Streaming CW-Pivacy

October 7, 2014

1 Streaming Setting 1

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. See (1) and (2) in Figure 1.

Hypothesis

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 any $(1-c)n$ 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 $(?\epsilon, ?\delta, \Delta, \Gamma)$ -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}. See (1) and (2) in Figure 1.

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

Proof for the base case

The base case is when $G(X) = (F(X_1), F(X_2))$. See (1) in Figure 1. 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.

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

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

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

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

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

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

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

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 $sim_F \circ \mathbf{alt}$ 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

$$Pr[F(X_1) \in S_1 \,|\, sim_F(\mathbf{alt}(X_2)) \in S2] \leq e^{\epsilon} Pr[sim_F(X_1) \in S_1 \,|\, sim_F(\mathbf{alt}(X_2)) \in S_2] + \delta \quad (4)$$

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

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

The problem when compose three queries
In this case, $G(X) = (F(X_1), F(X_2), F(X_3))$. See (2) in Figure 1.

For any set $S = (S_1, S_2, S_3)$, we can easily derive a similar equation as (2) above

$$Pr[G(X) \in S] \leq e^{\epsilon} Pr(sim_F(\mathbf{alt}(X_3)) \in S3, F(X_2) \in S_2, F(X_1) \in S_1) + \delta \quad (5)$$

But it appears we **cannot proceed** any further from here. It is hard to leverage a kind of independence between X_2, X_1 and X_3, X_2 simultaneously. This is because X_2 does not have a block that is independent with both X_1 and X_3; each row in X_2 is shared by either X_1 or X_3. For the general case when $t \geq 3$, databases in the middle could have the same problem.

The condition to make the composition work
Intuitively, when each X_i has at least one "private" block (not shared by any other X_j), the composition should work.

1. As we can see in (1) in Figure 1, the base case (when compose only two queries) satisfy this condition.

2. In the standard DP case, the mechanism is $F = f(X)$+noise, where f is the query and noise is independently generated at each time step. At each time step i, F can be seen as a query on a larger database $X_i = X \cup x_{i1}$, where x_{i1} corresponds to the database of the noise. Notice that X remains the same while x_{i1} is regenerated independently each time. In other words, each X_i owns a private block of x_{i1}. See (3) in Figure 1.

2 Streaming Setting 2

Problem setting
We use the same definitions and notations as before. Each X_i has n rows, or $1/c$ blocks. We construct each X_i such that its middle block is private. At each time step, $\frac{(1+c)n}{2}$ rows need to be generated independently. For a total of t time steps, X has size of $n + \frac{(1+c)(t-1)n}{2}$. See Figure 2.

Composition theorem
Let F be a query on each X_i. Assume F is $(\epsilon, \delta, \Delta_F, \Gamma)$-CW private with simulator sim_F, where Δ_F chooses the database according to some distribution and contains auxiliary information of any $(1-c)n$ rows of the database.

Let $G(X) = (F(X_1), F(X_2), \ldots, F(X_t))$ be the composite query that runs F at each time step. Then, G is $(t\epsilon, t\delta, \Delta, \Gamma)$ -CW private with simulator ${sim_F}^t$.

Extension

1. The private block of each X_i does not have to be in the middle. If so, the number of blocks that need to be generated at each step is different, though asymptotically remain the same. In other words, assume F is $(\epsilon, \delta, \Delta_F, \Gamma)$-CW private with a fraction of r of the database as auxiliary information, an asymptotical fraction of $(1 - r/2)$ rows need to be generated independently at each time step.

2. We can connect X_t back to X_1 to make the big database X circular. Then, X has size of $\frac{1+c}{2}nt$. See Figure 2. In the streaming setting, X can be seen as the local database at the client end. The client has

limited local memory, such that it has to replace the first streaming database X_1 after t queries.

Extra Notations. Let $X_{i\updownarrow}$ denote the set of blocks in X_i but the middle one. Let $x_{i\circ}$ denote the middle block of X_i. In Figure 2, $X_{i\updownarrow}$ contain the blues blocks while $x_{i\circ}$ is the white middle block in $X_2, \dots X_{t-1}$.

