Generic Probability Density Function Reconstruction for Randomization in Privacy-Preserving Data Mining

Vincent Y. F. Tan [1] See-Kiong Ng [2]

[1]Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

[2]Institute for Infocomm Research (I^2R), Singapore

Machine Learning and Data Mining MLDM 2007

Outline

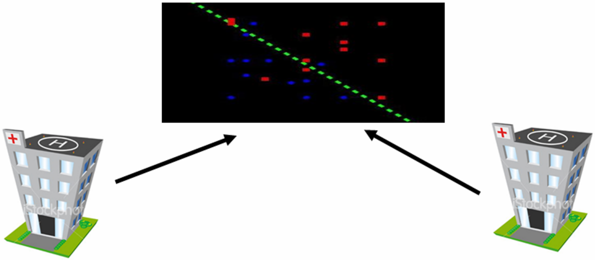
Outline

Outline

Outline

What is Privacy-Preserving Data Mining (PPDM)?

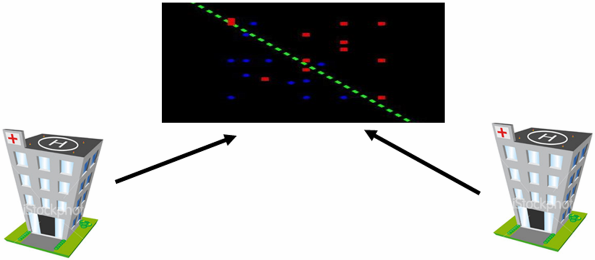
- Example: Two hospitals seek to construct a global classifier based on existing patient data.



- But patients' private data cannot be revealed.

What is Privacy-Preserving Data Mining (PPDM)?

- Example: Two hospitals seek to construct a global classifier based on existing patient data.



- But patients' private data cannot be revealed.

Randomization in PPDM

- Randomization
  1. Mask private data values by perturbing with noise.
  2. Task: To reconstruct the Probability Density Function (PDF) of the original dataset from the randomized data.
- Challenges: Two conflicting concerns.
  1. Confidentiality of the private information
  2. Utility of the aggregate statistics.

Randomization in PPDM

- Randomization
  1. Mask private data values by perturbing with noise.
  2. Task: To reconstruct the Probability Density Function (PDF) of the original dataset from the randomized data.
- Challenges: Two conflicting concerns.
  1. Confidentiality of the private information
  2. Utility of the aggregate statistics.

Randomization in PPDM

- Randomization
  1. Mask private data values by perturbing with noise.
  2. Task: To reconstruct the Probability Density Function (PDF) of the original dataset from the randomized data.

- Challenges: Two conflicting concerns.
  1. Confidentiality of the private information
  2. Utility of the aggregate statistics.

Randomization in PPDM

- Randomization
  1. Mask private data values by perturbing with noise.
  2. Task: To reconstruct the Probability Density Function (PDF) of the original dataset from the randomized data.
- Challenges: Two conflicting concerns.
  1. Confidentiality of the private information
  2. Utility of the aggregate statistics.

Randomization in PPDM

- Randomization
  1. Mask private data values by perturbing with noise.
  2. Task: To reconstruct the Probability Density Function (PDF) of the original dataset from the randomized data.
- Challenges: Two conflicting concerns.
  1. Confidentiality of the private information
  2. Utility of the aggregate statistics.

Related Work

- Agrawal et al. (2000) applied noise (e_i) to the true data (x_i) and transmit the sum $z_i = x_i + e_i$.
- Reconstruction of $f_X(x)$ (PDF of X) via EM.
- Kargupta et al. (2003) showed that such noise addition risk privacy breaches.
- We suggest a generic noise randomization model, to minimize privacy risk.
- We suggest a non-iterative PDF reconstruction algorithm.
- Other methods: k-anonymity (Sweeney, 2002), Secure Multi-Party Computation (Pinkas, 2002).

Related Work

- Agrawal et al. (2000) applied noise (e_i) to the true data (x_i) and transmit the sum $z_i = x_i + e_i$.
- Reconstruction of $f_X(x)$ (PDF of X) via EM.
- Kargupta et al. (2003) showed that such noise addition risk privacy breaches.
- We suggest a generic noise randomization model, to minimize privacy risk.
- We suggest a non-iterative PDF reconstruction algorithm.
- Other methods: k-anonymity (Sweeney, 2002), Secure Multi-Party Computation (Pinkas, 2002).

