An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining

Kun Liu, Chris Giannella, and Hillol Kargupta

University of Maryland Baltimore County (UMBC)
Baltimore, Maryland, USA

PKDD 2006, Berlin, Germany

Talk Outline

- Background
- Distance Preserving Perturbation
- Privacy Breach
- Known Input-Output Attack
- Known Sample Attack
- Conclusions

Background

- Application Scenario
  - Governmental and commercial organizations need to disseminate data for research or business-related applications.
  - Data owners are concerned about the privacy of their data, and not willing to release it in plain.
  - Data perturbation (randomization) strives to provide a solution to this dilemma.
- Existing Perturbation Approach
  - Additive noise perturbation, data condensation, data anonymization, data swapping, sampling, etc.
  - They do not preserve Euclidean distance of the original data exactly.

Distance Preserving Perturbation

- Dist. preserving perturbation
  $T : \mathbb{R}^n \to \mathbb{R}^n$ if $\forall x, y \in \mathbb{R}^n, \; \| x - y \| = \| T(x) - T(y) \|$

- Dist. preserving perturbation is equivalent to
  $x \in \mathbb{R}^n \to Mx + v$, for $M \in \mathrm{O}_n$ and $v \in \mathbb{R}^n$,
  where O_n is the set of all $n \times n$ orthogonal matrices.

- Dist. preserving perturbation with origin fixed
  $x \in \mathbb{R}^n \to Mx$, where $M \in \mathrm{O}_n$ ⟷ Orthogonal Transformation

