just beginning. We envision extensions of these techniques to other approximate inference algorithms, as well as their practical application to sensitive real-world data sets. Finally, we have argued that asymptotic efficiency is important in a privacy context, leading to an open question: how large is the class of private methods that are asymptotically efficient?

Acknowledgements

The work of K. Chaudhuri and J. Geumlek was supported in part by NSF under IIS 1253942, and the work of M. Welling was supported in part by Qualcomm, Google and Facebook. We also thank Mijung Park, Eric Nalisnick, and Babak Shahbaba for helpful discussions.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., Seaton, D. T., and Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Communications of the ACM*, 57(9):56–63.

Dimitrakakis, C., Nelson, B., Mitrokotsa, A., and Rubinstein, B. I. (2014). Robust and private Bayesian inference. In *Algorithmic Learning Theory (ALT)*, pages 291–305. Springer.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer.

Dwork, C. and Roth, A. (2013). The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407.

Dwork, C., Rothblum, G. N., and Vadhan, S. (2010). Boosting and differential privacy. In *The 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 51–60.

Goldwater, S. and Griffiths, T. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 744–751.

Huang, Z. and Kannan, S. (2012). The exponential mechanism for social welfare: Private, truthful, and nearly optimal. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 140–149. IEEE.

Husmeier, D., Dybowski, R., and Roberts, S. (2006). *Probabilistic modeling in bioinformatics and medical informatics*. Springer Science & Business Media.

Li, N., Qardaji, W., and Su, D. (2012). On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, pages 32–33. ACM.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *Foundations of Computer Science (FOCS), 2007 IEEE 48th Annual Symposium on*, pages 94–103. IEEE.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. (2013). Tuned models of peer assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining*, pages 153–160.

Salakhutdinov, R. and Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 880–887.

Song, S., Chaudhuri, K., and Sarwate, A. D. (2013). Stochastic gradient descent with differentially private updates. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 245–248. IEEE.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

van der Maaten, L., Chen, M., Tyree, S., and Weinberger, K. Q. (2013). Learning with marginalized corrupted features. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 410–418.

Wang, Y., Wang, Y.-X., and Singh, A. (2015a). Differentially private subspace clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1000–1008.

Wang, Y.-X., Fienberg, S. E., and Smola, A. (2015b). Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. *Proceedings of The 32nd International Conference on Machine Learning (ICML)*, pages 2493–2502.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 681–688.

Zhang, Z., Rubinstein, B., and Dimitrakakis, C. (2016). On the differential privacy of Bayesian inference. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*.