Vol.2010-DPS-142 No.3
Vol.2010-CSEC-48 No.3
2010/3/4

2パーティ秘匿回路計算を利用したプライバシー保護データ分析実験報告 (2) - 大学生の成績と生活実態との相関分析 -

柴田賢介[†1] 千田浩司[†1] 山本太郎[†1]
高橋克巳[†1] 金井敦[†2]

近年，個人のプライバシー情報や企業の機密情報を保護しながら利活用するニーズが高まっている．本研究では，プライバシー情報の利活用を実現する技術の 1 つである秘匿回路計算のシステムを実装し，本システムを用いて教育分野での実証実験を行なった．実験では，法政大学理工学部の学生約 110 名を対象として学生のプライベートに関するアンケートを実施し，これを本システムによって秘匿した状態で回収するとともに，学生の成績と突合せて秘匿回路計算による統計演算を行なった．実験の結果，本システムを用いることにより，従来教師が把握することが困難であった学生の生活行動をプライバシーを保護した状態で収集し，成績との相関関係の分析が可能であることを示した．

Experimental Trials on Privacy-preserving Data Analysis Using 2-party Secure Circuit Evaluation(2) - The Behavior Analysis of University Students -

Kensuke Shibata,[†1] Koji Chida,[†1] Taro Yamamoto,[†1]
Katsumi Takahashi[†1] and Atsushi Kanai[†2]

In recent years, there is a need for widespread utilization of privacy or confidential information as well as preserving one. One of the technologies which meets the need is secure function evaluation. In this paper, We evaluate our secure function evaluation system in the field of education. We collect and analyze information on daily life of students by privacy protection enquete system and show a correlation between lifestyle behaviour and academic performance.

1. はじめに

近年，ネットワークを介したサービスの中にはライフログとして蓄積された個人の行動履歴を収集／分析し，分析結果をレコメンデーション等に活用するといったものがある．従来，個人のプライバシー情報や企業において管理される顧客情報等の機密情報は厳重な保護の対象とされてきた．しかし，上記のようなサービスへのニーズが高まる中，情報を保護しつつも利活用することへの動きが促進されてきている．

プライバシー情報や機密情報の利活用のためには，プライバシー保護や企業情報漏洩対策の観点から，これらの情報を保護しつつ利活用を可能とする技術が必要となる．現在，PPDM(Privacy Preserving Data Mining) のための主要な技術として，匿名化，再構築計算，秘匿回路計算の 3 つの技術が検討されている[1)]．

匿名化の代表的な手法としては，k-匿名法[2)] がある．これは，任意の属性について，共通の組み合わせを持つレコードが少なくとも k 個以上存在するようデータ保護処理を施すものである．また，再構築計算[3)] はデータ提供の際に撹乱を施し，マイニングする際に撹乱されたデータベースからの再構築によって統計データを得る，確率的手法を用いた技術である．秘匿回路計算はデータを暗号化したまま計算を行なうことを可能とする技術であり，通常のマイニングと同等の精度と高いプライバシー保護性を持つ．例えば，複数の異なるデータに対し，秘匿した状態のままでデータを統合 (垂直統合) し，統計演算を実行するといったことが可能である．

秘匿回路計算は従来より計算効率が課題となっていたが，筆者らは，Yao が提案している 2 パーティ秘匿回路計算[4)] を応用した 2 パーティ秘匿回路計算プロトコル (以降 SCI 方式と呼ぶこととする) を提案し，これを「セキュア表計算システム」として実装した[5)]．本実装によると，計算を実行するための論理回路 1 ゲートあたりに要する計算時間は 11 μ sec 程度であり，10,000 件程度のデータを対象とした最大値計算やクロス集計といった演算については数分で実行可能となっており，計算効率については向上してきていると言える．

一方，教育の分野においては，教師は学生の生活行動を理解して，教育効果を上げたいと考えているものの，学生は教師に日常生活までは知られたくない，という思いがあり，生活

