LUGS: A Scalable Non-parametric Data Synthesizer for Privacy-Preserving Health Data Publiction

Abstract

This paper introduces a non-parametric data synthesizing algorithm to generate privacy-safe "realistic but not real" synthetic health data. The proposed algorithm synthesizes artificial records while preserving the statistical characteristics of the original data to the extent possible. The risk from "database linking attack" is quantified by an l-diversified data generation process. Moreover its algorithmic performance is optimized using Locality-Sensitive Hashing and parallel computation techniques to yield a linear-time algorithm that is suitable for Big Health data applications. We synthesize a public Medicare claim dataset using the proposed algorithm, and demonstrate multiple data mining applications and statistical analyses using the data. The synthetic dataset delivers results that are substantially identical to those obtained from the original dataset, without revealing the actual records.

1. Introduction

Synthetic data, generated from a certain random process, can address disclosure limitation issues in public use health data. Many health datasets contain privacy-sensitive and sometimes confidential information such as disease, payment, and treatment records. Revealing such health information is clearly disagreeable to many, and raises severe ethical and fiduciary issues. Instead, when created carefully, synthetic data can provide the required statistical information for various analyses in the healthcare domain without revealing person-specific data or person's identity. Developing appropriate algorithms to generate synthetic data is critical to meeting the growing need for well-grounded health informatics research.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Using traditional "parametric model-based" synthesizers, however, provides only a partial solution to such objectives (Reiter et al., 2006), as it introduces two open-ended issues, namely model selection problem and unquantifiable privacy-risk. The complexity of a synthetic model limits the answerable range of research questions; for example, a linear model synthetic data cannot address identifying quadratic relationships between covariates. However, most data mining research is based on retrospective analysis, where the research questions are posed *post hoc* and may be determined during the data exploration process. Thus, this type of synthetic data are not perfectly suited for data mining applications. Moreover, popular privacy metrics such as k-anonymity (Sweeney, 2002), l-diversity (Machanavajjhala et al., 2007), or ϵ-differential privacy (Dwork, 2006) cannot be directly applied to such model-based synthesizers (Abowd & Vilhuber, 2008). Recent reports on adversarial privacy attacks (Narayanan & Shmatikov, 2008; 2009) suggest that rigorous characterization of such risk is critical in data publishing.

Non-parametric synthetic data may be a remedy for the model selection issue. Moreover, if the generative process of such data adheres to a certain state of the art privacy metric, then the risk from synthetic data can be well characterized. However, such non-parametric schemes are rarely adopted in generating synthetic data in practice, due to its computational cost, and non-intuitive connections to privacy metrics. In this paper, we analyze the definition of privacy in the healthcare context, and derive its connection to probabilistic generative processes. We also propose a novel and practical algorithm for generating synthetic data, adapting and coupling: parallel computation and Locality Sensitive Hashing (LSH) techniques (Gionis et al., 1999) to address the computational challenges. This data type is dominant in many health records. Furthermore, for many numeric fields, binning, also known as histogramming or quantization, can be appropriately applied to transform data into categorical format, supporting the generality of our framework.

2. Preliminaries

The original table and synthetic table are denoted as $\mathcal{D}$ and $\mathcal{S}$, respectively. n and m represent the number of rows and columns in table $\mathcal{D}$, and $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ is a row of the table. $\mathbf{x}_{\setminus j}$ is a row with the value of x_j undetermined e.g. $\mathbf{x}_{\setminus 1} = (\cdot, x_2, \ldots, x_m)$. Random variables are expressed using capital letters, for example X for a scalar, and $\mathbf{X}$ for a vector. $\Pr(\mathbf{X})$ is a true probability distribution of the original data, and $\Pr_e(\mathbf{X} \mid \mathcal{D})$ represents an empirical probability mass function given data $\mathcal{D}$.

A categorical dataset $\mathcal{D}$ with indistinguishable rows, i.e. without any ID's, can be completely described by a contingency table. Thus, with enough number of samples, the contingency table can be said as the most precise and non-parametric approximation for the underlying joint distribution. We exchangeably use the notations for the normalized contingency table[1] and the empirical distribution $\Pr_e(\mathbf{X} \mid \mathcal{D})$.

The difference between the empirical and the true distribution decays as the number of data points increase (Csiszar & Shields, 2004). The empirical distribution approaches the true distribution exponentially fast as N increases. In practice, this empirical distribution can be treated as the underlying true distribution when the size of the data is huge.

