公開

研 究 終 了 報 告 書

「Web/IoT 横断的プライバシ保護データ解析基盤」

研究期間: 2019 年 10 月～2023 年 3 月
研 究 者: 清 雄一

1. 研究のねらい

監視カメラの映像や家電センサデータ、車載センサデータ等、人々の周りに存在する大量のセンサから IoT データが日々大量に生成されており、FP7/NICT による ClouT (Cloud of Things for Empowering Citizen's Clout in Smart Cities)プロジェクト等、国を超えて IoT 技術を活用する取り組みが積極的に進められている。日本では「個人情報の保護に関する法律」(いわゆる個人情報保護法)の改正(2017 年全面施行)や総務省によるオープンデータ戦略により、プライバシ保護処理が行われた IoT データ、個人データや統計データを組織間、またはオープンデータとして公開し共有することが促進されている。しかし経済や行政にとって利点がある一方、これらのデータが組み合わされることにより、意図しない個人特定や個人属性値漏洩が発生するリスクが拡大している。

プライバシ保護処理が施され共有されたデータは、個人情報保護法により他データとの突合が禁止されている。なぜなら、個々にプライバシ保護処理が施されていたとしても、突合により、個人特定や個人属性値漏洩のリスクがあるためである。しかし、法的な制限のみでなく、他データと突合されたとしても技術的に個人特定や個人属性値漏洩を防ぐ仕組みがなければ真の安全は達成されない。

このように、様々な人や組織が IoT データ及び Web 上のデータを横断的に活用した新たなサービスの構築・普及を考えており、今後これらのデータを流通させ、組み合せて活用していく制度やインフラが整っていくことが予想される。しかしながら、どこから個人のプライバシ情報が漏洩するかを予想することが困難になり、プライバシを保護する共通的で強固な枠組みの構築が重要な課題となる。本研究では特に、IoT に特有の、誤差を含むデータへのプライバシ保護、及び、データの組合せに対するプライバシ保護について現状の課題を明らかにするとともに、プライバシ保護技術の開発を行う。

2. 研究成果

(1)概要

研究成果は主に、誤差・欠損のあるデータのプライバシ保護データ解析・機械学習、データベースの組み合わせによる個人特定リスク解析、取得項目数増大に対応するプライバシ保護レベル自動決定、IoT センシングデータ収集と公開、医療 IoT データのプライバシ保護データ解析の 5 テーマに整理される。

テーマ共通として、プライバシの保護度合いを計測する指標としてローカル差分プライバシ (Local Differential Privacy) を採用した。この指標は Google や Apple 等多くの組織で採用されているほか、プライバシ保護データ解析分野で近年盛んに研究されている。本指標に基づく場合、個人を特定できる識別子を削除したり属性値を暗号化したりするだけでなく、属性値にノイズを加える必要がある。このノイズをいかに加えるか、また、ノイズが付加されたデータ集合から

いかに高精度に統計解析・機械学習を行うかが課題となる。本研究では特に IoT データを対象に研究を行った。IoT データの特徴として、誤差・欠損が多く含まれること、データの種類数が膨大であること、個人が認識しないまま複数個所でデータ収集される場合があることが挙げられる。これらの特徴にそれぞれ特化した手法を提案した。

(2)詳細

研究テーマ A「誤差・欠損のあるデータのプライバシ保護データ解析・機械学習」

IoT データには一般に観測誤差・欠損が含まれるが、これまでのほとんどのプライバシ保護技術は観測誤差・欠損を考慮していなかった。

まず観測誤差への対応について述べる。 図 1 に示すように、誤差を含む観測値にプライバシを保護するためのノイズを加えた際に、真値を統計的に分析するシナリオと、図 2 に示すように、ユーザ自身も知らない真値にノイズを加えて真値を統計的に分析するシナリオの 2 シナリオについて研究を行った。プライバシ保護の対象が観測値か真値かの違いはあるが、両シナリオとも保護すべき値をローカル差分プライバシ(LDP)で守るという点では同じである。