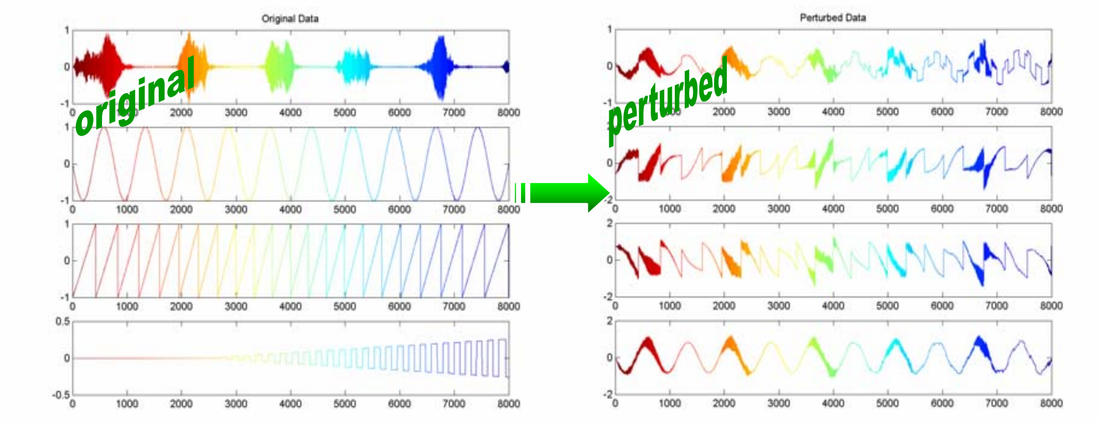
  Today's Talk

Dist. Preserving Perturbation for Privacy Preserving Data Mining

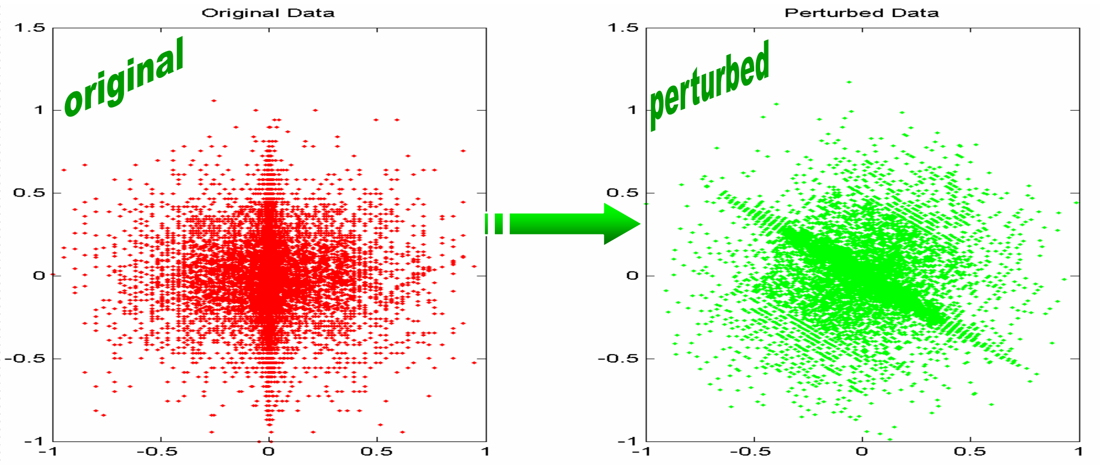
- Perturbation Model $Y = MX$
  - X: original private data with each column a record
  - Y: perturbed data
  - M: perturbation matrix
- Many data mining algorithms can be *efficiently* applied to the perturbed data and produce *exactly the same* results as if applied to the original data.
  - Clustering:[Oliveira04]
  - Classification: [Chen05]
  - Other related: [Liu06],[Mukherjee06], etc.

Dist. Preserving Perturbation Examples

Attributes



Records



Is Dist. Preserving Perturbation Secure?

- Attacker has No Prior Knowledge about Data
  - Very little can be done to accurately estimate X
- Two Types of Attacker's Prior Knowledge
  - Known Input-Output: The attacker knows some collection of linearly independent private data records and their corresponding perturbed version.
  - Known Sample: The attacker has a collection of independent data samples from the same distribution the original data was drawn.
- Two Types of Attack Techniques
  - Known Input-Output Attack: linear algebra, statistics
  - Known Sample Attack: principal component analysis

Privacy Breach

- Privacy Breach

  For any $\varepsilon > 0$, we say that an ε*-privacy breach* occurs if

$$\| \hat{x} - x_{\hat{i}} \| \leq \| x_{\hat{i}} \| \varepsilon$$

  where $\hat{x}$ is the attacker's estimate of $x_{\hat{i}}$, the $\hat{i}^{th}$ data tuple in X,

- Probability of Privacy Breach

$$\rho(x_{\hat{i}}, \varepsilon) = \text{Prob}\{ \| \hat{x} - x_{\hat{i}} \| \leq \| x_{\hat{i}} \| \varepsilon \}$$

  the probability that an ε*-privacy breach* occurs.

Known Input-Output Attack

$$[Y_{n\times k} \quad Y_{n\times(m-k)}] = M_{n\times n}[X_{n\times k} \quad X_{n\times(m-k)}]$$

KNOWN

- Assumption (can be relaxed): rank(X_{nxk})=k
- If k=n:
  - $M = Y_{n\times k} X^{-1}{}_{n\times k}$, $X_{n\times(m-k)} = M^T Y_{n\times(m-k)}$
  - Probability of privacy breach $\rho(x_{\hat{i}}, \varepsilon) = 1$ for $\varepsilon = 0$ and any $\hat{i}$.
  - The attacker has a perfect recovery of the private data.
- If k<n, what is going to happen?

Known Input-Output Attack

$$[Y_{n\times k} \quad Y_{n\times(m-k)}] = M_{n\times n}[X_{n\times k} \quad X_{n\times(m-k)}]$$

KNOWN

- If k<n, any matrix $\hat{M}$ in the set

  $$\Omega = \{\hat{M} \in O_n : \hat{M}X_{n\times k} = Y_{n\times k}\}$$

  can be the original perturbation matrix $M_{n\times n}$, where is O_n is the set of all nxn orthogonal matrices.