Obtaining a full contingency table is usually intractable especially with a large number of attributes. However, we can mimic sampling from a full contingency table using Gibbs sampling. Gibbs sampling is a prevalent estimation or sampling technique, when sampling from a joint distribution is intractable. A synthetic sample $\mathbf{x}_s$ from a joint distribution $\Pr_e(\mathbf{X} \mid \mathcal{D})$ can be Gibbs-sampled as follows:

$$
\begin{aligned}
x_1 &\sim \Pr_e(X_1 \mid \mathbf{x}_{\setminus 1}, \mathcal{D}) \\
x_2 &\sim \Pr_e(X_2 \mid \mathbf{x}_{\setminus 2}, \mathcal{D}) \\
&\vdots \\
x_m &\sim \Pr_e(X_m \mid \mathbf{x}_{\setminus m}, \mathcal{D})
\end{aligned}
$$

where $\Pr_e(X_i \mid \mathbf{x}_{\setminus i}, \mathcal{D})$ is a conditional frequency table derived from data $\mathcal{D}$. When the above cycle has reached an enough number iterations (burn-in period), the last sample $\mathbf{x}$ is equivalent to a random sample from the joint distribution. It is important to note that, in Gibbs sampling, the information about the joint distribution is distributed across its conditional distributions.

This brute-force Gibbs synthesizer, however, has three critical issues: a lack of convergence guarantee, computational inefficiency, and loose privacy guarantee. For a Gibbs sampler to converge, the Markov chain in a Gibbs sampler needs to be irreducible and aperiodic. In the brute-force Gibbs synthesizer, the Markov chain of the empirical conditional distributions needs to be verified to satisfy both conditions. Moreover, this brute-force Gibbs synthesizer is computationally expensive. The number of distinct conditional distributions exponentially grows with the number of features, thus pre-computing $\Pr_e(X_i \mid \mathbf{X}_{\setminus i}, \mathcal{D})$ is not desirable. Estimating $\Pr_e(X_i \mid \mathbf{X}_{\setminus i}, \mathcal{D})$ on the fly is not a smart choice, since the estimation needs a linear scan for every Gibbs iteration. Finally, synthetic samples from an empirical distribution always can always be found in the original data, as the support of an empirical distribution is the same as the support of the original data. In other words, the synthetic data are "realistic and real", not "realistic but not real".

[1] Normalized by the total row counts.

3. LUGS Algorithm

We now present an l-diversified uniformly smoothed Gibbs Ssynthesizer (LUGS). LUGS is an efficient non-parametric data synthesizer that meets the l-diversity principle, and is illustrated in Algorithm 1. The LUGS algorithm consists of three steps: estimation, perturbation, and sampling steps.

Algorithm 1 LUGS Algorithm

(Step 1) Estimate $\mathrm{P}_e(X_i \mid g(\mathbf{X}_{\setminus i}), \mathcal{D})$
(Step 2) Perturb $\mathrm{P}_e \to \mathrm{Q}$ s.t. $-E[\log \mathrm{Q}] \geq \log l$
(Step 3) Gibbs-sample synthetic data from the perturbed conditional distributions

3.1. Step 1: Estimation

Instead of $\Pr_e(X_i \mid \mathbf{X}_{\setminus i}, \mathcal{D})$, LUGS estimates an approximated distribution, $\mathrm{P}_e(X_i \mid g(\mathbf{X}_{\setminus i}), \mathcal{D})$, where $g(\mathbf{X}_{\setminus i})$ is a hash function such as MinHashing (Broder, 1997) and Locality-Sensitive Hashing (LSH) (Gionis et al., 1999; Indyk & Motwani, 1998). The hash function is employed to reduce the number of distinct conditional distributions and to access the conditional distributions faster.

We first reduce the hash key space using MinHashing, then perform a fast (approximate) nearest neighbor search using LSH. In this paper, we choose to use an LSH family $\mathcal{F}$ defined for a metric space $\mathcal{M} = (\mathbf{x}, \text{Hamming})$, but we note that other distance metrics are also applicable. MinHashing collision probability can be adjusted by its parameters, and the number of MinHashing keys can be made significantly smaller

than the size of the original database. Thus, even for a really large-scale database with high-dimensional features, LUGS can sample from its (approximated) conditional distributions. Note that there is a trade-off between the exact conditional distribution and the required memory size.

3.2. Step 2: Perturbation

The obtained conditional distributions are perturbed to meet a prescribed entropy l-diversity criterion for privacy:

Definition 1 (Entropy l-diversity (Machanavajjhala et al., 2007)). *A probability mass function* Q *is "entropy* l*-diverse" if*

$$-E[\log \mathrm{Q}] = -\sum_x \mathrm{Q}(x) \log \mathrm{Q}(x) \geq \log l$$

where $\log l > 0$.

