Streaming CW-Pivacy

October 6, 2014

Problem Setting

Let X be a database in the streaming setting. Let X_i represent the portion of X that is currently held at time step i. We assume that at each time step, a fraction of c of the database is replaced. We assume the oldest rows are always the ones replaced, and that X has row drown i.i.d. from some distribution D. Let n be the size of each X_i. This means that the first n rows of X constitute X_1, rows $cn+1$ through $cn+n$ constitute X_2, and so forth. X has size $n + cn(t-1)$, where t is the total number of time steps being considered. $1/c$ is the total number of time steps a given row will be present for.

We now consider a query $F : \mathcal{U}^n \to \mathbb{R}^d$ on each X_i. Let D_F be the distribution that draws a database of size n, with each row chosen i.i.d. from D. Let aux_F be F's auxiliary information, which consists of the first (or last) $(1-c)n$ rows of the database. Now, assume F is $(\epsilon, \delta, \Delta_F, \Gamma)$-CW private with some simulator sim_F, where Δ chooses the database according to D_F and the auxiliary information is aux_F as stated above.

Let $G(X) = (F(X_1), F(X_2), \dots, F(X_t))$ be the composite query that runs F at each time step. We show G is $(t\epsilon, t\delta, \Delta, \Gamma)$ -CW private.

Notations

For each X_i of size n, we represent it by $1/c$ blocks (each has cn rows). Namely, let $X_i = [x_{i1}, x_{i2}, \dots, x_{i\frac{1}{c}}]^\top$, where x_{ij} denote the jth block in X_i. Let $X_{i\downarrow k} = [x_{i(k+1)}, x_{i(k+2)}, \dots, x_{i\frac{1}{c}}]^\top$ and $X_{i\uparrow k} = [x_{i1}, x_{i2}, \dots, x_{i(\frac{1}{c}-k)}]^\top$. We use $X_{i\downarrow}$ to denote $X_{i\downarrow 1}$ and $X_{i\uparrow}$ to denote $X_{i\uparrow 1}$. Notice that $X_{i\downarrow k} = X_{i+k\uparrow k}$ (specifically $X_{i\downarrow} = X_{i+1\uparrow}$), which are the shared blocks between X_i and X_{i+k}.

Let $S = (S_1, S_2, \dots, S_t)$ be any set from $\mathbb{R}^{d\times t}$, where S_j is determined by the values of $(s_1, s_2, \dots, s_{j-1})$. Let S_{-t} denote set $(S_1, S_2, \dots, S_{t-1})$.

Proof
We prove it inductively.

1. The base case is when $G(X) = (F(X_1), F(X_2))$. For any set $S = (S_1, S_2)$, we have

$$Pr[G(X) \in S \,|\, \mathbf{priv}(X) = v]$$

$\mathbf{priv}(X) = v$ will be omitted from now on.

$$= Pr[F(X_2) \in S2 \,|\, F(X_1) \in S_1] \cdot Pr[F(X_1) \in S_1]$$

$$= (\Sigma_{s_1 \in S_1} Pr[F(X_2) \in S2 \,|\, F(X_1) = s_1] \cdot Pr[F(X_1) = s_1]) \cdot Pr[F(X_1) \in S_1] \quad (1)$$

We focus on $Pr[F(X_2) \in S2 \,|\, F(X_1) = s_1]$.

$$Pr[F(X_2) \in S2 \,|\, F(X_1) = s_1] = \Sigma_z Pr[F(X_2) \in S2 \,|\, F(X_1) = s_1, X_{2\uparrow} = z] \cdot Pr[X_{2\uparrow} = z]$$

$$= \Sigma_z Pr[F(X_{2\uparrow}, x_{2\frac{1}{c}}) \in S2 \,|\, F(x_{11}, X_{1\downarrow}) = s_1, X_{2\uparrow} = z] \cdot Pr[X_{2\uparrow} = z]$$

$$= \Sigma_z Pr[F(X_{2\uparrow}, x_{2\frac{1}{c}}) \in S2 \,|\, F(x_{11}, X_{2\uparrow}) = s_1, X_{2\uparrow} = z] \cdot Pr[X_{2\uparrow} = z]$$