- The attacker chooses one uniformly from Ω as an estimation of $M_{n\times n}$, uses that to recover other private data, and computes the probability of privacy breach.

Known Input-Output Attack

- Probability of Privacy Breach

$$\rho(x_{\hat{i}},\varepsilon) = \text{Prob}\{\|\hat{x} - x_{\hat{i}}\| \le \|x_{\hat{i}}\|\varepsilon\}$$
$$= \text{Prob}\{\|\hat{M}Mx_{\hat{i}} - x_{\hat{i}}\| \le \|x_{\hat{i}}\|\varepsilon\}$$
$$= \begin{cases} \frac{1}{\pi} 2\arcsin\left(\frac{\|x_{\hat{i}}\|\varepsilon}{2d(x_{\hat{i}}, X_{n\times k})}\right) & \text{if } \|x_{\hat{i}}\|\varepsilon < 2d(x_{\hat{i}}, X_{n\times k})\ ; \\ 1 & \text{otherwise.} \end{cases}$$

where $d(x_{\hat{i}}, X_{n\times k})$ is the distance of $x_{\hat{i}}$ from the column space of $X_{n\times k}$,

and $\hat{M}$ is uniformly chosen from $\Omega = \{\hat{M} \in O_n : MX_{n\times k} = Y_{n\times k}\}$.

Known Input-Output Attack

- Properties of the Probability of Privacy Breach
  - Attacker can compute the probability of privacy breach for a given private record and a relative error bound ε .
  - The larger the ε , the higher the probability of privacy breach.
  - The closer the private record is to the column space of the known records, the higher the probability of privacy breach.
  - The distance $d(x_{\hat{i}}, X_{n \times k})$ can be computed from the perturbed data.

Known Input-Output Attack Example

Private Data X:

X_1	X_2	X_3
25.0000	30.0000	45.0000
75.0000	90.0000	105.0000

X_1 -> Y_1 KNOWN

UNKNOWN

Perturbed Data Y:

Y_1	Y_2	Y_3
-42.0198	-50.4237	-68.5443
66.9652	80.3582	91.3875

- The distance of X_2 from the column space of X_1 is 0, therefore $\rho(x_2, \varepsilon) = 1$ for any ε.
- The distance of X_3 from the column space of X_1 is 9.4868, therefore $\rho(x_3, \varepsilon) = \frac{1}{\pi} 2\arcsin\left(\frac{\| x_3 \| \varepsilon}{2 \times 9.4868}\right)$, e.g. $\rho(x_3, 0.01) = 3.84\%$.

Known Sample Attack

- Assumptions
  - Each data record arose as an independent sample from some unknown distribution
  - The attacker has a collection of samples independently chosen from the same distribution
  - The covariance of the distribution has all distinct eigenvalues (holds true in most practical situations [Jolliffe02]).
- Attack Technique
  - Exploring the relationship between the principal eigenvectors of the original data and the principal eigenvectors of the perturbed data.

Known Sample Attack

- The principal eigenvectors of the original data have experienced the same distance preserving perturbation as the data itself.

  Let $Y = MX$, we have $Z_Y = MZ_X D$,

  where Z_Y is the eigenvector matrix of the covariance of Y;

  Z_X is the eigenvector matrix of the covariance of X;

  and D is a diagonal matrix with each entry on the diagonal ± 1.

- Z_Y can be computed from the perturbed data, Z_X can be estimated from the sample data. (See the paper for choice of D, details omitted.)
- Attacker uses Z_X, Z_Y and D to recover M, and therefore X.