Related Work

- Agrawal et al. (2000) applied noise (e_i) to the true data (x_i) and transmit the sum $z_i = x_i + e_i$.
- Reconstruction of $f_X(x)$ (PDF of X) via EM.
- Kargupta et al. (2003) showed that such noise addition risk privacy breaches.
- We suggest a generic noise randomization model, to minimize privacy risk.
- We suggest a non-iterative PDF reconstruction algorithm.
- Other methods: k-anonymity (Sweeney, 2002), Secure Multi-Party Computation (Pinkas, 2002).

Related Work

- Agrawal et al. (2000) applied noise (e_i) to the true data (x_i) and transmit the sum $z_i = x_i + e_i$.
- Reconstruction of $f_X(x)$ (PDF of X) via EM.
- Kargupta et al. (2003) showed that such noise addition risk privacy breaches.
- We suggest a generic noise randomization model, to minimize privacy risk.
- We suggest a non-iterative PDF reconstruction algorithm.
- Other methods: k-anonymity (Sweeney, 2002), Secure Multi-Party Computation (Pinkas, 2002).

Related Work

- Agrawal et al. (2000) applied noise (e_i) to the true data (x_i) and transmit the sum $z_i = x_i + e_i$.
- Reconstruction of $f_X(x)$ (PDF of X) via EM.
- Kargupta et al. (2003) showed that such noise addition risk privacy breaches.
- We suggest a generic noise randomization model, to minimize privacy risk.
- We suggest a non-iterative PDF reconstruction algorithm.
- Other methods: k-anonymity (Sweeney, 2002), Secure Multi-Party Computation (Pinkas, 2002).

Related Work

- Agrawal et al. (2000) applied noise (e_i) to the true data (x_i) and transmit the sum $z_i = x_i + e_i$.
- Reconstruction of $f_X(x)$ (PDF of X) via EM.
- Kargupta et al. (2003) showed that such noise addition risk privacy breaches.
- We suggest a generic noise randomization model, to minimize privacy risk.
- We suggest a non-iterative PDF reconstruction algorithm.
- Other methods: k-anonymity (Sweeney, 2002), Secure Multi-Party Computation (Pinkas, 2002).

Problem Definition + Notation

- PPDM framework: Randomization + Reconstruction.
- N original scalars $\{x_i\}_{i=1}^N$, drawn from IID random variables (RV) $\{X_i\}_{i=1}^N \sim f_X(x)$.

$$z_i = \mathcal{Z}(e_i, x_i), \qquad \forall i \in \{1, \ldots, N\}$$

- $\{e_i\}_{i=1}^N$ are realizations of IID noise RVs $\{E_i\}_{i=1}^N \sim f_E(e)$.
- E statistically independent of X.
- Task: *Given the randomized values* $\{z_i\}_{i=1}^N$ *and* $f_E(e)$*, estimate original PDF* $\hat{f}_X(x)$ *for arbitrary* $\mathcal{Z}(\cdot, \cdot)$.

Problem Definition + Notation

- PPDM framework: Randomization + Reconstruction.
- N original scalars $\{x_i\}_{i=1}^N$, drawn from IID random variables (RV) $\{X_i\}_{i=1}^N \sim f_X(x)$.

$$z_i = \mathcal{Z}(e_i, x_i), \qquad \forall i \in \{1, \ldots, N\}$$

- $\{e_i\}_{i=1}^N$ are realizations of IID noise RVs $\{E_i\}_{i=1}^N \sim f_E(e)$.
- E statistically independent of X.
- Task: *Given the randomized values $\{z_i\}_{i=1}^N$ and $f_E(e)$, estimate original PDF $\hat{f}_X(x)$ for arbitrary $\mathcal{Z}(\cdot,\cdot)$.*

Problem Definition + Notation

- PPDM framework: Randomization + Reconstruction.
- N original scalars $\{x_i\}_{i=1}^{N}$, drawn from IID random variables (RV) $\{X_i\}_{i=1}^{N} \sim f_X(x)$.

$$z_i = \mathcal{Z}(e_i, x_i), \qquad \forall i \in \{1, \ldots, N\}$$

- $\{e_i\}_{i=1}^{N}$ are realizations of IID noise RVs $\{E_i\}_{i=1}^{N} \sim f_E(e)$.
- E statistically independent of X.
- Task: *Given the randomized values* $\{z_i\}_{i=1}^{N}$ *and* $f_E(e)$, *estimate original PDF* $\hat{f}_X(x)$ *for arbitrary* $\mathcal{Z}(\cdot, \cdot)$.