Given $X_{2\uparrow} = z$ and $x_{2\frac{1}{c}}$ and x_{11} are i.i.d. generated, functions $F(X_{2\uparrow}, x_{2\frac{1}{c}})$ and $F(x_{11}, X_{2\uparrow})$ are independent with each other. Only S_2 depends on the value s_1. By the assumption that F is $(\epsilon, \delta, \Delta_F, \Gamma)$-CW private with simulator sim_F, we have

$$\leq \Sigma_z (e^\epsilon Pr[sim_F(\mathbf{alt}(X_2)) \in S2 \,|\, F(X_1) = s_1, X_{2\uparrow} = z] + \delta) \cdot Pr[X_{2\uparrow} = z]$$

$$= e^\epsilon Pr[sim_F(\mathbf{alt}(X_2)) \in S2 \,|\, F(X_1) = s_1] + \delta$$

By equation (1), we have

$$Pr[G(X) \in S] \leq e^\epsilon Pr(sim_F(\mathbf{alt}(X_2)) \in S2, F(X_1) \in S_1) + \delta$$

$$= e^\epsilon Pr[F(X_1) \in S_1 \,|\, sim_F(\mathbf{alt}(X_2)) \in S2] \cdot Pr[sim_F(\mathbf{alt}(X_2)) \in S2] + \delta \quad (2)$$

Similarly, $Pr[F(X_1) \in S_1 \,|\, sim_F(\mathbf{alt}(X_2)) \in S2] =$

$$\Sigma_{s_2 \in S_2} Pr[F(X_1) \in S_1 \,|\, sim_F(\mathbf{alt}(X_2)) = s_2] \cdot Pr[sim_F(\mathbf{alt}(X_2)) = s_2] \quad (3)$$

We focus on $Pr[F(X_1) \in S_1 \,|\, sim_F(\mathbf{alt}(X_2)) = s_2]$, which equals to

$$\Sigma_z Pr[F(X_1) \in S_1 \,|\, sim_F(\mathbf{alt}(X_2)) = s_2, X_{1\downarrow} = z] \cdot Pr[X_{1\downarrow} = z]$$

$= \Sigma_z Pr[F(x_{11}, X_{2\uparrow}) \in S_1 \,|\, sim_F(\mathbf{alt}(X_{1\downarrow}, x_{2\frac{1}{c}})) = s_2, X_{1\downarrow} = z] \cdot Pr[X_{1\downarrow} = z]$

Notice that $sim_F(\mathbf{alt}X_{1\downarrow}, x_{2\frac{1}{c}}))$ can be seen a composite function $sim_F \circ$ **alt** on $x_{2\frac{1}{c}}$. Given $X_{1\downarrow} = z$ and that $x_{2\frac{1}{c}}$ and x_{11} are i.i.d. generated, functions $F(x_{11}, X_{1\downarrow})$ and $sim_F(\mathbf{alt}(X_{1\downarrow}, x_{2\frac{1}{c}}))$ are independent with each other. Only S_1 depends on the value s_2. By the assumption that F is $(\epsilon, \delta, \Delta_F, \Gamma)$-CW private with simulator sim_F, we have

$\leq (e^\epsilon \Sigma_z Pr[sim_F(X_1) \in S_1 \,|\, sim_F(\mathbf{alt}(X_2)) = s_2, X_{1\downarrow} = z] + \delta) \cdot Pr[X_{1\downarrow} = z]$

$$= e^\epsilon Pr[sim_F(X_1) \in S_1 \,|\, sim_F(\mathbf{alt}(X_2)) = s_2] + \delta$$

By equation (2) and (3), we have

$Pr[F(X_1) \in S_1 \,|\, sim_F(\mathbf{alt}(X_2)) \in S2] \leq e^\epsilon Pr[sim_F(X_1) \in S_1 \,|\, sim_F(\mathbf{alt}(X_2)) \in S_2] + \delta$

and therefore

$Pr[G(X) \in S] \leq e^{2\epsilon} Pr[sim_F(\mathbf{alt}(X_1)) \in S_1, sim_F(\mathbf{alt}(X_2)) \in S_2] + 2\delta$

2. Assume it is true for $G_{t-1}(X) = (F(X_1), \ldots, F(X_{t-1}))$, we prove it holds for $G_t(X)$. For any set $S = S_1, S_2, \ldots, S_t$, we have

$$Pr[G_t(X) \in S] = Pr[F(X_t) \in S_t \,|\, G_{t-1}(X) \in S_{-t}] \cdot Pr[G_{t-1}(X) \in S_{-t}] \tag{4}$$