Known Sample Attack

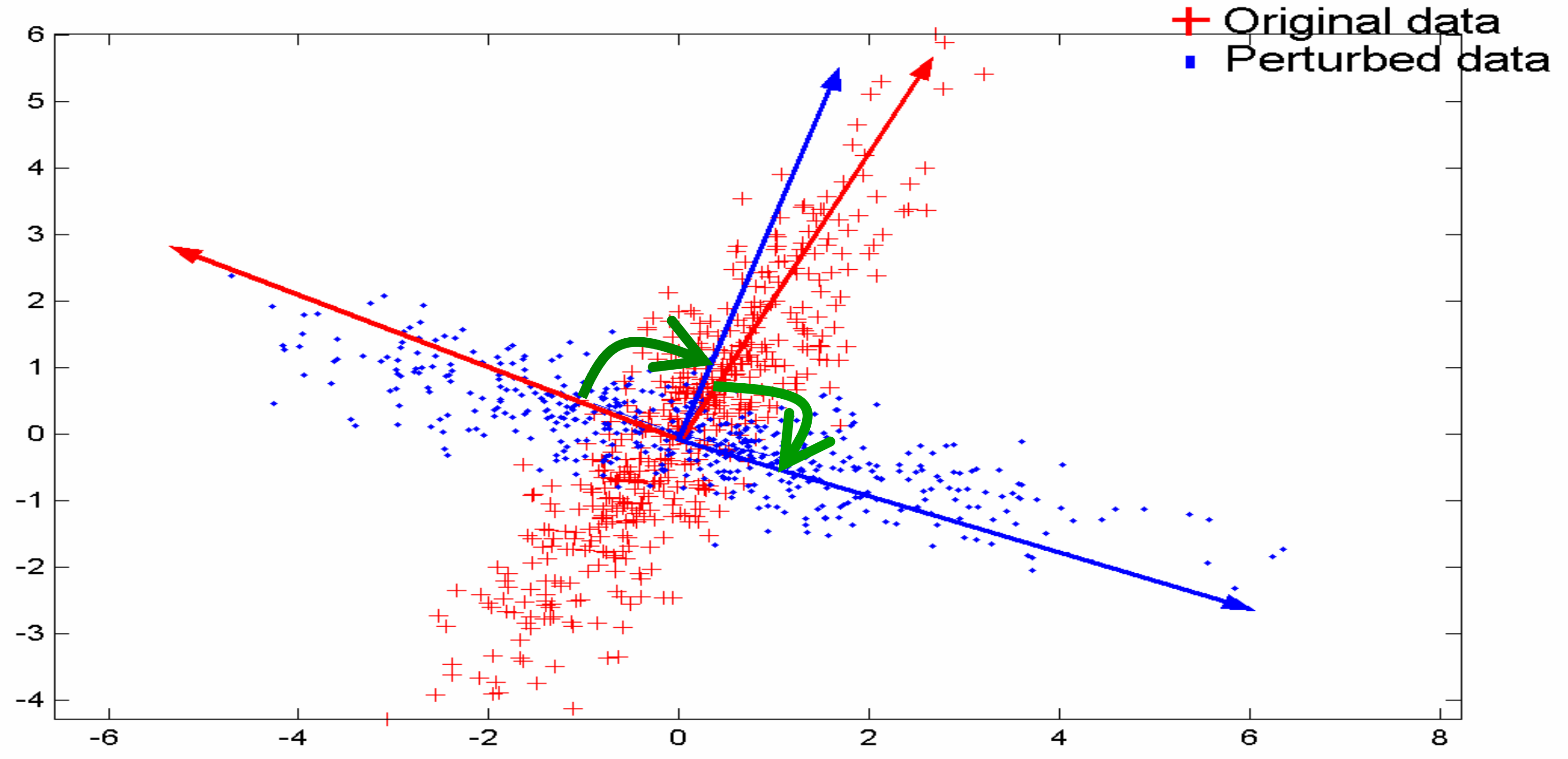


Fig. Relationship between original and perturbed principal eigenvectors.