We perturb the approximated probability distribution $\mathrm{P}_e(X_i \mid g(\mathbf{X}_{\setminus i}), \mathcal{D})$ to satisfy the l-diversity principle as follows:

$$\mathrm{Q}_\alpha(X_i \mid g(\mathbf{X}_{\setminus i})) = (1-\alpha)\mathrm{P}_e(X_i \mid g(\mathbf{X}_{\setminus i}), \mathcal{D}) + \alpha \mathrm{U}$$

where U is a uniform distribution over the range of X_i. As we want to keep the statistical properties of the original data to the extent possible, we choose the minimum α that satisfies the condition:

$$\alpha^* = \arg\min_\alpha \mathrm{Q}_\alpha(X_i \mid g(\mathbf{X}_{\setminus i})) \geq \log l$$
$$\mathrm{Q}(X_i \mid g(\mathbf{X}_{\setminus i})) = \mathrm{Q}_{\alpha^*}(X_i \mid g(\mathbf{X}_{\setminus i}))$$

Although we do not show details in this paper, perturbed distributions can be modified to satisfy other privacy metrics such as ϵ-differential privacy or Putterfish framework (Kifer & Machanavajjhala, 2012). This paper mainly focuses on the l-diversified synthetic data, and we leave further experiments and comparisons with other privacy metrics to future work.

3.3. Step 3: Sampling

LUGS samples are generated through a perturbed Gibbs sampler:

$$x_1 \sim \mathrm{Q}(X_1 \mid g(\mathbf{x}_{\setminus 1}))$$
$$x_2 \sim \mathrm{Q}(X_2 \mid g(\mathbf{x}_{\setminus 2}))$$
$$\vdots$$
$$x_m \sim \mathrm{Q}(X_m \mid g(\mathbf{x}_{\setminus m}))$$

This Markov chain converges to a unique stationary distribution, $\mathrm{Q}(\mathbf{X})$ if $\alpha > 0$, see Theorem 1.

Theorem 1 (Existence of $\mathrm{Q}(\mathbf{X})$). *If* $\mathrm{Q}(X_i|\mathbf{X}_{\setminus i}) > 0$ *(positivity condition), then the Markov chain of a perturbed Gibbs sampler is irreducible, thus there exists a unique stationary distribution* $\mathrm{Q}(\mathbf{X})$.

Proof. For any two states $\mathbf{x}$ and $\mathbf{x}'$, we have:

$$\Pr(\mathbf{X}^{(t+1)} = \mathbf{x} \mid \mathbf{X}^{(t)} = \mathbf{x}') > 0$$
$$\Pr(\mathbf{X}^{(t+1)} = \mathbf{x}' \mid \mathbf{X}^{(t)} = \mathbf{x}) > 0$$

where $\mathbf{X}^{(t+1)}$ represents the $(t+1)$th iteration sample of the Gibbs sampler. As $\mathbf{x}$ and $\mathbf{x}'$ intercommunicate, one of the following is true (Grimmett & Stirzaker, 2001): $\mathbf{x}$ and $\mathbf{x}'$ are both recurrent or $\mathbf{x}$ and $\mathbf{x}'$ are both null-recurrent. As the range of $\mathbf{X}$ is finite, we have at least one recurrent state, thus all the states of $\mathbf{X}$ are recurrent. Since the Markov chain is irreducible and its all the states are recurrent, the chain has a unique stationary distribution $\mathrm{Q}(\mathbf{X})$. □

The difference between $\Pr_e(\mathbf{X})$ and $\mathrm{Q}(\mathbf{X})$ is upper-bounded by the parameter $\log l$, and hence the difference between the true distribution $\Pr(\mathbf{X})$ and the LUGS-distribution $\mathrm{Q}(\mathbf{X})$ is also upper-bounded. Note that the perturbed Gibbs sampler generates artificial samples that are not in the original data. For these artificial samples, we find the nearest neighboring perturbed conditional distribution as follows:

$$\mathrm{Q}(X_i \mid g(\mathbf{X}^{ann}_{\setminus i})) \quad \text{s.t.} \quad \mathbf{X}^{ann}_{\setminus i} \approx \mathbf{X}_{\setminus i}$$

using LSH, then use this approximated nearest neighbor distirubion in the Gibbs sampling iterations.

3.4. Additional Step: Parallelization

The LUGS Algorithm is embarrassingly parallelizable with sub-sampling. Algorithm 2 illustrates the parallel-LUGS (PLUGS) algorithm. If each process in PLUGS satisfies the l-diversity principle, the collection of synthetic data also meet the l-diversity principle.

