Streaming CW-Pivacy

October 12, 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 2.

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 $G(X) = (F(X_1), F(X_2), \ldots, F(X_t))$ 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 $X_i = [x_{i1}, x_{i2}, \ldots, x_{i\frac{1}{c}}]^\top$, where x_{ij} denote the jth block in X_i. Let $X_{i\downarrow k} = [x_{i(k+1)}, x_{i(k+2)}, \ldots, x_{i\frac{1}{c}}]^\top$ and $X_{i\uparrow k} = [x_{i1}, x_{i2}, \ldots, x_{i(\frac{1}{c}-k)}]^\top$. 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 2.

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 $(S_1, S_2, \ldots, S_{t-1})$.

The cases & when it does (does not) work
The base case is when $G(X) = (F(X_1), F(X_2))$. See (1) in Figure 1.

1. In case (1), it is not hard to show $G(X)$ is $(2\epsilon, 2\delta, \Delta, \Gamma)$ -CW private, with simulator (sim_F, sim_F). Please refer to previous write-up for the proof. We can even prove $G(X)$ is $(2\epsilon, 2\delta, \Delta, \Gamma)$ -CW private, given the common blocks shared by X_1 and X_2.

2. In case (2) (we talked about it last Thursday), X_1 is a subset of X_2 with X_2 has a fraction of new blocks. It appears that this case **does not work**.

   Following the proof for the base case (or Adam's proof in page 46-47), we can get the follows: For any set $S = (S_1, S_2)$,

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

   But it appears that we still need to leverage the independence of $F(X_1)$ w.r.t. $sim_F(\mathbf{alt}(X_2))$. That is, we need to considering the following equation.

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

   Since every row in X_1 is also in X_2, we do not have any type of independence. In fact, we really need a fraction of rows in X_1 that are independent of X_2 to proceed from here.

3. Consider the case when compose three queries. See case (3) in Figure 1. Similarly, we get the following step.

   $$Pr[G_3(X) \in S] \leq e^{\epsilon} Pr(sim_F(\mathbf{alt}(X_3)) \in S3, F(X_2) \in S_2, F(X_1) \in S_1) + \delta$$

   Next, we have to consider the following conditional probability.

   $$Pr(sim_F(\mathbf{alt}(X_3)) \in S3, F(X_2) \in S_2, F(X_1) \in S_1) =$$

   $$Pr[G_2(X) \in (S_1, S_2) \,|\, sim_F(\mathbf{alt}(X_3)) \in S3] \cdot Pr[sim_F(\mathbf{alt}(X_3)) \in S3].$$

Now, we need to leverage a type of independence between $G_2(X)$ and $sim_F(\mathbf{alt}(X_3))$. The common blocks of X_3 and $X_1 \cup X_2$ are $X_{3\uparrow}$ (the blue and red blocks in (3) of the Figure 1). We need to consider the total conditional probability on the value of $X_{3\uparrow}$.

$$Pr[G_2(X) \in (S_1, S_2) \,|\, sim_F(\mathbf{alt}(X_3)) \in S3]$$

$$= \Sigma_z Pr[G_2(X) \in (S_1, S_2) \,|\, sim_F(\mathbf{alt}(X_3)) \in S3, X_{3\uparrow} = z] \cdot Pr[X_{3\uparrow} = z]$$

Given $X_{3\uparrow} = z$, $G_2(X)$ is independent with $sim_F(\mathbf{alt}(X_3))$. However, $G_2(X)$ is not CW-private given $X_{3\uparrow}$ as the auxiliary information (red+blue blocks). The problem here is: every block in X_2 is either shared with X_1 (blue+green blocks), or given in the auxiliary information (red blocks). Just like the case 2 above, we cannot leverage a type of independence between X_1 and X_2.

