様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １７ 日現在

機関番号：８２６２６
研究種目：基盤研究(B)（一般）
研究期間：2019～2021
課題番号：１９Ｈ０４１１３

研究課題名（和文）大規模なパーソナルデータに向けた局所型プライバシー保護技術の研究

研究課題名（英文）A Study on Locally Private Algorithms for Large-Scale Personal Data

研究代表者
村上　隆夫（Murakami, Takao）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・主任研究員

研究者番号：８０５８７９８１

交付決定額（研究期間全体）：（直接経費）　　13,300,000円

研究成果の概要（和文）：本研究では，長期間にわたる時系列データ（位置情報など）やソーシャルグラフデータといった相関のある大規模なパーソナルデータに対して，ユーザ自身が加工を施してサービス事業者に送信する局所型プライバシー保護技術で成果を上げた．具体的には，LSH（Locality Sensitive Hashing）を用いた局所型プライバシー保護技術，位置情報生成技術，有用性の理論的保証を持つグラフLDP（Local Differential Privacy）技術を確立した．また，センシティブデータに対してのみLDPと同等の安全性を保証する安全性指標ULDP（Utility-Optimized LDP）も確立した．

研究成果の学術的意義や社会的意義
従来の局所型プライバシー保護技術のほとんどは，各データが独立であると仮定しており，長期間にわたる時系列データ（位置情報など）やソーシャルグラフデータのような相関を持ったパーソナルデータには適用できない．本研究での成果は，このようなデータに対しても安全性や有用性の理論的保証を与え，ユーザにプライバシーの観点での安心感を与えつつ，パーソナルデータの利活用促進を加速させることが可能となる．

研究成果の概要（英文）：In this work, we studied locally private algorithms for large-scale personal data, such as time-series data (e.g., location traces) and social graph data. Specifically, we proposed a locally private algorithm based on LSH (Locality Sensitive Hashing), a location trace synthesizer, and graph LDP (Local Differential Privacy) algorithms with utility guarantees. We also proposed a privacy notion called ULDP (Utility-Optimized LDP), which provides privacy guarantees equivalent to LDP for only sensitive data.

研究分野：プライバシー保護

キーワード：局所型プライバシー　差分プライバシー　時系列データ　グラフデータ　安全性指標

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年，IoT（Internet of Things）の普及に伴い，スマートフォン，カーナビ，タブレット端末，スマートメータなどから大量のパーソナルデータ（個人に関する情報）を収集し，利活用することが期待されている．例えば，数か月～数年間にわたる位置情報や電力使用量を収集し，人気のある観光地，道路交通情報，電力使用パターンなどを分析するクラウドセンシングが研究されている．また，年齢，結婚状況，収入，学歴，製品満足度などの多数（例えば 4～5 個以上）の属性データを収集し，元データの分布推定や相関分析などを行って，ターゲット顧客を明確にすることもできる．

しかし，その一方でプライバシーの侵害が懸念されている．例えば，公開位置情報を利用した空き巣事件やストーカー事件などが実際に起きている．また，電力使用量から使用機器や睡眠時間帯などの生活パターンが推測される恐れもある．さらには，自宅，生活パターン，結婚状況，収入などの様々なパーソナルデータが漏洩し，個人と紐づくことで「個人プロファイル」が作成され，売買される恐れもある．

これに対して近年，ユーザが自身のパーソナルデータに対してノイズを加えるなどの加工を施してサービス事業者に送信し，サービス事業者側でデータ解析を行う局所型プライバシー（Local Privacy）の研究が活発化している．このモデルでは，サービス事業者には元のデータは送られないため，不正アクセスなどによってサービス事業者から元データが漏洩する恐れがない．これは，情報漏洩の事故が多発している近年において（例えば，2015 年 5 月に日本年金機構から約 125 万件の氏名と基礎年金番号が漏洩），必要不可欠な要件である．このモデルにおける代表的な安全性の評価指標は，局所型差分プライバシー（LDP: Local Differential Privacy）（文献①）である．LDP は「どのような攻撃者が加工済みデータを入手しても，元データに関する情報をほとんど得ることができない」という安全性を数理的に保証するもので，Google が LDP を満たす技術「RAPPOR」を Chrome に実装し，ユーザからブラウザのスタートページを収集するなどの実用化を進めている（文献②）．