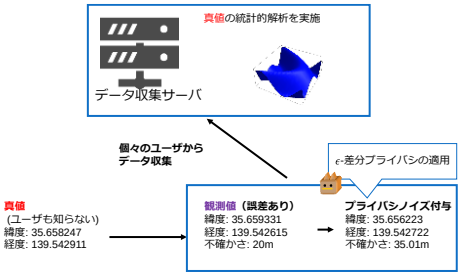

図 1: 観測値を差分プライバシで保護して真値を分析

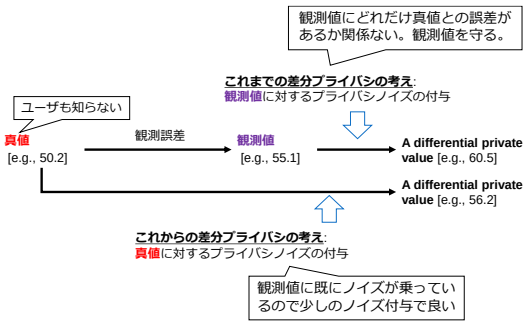

図 2: 真値を差分プライバシで保護して真値を分析

本研究では、議論をシンプルにするためいずれのシナリオも観測誤差の分布がガウス分布であることを仮定したが、他の分布であっても提案手法は適用可能である。提案手法が LDP を満たすことを数学的に証明し、実データを用いた実験により既存手法と比べてデータ分析精度がシナリオ 1 では 50-80%程度、シナリオ 2 では 30-40%程度の精度向上が確認できた。

次に欠損値への対応について述べる。IoT 環境では各個人から多数の属性情報を取得できるが、欠損も多い。特に医療分野の IoT (IoMT: Internet of Medical Things) 環境下では、プライバシ保護処理を施したとしてもユーザがデータの提供を拒むことも想定される。このような場合、多数の属性値を同時に分析することが困難となる。さらに取得できた値には大きな LDP ノイズが乗っている。本研究では、LDP ノイズを考慮した上で各属性ペアの共分散及び各属性のデータ値の確率密度関数を導出し、LDP ノイズを緩和した生成モデルを構築する手法を提案した。構築した生成モデルから真の属性値集合をサンプリングし、その結果をデータ解析や機械学習に利用することができる。COVID-19 データを含む実データで実験を行い、欠損値を考慮しない手法と比べて 50-80%の精度向上が実現できることを示した。

また、このように誤差・欠損値のあるデータから機械学習を行う場合、その精度は低下する。機械学習で入力データを識別しその出現頻度をカウントするシナリオ（例：来場者の年齢の分布）において、機械学習の精度が低い場合でも、カウント結果の精度を向上させる手法を提案した。画像データを基に実験を行い、約 60%精度向上することが確かめられた。これらの研究成果は、IEEE Transactions on Dependable and Secure Computing, IEEE Access, IEEE Internet of Things Journal 等に掲載された。

研究テーマ B「データベースの組み合わせによる個人特定リスク解析」

統計解析や機械学習を行う多くの研究者が個人データを利用できるようにするためには、個人を特定できる情報を排除する必要がある。LDP に基づいてデータを収集し、かつサンプリングを行うことによりプライバシが保護されると考えられているが、本研究では、これらの手法を適用しても、データによっては再識別されるリスクが非常に大きいことを示す。具体的には、サンプリングされた LDP データベースに対し、ある属性値を持つ人が母集団に何人いるかを推定するアルゴリズムを提案した（図 3）。