Problem Definition + Notation

- PPDM framework: Randomization + Reconstruction.
- N original scalars $\{x_i\}_{i=1}^N$, drawn from IID random variables (RV) $\{X_i\}_{i=1}^N \sim f_X(x)$.

$$z_i = \mathcal{Z}(e_i, x_i), \qquad \forall i \in \{1, \ldots, N\}$$

- $\{e_i\}_{i=1}^N$ are realizations of IID noise RVs $\{E_i\}_{i=1}^N \sim f_E(e)$.
- E statistically independent of X.
- Task: *Given the randomized values* $\{z_i\}_{i=1}^N$ *and* $f_E(e)$, *estimate original PDF* $\hat{f}_X(x)$ *for arbitrary* $\mathcal{Z}(\cdot,\cdot)$.

Problem Definition + Notation

- PPDM framework: Randomization + Reconstruction.
- N original scalars $\{x_i\}_{i=1}^N$, drawn from IID random variables (RV) $\{X_i\}_{i=1}^N \sim f_X(x)$.

$$z_i = \mathcal{Z}(e_i, x_i), \qquad \forall i \in \{1, \ldots, N\}$$

- $\{e_i\}_{i=1}^N$ are realizations of IID noise RVs $\{E_i\}_{i=1}^N \sim f_E(e)$.
- E statistically independent of X.
- Task: *Given the randomized values* $\{z_i\}_{i=1}^N$ *and* $f_E(e)$, *estimate original PDF* $\hat{f}_X(x)$ *for arbitrary* $\mathcal{Z}(\cdot,\cdot)$.

Estimate $f_Z(z)$ via Parzen Windows

- The Parzen-Window approximation of the PDF of the perturbed samples $\{z_i\}_{i=1}^N$ is

$$\hat{f}_Z(z) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{N}(z, z_i, \sigma_p^2).$$

- Quality of estimator depends largely on N and σ_p.
- Choose σ_p via a cross-validation scheme.

Estimate $f_Z(z)$ via Parzen Windows

- The Parzen-Window approximation of the PDF of the perturbed samples $\{z_i\}_{i=1}^N$ is

$$\hat{f}_Z(z) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{N}(z, z_i, \sigma_p^2).$$

- Quality of estimator depends largely on N and σ_p.
- Choose σ_p via a cross-validation scheme.

Estimate $f_Z(z)$ via Parzen Windows

- The Parzen-Window approximation of the PDF of the perturbed samples $\{z_i\}_{i=1}^N$ is

$$\hat{f}_Z(z) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{N}(z, z_i, \sigma_p^2).$$

- Quality of estimator depends largely on N and σ_p.
- Choose σ_p via a cross-validation scheme.

Illustration of Parzen Windows

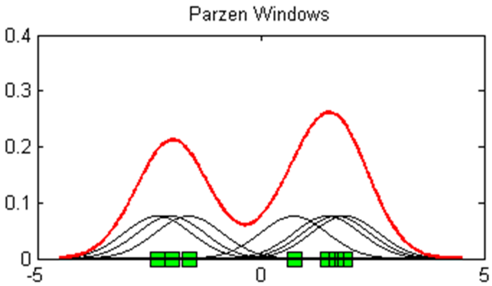


Figure: Illustration of Parzen-Windows for estimation of the multimodal PDF.