しかし，従来の局所型プライバシー技術では，一人のユーザが数多くのデータ（例えば 100 個以上の位置情報）を送信する場合に安全性を保証できない問題を抱えている．具体的には，LDP では各データは独立であると仮定するが，一人のユーザが持つ複数のデータは一般に独立ではなく，データ同士に相関がある．例えば，位置情報や電力使用量などの時系列データにおいて，ある時刻のデータは過去のデータに強く依存する．また，結婚状況と収入などの属性データ同士にも相関がある．データ同士に相関がある場合，LDP を満たす加工済みデータから元データを推定する攻撃も存在する．例えば，長期間に渡って似た値をとる時系列データの各々に LDP を満たすようにノイズを加えても，ノイズ付きデータの平均を求める平均化攻撃（文献②）により，元データが推定されるリスクがある．

２．研究の目的

本研究では，パーソナルデータの利活用促進を加速させるため，長期間にわたる時系列データや数多くの属性データなどの，相関を持った大規模なパーソナルデータ向けの局所型プライバシー保護技術を確立することを目的とする．

３．研究の方法

上記の目的を達成するため，大規模なパーソナルデータに対して次元圧縮を行った上で加工を施すことで，相関を用いた攻撃を防ぎ，安全性を高める技術を確立する．次元圧縮の方法としては，例えば類似度の大きいデータを同一のバイナリ系列に変換する LSH（Locality Sensitive Hashing）（文献③）などが考えられる．また，LDP 以外の安全性指標を用いることも検討する．例えば，位置情報に対しては PD（Plausible Deniability）（文献④⑤）などが考えられる．PD は加工データを得た攻撃者に対して，その加工データの基となるパーソナルデータの候補が多数存在することを保証する指標である．さらには，従来の安全性指標に捕らわれず，新しい局所型の安全性指標を確立することも検討する．このような次元圧縮と適切な安全性指標を基に，有用性・安全性の高い技術を確立し，実験的に有効性を示す．

４．研究成果

(1) LSH に基づく局所型プライバシー保護技術

まず，LSH を用いてパーソナルデータに対する次元圧縮を行う局所型プライバシー保護技術を確立した．本技術は，LSH を用いてデータをバイナリ系列に変換した後，RR（randomized response）を各要素に適用することでデータの加工を行う．本技術が，差分プライバシーを入力データ同士の任意の距離尺度に一般化した XDP(extended DP)を満たす（厳密には，concentrated XDP と probabilistic XDP を満たす）ことを証明した．本技術は例えば，映画レビューや位置情報を基に，rating vector や visit-count vector（各 Point-of-Interest の訪問回数から構成されるベクトル）の angular distance の小さいユーザを推薦する friend matching に適用可能である．

映画レビューや位置情報の実データを用いた評価実験により，本技術の有効性を示した．この成果は，情報セキュリティ分野における難関国際会議の ESORICS'21 に採択された．

(2) プライバシー保護型位置情報生成技術

また，プライバシーを保護しつつ，大規模な位置情報の人工データを生成する技術を確立した．本技術は，トレース（位置情報の時系列データ）に関する様々な統計情報（時間帯ごとの人口分布，遷移行列，学生・通勤者などある特定の特徴をもったユーザの行動パターンなど）を visit-count tensor と transition-count tensor という二つのテンソルでモデル化し，これらを同時に分解する MTF（Multiple Tensor Factorization）によって，元の位置情報の次元圧縮を行う．その後，MH（Metropolis Hastings）法を用いて合成トレースを生成する．安全性指標としては PD を用い，各合成トレースに対して PD を満たすかどうかの検証を行い，PD を満たす合成トレースのみを出力する．実データを用いた網羅的な評価実験により，本技術が有用性，安全性，スケーラビリティの 3 つの観点で，従来の方式よりも優れていることを示した．この成果は，プライバシー分野のトップ国際論文誌 PoPETs'21 に採択された．