By inductive assumption, we have

$$Pr[G_{t-1}(X) \in S_{-t}] \leq$$

$$e^{(t-1)\epsilon} Pr[(sim_F(\mathbf{alt}(X_1)), \ldots, sim_F(\mathbf{alt}(X_{t-1}))) \in S_{-t}] + (t-1)\delta$$

We focus on $Pr[F(X_t) \in S_t \,|\, G_{t-1}(X) \in S_{-t}]$, which equals to

$$= \Sigma_z Pr[F(X_t) \in S_t \,|\, G_{t-1}(X) \in S_{-t}, X_{t\uparrow} = z] \cdot Pr[X_{t\uparrow} = z]$$

For any database X_i where $1 \leq i \leq t-1$, that shares no common block with X_t, $F(X_t)$ is independent with $F(X_i)$. For the simplicity of analysis, we assume every X_i has shared at least one common block with X_t.

$$Pr[F(X_t) \in S_t \,|\, G_{t-1}(X) \in S_{-t}, X_{t\uparrow} = z] =$$

$Pr[F(X_{t\uparrow}, x_{t\frac{1}{c}}) \in S_t \,|\, F(x_{11}, \ldots, x_{1(t-1), X_{1\downarrow t-1}}), \ldots, F(x_{(t-1)1}, X_{t-1\downarrow}) \in S_{-t}, X_{t\uparrow} = z]$

Given $X_{t\uparrow} = z$, each $F(X_i)$ can be seen as a function on $X_{i\uparrow t-i} = [x_{i1}, \ldots, x_{i(t-i)}]^\top$. Notice $x_{ij} = x_{i'j'}$ as long as $i + j = i' + j'$. Given

$X_{t\uparrow} = z$, $G_{t-1}(X)$ is a function on $X_{1\uparrow t-1} = [x_{11}, \ldots, x_{1(t-1)}]^\top$. Since $x_{t\frac{1}{c}}$ is independent with each block in $X_{1\uparrow t-1}$, functions $F(X_{t\uparrow}, x_{t\frac{1}{c}})$ and $G_{t-1}(X)$ are independent with each other, given $X_{t\uparrow} = z$. Only S_t depends on the value of $G_{t-1}(X)$. By the assumption that F is $(\epsilon, \delta, \Delta_F, \Gamma)$-CW private with simulator sim_F, we have

$$\leq \Sigma_z(e^\epsilon Pr[sim_F(\mathbf{alt}(X_t)) \in S_t \,|\, G_{t-1}(X) \in S_{-t}, X_{t\uparrow} = z] + \delta) \cdot Pr[X_{t\uparrow} = z]$$

$$= e^\epsilon Pr[sim_F(\mathbf{alt}(X_t)) \in S_t \,|\, G_{t-1}(X) \in S_{-t}] + \delta$$

By equation (4), we have

$$Pr[G_t(X) \in S] \leq e^\epsilon Pr[sim_F(\mathbf{alt}(X_t)) \in S_t, G_{t-1}(X) \in S_{-t}] + \delta \quad (5)$$

We now focus on $Pr[sim_F(\mathbf{alt}(X_t)) \in S_t, G_{t-1}(X) \in S_{-t}]$, which equals to

$$= Pr[F(X_{t-1}) \in S_{t-1} \,|\, sim_F(\mathbf{alt}(X_t)) \in S_t, G_{t-2}(X) \in S_{-(t-1)}]$$

$$\cdot Pr[sim_F(\mathbf{alt}(X_t)) \in S_t, G_{t-2}(X) \in S_{-(t-1)}]$$

Similarly, we can show

$$Pr[F(X_{t-1}) \in S_{t-1} \,|\, sim_F(\mathbf{alt}(X_t)) \in S_t, G_{t-2}(X) \in S_{-(t-1)}]$$

$$\leq e^\epsilon Pr[sim_F(\mathbf{alt}(X_{t-1})) \in S_{t-1} \,|\, sim_F(\mathbf{alt}(X_t)) \in S_t, G_{t-2}(X) \in S_{-(t-1)}] + \delta$$