Known Sample Attack Experiments

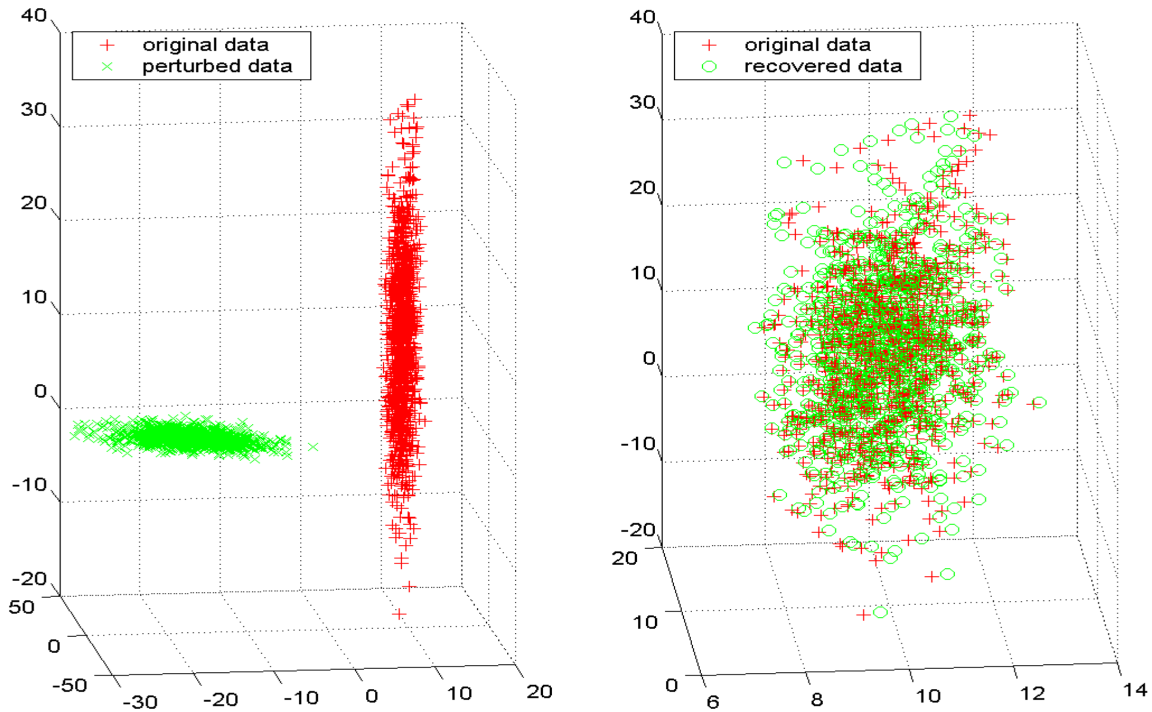


Fig. Known sample attack for 3D Gaussian data with 10,000 private tuples. The attacker has 2% samples from the same distribution. The average relative error of the recovered data is 0.0265 (2.65%).

