RAPPOR における差分プライバシー評価

Hirotsugu Seike

2025 年 12 月 7 日

1　導入

RAPPOR [1] は, Google の U. Erlingsson らによって 2014 年に提案された, クライアントが保持する値 v をローカル差分プライバシーを保証したまま収集し, 統計的に分析するためのアルゴリズムである. 値 v から得られる h 個のハッシュを用いたブルームフィルタの出力を B として, $B = (B_1, ..., B_m)$ で定義する (m は十分に大きいとして, h 個のビットの値が 1 とする). このビット列 B に対して, 以下の Permanent Randomized Response (PRR) 処理を施したビット列を $B' = (B'_1, ..., B'_m)$ を次式で定義する.

$$B'_i = \begin{cases} 1, & \text{with probability } \frac{f}{2}, \\ 0, & \text{with probability } \frac{f}{2}, \\ B_i, & \text{with probability } 1 - f. \end{cases} \tag{1}$$

また, ビット列 B' に対して, 以下の Instantaneous Randomized Response (IRR) 処理を施したビット列 $S = (S_1, ..., S_m)$ を次式で定義する (S の全てのビットは初期値で 0).

$$P(S_i = 1) = \begin{cases} q, & \text{if } B'_i = 1, \\ p, & \text{if } B'_i = 0. \end{cases} \tag{2}$$

ユーザはビット列 B' は固定して, (2) 式の処理で得られるビット列 S を都度サーバーに送信する (図. 1 を参照).

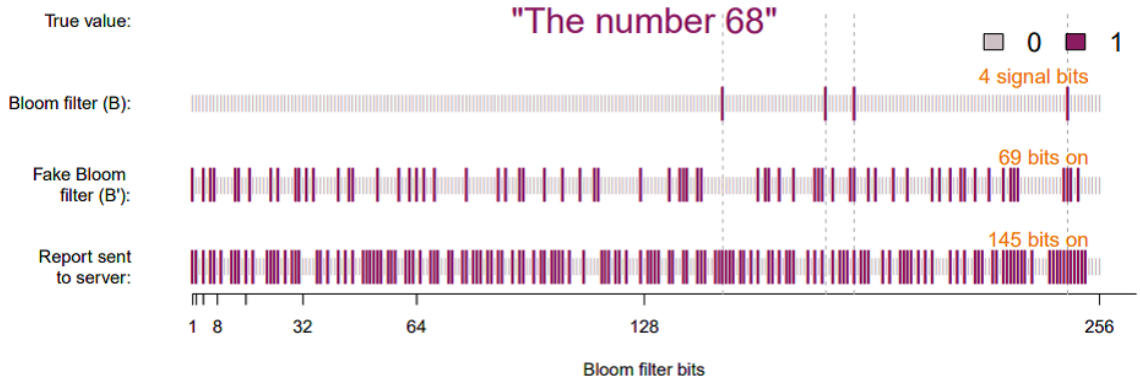

図 1: RAPPOR アルゴリズムにより送信されるビット列 S 生成の流れ ([1] の Fig. 1 より引用).

2 RAPPOR アルゴリズムのローカル差分プライバシーの評価

以下, 大文字を確率変数, 小文字を実現値と考える.

2.1 PRR に関する ϵ_∞-差分プライバシー

クライアントの保持する v に対して, 最終的にサーバーにレポートするビット列が s となる確率は次式で与えられる (ハッシュ関数計算は決定論的に実行されるため, $P(B|v) = 1$).

$$\begin{aligned} P(S = s|V = v) &= P(S = s|B', B, v) \cdot P(B'|B, v) \cdot P(B|v), \\ &= P(S = s|B') \cdot P(B'|B). \end{aligned} \tag{3}$$

ここで, B' が与えられたとき, S は B に依存しない確率変数であるため, 上式の確率 $P(S = s|B')$ は, 差分プライバシーに与える影響はない. 従って, $P(B'|B)$ について論ずれば十分である.