Proof

For any set $S = S_1, S_2, \dots, S_t$, we have

$$Pr[G_t(X) \in S] = Pr[F(X_t) \in S_t \,|\, G_{t-1}(X) \in S_{-t}] \cdot Pr[G_{t-1}(X) \in S_{-t}] \quad (6)$$

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

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

Given $X_{t\updownarrow} = z$, $F(X_t)$ is a function on $x_{t\circ}$ and $G_{t-1}(X)$ is a function on blocks before x_{t1}. Since every block is generated independently, functions $F(X_t)$ and $G_{t-1}(X)$ are independent with each other. 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

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

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

Combined with equation (6), we have

$$Pr[G_t(X) \in S] \leq e^\epsilon Pr[sim_F(\mathbf{alt}(X_t)) \in S_t, G_{t-1}(X) \in S_{-t}] + \delta \quad (7)$$

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

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

$$\cdot Pr[sim_F(\mathbf{alt}(X_t)) \in S_t, G_{t-2}(X) \in S_{-(t-1)}] \quad (8)$$

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

$$= \Sigma_{z'} Pr[F(X_{t-1}) \in S_{t-1} \,|\, sim_F(\mathbf{alt}(X_t)) \in S_t, G_{t-2}(X) \in S_{-(t-1)}, X_{t-1\updownarrow} = z']$$

$$\cdot Pr[X_{t-1\updownarrow} = z']$$

Given $X_{t-1\updownarrow} = z'$, $F(X_{t-1})$ is a function on $x_{t-1\circ}$, $sim_F(\mathbf{alt}(X_t))$ is a function on blocks after $x_{t-1\frac{1}{c}}$ and $G_{t-2}(X)$ is a function on blocks before $x_{(t-1)1}$.

Since every block is generated independently, functions $F(X_{t-1})$ is independent with both $sim_F(\mathbf{alt}(X_t))$ and $G_{t-2}(X)$. Only S_{t-1} depends on the value of $G_{t-2}(X)$ and $sim_F(\mathbf{alt}(X_t))$. By the assumption that F is $(\epsilon, \delta, \Delta_F, \Gamma)$-CW private with simulator sim_F, we have

$$\leq \Sigma_{z'}(e^{\epsilon} Pr[sim_F(\mathbf{alt}(X_{t-1})) \in S_{t-1} \mid sim_F(\mathbf{alt}(X_t)) \in S_t, G_{t-2}(X) \in S_{-(t-1)}, X_{t-1\updownarrow} = z'] + \delta)$$

$$\cdot Pr[X_{t-1\updownarrow} = z']$$

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

Combined with equation (8), we have

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

$$\leq e^{\epsilon} Pr[sim_F(\mathbf{alt}(X_t)) \in S_t, sim_F(\mathbf{alt}(X_{t-1})) \in S_{t-1}, G_{t-2}(X) \in S_{-(t-1)}] + \delta \tag{9}$$

By equation (7) and equation (9), we have

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

Repeat the same procedure on $F(t-2), \ldots, F(1)$, we can show

$$Pr[G_t(X) \in S] \leq e^{t\epsilon} Pr[sim_F(\mathbf{alt}(X_t)) \in S_t, \ldots, sim_F(\mathbf{alt}(X_1)) \in S_1] + t\delta \quad \blacksquare$$