Algorithm 2 PLUGS Algorithm

(Step 0) Partition $\mathcal{D}$ into $\{\mathcal{D}^p\}$ and distribute to parallel processes
(Step 1) Estimate $\mathrm{P}_e(X_i \mid g(\mathbf{X}_{\setminus i}), \mathcal{D}^p)$
(Step 2) Perturb $\mathrm{P}_e \to \mathrm{Q}$ s.t. $-E[\log \mathrm{Q}] \geq \log l$
(Step 3) Gibbs-sample synthetic data from the perturbed conditional distributions
(Step 4) Combine and mix synthetic data $\{\mathcal{S}^p\}$ from the parallel processes

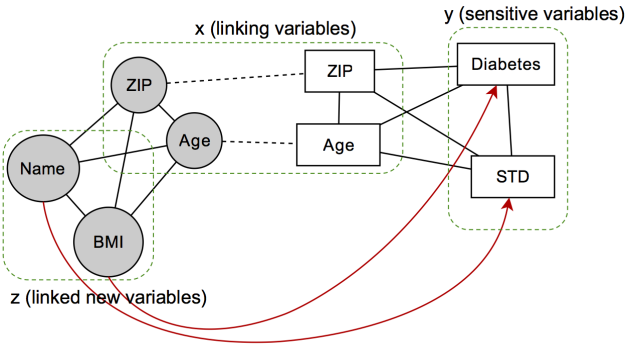

Figure 1. A link attack example. Two datasets can be linked based on Age and ZIP code. Sensitive values such as Diabetes and STD can be revealed using the new linked variables Name and BMI.

4. Privacy Analyses

The "realistic but not real" principle can be achieved through rigorous analyses on the definition of "privacy" in the healthcare context. Unlike other scientific measurement datasets, healthcare datasets in general contain the information about real people, such as patients and physicians. Thus, the definition of data privacy in healthcare datasets can be narrowed down. We define the goal of privacy protection in the healthcare domain is to protect the identity of the people, and to protect their sensitive information.

The attacks on published datasets can be categorized as follows:

- **Identity attack**: identifying "who"
- **Feature attack**: identifying "what"

Although many other types of attacks may be threats to uncover private information, such attacks are not directly related to our definition of privacy protection. For example, an attacker can link two datasets and find out new information for a specific row, without knowing a person's identity. In this case, as the attacker has no clue about the identity of the records, this attack does not infringe on privacy by our definition. Again, it is possible that even such information can be useful for later identity and feature attacks (transitive attacks). We note that the uncertainty level of such information may decrease as more linkable datasets become available.

In this paper, we analyze the effect of database linking on feature attacks. Consider a dataset with two features: X (non-sensitive) and Y (sensitive). When publishing the dataset, a typical strategy is to noise Y to be $\tilde{Y}$. However, if another dataset is linked with the original dataset based on X, the linked information may reduce the uncertainty about the sensitive field. Figure 1 illustrates this linking scenario. Theorem 2 illustrates this information gain for attackers by linking two datasets:

Theorem 2 (Link Gain)**.** *Suppose that two datasets* $\mathcal{D}_Z = \{Z, X\}$ *and* $\mathcal{D}_Y = \{(\tilde{Y}, X)\}$ *are linked based on* X*, then:*

$$\frac{\Pr(\tilde{Y} = Y \mid X, Z)}{\Pr(\tilde{Y} = Y \mid X)} = \frac{\Pr(Z \mid X, \tilde{Y} = Y)}{\Pr(Z \mid X)} = \Lambda \qquad (1)$$

where Λ *is the "link gain ratio" by linking two datasets.*

Proof. Equation (1) is a direct result by applying Bayes' theorem. □

Theorem 2 states two competing objectives when publishing data: Λ needs to be low and $\Pr(\tilde{Y} = Y \mid X)$ needs to be low. As can be seen, Λ increases as $\Pr(\tilde{Y} = Y \mid X)$ decreases, and $\Pr(\tilde{Y} = Y \mid X)$ increases as Λ decreases. Furthermore, Λ can be arbitrarily large when $\Pr(Z \mid X) \approx 0$, where we have no control over Z and $\mathcal{D}_Z$. Thus, lowering Λ is very difficult. LUGS addresses this issue by perturbing every variable in a dataset i.e. including linking variables. In other words, instead of directly lowering Λ, we inject uncertainty on linking.

The perturbation step in LUGS can be interpreted from the information theory context. The mutual information between X_i and $\mathbf{X}_{\setminus X_i}$ is as follows:

$$I(X_i; \mathbf{X}_{\setminus i}) = H(X_i) - H(X_i|\mathbf{X}_{\setminus i}) \qquad (2)$$

where $I(X_i; \mathbf{X}_{\setminus i})$ is Shannon Mutual Information between X_i and $\mathbf{X}_{\setminus i}$. Uniformly smoothing $\Pr_e(X_i \mid \mathbf{X}_{\setminus i}, \mathcal{D})$ increases the conditional entropy $H(X_i|\mathbf{X}_{\setminus i})$, decreasing $I(X_i; \mathbf{X}_{\setminus i})$ as a result. Thus, LUGS weakens the link between X_i and $\mathbf{X}_{\setminus i}$ by increasing the conditional entropy. On the other hand, traditional data publishing methods such as k-anonimity and l-diversity decreases the entropy of $H(X_i)$ by generalization or suppression of the feature values. However, both LUGS and traditional data publishing methods try to reduce the mutual information between features.

