Randomized Response に対するポイズニング攻撃の調査

武田 花 †

明治大学総合数理学部　先端メディアサイエンス学科　菊池研究室 †

1　はじめに

データの収集や分析は，企業の経営状況の把握や新たなビジネスチャンスの発掘に繋がる．そのため企業の発展において必要不可欠なものとなっている．しかし，近年多くの漏洩事件が起こっていることも事実である．2022 年 9 月に株式会社ニトリホールディングスが約 13 万件のニトリメンバーズユーザの会員情報等の情報漏洩，2022 年 3 月には森永製菓株式会社が，展開する通信販売事業の顧客情報が漏洩した可能性があると発表した [5] [6]．データ収集の際個人情報をデータ収集者に依存することの課題が浮彫となっている．

このような課題に対処した，個人情報保護の技術の 1 つに局所差分プライバシー [1] がある．局所差分プライバシーはデバイスからデータを収集する際に確率的にノイズを付与し，入力値のデータを一定の確率で摂動する技術である．この技術により，データの収集者ですらユーザの真の値が分からなくなるため，強いプライバシーレベルが保証される．

しかし，この局所差分プライバシーは摂動結果を意図的に操作するポイズニング攻撃に対して脆弱であることが知られている [2] [3] ．そこで，本研究では局所差分プライバシーの代表的なアルゴリズムの 1 つである Randomized Response(RR)[2] に対し，ポイズニング攻撃の 1 つである Random Item Attack(RIA)[2] と Maximal Gain Attack(MGA)[2] の 2 種類の攻撃方法を乃木坂 46 のメンバーの知名度に適用し，出力結果を比較する．乃木坂 46 のメンバーの知名度に攻撃方法を適用した理由は，大人数の女性アイドルグループでは全員のメンバーを等しく応援するよりも，特に応援するメンバーを何人か作る傾向にあり，自分の応援しているメンバーが人気になってほしいとファンは考えるため実際の人気度，知名度をよりも高く見せたい心理がポイズニングの仕組みと通ずるものがあると考えたからである．作成したシステム，データ，出力結果を報告する．

2　準備

2.1　基本定義

各ユーザが自身のデータにノイズを付与し，その結果をデータの収集者に送信する．収集者は各ユーザから得られた結果を集計し，度数，平均値，推定誤差を求める．ここで，ユーザ数を n とし，ユーザの集合を $U = \{u_1, u_2, ..., u_n\}$ とする．取り扱う d 種類の離散値の集合を $K = \{k_1, k_2, ..., k_d\}$ とする．プライバシー費用を ε, ある入力を t に対してランダムアルゴリズム A を適用することを $A(t,\varepsilon)$ とする．

2.2　局所差分プライバシー

データ収集者がユーザからデータを集める際，摂動後のデータから真のユーザの入力値の推定不可能であることが保証される．任意の異なる 2 つの入力に対して，A の出力が同一にならない確率に差がないことを保証している．ランダムアルゴリズム M に対して局所差分プライバシーは以下のように定義される．

定義 1．（局所差分プライバシー）

K を入力の集合, Z を出力の集合とする．A を入力 $t \in K$ に対して $z \in Z$ を出力するランダムアルゴリズムとする．任意の 2 つの入力 $t,t' \in D$ と任意の出力 $z \in Z, \varepsilon \geq 0$ に対して，

$\forall_z \in Z : Pr[A(t,\varepsilon) = z] \leq e^{\varepsilon} Pr[A(t',\varepsilon) = z]$

が成立する時，ランダムアルゴリズム A は局所差分プライバシーを満たすという．

2.3　Randomized Response(RR)[2]

離散値データの局所差分プライバシーアルゴリズムに Randomized Response(RR)[2] がある．RR では，ユーザの持つ値 x を入力値とし，確率 $p=\frac{e^{\varepsilon}}{e^{\varepsilon}+1}$ で真の値を出力し，確率 $q=\frac{1}{e^{\varepsilon}+1}$ で d 種類の値の集合 K 中から入力値 k を除く $k' \in K - k$ を出力する．値の頻度推定は以下の式で行う期待値を E とする．

$$E = \frac{f'_x(d-1) - n(1-p)}{dp - 1}$$

2.4 Random Item Attack(RIA)[2]

各不正ユーザはランダムにアイテム $t \in K$ を選択し，選択された入力 t に RR を適用し，出力された結果 z をサーバに送信する．

2.5 Maximal Gain Attack(MGA)

RR を適用せず，出力 z を任意の値に置換する．

3 実験

3.1 実験方法

3.1.1 実験 1

ε の値と不正ユーザの数 n を変化させ，作成したデータセットに RR アルゴリズムを適用する．RIA 攻撃は計 100 回ずつ，MGA 攻撃は K の各値につき 30 回ずつ実験を行う．本実験では，林の知名度を操作する．
この一連の流れを行うプログラムを Python を用いて実装した．データの入力，攻撃，摂動，出力までの流れを図 1 に示す．