尚，本技術は，国内における位置情報の匿名加工と再識別のコンテスト PWSCup 2019 における，人工データ生成法の一部として実際に使用された．研究代表者は，PWSCup 2019 の実行委員長を務め，コンテスト論文を国内会議 CSS'19 において発表した．

(3) 新しい局所型の安全性指標

さらに，パーソナルデータが取り得る値が sensitive と non-sensitive に分けられる場合において，センシティブデータに対してのみ LDP と同等の安全性を保証する新しい局所型の安全性指標「ULDP（Utility-Optimized LDP)」を確立した．ULDP を満たすデータ加工メカニズムとして，RR や RAPPOR を拡張した「Utility-Optimized RR」，「Utility-Optimized RAPPOR」を確立した．元データの分布推定問題において，これらのメカニズムの有用性が，LDP を満たす RR，RAPPOR と比べてオーダーレベルで分布推定誤差を改善できることを理論・実験の両面で示した．この成果は，情報セキュリティ分野のトップ国際会議 USENIX Secuity'19 に採択された．

(4) グラフ LDP

尚，「1．研究開始当初の背景」では「一人のユーザが持つ複数のデータ間の独立性」が仮定できないケースを考えていたが，これとは別に「複数のユーザ間でのデータの独立性」が仮定できないようなデータも存在する．例えば，ノードがユーザ，枝が友達関係を表すソーシャルグラフにおいては，一つの枝を二人のユーザが共有するため，各ユーザの友達リスト（隣接行列の各行）は独立ではない．従来では，各データが独立であるという仮定の下で，有用性（元データの分布推定誤差など）の理論的保証を持つ LDP 技術が提案されているが（文献⑥），グラフデータにおいては有用性の理論的保証を持つ LDP 技術は提案されていない．

そこで，グラフデータにおいて有用性の理論的保証を持つ LDP を確立した．具体的には，グラフ内にある部分グラフ（k-stars，triangles など）を数え上げる subgraph counting に着眼し，LDP を満たしながら部分グラフ数を推定する問題に取り組んだ．大規模なグラフデータに対して有効なアルゴリズムを確立し，推定誤差の upper-bounds と lower-bounds を示し，実験的に有効性を示した．この成果は情報セキュリティ分野のトップ国際会議 USENIX Security'21 に採択された．さらに，ここで確立した triangle counting のアルゴリズムと比べて，ダウンロード時の通信量の大幅な削減（例：400Gbits から 160Mbits）を実現する新しいアルゴリズムも確立した．この成果は，USENIX Security'22 に採択された．また，(3)の USENIX Secuity'19 の内容と(4)の USENIX Security'21 の内容を纏めて，国際会議 IWSEC'21 の keynote talk を行った．

(5) プライバシーのリスク評価

最後に，ユーザのパーソナルデータに対して加工を施さない場合のプライバシーリスクがどれくらいあるかの評価を定量的に行った．ここでは深層学習システムに着眼し，プライバシーのリスク評価を行った．特に，加工側が転移学習を利用して転移後のモデルの一部を開示した場合に，攻撃者も転移学習を利用することで高精度なメンバーシップ推定攻撃ができることを実験的に示した．本成果は国際会議 IJCNN'21 に採択された．

＜引用文献＞

① J.C. Duchi, et al., “Local Privacy and Statistical Minimax Rates,” Proceedings of the IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS’13), pp.429-438, 2013.
② U. Erlingsson et al., “RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response,” Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS’14), pp.1054-1067, 2014.
③ J. Wang et al., “Learning to Hash for Indexing Big Data -- A Survey,” Proceedings of the IEEE, vol.104, no.1, pp.34-57, 2015.
④ V. Bindshaedler and R. Shokri, “Synthesizing Plausible Privacy-Preserving Location