Estimate $f_X(x)$ via Quadratic Programming (QP)

- Applying (i) the theory of transformation of RVs and (ii) discretizing the space (See our paper).
- A QP can be formulated.

$$\min_{\mathbf{f}_X \in \mathcal{C}} \quad J(\mathbf{f}_X) = \frac{1}{2}\mathbf{f}_X^{\mathrm{T}}\mathbf{H}\mathbf{f}_X + \mathbf{h}^{\mathrm{T}}\mathbf{f}_X,$$

- Constraints are given by

$$\mathbf{f}_X \geq \mathbf{0}_{N_X \times 1}, \quad \sum_{n\Delta x \in \mathcal{D}_X} f_X(n\Delta x) = \frac{1}{\Delta x}.$$

- Natural question: Is it a convex program?

Estimate $f_X(x)$ via Quadratic Programming (QP)

- Applying (i) the theory of transformation of RVs and (ii) discretizing the space (See our paper).
- A QP can be formulated.

$$\min_{\mathbf{f}_X \in \mathcal{C}} \quad J(\mathbf{f}_X) = \frac{1}{2}\mathbf{f}_X^{\mathrm{T}}\mathbf{H}\mathbf{f}_X + \mathbf{h}^{\mathrm{T}}\mathbf{f}_X,$$

- Constraints are given by

$$\mathbf{f}_X \geq \mathbf{0}_{N_X \times 1}, \qquad \sum_{n\Delta x \in \mathcal{D}_X} f_X(n\Delta x) = \frac{1}{\Delta x}.$$

- Natural question: Is it a convex program?

Estimate $f_X(x)$ via Quadratic Programming (QP)

- Applying (i) the theory of transformation of RVs and (ii) discretizing the space (See our paper).
- A QP can be formulated.

$$\min_{\mathbf{f}_X \in \mathcal{C}} \quad J(\mathbf{f}_X) = \frac{1}{2}\mathbf{f}_X^{\mathrm{T}}\mathbf{H}\mathbf{f}_X + \mathbf{h}^{\mathrm{T}}\mathbf{f}_X,$$

- Constraints are given by

$$\mathbf{f}_X \geq \mathbf{0}_{N_X \times 1}, \quad \sum_{n\Delta x \in \mathcal{D}_X} f_X(n\Delta x) = \frac{1}{\Delta x}.$$

- Natural question: Is it a convex program?

Convexity

- Cost function and constraint set $\mathcal{C}$ are convex.

$$\mathcal{C} = \left\{ \mathbf{f}_X \,\middle|\, \mathbf{f}_X \geq \mathbf{0}, \sum_{n=1}^{N_x} [\mathbf{f}_X]_n = \frac{1}{\Delta x} \right\}$$

- $\Rightarrow$ Convex Programming/Optimization.
- Necessary conditions are sufficient conditions for optimality.

$$[\mathbf{f}_X^*]_i > 0 \Rightarrow \frac{\partial J(\mathbf{f}_X^*)}{\partial [\mathbf{f}_X]_i} < \frac{\partial J(\mathbf{f}_X^*)}{\partial [\mathbf{f}_X]_j} \quad \forall j.$$

Convexity

- Cost function and constraint set $\mathcal{C}$ are convex.

$$\mathcal{C} = \left\{ \mathbf{f}_X \;\middle|\; \mathbf{f}_X \geq \mathbf{0}, \sum_{n=1}^{N_x} [\mathbf{f}_X]_n = \frac{1}{\Delta x} \right\}$$

- $\Rightarrow$ Convex Programming/Optimization.
- Necessary conditions are sufficient conditions for optimality.

$$[\mathbf{f}_X^*]_i > 0 \Rightarrow \frac{\partial J(\mathbf{f}_X^*)}{\partial [\mathbf{f}_X]_i} < \frac{\partial J(\mathbf{f}_X^*)}{\partial [\mathbf{f}_X]_j} \quad \forall j.$$

Convexity

- Cost function and constraint set $\mathcal{C}$ are convex.

$$\mathcal{C} = \left\{ \mathbf{f}_X \;\middle|\; \mathbf{f}_X \geq \mathbf{0}, \sum_{n=1}^{N_x} [\mathbf{f}_X]_n = \frac{1}{\Delta x} \right\}$$

- $\Rightarrow$ Convex Programming/Optimization.
- Necessary conditions are sufficient conditions for optimality.

$$[\mathbf{f}_X^*]_i > 0 \Rightarrow \frac{\partial J(\mathbf{f}_X^*)}{\partial [\mathbf{f}_X]_i} < \frac{\partial J(\mathbf{f}_X^*)}{\partial [\mathbf{f}_X]_j} \quad \forall j.$$