5. Empirical Studies

In this section, we demonstrate the impact of LUGS algorithm using Medicare Claims records for both descriptive analysis and predictive modeling. Centers for Medicare and Medicaid Services (CMS) provides several public-use data files (PUF), such as inpatient

claims, line-items, drug events, etc. For our experiment, we use "BSA Inpatient Claims PUF"[2] describing Basic Stand Alone Inpatient records. The data file contains seven variables: ID, Gender, Age, DRG (drug code), ICD-9 (procedure code), Length (the length of stay), and Amount (payment), and has 15K rows. Age, Length, and Amount variables are originally numeric records, but CMS has categorized them into five quantiles. Note that data re-coding methods such as k-anonymity and l-diversity cannot be directly compared to LUGS, as their feature granularities are different. To the best of our knowledge, *LUGS is the first privacy-safe data synthesizer, which adheres to the rigorous privacy metric, "synthetic l-diversity".*

5.1. Sample Path and Marginal Distribution

We first show a sample path of Gibbs samples. Table 1 illustrates a random sample from the synthetic process. Figure 2 shows the first 300 Gibbs iterations of one synthetic sample. As can be seen, the Gibbs sample traverses over all the possible combinations of the feature space. From Figure 3, we can observe that the autocorrelations of the Gibbs samples are less than 0.1 after 5 iterations, suggesting that the LUGS samples are reasonably converged to a stationary distribution.

Table 1. Gibbs Samples over Iteration

(t)	Gender	Age	DRG	ICD9	DAYS	AMT
1	2	4	147	81	3	4
2	1	1	158	39	1	2
3	2	6	88	45	2	4
			⋮			

Figure 4 and 5 shows marginal histograms of ICD-9 procedure and drug codes over different levels of privacy metrics ($\log l$). We can observe uniform smoothing effects from the LUGS-generated synthetic data; rare data points are amplified.

5.2. K-way Correlation

K-way correlation ($K > 2$) between categorical variables can be visualized using log-linear analysis (Simkiss et al., 2012). Log-linear analysis can be viewed as a multi-way chi-square independence test. For example, if Age (i), Gender (j), and Amount (k) variables are cross-tabulated, then each cell frequency

[2]http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/BSAPUFS/Inpatient_Claims.html

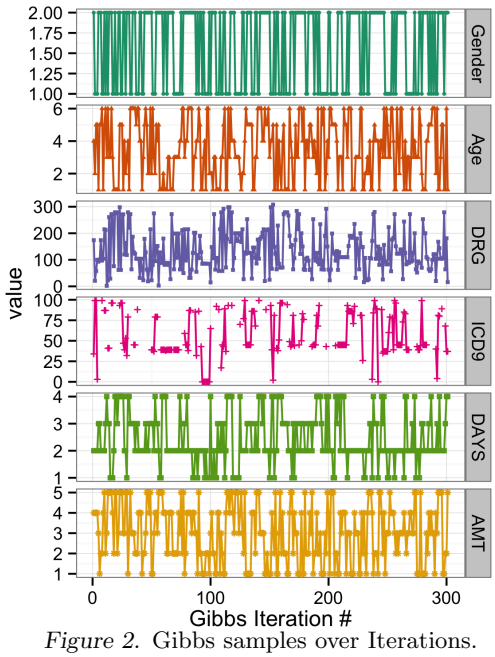

Figure 2. Gibbs samples over Iterations.

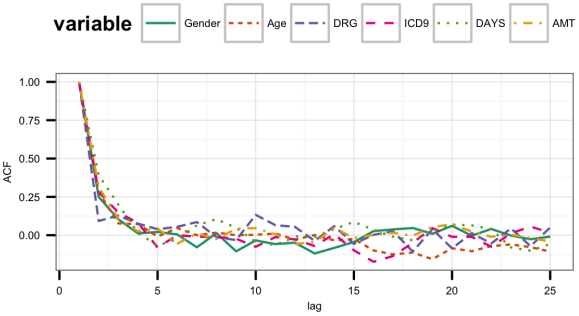

Figure 3. Auto-correlation Functions of Gibbs Samples.

F_{ijk} can be fully modeled as follows:

$$\log F_{ijk} = \underbrace{\lambda_0}_{\text{offset}} + \underbrace{\lambda_i + \lambda_j + \lambda_k}_{\text{independent effect}} + \underbrace{\lambda_{ij} + \lambda_{jk} + \lambda_{ik}}_{\text{2-way interaction terms}} + \underbrace{\lambda_{ijk}}_{\text{3-way interaction term}}$$

However, if the features are independent of each other, then the full description can be simplified:

$$\log F_{ijk} = \lambda_0 + \lambda_i + \lambda_j + \lambda_k \quad (3)$$

Log-linear analysis basically compares the log-likelihood values from these two models. Age, Gender, and Amount variables, are cross-tabulated, and their Pearson Chi-square (χ^2) statistics for each l-diversified synthetic dataset are plotted in Figure 6. The l-diversified synthetic datasets show less correlation among variables.