†1 日本電信電話株式会社 NTT 情報流通プラットフォーム研究所
NTT Information Sharing Platform Laboratories
†2 法政大学 理工学部
Faculty of Science and Engineering, Hosei University

Vol.2010-DPS-142 No.3
Vol.2010-CSEC-48 No.3
2010/3/4

行動にまで踏み込んだ分析が行なえていないという課題があった．そこで本研究では，教育分野における上記の課題と秘匿回路計算との親和性が高いと考え，セキュア表計算システムを用い，教育分野を対象とした実証実験を行なった．実験では，学生の日常生活等，プライベートに関するアンケートの回答を秘匿した状態で回収し，これと併せて実施された成績に影響しない中間実力チェック (以降実力チェックと呼ぶ) の成績データとの垂直統合を行ない，秘匿回路計算によって平均値やクロス集計等の統計演算を実行した．これにより教師は学生の日常生活に関する情報を一切閲覧することなく，学生の生活実態と成績との相関分析を行なうことが可能となる．

本論文では，まず 2 章において SCI 方式ならびにセキュア表計算システムの概要について述べ，3 章では本実験において使用したシステムと実験の内容を示すとともに，4 章において得られた実験結果について述べる．5 章では，今回の実験後に行なった学生へのアンケートの結果から，プライバシー情報公開への抵抗感を秘匿回路計算技術によってどの程度低減できたのかについて述べる．

2. セキュア表計算システム

本章では，実証実験において学生のプライバシーを保護しながら生活実態と成績との相関分析を可能とする秘匿回路計算技術 (SCI 方式) と，その実装であるセキュア表計算システムについて概説する．

SCI 方式は，秘匿回路計算の一方式として Yao が提案している 2 パーティ秘匿回路計算[4] に対し，1-out-of-2 oblivious transfer プロトコル[6] を用いることなく秘匿回路計算を実現するものである．

Yao の 2 パーティ秘匿回路計算は，目的の演算を実行する論理回路を一方のパーティが乱数化し，もう一方のパーティが乱数化された回路を用いて目的の演算を行なうプロトコルである．本プロトコルはクライアントーサーバ型であり，クライアントへ入力 α，サーバへは関数 f を入力とすると，クライアントは f に関する情報を，サーバは α に関する情報を得ることなく $f(\alpha)$ を求めることができる．関数 f は，論理ゲート $g : \{0,1\} \times \{0,1\} \rightarrow \{0,1\}$ を組み合わせた論理回路である．

筆者らは，Yao の 2 パーティ秘匿回路計算をベースとし，より簡易な手法を用いて秘匿回路計算を実現するプロトコルを提案している．SCI 方式のシステム構成図を図 1 に示す．処理は以下の手順で行なわれる．

(1) 個々の Client において，回路の入力ワイヤ i について，0 と 1 に対応する 2 つの乱数値 $W_i^0, W_i^1 \in \{0,1\}^\kappa$ ならびにランダムビット c_i を生成する．

(2) 個々の ClientC_ν が保持する生データ $\alpha_\nu \in \{0,1\}^l$ を 2 つの断片 s_ν，t_ν に分割し，代表 Client へ一旦送付する．代表 Client は，すべての Client からの s_ν，t_ν を収集し，s_ν を Proxy へ送付する．また，t_ν の個々のビット b に対応する $< W_i^b, b \oplus c_i >$ を Server へ送付する．なお，$s_\nu \in_R \{0,1\}^l$，$t_\nu = \alpha_\nu \oplus s_\nu$ である．

(3) Proxy は s_ν を用いて関数 f に対応する乱数化回路 T_g を生成し，この乱数化回路を Server へ送信する．

(4) Server は T_g と $< W_i^b, b \oplus c_i >$ から $f(\alpha_1, ..., \alpha_n)$ を得る．