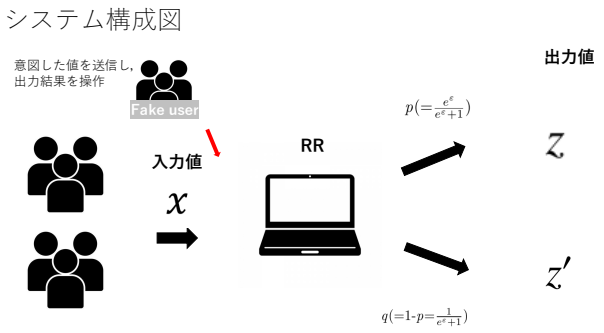

図 1　システム構成図

3.1.2 実験 2

RR の推定精度調査実験では，ε の値と真の値 i を変化させ，作成したデータセットに RR アルゴリズムを適用する．各値につき，5 回ずつ実験を行う．この一連の流れを行うプログラムを Python を用いて実装した．データの入力，摂動，出力までの流れを図 2 に示す．

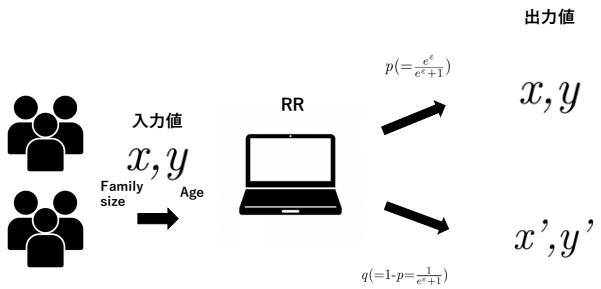

図 2　システム構成図

3.2 目的

本研究では，ユーザが持つデータを RR により摂動する時，どの様な条件のポイズニング攻撃における影響が大きいか明らかにする事，RandomizedResponse(RR) における推定誤差が最も小さくなる秘匿化度合 ε の調査との比較を目的とする．RR の調査ではポイズニング攻撃は行わない．

3.3 データセット

3.3.1 実験 1

データセットには, 本研究で実施したアンケートを使用する．アンケートは GoogleForm によって作成した．乃木坂 46 のメンバー 29 名に対する知名度を調査している．名前と顔写真を見て，メンバーを知らない場合は 0，知っている場合は 1 に回答する．被験者は明治大学総合数理学部先端メディアサイエンス学科菊池研究室に所属する学部 2,3,4 年生を対象とし，期日までに回答が得られたのは n=13 名であった．表 1 にアンケートの結果を示す．

3.3.2 実験 2

RR の推定精度に調査に使用したデータセットは onlinedeliverydata[4] を使用する．本実験では age,familysize のカラムを使用する．このデータはオンラインデリバリーを利用した人の世帯数，年齢，性別等の個人情報を含む．ユーザ数 n=389 人分のデータを使用する．

表 1　アンケート結果

メンバー名	知らない:0	知っている:1
白石	0	13
生田	1	12
齋藤	1	12
生駒	2	11
秋元	2	11
山下	4	9
与田	6	7
堀	7	6
新内	8	5
遠藤	8	5
賀喜	8	5
中西	8	5
樋口	9	4
北野	9	4
梅澤	9	4
久保	9	4
林	9	4
井上	9	4
岩本	10	3
筒井	10	3
池田	10	3
鈴木	11	2
伊藤	11	2
中村	11	2
金川	11	2
松尾	11	2
一ノ瀬	11	2
奥田	11	2
小川	11	2
平均	9.48	6.2
最大	11	13
最小	0	2
標準偏差	3.51	3.51