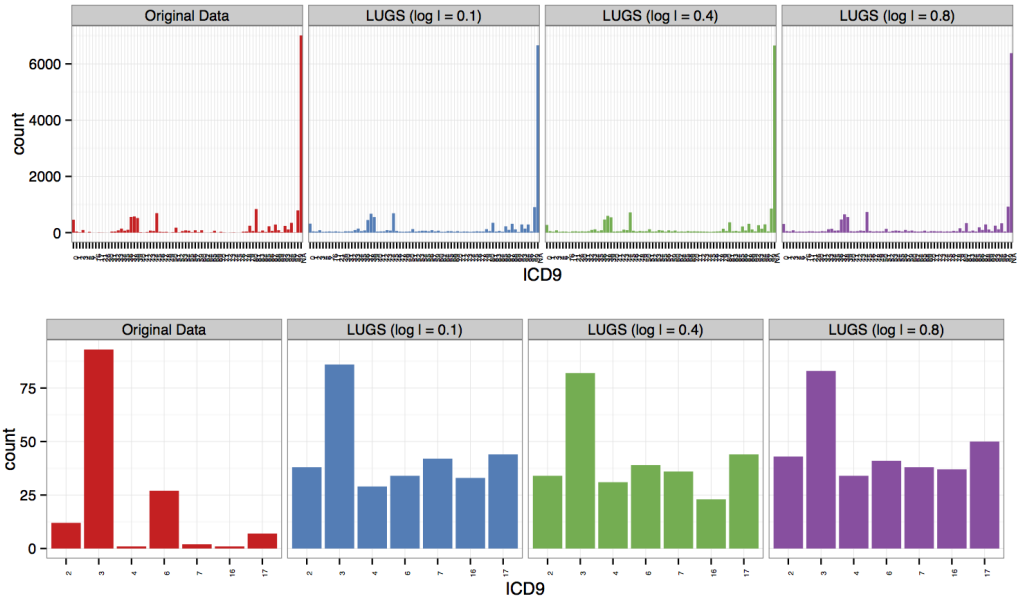

Figure 4. Marginal Histogram of ICD-9 Procedure Code vs. $\log l$. The full range of drug codes (top) and the first seven categories (bottom).

5.3. Missing Value Imputation

As a side effect of the l-diversified generative process, generated synthetic data tend to have less missing values. Figure 6 shows missing value percentages for each synthetic data set. While the original data ($\log l = 0.0$) exhibits $\approx 8\%$ missing rate, the generated synthetic data have less than 6% missing values.

5.4. Purely Artificial Records

LUGS generates "realistic but not real" synthetic data (non-support region records). The generated samples are "not real", since its l-diversified distributions have positive probabilities for any possible record combinations. Purely artificial records are the records that cannot be found in the original data. Figure 6 illustrates the ratios of these artificial records for each synthetic dataset. We emphasize that even with very small $\log l$ values, more than 60% of synthetic records are filled with purely artificial records.

5.5. Effect on Predictive Models

Privacy preserving synthetic data can be publicly disclosed to answer various types of data mining research questions. If the statistical properties of the original data are well preserved in the LUGS-generated data, then the data mining results from the synthetic data should be significantly identical to those from the original data. We demonstrate illustrative results of simple predictive modeling by arbitrarily setting Amount as a dependent variable:

$$\underbrace{\text{AMT}}_{y\text{: dependent variable}} \sim \underbrace{\text{Age} + \text{Gender} + \text{DRG} + \text{ICD9}}_{X\text{: independent variables}}$$

Dummy coding is applied to the categorical variables such as DRG and ICD9, resulting in 352 dependent variables. We apply the Lasso regression (Friedman et al., 2010) to deal with this high dimensionality, then measure Mean Square Errors (MSE) using the original data:

$$\beta^* = \min_{\beta} \|\mathbf{y}_{\text{synth}} - \mathbf{X}_{\text{synth}}\beta\|^2 + \lambda\|\beta\|$$
$$\text{MSE} = \|\mathbf{y}_{\text{orig}} - \mathbf{X}_{\text{orig}}\beta^*\|^2$$

Figure 7 (top) shows the measured MSE values over different λ and $\log l$ values. As can be seen, MSE values increase with the $\log l$ value, but their deviation from the original MSE is negligible compared to the spread of values at each privacy setting, $\log l$. We next show a classification example setting AMT > 3 as "positive" and AMT ≤ 3 as "negative". Logistic regression with the Lasso penalty is used. Figure 7 (bottom) shows the measured Area Under Receiver Operating Characteristic (AUC) curves. As in the regression example, classification results are very robust and consistent with the synthetic data.