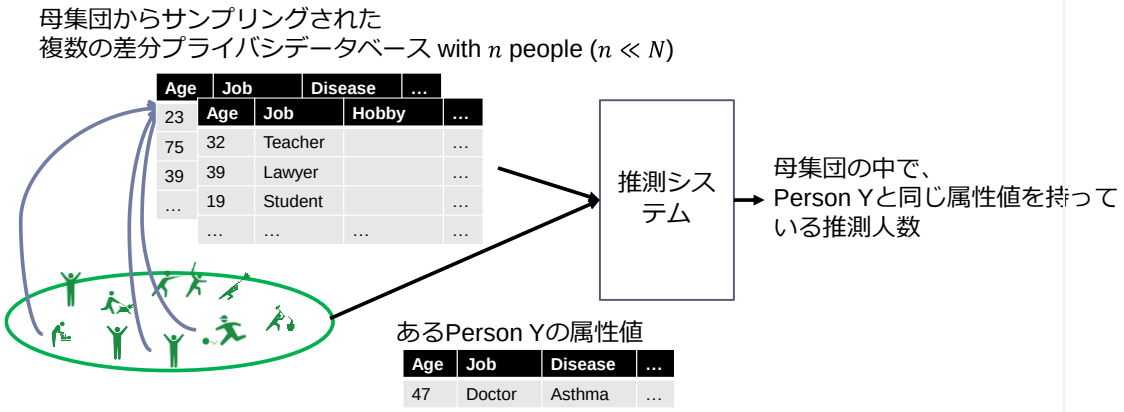

図 3 推定人数=1 の場合 Person Y と同じ属性値を持つ人が母集団に他に存在しない

推定された数が 1 であれば、その属性値を持つ人が母集団に 1 人しか存在しない確率が高く、再識別される確率が高いことを意味する。そのため図 3 における Person Y のデータを扱う際は他の人のデータより慎重な扱いが求められる。さらに、単一のデータベースによる識別リスクよりも、複数のデータベースによる識別リスクが高まることを示した。

本研究は、研究テーマ D で取得した個人データを用いて実験を行った。この成果は IEEE Open Journal of the Computer Society に掲載された。

研究テーマ C「取得項目数増大に対応するプライバシ保護レベル自動決定」

LDPを適用する際は、保護するプライバシの度合い（プライバシ損失）を事前に決める必要がある。しかし、GDPR などで再識別が大きなリスクとされる現在、各データに対してプライバシ損失をどの値に設定すべきかこれまで確立された方法論は無かった。本論文では、再識別を防止するために必要なプライバシ損失を、データ収集時に適応的に導出することにより、データ活用と保護を高いレベルで同時に実現する手法を提案した。

具体的には、人々からLDPの下でデータを収集して属性値のヒストグラムを生成するシナリ

オを想定する。このとき、攻撃者が収集したデータから、任意の人物 A のデータを当てるときの正解確率をγ/n以下にすることを目標とする。ここでnはデータ数であり、γは 1 以上n以下の任意の値である。攻撃者は人物 A の真のデータを把握していると想定する。各属性値を持つ人数は事前には不明であり、データを収集しながら当該人数を推測していく必要がある。データ収集初期では全ての属性値に対して worst-case で保護する。データ収集中盤で、各属性値の人数を推測する。このとき、1-alpha 信頼区間での下限値を計算する。この下限値に基づいてデータ収集後半ではプライバシ損失を設定する。この成果は IEEE Internet of Things Journal に掲載された。

研究テーマ D「IoT センシングデータ収集と公開」

プライバシ情報を含むIoTデータを対象に研究を行う場合、そのデータを収集することが困難である。そのため本研究では、マンションの 1 室を借りて各被験者に 2 週間程度住んでもらい、そのときの生体情報、センシング情報等を収集した。図 4 にセンシングシステムの概要を示す。

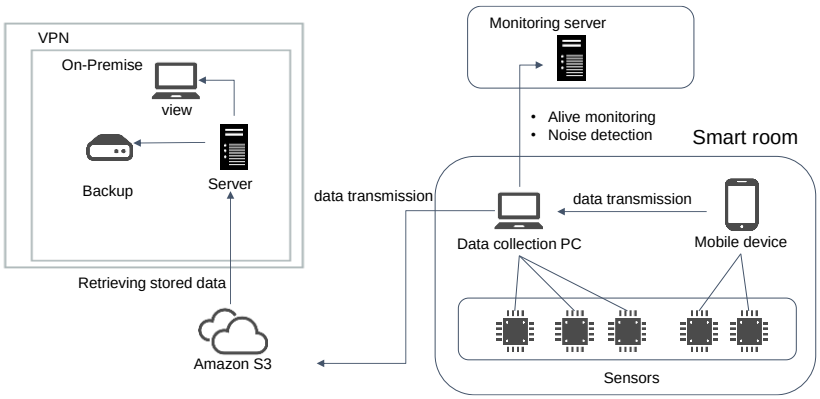

図 4 センシングシステム概要

被験者の行動パターンを正確に把握するため、玄関、電子レンジ、冷蔵庫、シャワー室等、扉やドアがある箇所には開閉センサを取り付けた。開閉センサシステムは図 5 に示すように自作した。