Traces," Proceedings of the 37th IEEE Symposium on Security and Privacy, pp.546-553, 2016.
⑤ V. Bindshaedler et al., "Plausible Deniability for Privacy-Preserving Data Synthesis," Proceedings of the VLDB Endowment, vol.10, no.5, pp.481-492, 2017.
⑥ P. Kairouz et al., "Discrete Distribution Estimation under Local Privacy," Proceedings of the 33rd International Conference on Machine Learning (ICML'16), pp.2436-2444, 2016.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Takao Murakami, Koki Hamada, Yusuke Kawamoto, Takuma Hatano	４．巻 2
２．論文標題 Privacy-Preserving Multiple Tensor Factorization for Synthesizing Large-Scale Location Traces with Cluster-Specific Features	５．発行年 2021年
３．雑誌名 Proceedings on Privacy Enhancing Technologies (PoPETs)	６．最初と最後の頁 5-26
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計8件（うち招待講演　2件／うち国際学会　6件）

１．発表者名 Jacob Imola*, Takao Murakami*, Kamalika Chaudhuri (*: equal contributions)
２．発表標題 Communication-Efficient Triangle Counting under Local Differential Privacy
３．学会等名 Proceedings of the 31st USENIX Security Symposium (USENIX Security 2022)（国際学会）
４．発表年 2022年

１．発表者名 Natasha Fernandes*, Yusuke Kawamoto*, Takao Murakami* (*: equal contribution)
２．発表標題 Locality Sensitive Hashing with Extended Differential Privacy
３．学会等名 Proceedings of the 26th European Symposium on Research in Computer Security (ESORICS 2021)（国際学会）
４．発表年 2021年

１．発表者名 Takao Murakami
２．発表標題 Toward Accurate Data Analysis under Local Privacy
３．学会等名 The 16th International Workshop on Security (IWSEC 2021), Keynote Talk, 2021.（招待講演）（国際学会）
４．発表年 2021年

１．発表者名
Jacob Imola*, Takao Murakami*, Kamalika Chaudhuri (*: equal contributions)

２．発表標題
Locally Differentially Private Analysis of Graph Statistics

３．学会等名
Proceedings of the 30th USENIX Security Symposium (USENIX Security 2021)（国際学会）

４．発表年
2021年

１．発表者名
Takao Murakami

２．発表標題
Locally Differentially Private Analysis of Graph Statistics

３．学会等名
IT-Security & Privacy Colloquium, the University of Luebeck（招待講演）

４．発表年
2021年

１．発表者名
Seira Hidano, Takao Murakami, Yusuke Kawamoto

２．発表標題
TransMIA: Membership Inference Attacks Using Transfer Shadow Training

３．学会等名
Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN 2021)（国際学会）

４．発表年
2021年

１．発表者名
Takao Murakami, Yusuke Kawamoto

２．発表標題
Utility-Optimized Local Differential Privacy Mechanisms for Distribution Estimation

３．学会等名
Proceedings of the 28th USENIX Security Symposium (USENIX Security 2019)（国際学会）

４．発表年
2019年

１．発表者名 村上隆夫，荒井ひろみ，井口誠 ，小栗秀暢，菊池浩明，黒政敦史，中川裕志，中村優一，西山賢志郎，野島良，波多野卓磨，濱田浩気，山岡裕司，山口高康，山田明，渡辺知恵美
２．発表標題 PWS Cup 2019: ID識別・トレース推定に強い位置情報の匿名加工技術を競う
３．学会等名 コンピュータセキュリティシンポジウム2019（CSS 2019）
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	日野　英逸 (Hino Hideitsu) (10580079)	統計数理研究所・モデリング研究系・教授 (62603)	
研究分担者	清　雄一 (Sei Yuichi) (20700157)	電気通信大学・大学院情報理工学研究科・准教授 (12612)	
研究分担者	松田　隆宏 (Matsuda Takahiro) (60709492)	国立研究開発法人産業技術総合研究所・情報・人間工学領域・主任研究員 (82626)	
研究分担者	川本　裕輔 (Kawamoto Yusuke) (60760006)	国立研究開発法人産業技術総合研究所・情報・人間工学領域・主任研究員 (82626)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
米国	University of California, San Diego			
オーストラリア	Macquarie University			