現在の実装において実行が可能な演算種別は以下の 11 種類である．

- 最大値／最小値／中央値／クロス集計／平均値／最頻値／分散／加算／減算／乗算／平方算

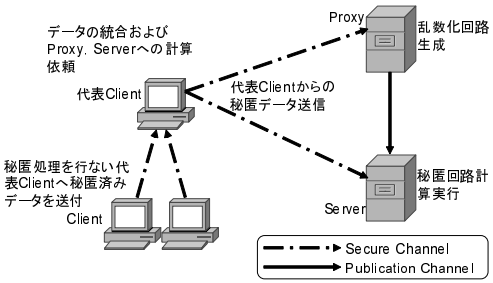

図 1 SCI 方式
Fig. 1 Two-party Secure Function Evaluation System in CSS2009

Proxy と Server は Linux 上で動作するサーバプログラムとなっており，乱数化回路の生成と秘匿回路計算の実行を行ない，計算結果を代表 Client へ返却する．

なお，図 1 において，Client から Proxy および Server へのデータの送信にはセキュアチャネルを用いているが，本実装においてはこれを PKI によって実現しており，個々の Client は Proxy および Server の公開鍵を用いて送付するデータを暗号化した上で代表 Client に送信している．

なお，代表 Client において実行される秘匿されたデータの統合に際しては，データの提供は行なわず，統合のみを実行する分析者として配置することも可能である．また，代表 Client(もしくはデータ統合のみを行なう分析者) が秘匿回路計算サーバの 2 パーティのうち一方 (Proxy もしくは Server) を兼ねる，といった構成も可能となっている．今回の実証実

Vol.2010-DPS-142 No.3
Vol.2010-CSEC-48 No.3
2010/3/4

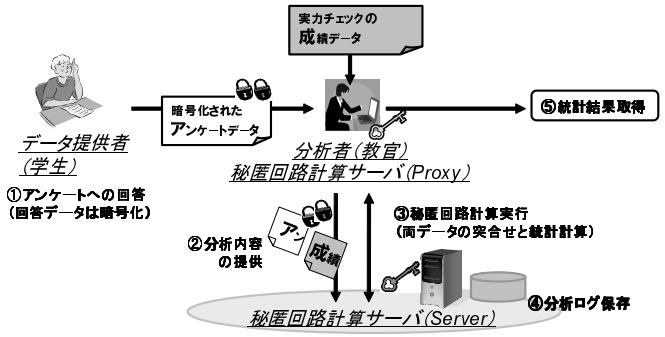

図 2 実証実験のシステム構成およびプレイヤ
Fig. 2 Evaluation System in the Field of Education

験では，代表 Client はデータ提供を行なわない分析者となっており，この分析者は Proxy も兼ねるプレイヤ構成となっている．次章において，実験時のプレイヤおよびシステムに関して述べることとする．

3. 実証実験の概要

本章では，秘匿回路計算の実証実験として行なった学生へのプライバシー保護アンケートおよびそのアンケート結果を用いた分析の内容について述べる．

3.1 実験に用いたアンケートシステム

今回の実験におけるプレイヤと，プライバシー保護アンケートおよび秘匿回路計算を実行するシステムの構成を図 2 に示す．本実験におけるプレイヤは，アンケート回答 (データ提供) および実力チェックへの回答を行なう法政大学理工学部の学生約 110 名，分析者かつ Proxy として法政大学理工学部金井研究室，そして Server を NTT 情報流通プラットフォーム研究所が担当した．以下に，アンケートの回収および秘匿回路計算による分析の手順を示す．

(1) データ提供者である学生は，学内に設置したアンケート Web サイトへアクセスし，日常の生活やプライベートに関するアンケートに Web 上で回答する．
(2) アンケートの回答データに対し，データ提供者の端末上 (ブラウザ) で秘匿処理を行なう．秘匿処理は Javascript によって実行され，秘匿済み回答データを Web サーバへ送信する．(但し，ID となる学籍番号については，平文で送信される)
(3) 複数の学生からの秘匿済み回答データが Web サーバにおいて蓄積される．
(4) 分析者は別途収集した学生の成績データを用いて，ID で名寄せを実行し，データの垂直統合および成績ランク順での秘匿済み回答データのグループ分けを行なう．
(5) 分析者は得られた秘匿済み回答データに対し，秘匿回路計算を実行して各種演算や集計を行ない，各成績ランク毎の各回答結果の平均値等を求める．