Known Sample Attack Experiments

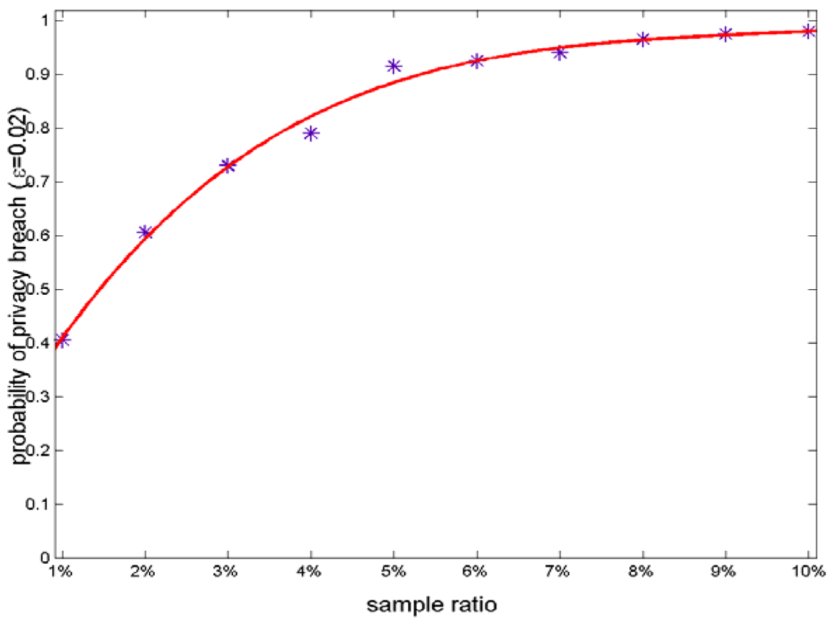


Fig. Probability of privacy breach w.r.t. attacker's sample size. The relative error bound ε is fixed to be 0.02. (3D Gaussian data with 10,000 private tuples.)

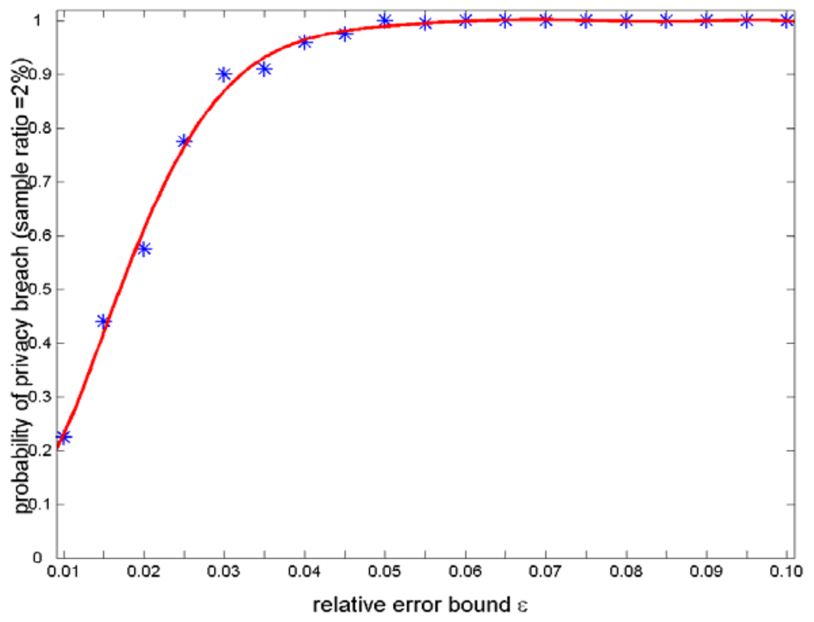


Fig. Probability of privacy breach w.r.t. the relative error bound ε. The sample ratio is fixed to be 2%. (3D Gaussian data with 10,000 private tuples.)

