RANDOMIZED RESPONSEに対するポイズニング攻撃の調査

総合数理学部　先端メディアサイエンス学科

菊池研4年　武田　花

乃木坂問題

- 推しメンの知名度を上げたい！
  - ポイズニング攻撃を用いて知名度を操作
- 林の知名度を操作
- 有名なメンバーを知らなくて恥ずかしい，推しメンの印象に悪影響を与えたくないから自分の回答を隠したい等，様々な理由で自分の回答を隠したい
  - 局所差分プライバシーを用いてプライバシーを守る

局所差分プライバシー　（RR）

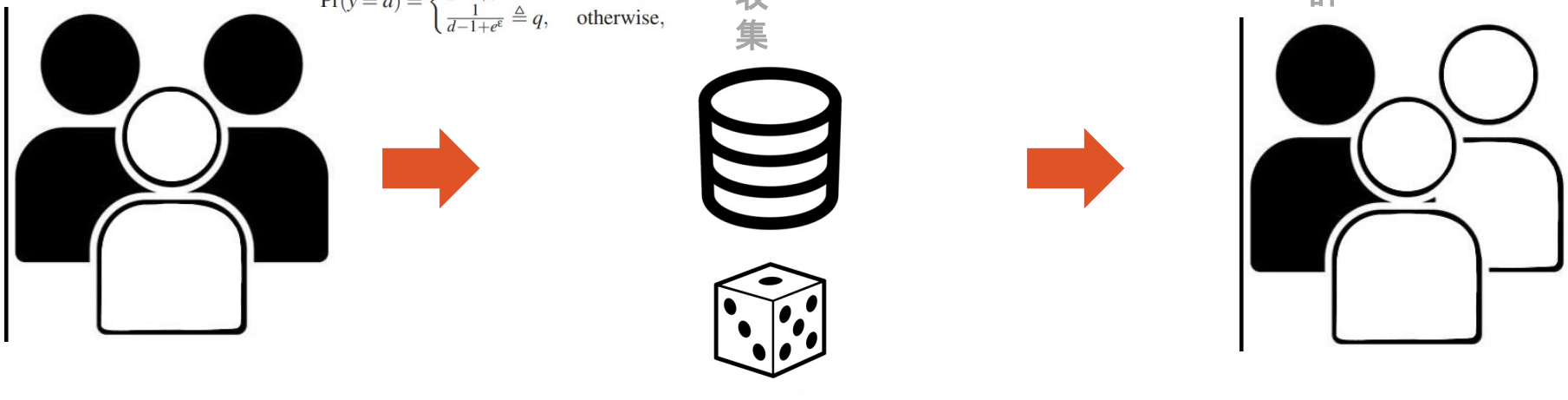
- 値をランダムに摂動してから収集者に送信
  ➡真の値が本人しか分からないため，プライバシーが保護される
- 秘匿化度合は自分で設定する（低いほど安全性が高い）
- d:値の種類（今回の実験ではo:知らない　1:知っている　なので、d=2)

値を摂動して送信

$$\Pr(y=a)=\begin{cases}\frac{e^{\varepsilon}}{d-1+e^{\varepsilon}} \triangleq p, & \text{if } a=v,\\ \frac{1}{d-1+e^{\varepsilon}} \triangleq q, & \text{otherwise,}\end{cases}$$

収集

集計



RR

- LDPアルゴリズムの１つ

$$\Pr(y=a)=\begin{cases}\frac{e^{\varepsilon}}{d-1+e^{\varepsilon}} \triangleq p, & \text{if } a=v,\\ \frac{1}{d-1+e^{\varepsilon}} \triangleq q, & \text{otherwise,}\end{cases}$$

- P:維持確率　Q:遷移確率　ε:秘匿化度合
- この確率に従いランダムに値を摂動していく
- 秘匿化度合は自分で設定する（低いほど安全性が高い）
- d:値の種類（今回の実験ではo:知らない　1:知っている　なので、d=2)