Privacy Loss and Information Loss

- Quantify privacy loss using mutual information (Agrawal et al. 2000).

$$\mathcal{P}(X|Z) \triangleq 1 - 2^{-I(X;Z)}.$$

- $0 \leq \mathcal{P}(X|Z) \leq 1$.
- Information Loss is a measure of the accuracy of the PDF reconstruction algorithm using

$$\mathcal{I}(f_X, \hat{f}_X) \triangleq \frac{1}{2}\mathbf{E}\left[\int_{\mathcal{D}_X} \left| f_X(x) - \hat{f}_X(x) \right| dx\right],$$

- $0 \leq \mathcal{I}(f_X, \hat{f}_X) \leq 1$.

Privacy Loss and Information Loss

- Quantify privacy loss using mutual information (Agrawal et al. 2000).

$$\mathcal{P}(X|Z) \triangleq 1 - 2^{-I(X;Z)}.$$

- $0 \leq \mathcal{P}(X|Z) \leq 1.$
- Information Loss is a measure of the accuracy of the PDF reconstruction algorithm using

$$\mathcal{I}(f_X, \hat{f}_X) \triangleq \frac{1}{2}\mathbf{E}\left[\int_{\mathcal{D}_X} \left| f_X(x) - \hat{f}_X(x) \right| \, dx\right],$$

- $0 \leq \mathcal{I}(f_X, \hat{f}_X) \leq 1.$

Experimental Setup/Data

- Multiplicative and additive randomization models used.
- $N = 500$.
- Original PDF $f_X(x)$ is Gaussian.
- Noise is Uniform.
- Varied σ_e to get different Privacy Loss/Info Loss points.

Experimental Setup/Data

- Multiplicative and additive randomization models used.
- $N = 500$.
- Original PDF $f_X(x)$ is Gaussian.
- Noise is Uniform.
- Varied σ_e to get different Privacy Loss/Info Loss points.

Experimental Setup/Data

- Multiplicative and additive randomization models used.
- $N = 500$.
- Original PDF $f_X(x)$ is Gaussian.
- Noise is Uniform.
- Varied σ_e to get different Privacy Loss/Info Loss points.

Experimental Setup/Data

- Multiplicative and additive randomization models used.
- $N = 500$.
- Original PDF $f_X(x)$ is Gaussian.
- Noise is Uniform.
- Varied σ_e to get different Privacy Loss/Info Loss points.

Experimental Setup/Data

- Multiplicative and additive randomization models used.
- $N = 500$.
- Original PDF $f_X(x)$ is Gaussian.
- Noise is Uniform.
- Varied σ_e to get different Privacy Loss/Info Loss points.

Tradeoff curves

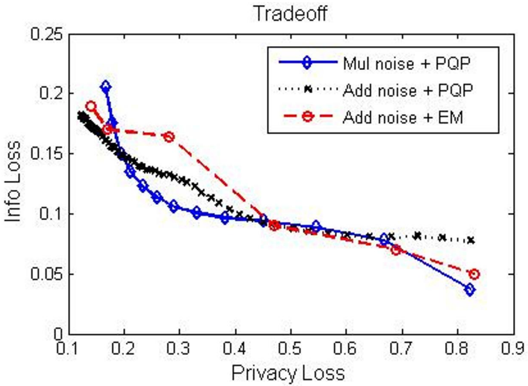


Figure: Our PDF reconstruction algorithm (‘PQP’) performs just as well as EM but has the added bonus of being a generic, non-iterative reconstruction method.

US Housing Dept data

- Real data obtained from The U.S. Department of Housing and Urban Development's (USDHUD's).
- Median income of all the counties in the 50 states in the U.S in 2005.
- Multiplicative and additive randomization models used.
- $N = 3195$.
- Histogram with 75 bins.
- Noise is Uniform.
- Privacy loss is kept constant at $\mathcal{P}(X|Z) = 0.330$.

US Housing Dept data

- Real data obtained from The U.S. Department of Housing and Urban Development's (USDHUD's).
- Median income of all the counties in the 50 states in the U.S in 2005.
- Multiplicative and additive randomization models used.
- $N = 3195$.
- Histogram with 75 bins.
- Noise is Uniform.
- Privacy loss is kept constant at $\mathcal{P}(X|Z) = 0.330$.