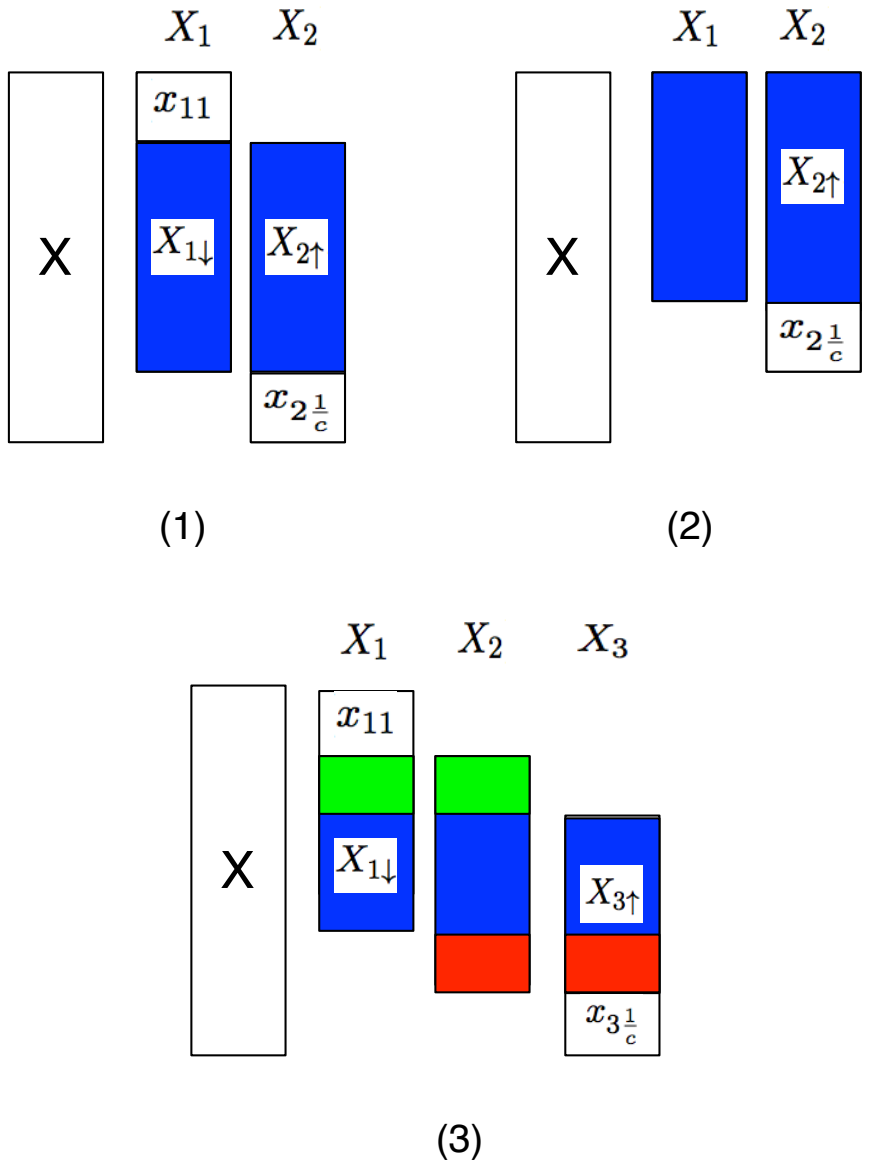

Figure 1: Blocks in color are shared by more than one databases. Blocks in white are owned by only one database. (1) Composition of $F(X_1)$ and $F(X_2)$. It has "good" independence since each X_i owns one block. (2) Composition of $F(X_1)$ and $F(X_2)$ with X_1 to be a part of X_2. (3) Composition of queries on three streaming databases. Cannot leverage "good" independence.