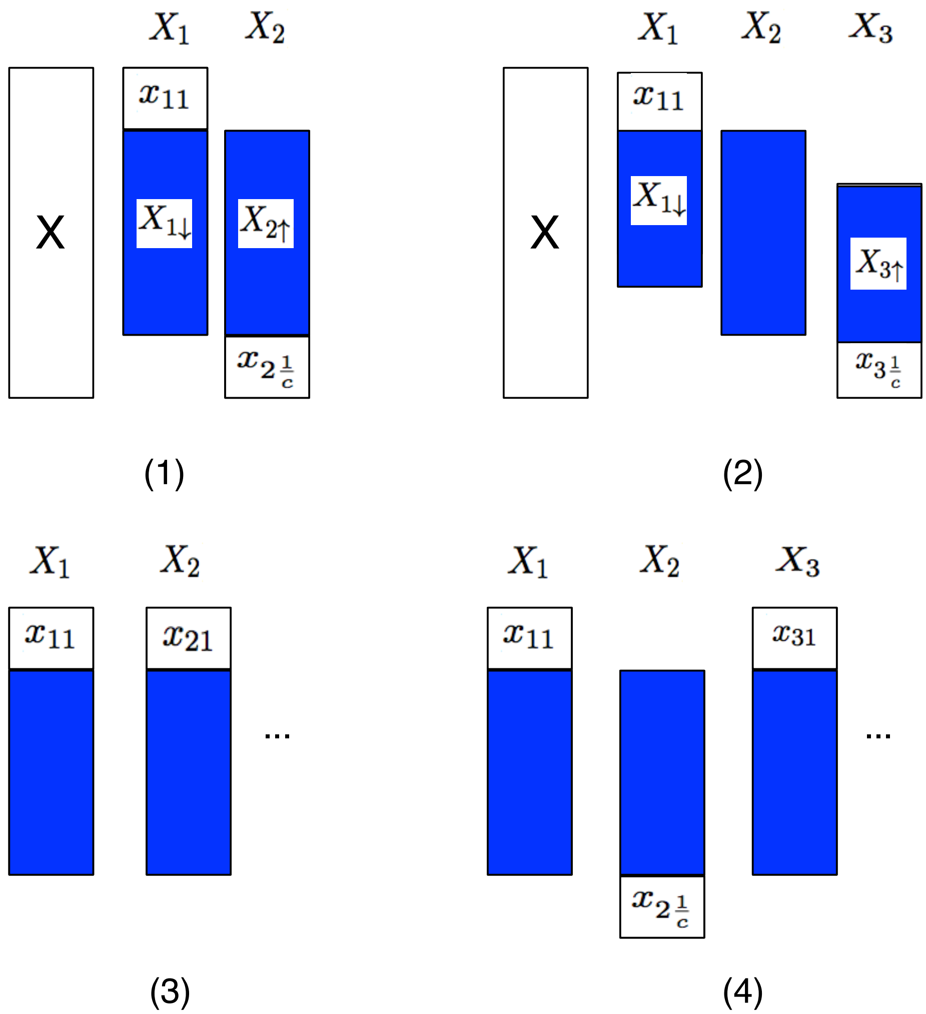

Figure 1: Blocks in blue are shared by more than one databases. Blocks in white are owned by only one database. (1) Composition of $F(X_1)$ and $F(X_2)$. It has "good" independence since each X_i owns one block. (2) Composition of queries on three streaming databases. Cannot leverage "good" independence, because each block in X_2 is shared by two databases. (3) Composition of queries in the standard DP case. The blue blocks refer to the query while the white blocks refer to the independent noise. (4) It works as long as each streaming database owns one block of "private" block.

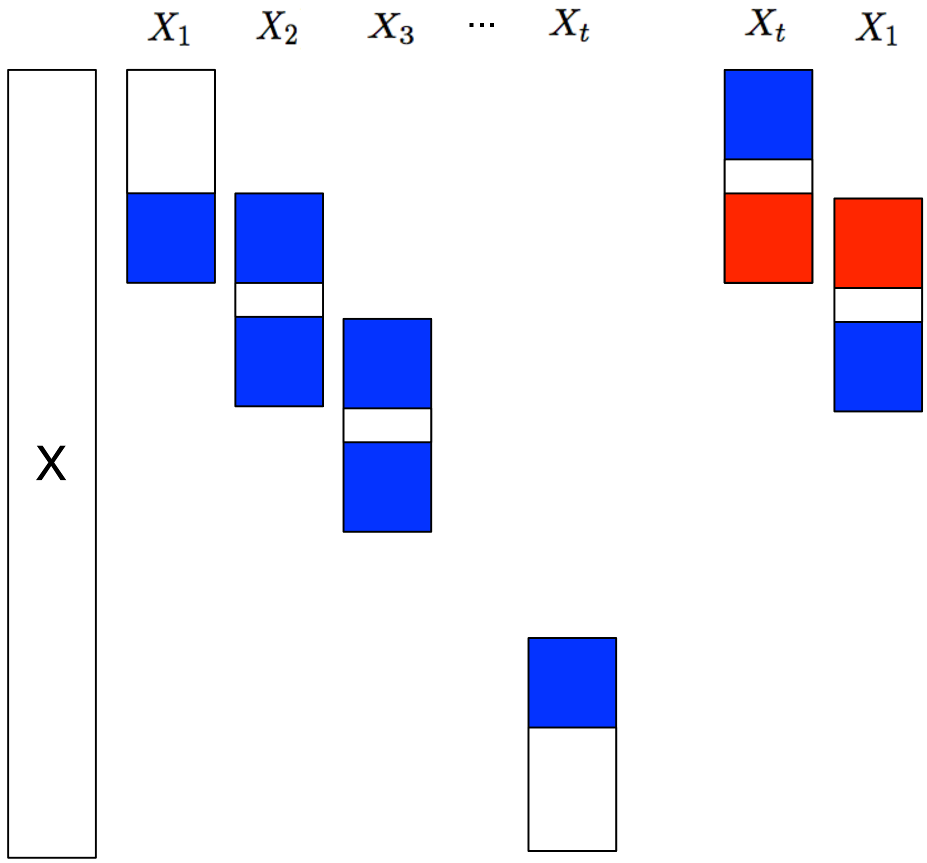

Figure 2: Blocks in blue are shared by more than one databases. Blocks in white are owned by only one database. Construct streaming databases such that each X_i owns a block of cn rows by itself. In this figure, the middle block is always private. At each time step, $\frac{1+c}{2}n$ rows need to be generated independently. We can connect X_t back to X_1 to make the big database X circular. Now, X has size of $\frac{1+c}{2}nt$