ポイズニング攻撃①

MGA

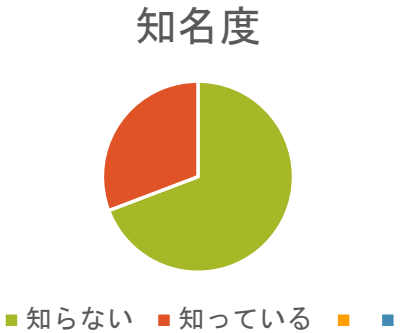
ユーザー

暗号化

データベース

統計情報の
取得

データ利用者



攻撃者

攻撃者が望む結果を送信

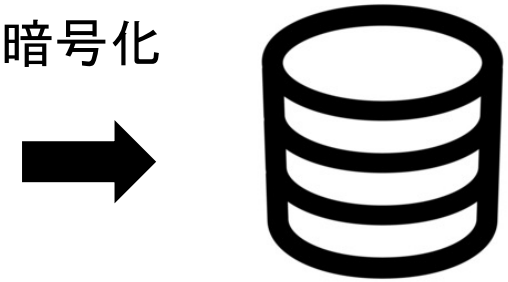
ポイズニング攻撃②

RIA

データ利用者

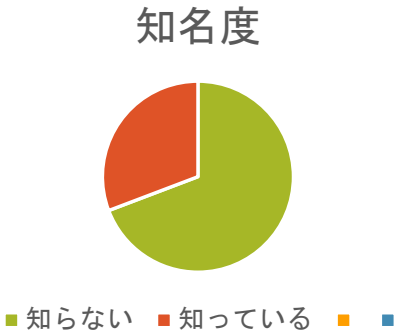
ユーザー

暗号化



データベース

統計情報の
取得

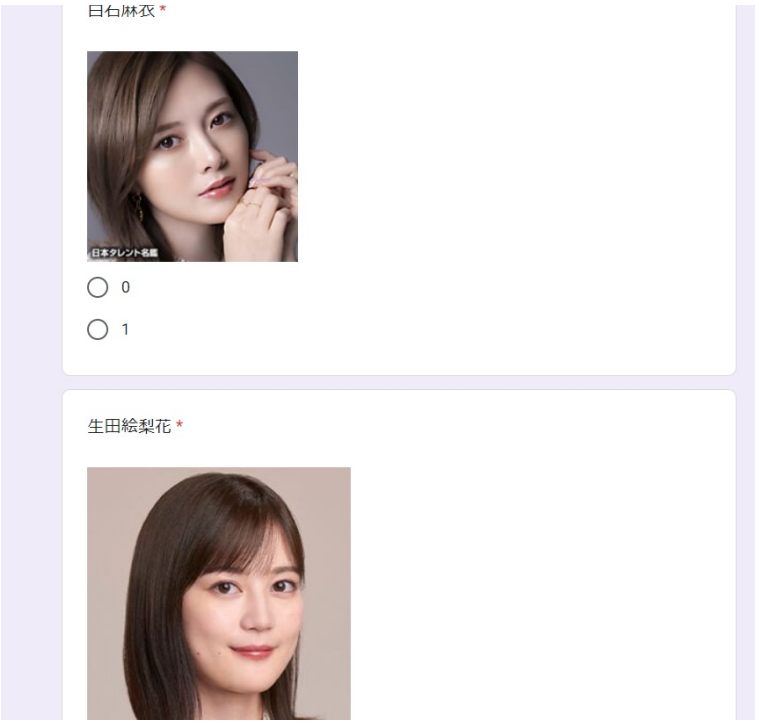


攻撃者

攻撃者がランダムにアイテムを
選択

データセット

- Google Formで作成
- 乃木坂46のメンバー29名（OG含む）についての知名度調査
- 画像+名前を見て知らない場合:0，知っている場合:1を選択
- 菊池研学部2，3，4年生対象
- 期日までに回答が得られた13名分のデータを実験に使用
- MGA攻撃，RIA攻撃それぞれで知名度を上げる攻撃，下げる攻撃を行う
- MGA攻撃は30回，RIA攻撃では100回ずつ試行