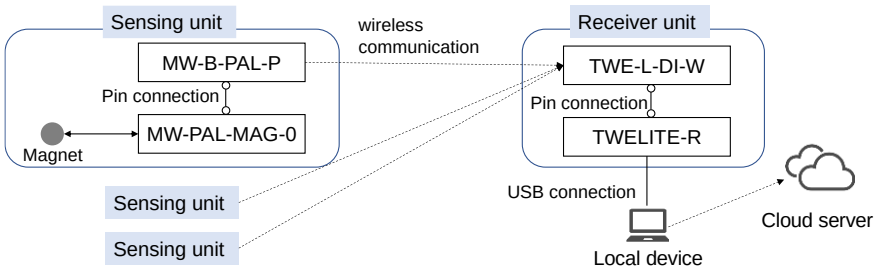

図 5 開閉センサシステム

被験者は合計 34 名となった。改めて被験者に同意を取った上で、取得したデータの一部（23 名分）は Mendeley Data に公開しており、1600 回以上ダウンロードされている。この成果は Data in Brief (Elsevier)に掲載された。

研究テーマ E「医療 IoT データのプライバシ保護データ解析」

病気のなりやすさ等を解析する際に、カイ二乗検定などの検定が行われる。この際、カイ二乗検定の結果をそのまま公開すると、解析対象データのセンシティブな情報漏洩につながるリスクが指摘されている。また、COVID-19 等のように新しい病気が登場した場合、多くのサンプルを得ることが難しい。そこで、少ないサンプルを対象に、差分プライバシで保護した上で高い精度でカイ二乗検定を行う手法を開発した。

また、IoT ベースのヘルスケアサービスシステムのための新しいプライバシ保護データ収集スキームを提案した。提案方式は、属性開示、個人情報開示、会員情報開示、感度攻撃、類似度攻撃、歪度攻撃などのプライバシ攻撃に対して効率的に対処することが可能である。提案方式の有効性と効率性は、理論解析と実験解析により証明した。さらに、医療データをブロックチェーンを活用して安全に保存、解析するための手法を体系的に整理しサーベイ論文として取りまとめた。

これらの成果は BioData Mining (BioMed Central) や Peer-to-Peer Networking and Applications (Springer) に掲載された。サーベイ論文は Journal of Network and Computer Applications (Elsevier) に掲載された。

3. 今後の展開

海外研究者については、Manipal Academy of Higher Education 所属の Prof. J. Andrew Onesimu との共同研究を開始しており、IEEE Transactions 等複数の成果を出し始めている。また、University of Lagos 所属の Dr. Agbotiname Lucky Imoize との共同研究を開始しており、IET や CRC へ複数の著作物を執筆する等活動を開始している。今後彼らとの共同研究を通じて、海外における人々のプライバシに対する考え方の解析や、実際にスマートルームにおいてデータ収集を行うことで、日本だけでなく世界におけるプライバシデータ収集・分析を行うことを計画している。

また、株式会社三菱総合研究所の奥村主席研究員と共同研究を行っており、適宜、生じた案件に対して議論を行っている。今後も協力関係を引き続き築いていく計画である。

4. 自己評価

・研究目的の達成状況

IoT データに特有の誤差・欠損に対応して、ローカル差分プライバシに基づいてデータを収集・解析する手法の開発をメインのテーマに設定していた。本テーマに関してはその研究成果が複数の IEEE トランザクション/ジャーナルに採択されており、期待どおりの成果が出たと判断している。一方、マンションの 1 室を借りて個人の生活状況を収集する実験については、研究の開始直後に新型コロナウイルス感染症のパンデミックが起こり、当初予定していたような大規模な広報活動を行うことができなかったのが心残りである。しかし 30 名を超える被験者に協力をいただき、結果的には計画どおりの成果を得ることができた。

ローカル差分プライバシデータを用いて高精度の機械学習モデルを生成する手法について、決定木や k-nearest neighbor、深層学習など一般的な手法を提案することができたものの、深層学習に特化した手法としてはまだ十分な精度が出たとは言い難い。また本研究では主に数値

データやカテゴリデータを対象としており、画像・動画データへの取り組みが不十分である。この点は将来課題としたい。

・研究の進め方（研究実施体制及び研究費執行状況）

さきがけの研究を開始して以降、海外研究者や民間企業の研究者との共同研究を開始した。多くの共同研究について自分が主導することができており、さきがけ研究の理念に従った成果を出せていると認識している。

研究費については、主にマンションの家賃や被験者謝金として利用させていただいた。機械学習を行う際は、もともと保有していた DGX-1 を利用することにより、ハードウェアの購入費用を大幅に抑えることができた。

