修士論文 要旨　2018 年 3 月修了

パブリッククラウドにおける暗号化方式の安全性評価

複雑理工学専攻 166093　小野澤 綜大
指導教員　國廣 昇 准教授

1　はじめに

パブリッククラウドの普及により個人情報をクラウドストレージで管理するケースが増えている．情報セキュリティ上の観点から個人情報を安全に管理するために暗号が用いられている．暗号化してデータを保存することでシステムに欠陥があったとしてもデータが暗号化されている限り情報漏洩を防ぐことができる．しかし，データを常に暗号化している場合，クラウドを利用するメリットが大きく損なわれる．なぜなら，データを利用する際に全てのデータをダウンロードし手元で復元する必要あるためである．

そこで，Property-Preserving Encryption (PPE) を暗号化に用いる方法がある．PPE とは平文のある特徴を暗号文に保存したまま暗号化する暗号化方式の総称である．ただし，平文とは暗号化される前のデータを指している．PPE で用いられる暗号化方式の例として順序保存暗号 (OPE)[1] や決定的な暗号 (DTE) がある．OPE は平文の大小関係が暗号文においても成立する暗号化方式で DTE は平文と暗号文が一対一で対応する暗号化方式である．しかし，これらの暗号化方式は機能性の高さと引き換えに安全性が既存手法に比べて弱くなるという問題点がある．

ACM CCS2015 で Naveed らは OPE の順序を保つ性質に注目して Cumulative 攻撃を提案した．Cumulative 攻撃は暗号文が十分多くある時に暗号文が復元できることを実験により示した．本研究では Naveed らの研究で提案された Cumulative 攻撃の改良を行う．

2　Cumulative 攻撃 [2]

この節では OPE に対する攻撃である Cumulative 攻撃の説明をする．まず，はじめに OPE について簡単に確認する．OPE は順序関係を保ったまま平文を暗号化する暗号化方式である．すなわち $m_1 \leq m_2$ の時，m_1, m_2 に対応する暗号文を c_1, c_2 とすると $c_1 \leq c_2$ が成り立つ．また，$m_1 = m_2$ の時，$c_1 = c_2$ も満たす．

Naveed らの提案した Cumulative 攻撃 [2] では補助的なデータを用いる．補助的なデータとは暗号化データと分布が類似するデータのことである．例えば，暗号化データがある病院の患者の年齢データを暗号化したものとする．患者の年齢データの分布は国勢調査などに載っている年齢の分布と類似している．よって，ある病院の患者の暗号化された年齢データは国勢調査などの公開されているデータと照らし合わせることで推測が可能である．これが Naveed らのアプローチである．ただし，我々の提案手法もここは同様である．

このアプローチに基づいた Naveed らの Cumulative 攻撃を説明する．暗号化データ $\boldsymbol{c}$，補助的なデータ $\boldsymbol{z}$ を得ている状況を考える．Cumulative 攻撃は次の目的関数を最小化する n 次置換行列 P を求めることにより復元結果を得る．

$$\sum_{i=1}^{|\mathbb{M}|} \left(\|\mathbf{Hist}\left(\boldsymbol{c}\right)_i - \langle \boldsymbol{P}_i, \mathbf{Hist}\left(\boldsymbol{z}\right)\rangle\|^2 + \|\mathbf{CDF}\left(\boldsymbol{c}\right)_i - \langle \boldsymbol{P}_i, \mathbf{CDF}\left(\boldsymbol{z}\right)\rangle\|^2 \right) \tag{1}$$

$\boldsymbol{P}_i$ は行列 P の i 行目のベクトルであり，$\mathbf{Hist}\left(\boldsymbol{c}\right)_i$ はヒストグラムの i 番目の成分，$\mathbf{CDF}\left(\boldsymbol{c}\right)_i$ はヒストグラムの i 番目までの成分の和である．$\langle \boldsymbol{P}_i, \mathbf{Hist}\left(\boldsymbol{z}\right)\rangle$ は $\boldsymbol{P}_i$ と $\mathbf{Hist}\left(\boldsymbol{z}\right)$ の内積を表している．ヒストグラムは頻度情報の当てはまりの良さを,CDF は順序情報の当てはまりの良さを定量化している．

$\boldsymbol{P}_i$ の j 番目の成分が 1 のとき，目的関数の i 番目の項は

$$\|\mathbf{Hist}\left(\boldsymbol{c}\right)_i - \mathbf{Hist}\left(\boldsymbol{z}\right)_j\|^2 + \|\mathbf{CDF}\left(\boldsymbol{c}\right)_i - \mathbf{CDF}\left(\boldsymbol{z}\right)_j\|^2$$

と変形することができる．これは，i 番目に小さい暗号文を j 番目に小さい平文に対応させた時の誤差を表している．すなわち，目的関数を最小にする置換行列 P を求めることができれば，誤差が最小となる暗号文と平文の対応関係が得られることになる．Cumulative 攻撃は LSAP に帰着させることで式 (1) を最適化する．