6. Discussions

We proposed a novel data synthesizer that protects sensitive information by adhering to a prescribed l-

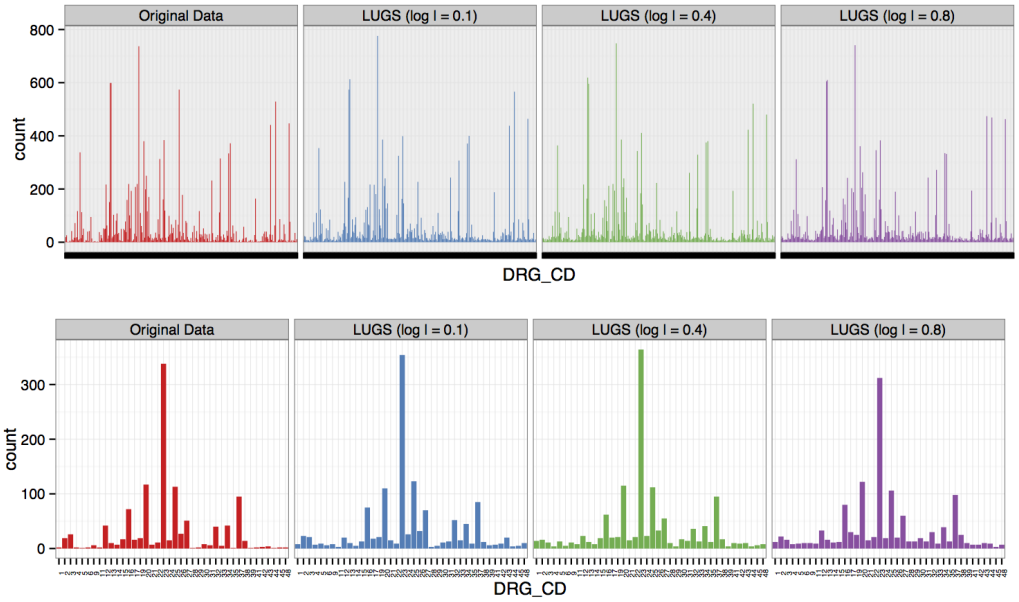

Figure 5. Marginal Histogram of Drug Code vs. $\log l$. The full range of drug codes (top) and the first 40 categories (bottom).

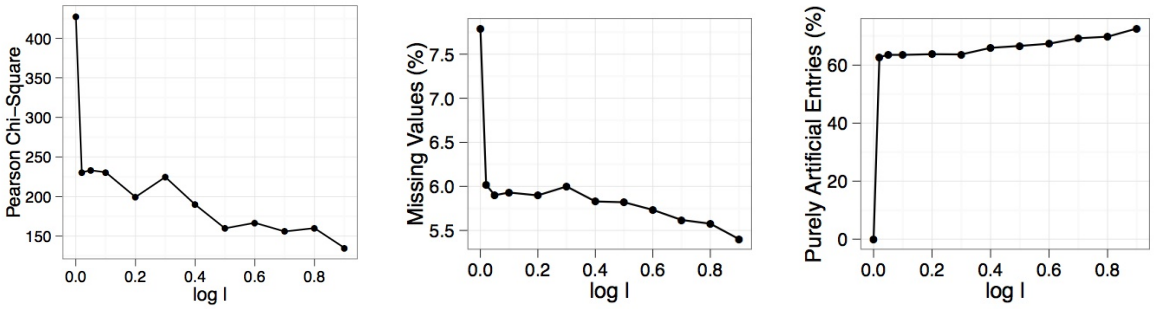

Figure 6. Pearson Chi-Square Statistics from Log-linear Analysis (left), Missing Value Ratio (center), and % of Purely Artificial Entries (right) in Synthetic Data as function of $\log l$. 3-way correlation decreases, as $\log l$ increases.

diversity privacy metric. The proposed algorithm, LUGS, scales linearly with respect to the size of the data. Furthermore, LUGS is a non-parametric and non-model based technique. This property assures that a wide range of data mining algorithms can be applied without considering the generative process of the synthetic data. Many health datasets are not available because of privacy restrictions such as HIPAA privacy regulation. We believe that the proposed solution can be an alternative way to facilitate collaborative healthcare research efforts.

References

Abowd, John M. and Vilhuber, Lars. How protective are synthetic data? *Privacy in Statistical Databases*, 5262:239–246, 2008.

Broder, Andrei Z. On the resemblance and containment of documents. In *Compression and Complexity of Sequences*, pp. 21–29, 1997.