3.2 学生へのアンケートおよび分析項目

本実証実験において，学生の日常生活に関するアンケートでは，以下の質問を実施した．質問は 4 種類のカテゴリに分けており，生活／対人関係／情報行動／勉強に関するものとなっている．また，日常的な行動と，今回相関分析の対象となった実力チェックの前日の行動についてそれぞれ質問を行なっており,「普段の勉強時間と実力チェック前日の勉強時間との比較」といった分析が可能となっている．

- 生活に関するアンケート
  - 日常的な行動
    - 一人暮らしか，実家に住んでいるか
    - 通学時間
    - 睡眠時間
    - バイトに使っている時間
    - 運動をしている時間
    - 新聞を読んでいるか？
    - 読書量
    - よく読む本のジャンル
  - 実力チェック前日の行動
    - 実力チェック前日の睡眠時間
    - 実力チェックに臨んだ時の気分
- 対人関係に関するアンケート
  - 日常的な行動
    - 一月あたりの合コンの回数
    - 彼氏／彼女の有無
    - 友達の数
    - 大学に来たときに何人と話をしているか
- 情報行動に関するアンケート

Vol.2010-DPS-142 No.3
Vol.2010-CSEC-48 No.3
2010/3/4

  - 日常的な行動
    - ゲームをしている時間
    - 携帯電話を使っている時間
    - テレビを見ている時間
    - 送信しているメールの数
    - マンガを読んでいる時間
    - マンガ以外の本を読んでいる時間
- 勉強に関するアンケート
  - 日常的な行動
    - 普段の勉強時間
    - 自己啓発に使っている時間
    - 講義の理解度はどの程度か？
    - 講義は面白いか？
    - 普段勉強している場所
    - 講義でいつも座っている場所
    - 講義で居眠りした回数
    - 遅刻／欠席した回数
    - 授業時間中の私語の割合
  - 実力チェック前日の行動
    - 実力チェック前日の勉強時間

データ提供者 (アンケート回答者となる学生) に対しては，秘匿回路計算技術とアンケートシステムに関する概説を事前に行ない，個々の学生の回答結果については閲覧できない状態で統計データのみを取得できることを説明した上でアンケートを実施した．

アンケートとの相関分析に用いる成績データに関しては，今回アンケートの回答者となった学生約 110 名が受講している講義の中で行なわれた実力チェックの結果を利用している．成績の人数分布は表 1 に示すとおりであり，成績が良かった順に評価 5〜1 の 5 グループに分割している．3.1 節の処理手順において述べたとおり，Web サーバにおいて蓄積された学生からの秘匿済み回答データについて，平文で管理されている ID(学籍番号) を用いて成績データとの名寄せおよび 5 段階評価でのグルーピングを行ない，それぞれのグループに対して秘匿回路計算を実行する．これにより，例えば「成績が良かった学生 (評価 5 の学生) の普段の勉強時間は平均 n 時間」といった分析が可能となる．