Known Sample Attack Experiments

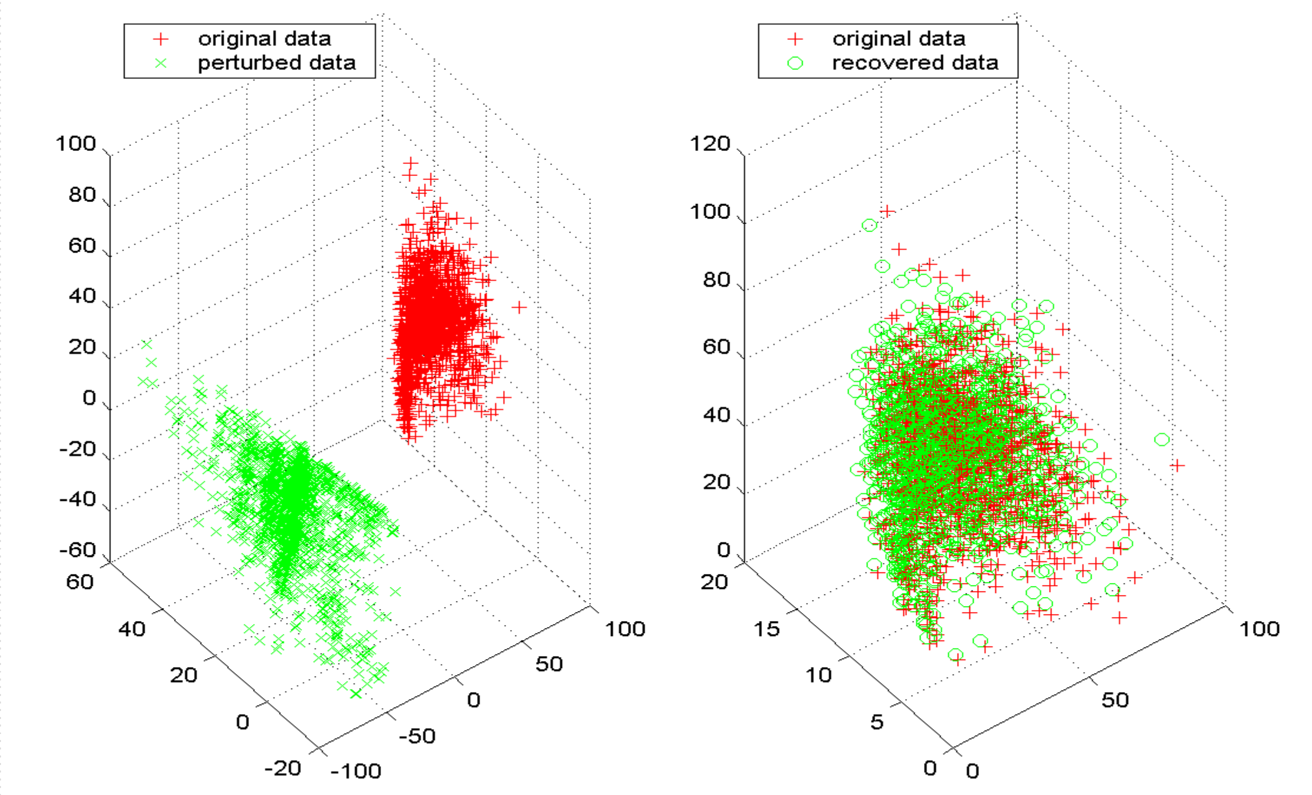


Fig. Known sample attack for Adult data with 32,561 private tuples. The attacker has 2% samples from the same distribution. The average relative error of the recovered data is 0.1081 (10.81%).