3.4　評価方法

n 人のユーザから出力されたデータに対し，各メンバーの認知頻度を推定する．推定値 E_i と真の頻度 f_i の差の絶対値の平均で精度を求める．

RR の推定精度の調査では，出力されたデータから度数と平均値を推定する．推定値 E_i と度数の差の絶対値の和から局所差分プライバシーアルゴリズム RR の精度を求める．また，f_x を真の値の総数とした時，アルゴリズムの精度 S は

$$S = \frac{|f_x - E|}{n}$$

で定める．

3.5　実験結果

MGA 攻撃と RIA 攻撃を，プライバシー費用 ε=0.5,0.8,1.0，不正ユーザ=3,5,10 の場合について適用した時の推定誤差を, 図 3, 図 4, 図 5, 図 6 に示す．図 3, 図 5 は知名度を増やす操作を行った場合，図 4, 図 6 は知名度を減らす操作を行った場合についての実験結果を示している．RR の推定精度の調査実験ではユーザ数 n=389，ε=0.1,0.5,1.0 の場合について，最もアルゴリズムの精度が高くなった条件の結果を以下の表 2, 図 7 に, 次いで精度が良かった結果を図 8 に示す．最も精度が低くなった条件の結果を表 3，図 9 に, 次いで精度が悪かった結果を図 10 に示す．

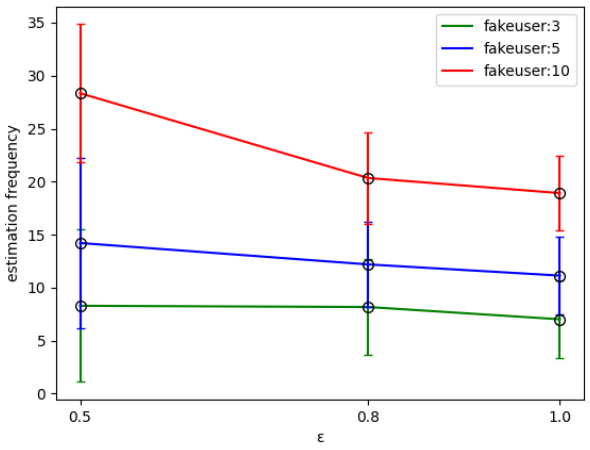

図 3　MGA による推定誤差 (知名度増)

表 2　平均絶対誤差 MAE:age=28, ε=1.0

fs	1 回目	2 回目	3 回目	4 回目	5 回目
1	52.27	51.85	52.04	51.72	51.62
2	752.62	757.28	750.13	752.84	750.39
3	559.84	561.00	563.62	557.96	562.01
4	0	0	0	0	0
5	241.63	243.48	241.59	241.65	241.67
6	57.26	56.14	56.39	57.24	57.42

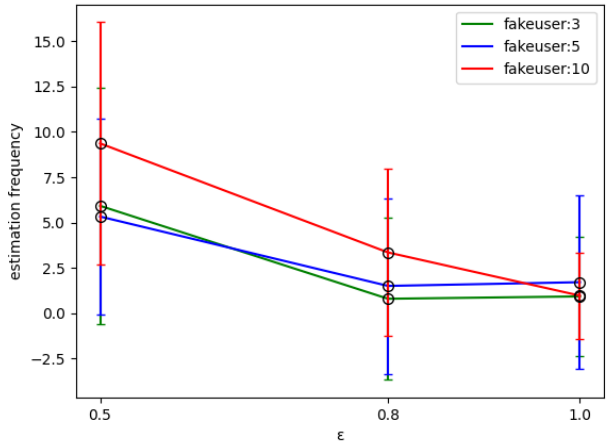

図 4　MGA による推定誤差 (知名度減)

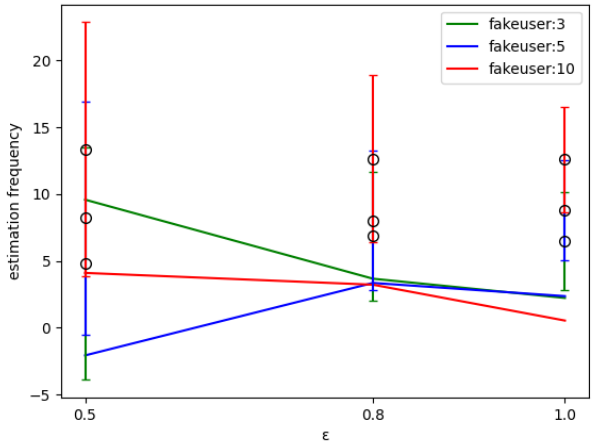

図 6　RIA による推定誤差 (知名度減)