US Housing Dept data

- Real data obtained from The U.S. Department of Housing and Urban Development's (USDHUD's).
- Median income of all the counties in the 50 states in the U.S in 2005.
- Multiplicative and additive randomization models used.
- $N = 3195$.
- Histogram with 75 bins.
- Noise is Uniform.
- Privacy loss is kept constant at $\mathcal{P}(X|Z) = 0.330$.

US Housing Dept data

- Real data obtained from The U.S. Department of Housing and Urban Development's (USDHUD's).
- Median income of all the counties in the 50 states in the U.S in 2005.
- Multiplicative and additive randomization models used.
- $N = 3195$.
- Histogram with 75 bins.
- Noise is Uniform.
- Privacy loss is kept constant at $\mathcal{P}(X|Z) = 0.330$.

US Housing Dept data

- Real data obtained from The U.S. Department of Housing and Urban Development's (USDHUD's).
- Median income of all the counties in the 50 states in the U.S in 2005.
- Multiplicative and additive randomization models used.
- $N = 3195$.
- Histogram with 75 bins.
- Noise is Uniform.
- Privacy loss is kept constant at $\mathcal{P}(X|Z) = 0.330$.

US Housing Dept data

- Real data obtained from The U.S. Department of Housing and Urban Development's (USDHUD's).
- Median income of all the counties in the 50 states in the U.S in 2005.
- Multiplicative and additive randomization models used.
- $N = 3195$.
- Histogram with 75 bins.
- Noise is Uniform.
- Privacy loss is kept constant at $\mathcal{P}(X|Z) = 0.330$.

Reconstructed Histograms

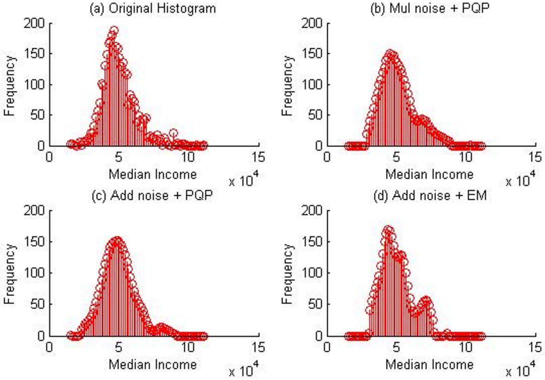


Figure: Comparison among different randomization / reconstruction schemes.

Summary and Advantages

- Devised a novel PDF reconstruction algorithm for privacy-preserving data mining.
- Our non-iterative algorithm eliminated the common need for the iterative EM algorithm.
- Our reconstruction method is also generic i.e. for all randomization models $z_i = \mathcal{Z}(e_i, x_i)$.

Summary and Advantages

- Devised a novel PDF reconstruction algorithm for privacy-preserving data mining.
- Our non-iterative algorithm eliminated the common need for the iterative EM algorithm.
- Our reconstruction method is also generic i.e. for all randomization models $z_i = \mathcal{Z}(e_i, x_i)$.

Summary and Advantages

- Devised a novel PDF reconstruction algorithm for privacy-preserving data mining.
- Our non-iterative algorithm eliminated the common need for the iterative EM algorithm.
- Our reconstruction method is also generic i.e. for all randomization models $z_i = \mathcal{Z}(e_i, x_i)$.

Further Work

- Different privacy loss metrics address different problems.
- Does a fundamental relation between the privacy loss and information loss exist?

- I would like to thank the support of the Agency for Science, Technology and Research (A*STAR), Singapore.

- http://web.mit.edu/vtan/www.

Further Work

- Different privacy loss metrics address different problems.
- Does a fundamental relation between the privacy loss and information loss exist?

- I would like to thank the support of the Agency for Science, Technology and Research (A*STAR), Singapore.
- http://web.mit.edu/vtan/www.

Further Work

- Different privacy loss metrics address different problems.
- Does a fundamental relation between the privacy loss and information loss exist?

- I would like to thank the support of the Agency for Science, Technology and Research (A*STAR), Singapore.

- `http://web.mit.edu/vtan/www`.