Known Sample Attack Experiments

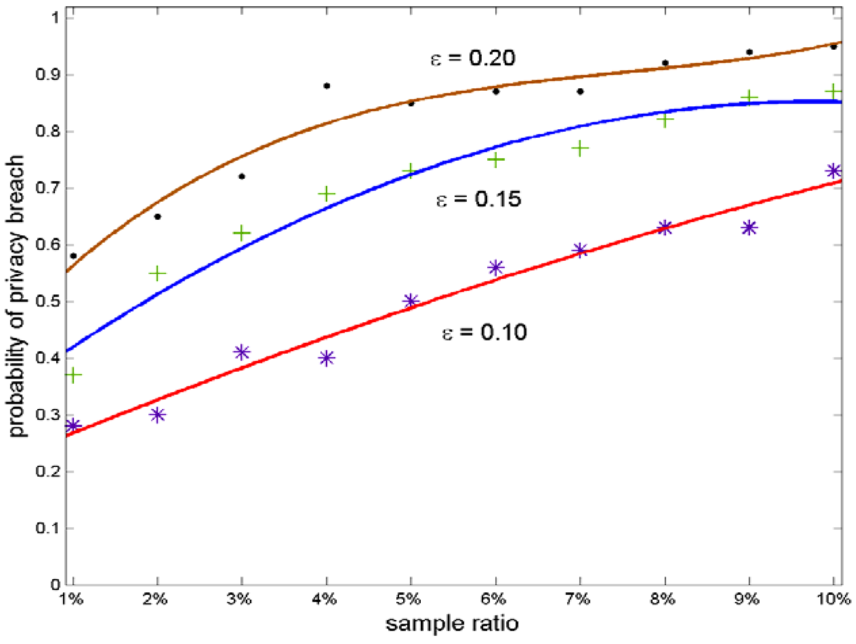


Fig. Probability of privacy breach w.r.t. attacker's sample size. The relative error bound ε changes from 0.10 to 0.20. (Adult data with 32,561 private tuples)

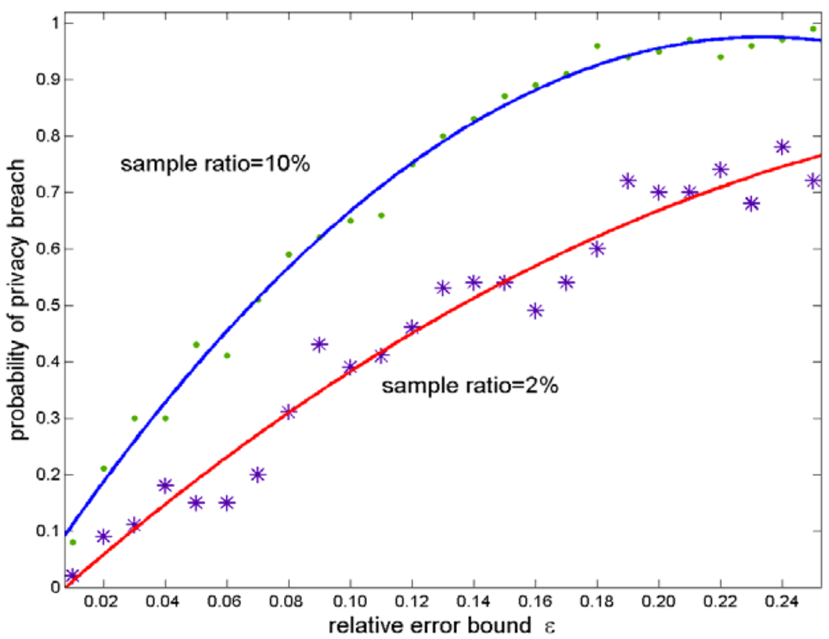


Fig. Probability of privacy breach w.r.t. the relative error bound ε. The sample ratio is fixed to be 2% and 10%. (Adult data with 32,561 private tuples.)

Effectiveness of Known Sample Attack

- Covariance Estimation Quality
  - Larger sample size gives attacker better recovery
  - Robust covariance estimator helps to downweight the influence of outliers
- PDF of the Data
  - The greater the difference between any pair of eigenvalues of the covariance, the higher the probability of privacy breach
- More details can be found in the extended version of this paper.

Conclusions

- Dist. Preserving Perturbation
  - Perturbed data preserves Euclidean distance/inner product exactly
  - Vulnerable to Known Input-Output Attack
  - Vulnerable to Known Sample Attack
- Possible Remedy?
  - Random projection [Liu06]

References

[Liu06] K. Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 18, no. 1, pp. 92–106, January 2006.

[Mukherjee06] S. Mukherjee, Z. Chen, and A. Gangopadhyay, "A privacy preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms," *The VLDB Journal*, p. to appear, 2006.

[Chen05] K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation," in *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)*, Houston, TX, pp. 589–592, November 2005.

[Oliveira04] S. R. M. Oliveira and O. R. Zaïane, "Privacy preservation when sharing data for clustering," in *Proceedings of the International Workshop on Secure Data Management in a Connected World*, Toronto, Canada, pp. 67–82, August 2004.

[Jolliffe02] I. T. Jolliffe, Principal Component Analysis, 2nd ed., ser. Springer Series in Statistics. Springer, 2002.

Questions