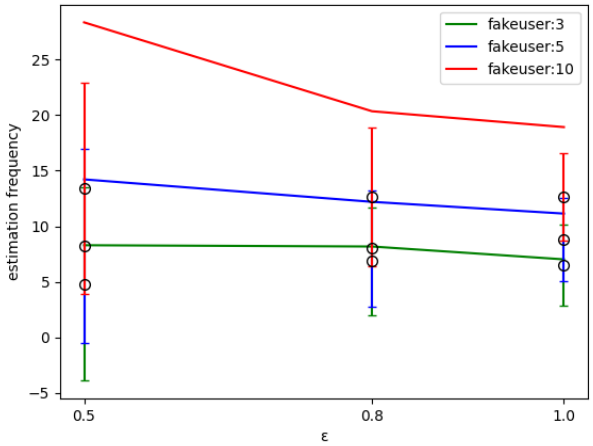

図 5　RIA による推定誤差 (知名度増)

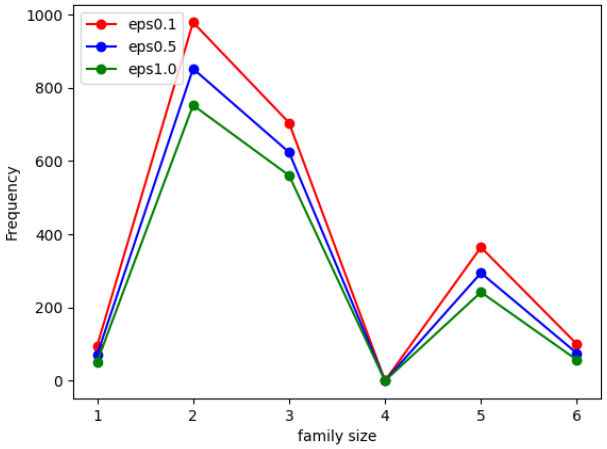

図 7　age=28

表 3　平均絶対誤差 MAE:age=25, ε=0.1

fs	1 回目	2 回目	3 回目	4 回目	5 回目
1	484.22	473.51	478.47	475.96	478.47
2	2123.37	2123.64	2119.02	2121.46	2117.59
3	3880.91	3888.99	3875.77	3862.09	3872.32
4	901.56	911.61	905.91	909.31	906.75
5	199.29	243.26	241.89	244.50	243.43
6	298.94	298.11	302.51	299.94	300.63

3.6　評価

実験結果より，誤差が最も大きいのは ε=0.5, 不正ユーザ=10 の場合に知名度を増やす MGA 攻撃を行った時で，28.3 となった．反対に，最も誤差が小さいのは ε=0.5，不正ユーザ=3 の場合に知名度を減らす RIA 攻撃を行った時で，2.5 となった．知名度を増やす攻撃の場合の方が誤差が大きくなる傾向があるのは，本実験で知名度を操作したメンバー（林瑠奈）の元の知名度が知らない人の割合の方が多いからであると考えられる．どちらの攻撃の場合でも不正ユーザ数を増やすほど誤差が大きくなることが確認できた．一般にプライバシー費用 ε が大きくなるほど誤差が小さくなる．しかし，RIA 攻撃の場合は, 偽ユーザ数によってプライバシー費用 ε と誤差が反比例しなかった．プライバシー費用 ε と誤差に相関関係が見られなかった．また，RIA 攻撃よりも，MGA 攻撃の方がプライバシー費用 ε を変化させた時の変化が大きい．

RR の推定精度調査の結果は，精度が最も高いのは age=28,familysize=1,ε=1.0 の場合に 51.9 となった．ユーザ数 n=0 の場合は精度が 0 になるため，除外している．最も精度が低いのは age=25,familysize=3,ε=0.1