回答結果

- 知名度が最も高い:白石

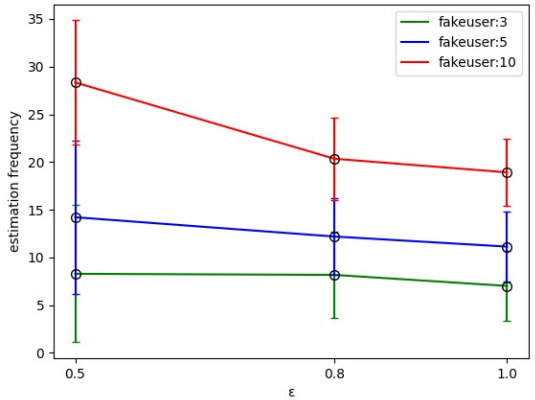
- 知名度が最も低い:鈴木，伊藤，中村，金川，松尾
一ノ瀬，奥田，小川

意外性はそこまでなかった、個人的には自分の推し（林）の知名度が思ったより高かった

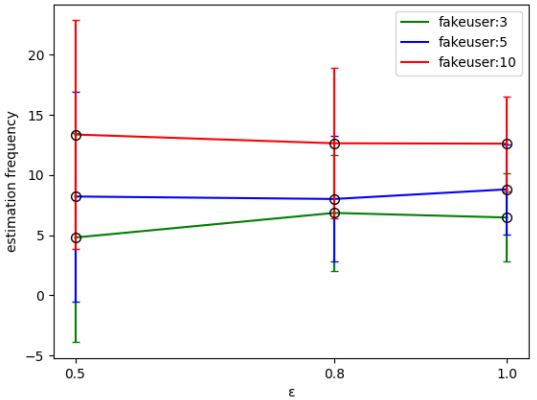
表1　アンケート結果

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

実験結果①
知名度を上げる攻撃



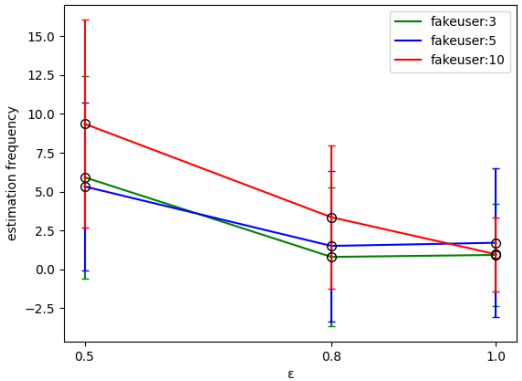
MGA



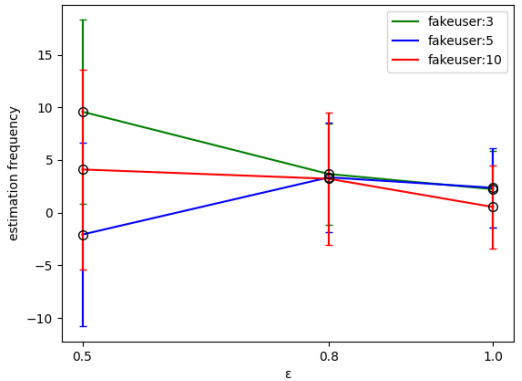
RIA

MGA攻撃の方がRIA攻撃に比べ，平均的に誤差が大きい
RIA攻撃はプライバシー費用εが上がると誤差が小さくなる一般的な傾向が見られない

実験結果②
知名度を下げる攻撃



MGA



RIA

・どちらの結果でもプライバシー費用εと誤差の一般的な関係は観測出来なかった
・知名度を操作したメンバー（林）の知名度が元から低いため、知名度を上げる攻撃に比べ差が出なかった
・知名度を下げた場合も，上げた場合と同様にプライバシー費用εと誤差の関係が一般に予想される結果とは異なった.

まとめ

- RIA攻撃の場合最も大きい誤差はε=0.5,fakeuser=10の条件で知名度を上げる操作を行った時13.4，MGA攻撃の場合最も大きい誤差はε=0.5,fakeuser=10の条件で知名度を上げる操作を行った時28.3のため，MGA攻撃による影響の方が大きいことが明らかになった

- RIA攻撃の場合，知名度を上げる攻撃と下げる攻撃のどちらにおいても一般的にはプライバシー費用εが大きくなると誤差が小さくなるが，その傾向が見られなかった。