何らかの確率的な処理を施す機構 A が, 値 v_1, v_2 について, ϵ 差分プライバシーを満たすとは, 次の不等式が成立することである ([2] を参考. R は機構 A の出力により構成される任意集合).

$$\frac{P(A(v_1) \in R)}{P(A(v_2) \in R)} \leq e^\epsilon. \tag{4}$$

従って, ブルームフィルタの入力を B_1, B_2, ビット列 B' が取りうる任意の集合を R^* として以下のようにして, 確率比の上限を求める.

$$\begin{aligned} RR_\infty &= \frac{P(B' \in R^*|B = B_1)}{P(B' \in R^*|B = B_2)}, \\ &= \frac{\sum_{B'_j \in R^*} P(B' = B'_j|B = B_1)}{\sum_{B'_j \in R^*} P(B' = B'_j|B = B_2)}, \\ &\leq \max_{B'_j \in R^*} \frac{P(B' = B'_j|B = B_1)}{P(B' = B'_j|B = B_2)}. \end{aligned} \tag{5}$$

ここで, B_i がビット列であるため離散的な総和で確率を表現できること, および付録 A の不等式が成立することによって, 上限が計算できていることに注意する. (5) 式の右辺が最大となるケースは, 2 つのブルームフィルタのハミング距離が $2h$ になる時である. また, (1) 式より, (6), (7) 式が成立し, $1 - f/2 > f/2$ である.

$$P(b'_i = 1|b_i = 1) = \frac{1}{2}f + 1 - f = 1 - \frac{1}{2}f, \tag{6}$$

$$P(b'_i = 1|b_i = 0) = \frac{1}{2}f. \tag{7}$$

従って, (8) 式が成立する.

$$\max_{B'_j \in R^*} \frac{P(B' = B'_j|B = B_1)}{P(B' = B'_j|B = B_2)} = \frac{(1 - f/2)^h}{(f/2)^h} \cdot \frac{(1 - f/2)^h}{(f/2)^h}. \tag{8}$$

右辺の第一因子は, ビット列 B_2 の h 個の 0 が 1, ビット列 B_1 の同じ位置のビットが 1 から 1 になる確率. 右辺の第二因子は, ビット列 B_2 の h 個の 1 が 0, ビット列 B_1 の同じ位置のビットが 0 から 0 になる確率と捉えればよい. つまり, $RR_\infty = ((1 - f/2)/(f/2))^{2h}$ となるため, $\epsilon_\infty = 2h \ln((1 - f/2)/(f/2))$ となる.

2.2 IRR に関する差分プライバシー

ビット列の i 番目の要素 S_i, B_i について, 次式が成立する.

$$q^* = P(S_i = 1|B_i = 1) = \frac{1}{2}f(p+q) + (1-f)q, \tag{9}$$

$$p^* = P(S_i = 1|B_i = 0) = \frac{1}{2}f(p+q) + (1-f)p \tag{10}$$

従って, PRR の時と同様な議論をして, 1 回の IRR が保証する差分プライバシーは次式で表現できる (ここで, $q^* > p^*$ (即ち $q > p$) となるようにパラメータを設定する).

$$\begin{aligned} RR_1 &= \frac{P(S \in R|B = B_1)}{P(S \in R|B = B_2)}, \\ &= \frac{\sum_{S_j \in R} P(S = S_j|B = B_1)}{\sum_{S_j \in R} P(S = S_j|B = B_2)}, \\ &\leq \max_{S_j \in R} \frac{P(S = S_j|B = B_1)}{P(S = S_j|B = B_2)}, \\ &= \left[\frac{q^*(1-p)^*}{p^*(1-q^*)}\right]^h. \end{aligned} \tag{11}$$

つまり, $\epsilon_1 = h\ln((q^*(1-p)^*)/(p^*(1-q^*)))$. これが 1 回の IRR を実施した際の差分プライバシー保障である. IRR を繰り返すたびに, この差分プライバシーの上限値は大きくなるが, その上限は ϵ_∞ で抑えられている.

参考文献

[1] U. Erlingsson, V. Pihu and A. Korolova, "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response," *arXiv:1407.6981*, 2014.

[2] C. Dwork, "A firm foundation for private data analysis," in *Comm. ACM 54(1)*, pp.86-95, 2011.

付録 A Max-Dominance Inequality

次の不等式が成立する. 証明は, 分子を $\sum_i b_i \cdot (a_i/b_i)$ と考えて, b_i に関する重み付き平均を取っていると考えれば, 明らかである (常に最大の重み $M = \max_i a_i/b_i$ を取るケースを考えれば, 等式が成立する).

$$\frac{\sum_i a_i}{\sum_i b_i} \leq \max_i \frac{a_i}{b_i}. \tag{12}$$