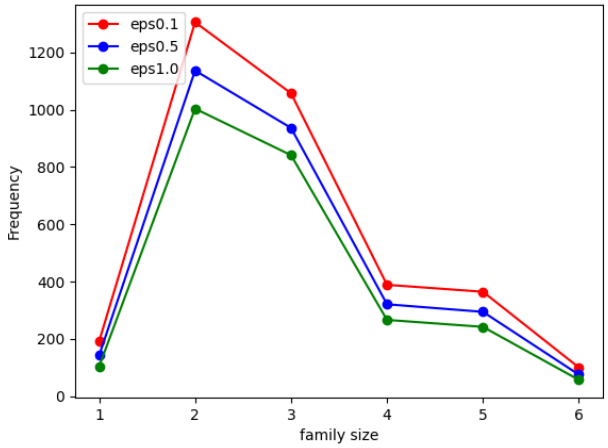

図 8　age=21

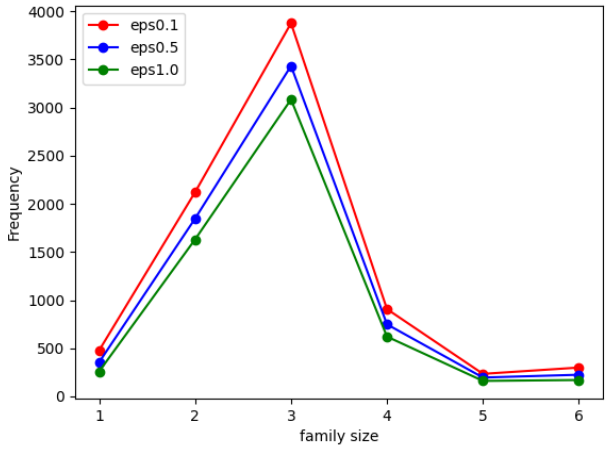

図 9　age=25

の場合に 3876.02 となった．アルゴリズムの推定精度を調査する場合にもプライバシー費用 ε が大きくなる場合に誤差が小さくなることが確認出来た．ユーザ数が多い場合の条件において精度が低く，ユーザ数が少ない場合の条件の場合に比較的精度が高くなった．アルゴリズムの精度にデータセットの数が関係することが考察出来た．また，どのプライバシー費用 ε の場合にも 1 秒以内に実行完了すること，5 回の試行でどの条件においても結果の幅が少なくなったことから，本研究で作成したアルゴリズムの安定性を確認出来た．

4　おわりに

本稿では，局所差分プライバシーアルゴリズムの 1 つである Randomized Response(RR) に対するポイズニング攻撃の調査を行った．実験の結果から，RIA より MGA 攻撃の方がポイズニングで生じる誤差が大きいことが明らかになった．また，RIA 攻撃の場合はプライバシー費用 ε と誤差の関係で一般に予想される結果が得られなかった．そのため，今後は，RIA 攻撃においてプライバシー費用 ε が誤差と反比例しない理由についてシステムの再実装と考察を課題とする．

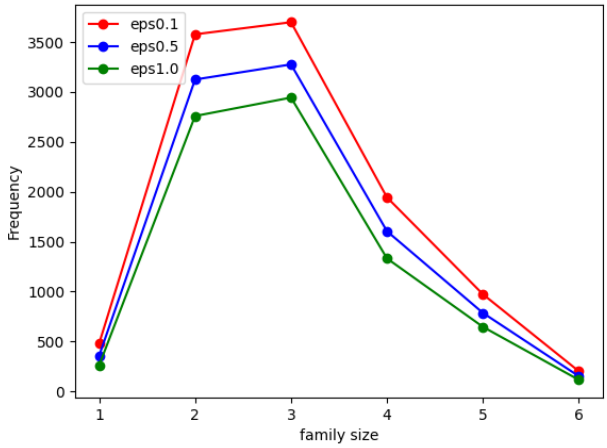

図 10　age=23

参考文献

[1] Tianhao Wang, Jeremiah Blocki, and Ninghui Li, ”Locally Differentially Private Protoclos for Frequency Estimation”,The Proceedings of the 26th USENIX Security Symposium,August , pp.729-746, 2017.

[2] Xiaoyu Cao, Jinyuan Jia, Neil Zhenqiang Gong ”Data Poisoning Attacks to Local Defferential Privacy Protocols”,The Proceedings of the 30th USENIX Security Symposium,August, pp.947-964, 2021

[3] Yongji Wu, Xiaoyu Cao, and Neli Zhenqiang Gong, ”Poisoning Attacks to Local Differential Privacy Protocols for Key-Value Data”, The Proceedings of the 31st USENIX Security Symposium, August, pp.519-536, 2022.

[4] Kaggle ”Online Food Delivery Preferences-Bangalore region”. (https://www.kaggle.com/datasets/benroshan/online-food-delivery-preferencesbangalore-region),(参照 2022-12)

[5] ” 不正アクセス発生による個人情報流出の可能性のお知らせとお詫

び”,(https://www.morinaga.co.jp/company/newsrelease/detail.php?no=2178)(参照 2022-5)
[6] ”「ニトリアプリ」への不正アクセスによる個人情報流出の可能性に関するお詫びとお知らせ”,(https://www.nitorihd.co.jp/news/items/2cdbc59fa4c3ffe4da790cc1cfe85200.pdf)(参照 2022-5)