表 1 実力チェックの成績分布
Table 1 Score Distribution of Midterm Examination

評価	5(良い)	4	3	2	1(悪い)
人数分布	約 10 名	約 20 名	約 50 名	約 20 名	約 10 名

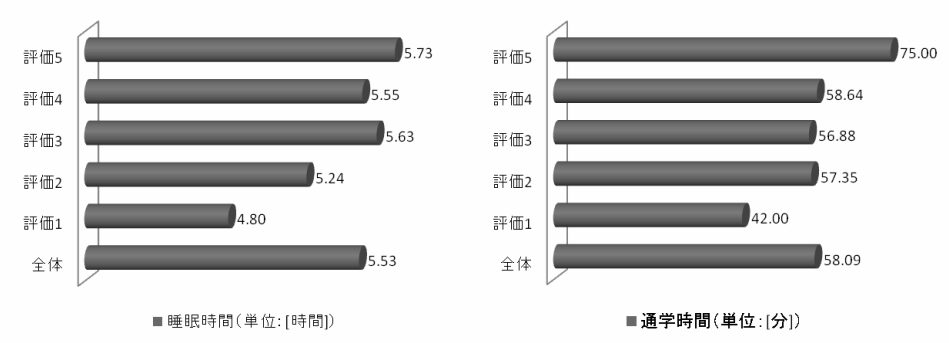

図 3 実験結果 1：睡眠時間および通学時間と成績との相関
Fig. 3 Correlation Between Academic Performance and Commuting Time/Sleeping Hours

4. 実験結果

本章では，3 章において示した学生へのプライバシー保護アンケートと成績との相関分析を行なった結果のうち，興味深かったものについて示す．

まず，図 3 として，睡眠時間と成績との相関，および通学時間と成績との相関を示す．すべての学生の平均睡眠時間が 5.53 時間であるのに対し，成績が良かった学生の方が若干睡眠時間が長く，逆に成績が悪かった学生の睡眠時間が短いという結果が得られている．同様に，成績が良かった学生の方が通学時間が長いという傾向も得られている．

次に，普段の勉強時間と成績との相関，および実力チェック前日の勉強時間と成績との相関を示したのが図 4 である．成績が良かった学生は普段の勉強時間が長く，実力チェック前日の勉強時間が短くなっているが，成績が悪かった学生は一夜漬けの傾向が強いと言える．

本実験においてアンケートを行なった学生が受講していた講義はアンケート実施時点で 6 回行なわれていたが，その 6 回の中での欠席／遅刻の回数と成績との相関を示したのが図 5 である．成績が良いほど欠席／遅刻の回数が少ないという傾向が顕著に出ていることが分かる．

Vol.2010-DPS-142 No.3
Vol.2010-CSEC-48 No.3
2010/3/4

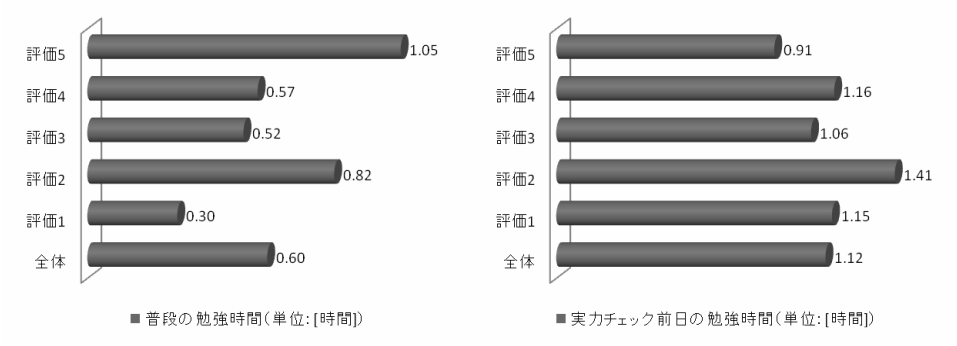

図 4 実験結果 2：普段の勉強時間および実力チェック前日の勉強時間と成績との相関
Fig. 4 Correlation Between Academic Performance and Study Time

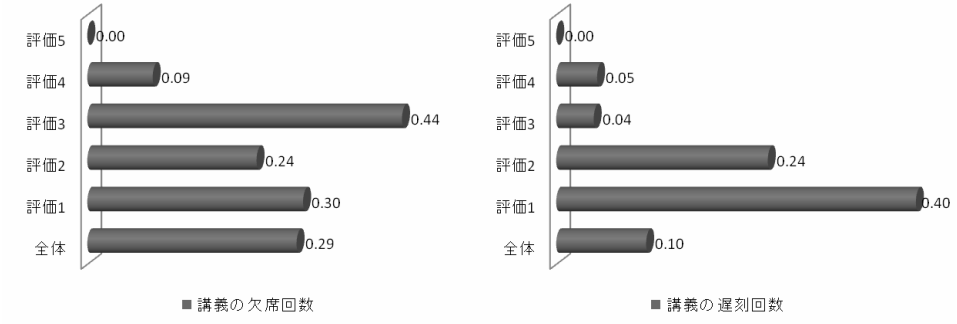

図 5 実験結果 3：講義への欠席／遅刻の回数と成績との相関
Fig. 5 Correlation Between Academic Performance and Attendance Rate