・研究成果の科学技術及び社会・経済への波及効果

個人のプライバシを守りながらデータを収集・解析することは社会的な要請であり、今後は個人情報保護法や GPDR 等、法律との関係性をより明確にしながら研究を行っていく必要があると考えられる。

5. 主な研究成果リスト

(1) 代表的な論文（原著論文）発表

研究期間累積件数：ジャーナル28件、国際会議論文17件

1. Yuichi Sei, J. Andrew Onesimu, Hiroshi Okumura, Akihiko Ohsuga: Privacy-Preserving Collaborative Data Collection and Analysis with Many Missing Values, IEEE Transactions on Dependable and Secure Computing, Vol.20, No.3, pp.2158-2173, 2023
IoT データには欠損が多く含まれる。しかしプライバシを保護しながらデータを収集する既存手法は欠損値が考慮されていない。本論文では欠損値を考慮した、プライバシ保護データ収集・解析手法を提案する。ローカル差分プライバシに基づいて収集されたデータに対し、サーバは期待値最大化法とガウシアンコピュラ法に基づき、多値属性分析に適した生成モデルを構築する。COVID-19 データを含む実データで実験を行い、欠損値を考慮しない手法と比べて 50-80%の精度向上が実現できることを示した。
2. Yuichi Sei, Akihiko Ohsuga: Private True Data Mining: Differential Privacy Featuring Errors to Deal with Internet-of-Things Data, IEEE Access, Vol.10, pp.8738-8757, 2022
IoT 環境ではセンシングエラー等により真の値とは異なる値が得られることが多い。既存研究では、誤差を含む測定値を保護することに着目しており、真の値を保護することはこれまで考慮されていなかった。本論文では、データ所有者も知らない真の値に対してローカル差分プライバシを適用することにより、データの有用性を既存手法よりも大幅に高める手法を提案した。実データに基づく実験により、精度を 30-50%程度向上させることができた。
3. Takao Murakami*, Yuichi Sei*: Automatic Tuning of Privacy Budgets in Input-Discriminative Local Differential Privacy, IEEE Internet of Things Journal, 2023 (*equally contributed)
ローカル差分プライバシを適用する際は、保護するプライバシの度合い（プライバシ損

失)を事前に決める必要がある。しかし、GDPR などで再識別が大きなリスクとされる現在、各データに対してプライバシ損失をどの値に設定すべきかこれまで不明であった。本論文では、再識別を防止するために必要なプライバシ損失を、データ収集時に適応的に導出することにより、データ活用と保護を高いレベルで同時に実現する手法を提案した。

(2)特許出願

研究期間全出願件数:0 件(特許公開前のものも含む)
ただし、研究期間前に出願した案件について、研究期間中の研究成果を踏まえて手続補正を実施した。

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

・基調講演

[1] Yuichi Sei: Machine Learning on Differentially Private Data, International Conference on AI and Machine Learning, 2022 (Keynote speech)

・受賞

[1] 清雄一: 船井学術賞, 2022

[2] Satoru Mizusawa and Yuichi Sei: 4th IEEE International Conference on Computing, Electronics & Communications Engineering (iCCECE), Best Paper Award, 2021

[3] Yuichi Sei: IPSJ/IEEE Computer Society Young Computer Researcher Award, 2021

・著作物

[1] Yuichi Sei: Privacy-Preserving Data Collection and Analysis for Smart Cities, chapter in book "Human-Centered Services Computing for Smart Cities", IEICE Monograph, Springer (招待あり)

[2] Yuichi Sei, Akihiko Ohsuga, Agbotiname Lucky Imoize: Statistical Test with Differential Privacy for Medical Decision Support Systems, chapter in book "Explainable Artificial Intelligence in Medical Decision Support Systems", The Institution of Engineering and Technology (IET)

[3] Yuichi Sei, J. Andrew Onesimu, Akihiko Ohsuga, Agbotiname Lucky Imoize: Local Differential Privacy for Artificial Intelligence of Medical Things, chapter in book "Handbook of Security and Privacy of AI Enabled Healthcare Systems and Internet of Medical Things", CRC Press

[4] Yuichi Sei, Akihiko Ohsuga, Agbotiname Lucky Imoize: A Lightweight Algorithm for Detection of Fake Incident Reports in Wireless Communication Systems, chapter in book "Security and Privacy Schemes for Dense 6G Wireless Communication", The Institution of Engineering and Technology (IET)