LSAP は n 次元コスト行列 $C = (c_{ij}), (0 \leq c_{ij})$ に対して，次のように定式化される．

$$\begin{aligned}
&\text{minimize} \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}X_{ij} \\
&\text{subject to} \sum_{i=1}^{n} X_{ij} = 1, 1 \leq j \leq n \\
&\sum_{j=1}^{n} X_{ij} = 1, 1 \leq i \leq n \\
&X_{ij} \in \{0,1\}, 1 \leq i, j \leq n
\end{aligned}$$

LSAP はハンガリアンアルゴリズムによって $O\left(n^3\right)$ の計算時間で最適解を求めることができる．

Cumulative 攻撃はコスト行列 $C = (c_{ij})$ を

$$c_{ij} = \|\mathbf{Hist}\,(\boldsymbol{c})_i - \mathbf{Hist}\,(\boldsymbol{z})_j\|^2 + \|\mathbf{CDF}\,(\boldsymbol{c})_i - \mathbf{CDF}\,(\boldsymbol{z})_j\|^2$$

と設定することにより，LSAP に帰着させている．ここで，c_{ij} は暗号文 c_i を平文 m_j に対応させたときの暗号文と平文のヒストグラム，累積分布関数の差の二乗和を表している．

3 提案手法

Cumulative 攻撃は暗号化データが少ない時に復元結果が悪くなる問題点がある．これはヒストグラムのノイズが大きくなり復元結果がヒストグラムのノイズに引っ張られるのが原因である．これによって，CDF による順序の評価よりも頻度の情報を優先してしまい本来の順序と矛盾した結果が得られてしまう．

そこで，LSAP を新たな最適化問題に変更することでこれらの問題を解決する．新たな最適化問題は，LSAP を改良し，順序の制約を新たに設けることにする．

既存手法では LSAP に帰着させ，式 (1) を最小とする P を求めていた．その一方で，我々の提案手法は順序の制約を持つ最適化問題に帰着させ，式 (1) を最小とする P を求める．次の式によって順序の制約を定義する．

$$X_{ij} = 1 \text{ ならば } X_{i'j'} \neq 1\,(i < i', j' < j) \qquad (2)$$
$$X_{ij} = 1 \text{ ならば } X_{i'j'} \neq 1\,(i' < i, j < j') \qquad (3)$$

式 (2) は，暗号文 c_i を平文 m_j に対応させた時，c_i より小さい暗号文は m_j より大きい暗号文に対応しない制約を表している．式 (3) は，暗号文 c_i を平文 m_j に対応させた時，c_i より大きい暗号文は m_j より小さい暗号文に対応しない制約を表している．この制約を加えた問題に帰着させることで式 (1) の最適化を行うのが提案手法である．

4 実験

Naveed らは，暗号文空間に対して，攻撃者が取得した暗号化データが十分多い時には，Cumulative 攻撃により生データを復元できることが示されている．今回の実験では，Cumulative 攻撃では復元が困難な暗号文の種類が多く暗号化データが少ない状況下に対して実験を行う．実験では，UCI Machine Learning Repository Adult Data Set[3] の 48842 件を用いる．このデータセットは訓練データ 32561 件とテストデータ 16281 件から構成される．要項では補助的なデータは訓練データを用いて，暗号化データはテストデータから 1000 件サンプルした場合の実験のみについて説明する．他のサンプル数に対する実験は本文を参照されたい．

要項では復元結果と生データを編集距離によって評価した結果を示す．Naveed らは異なる手法で評価を行っている．本研究では Naveed らの方法による評価も行っている．その評価による比較は本文を参照されたい．

編集距離とは文字列の類似度を測る際に用いられる評価基準である．今回は復号結果と生データをそれぞれ文字列とみなして編集距離によりこの 2 つの結果の類似度を測る．復元結果に数値の挿入，削除，置き換えの操作を行い生データと一致するまでの操作の回数を評価とする．

図 1 は既存手法と提案手法を編集距離を比較を示している．出力結果と正しい結果の編集距離の平均は既存手法は 14.65，提案手法は 4.266 である．提案手法の出力は既存手法よりも正しい結果との類似度が高いことがわかる．提案手法は既存手法よりも性能が向上していることがわかる．

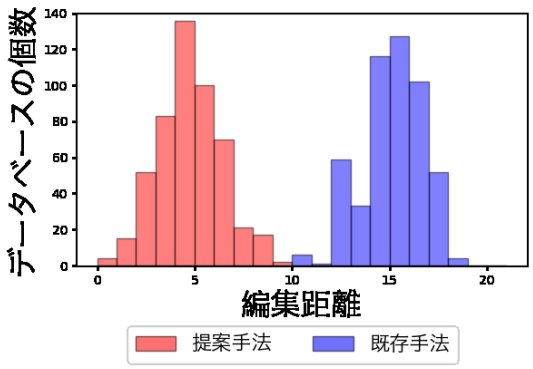

図 1. 編集距離による提案手法と既存手法の比較

参考文献

[1] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. Order preserving encryption for numeric data. In Proc. of SIGMOD, pp. 563–574, 2004.
[2] M. Naveed, S. Kamara, and C. V. Wright. Inference attacks on property-preserving encrypted databases. In Proc. of ACM CCS 2015, pp. 644–655, 2015.
[3] UCI Machine Learning Repository: Adult Data Set https://archive.ics.uci.edu/ml/datasets/adult