最後に，講義の際に普段座っている座席の位置と成績との相関を示したのが図 6 である．こちらも欠席／遅刻の回数の結果と同様，成績が良いほど教室の前方に座って講義を聴講している傾向が顕著に見られている．

5. 考　　察

前章では，学生の日常生活と成績の相関分析に関する分析結果を示した．分析により，学生の日常生活と成績との間には，いくつかの相関が見られるとの結果が得られている．本章では，学生の日常生活に関するアンケートとは別に，本実験の後にデータ提供者である学生へ匿名での事後アンケートを実施した結果について述べる．

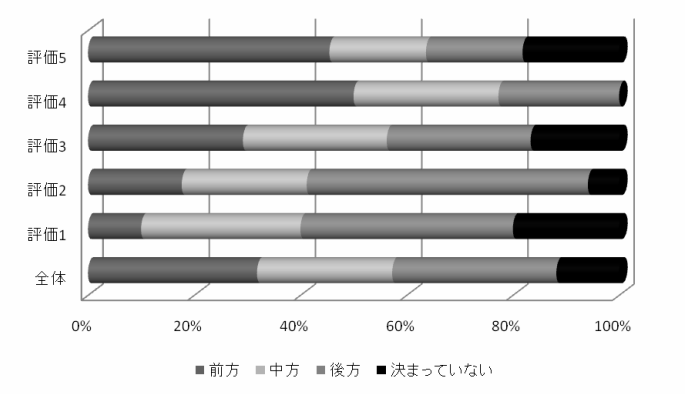

図 6 実験結果 4：講義の際に座っている場所と成績との相関
Fig. 6 Correlation Between Academic Performance and Seating Position

今回実施したアンケートでは，すべての質問に「回答したくない」という選択肢を設け，データ提供者である学生が「たとえ暗号化されていたとしてもこの質問には回答したくない」と感じた場合への対処を行なった．但し，「回答したくない」という選択肢を選んだことも秘匿されるため，どの学生が回答をしなかったのかを特定することはできない．

図 7 として，アンケートにおける 30 項目のうち，「回答したくない」を選んだ学生が 2 名以上いた質問を挙げている．「彼氏／彼女の数」が今回の質問の中で答えにくいものであることは自明であるが，その他に挙がっている質問には，講義に関係するものが多く含まれている．これは，たとえ暗号化されているとの説明を受けても，「講義の成績に影響するのではないか」という不安が拭いきれていないためであると思われる．

実験後に実施した事後アンケートでは，秘匿回路計算を用いることにより，学生が自身のプライバシー情報を公開する際の抵抗感がどの程度緩和できているかについて確認した．実験においてはほとんどの学生が「回答したくない」を選択することなく，すべての質問に答えているが，すべての質問に回答した理由については，図 8 のような結果が得られている．

今回のアンケートでの質問は「公開しても問題ない程度だった」と回答した学生が多数を占めているが，15%程度の学生が「秘匿回路計算で暗号化されているので安心したから」を