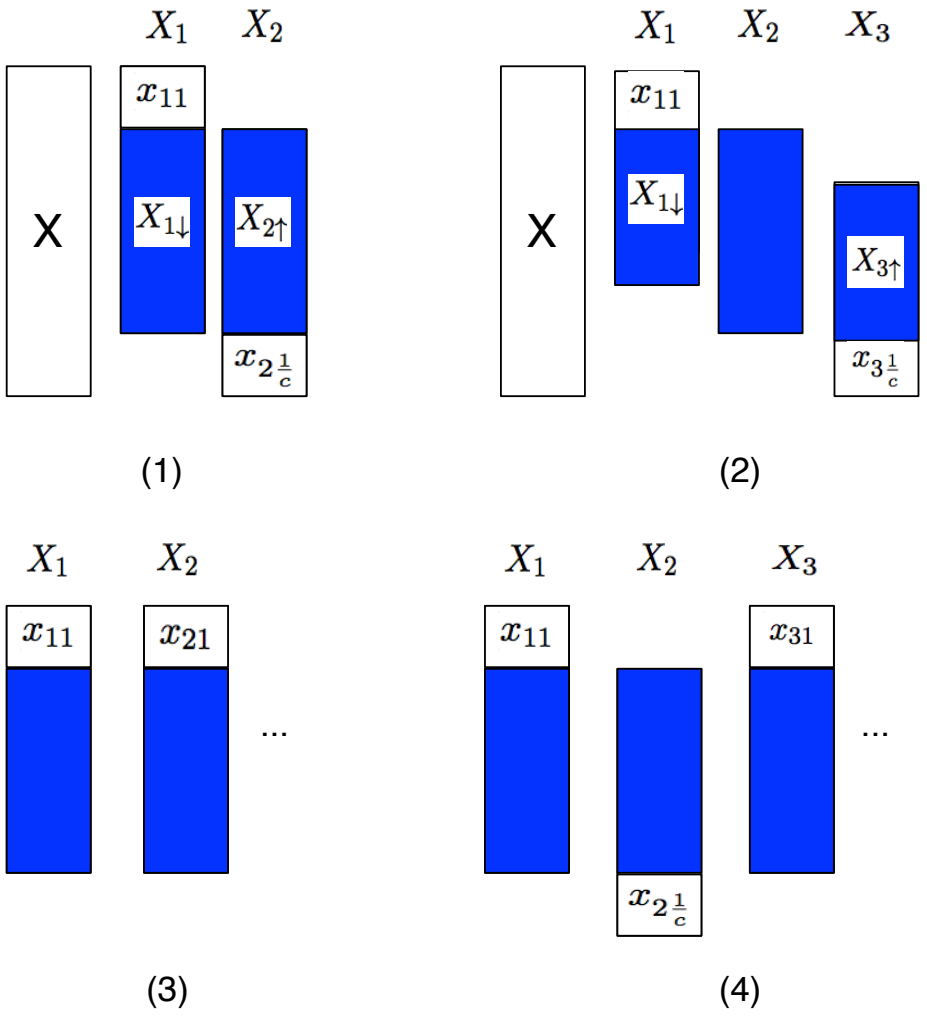

Figure 2: Blocks in blue are shared by more than one databases. Blocks in white are owned by only one database. (1) Composition of $F(X_1)$ and $F(X_2)$. It has "good" independence since each X_i owns one block. (2) Composition of queries on three streaming databases. Cannot leverage "good" independence, 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.

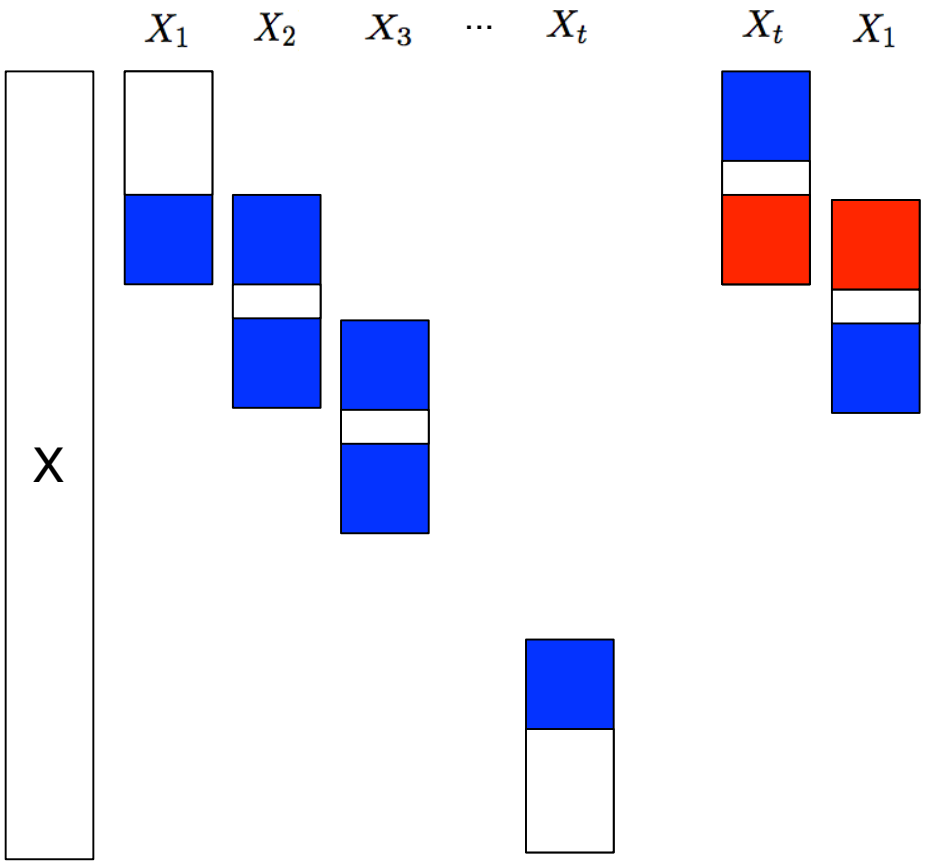

Figure 3: Blocks in blue are shared by more than one databases. Blocks in white are owned by only one database. Construct streaming databases such that each X_i owns a block of cn rows by itself. In this figure, the middle block is always private. At each time step, $\frac{1+c}{2}n$ rows need to be generated independently. We can connect X_t back to X_1 to make the big database X circular. Now, X has size of $\frac{1+c}{2}nt$