Csiszar, Imre and Shields, Paul C. *Information Theory and Statistics: A Tutorial*. Now Publishers Inc., 2004.

Dwork, Cynthia. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming*, volume 4052, pp. 1–12, 2006.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 2010.

Gionis, A., Indyk, P., and Motwani, R. Similarity search in high dimensions via hashing. In *Proceedings of the 25th Very Large Database*, 1999.

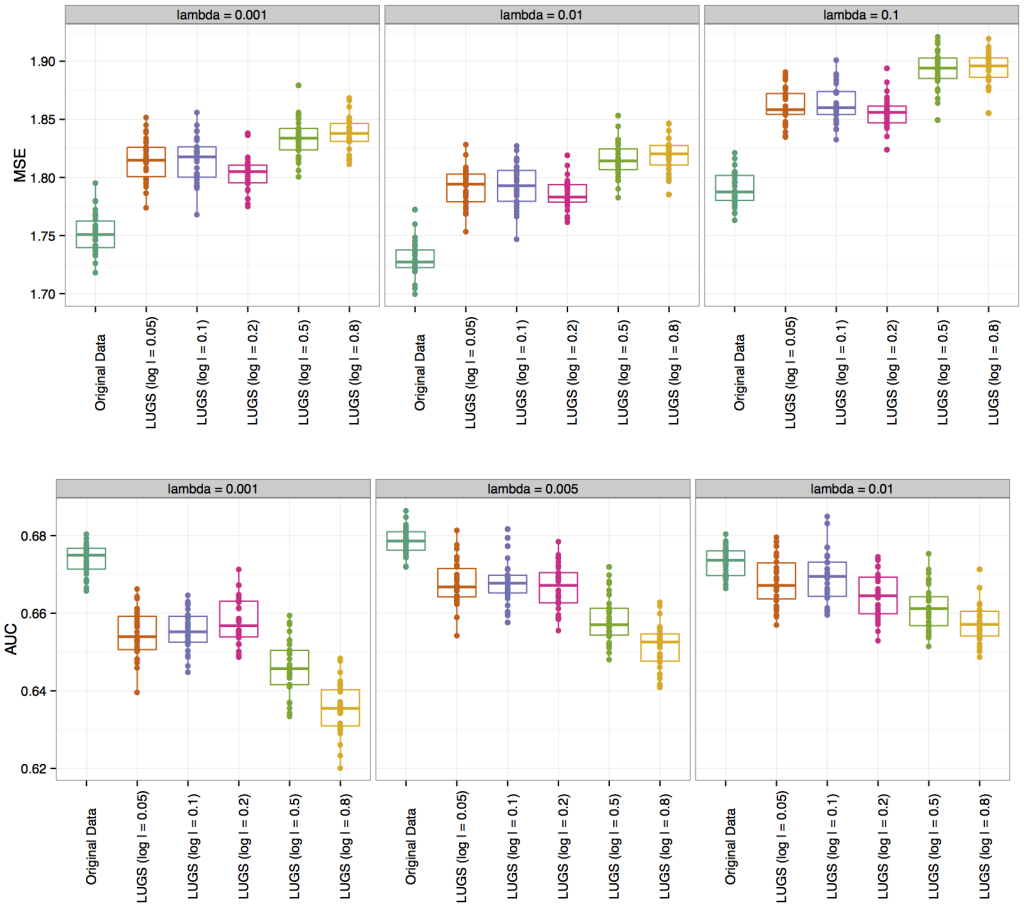

Figure 7. MSE (top) and AUC (bottom) vs. $\log l$ and λ (regularization parameter).

Grimmett, Geoffrey and Stirzaker, David. *Probability and Random Processes*, chapter 3.7, pp. 67. Oxford, third edition, 2001.

Indyk, P. and Motwani, R. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of 30th Symposium on Theory of Computing*, 1998.

Kifer, Daniel and Machanavajjhala, Ashwin. A rigorous and customizable framework for privacy. In *ACM Symposium on Principles of Database Systems*, 2012.

Machanavajjhala, Ashwin, Kifer, Daniel, Gehrke, Johannes, and Venkitasubramanian, Muthuramakrishnan. l-diversity: Privacy beyond k-anonymity. *Transactions on Knowledge Discovery from Data*, 1, 2007.

Narayanan, Arvind and Shmatikov, Vitaly. Robust de-anonymization of large sparse datasets. In *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 111–125, 2008.

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

Reiter, Jerome P., Raghunathan, Trivellore E., and Kinney, Satkartar K. The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32:143–149, 2006.

Simkiss, D., Ebrahim, G. J., and Waterson, A. J. R. Chapter 14: Analysing categorical data: Log-linear analysis. *Journal of Tropical Pediatrics*, 2012.

Sweeney, Latanya. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10:557–570, October 2002. ISSN 0218-4885.