Vol.2010-DPS-142 No.3
Vol.2010-CSEC-48 No.3
2010/3/4

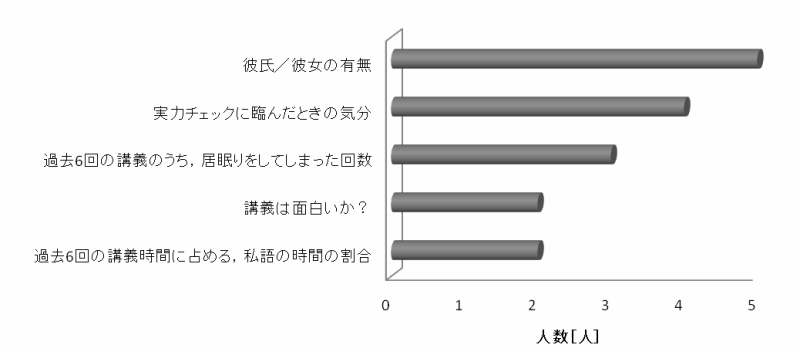

図 7 「回答したくない」が選ばれた質問とその人数
Fig. 7 Unanswerd Questions

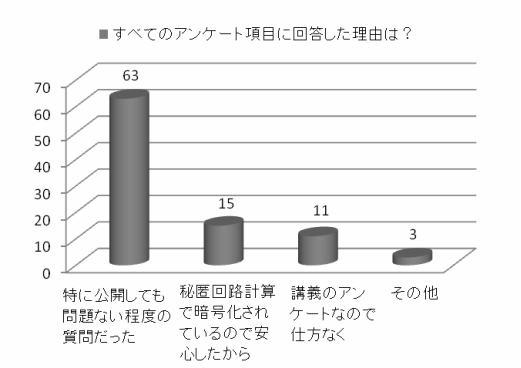

図 8 事後アンケートの結果
Fig. 8 Result of Posteriori Enquete

選んでおり，秘匿回路計算を使うことにより，若干ではあるがアンケートの回答率を向上させる効果が得られていると言える．

教育分野において，教師が学生の生活行動を把握することが難しく，教育効果を向上させるための踏み込んだ分析が困難であるとの課題を 1 章において述べた．今回の実証実験により，学生が自身のプライバシーを公開したくないという心理的な抵抗感をある程度低減できているとともに，教師が学生の生活実態の情報を一切閲覧することなく，生活行動と成績との相関分析を実施することが可能であることを示すことができた．

6. まとめと今後の課題

本論文では，筆者らが提案する秘匿回路計算技術の実証実験として，学生約 110 名を対象として学生のプライベートに関するアンケートを実施し，これを秘匿した状態で収集するとともに，学生の成績と突合せて秘匿回路計算による統計演算を行なった．実験の結果，秘匿回路計算により，教師は学生のプライベートに関する情報を一切閲覧することなく，生活実態と成績との相関分析を行なうことが可能であることを示した．

今回の実験により，データを垂直統合することによる統計分析に秘匿回路計算が有用であるということが示された．実験では，アンケートで得た日常の行動に関する情報と，成績という属性データの相関分析により，興味深い結果が得られている．日常の行動に関しては，今後携帯端末やセンシング技術の発達により，より詳細なライフログデータが取得できる可能性がある．今後ますます増大し重要視されるであろうライフログデータに対し，秘匿回路計算技術によるプライバシー保護の効果を適用することで，よりセンシティブなデータを利活用することが可能となるであろう．

参 考 文 献

1) 高橋克巳, 廣田啓一, 千田浩司, 五十嵐大: プライバシー保護データ活用技術の現状と課題, コンピュータセキュリティシンポジウム 2009 論文集, pp.757–762 (2009).
2) Samarati, P. and Sweeney, L.: Generalizing data to provide anonymity when disclosing information (abstract), *Proc. of the 17th ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems*, p.188 (1998).
3) Agrawal, R. and Srikant, R.: Privacy-preserving data mining, *Proc. of the ACM SIGMOD Conference on Management of Data*, pp.439–450 (2000).
4) A.C.Yao.: How to generate and exchange secrets, *Proc. of FOCS '86*, pp.162–167 (1986).
5) 柴田賢介, 千田浩司, 五十嵐大, 山本太郎, 高橋克巳: 表計算ソフトをフロントエンドとした委託型 2 パーティ秘匿回路計算システム, コンピュータセキュリティシンポジウム 2009 論文集, pp.625–630 (2009).
6) Even, S., Goldreich, O. and Lempel, A.: A randomized protocol for signing contracts., *Communications of the ACM 28(6)*, pp.637–647 (1985).