APPENDIX

A. PROOF OF THEOREM 2

PROOF. For item V_i, let c_i' denote the true count computed by $\mathbf{q_1}$ from the sample D_R. Therefore, the noisy count $\tilde{c}_i$ is derived by adding a Laplace noise to c_i' as follows:

$$\tilde{c}_i = c_i' + \nu_i \ , \tag{12}$$

$$\nu_i \sim Laplace(0, l/\epsilon_1) \ . \tag{13}$$

The MSE of $\tilde{c}_i$ can be re-written as:

$$\begin{aligned} MSE(\tilde{c}_i) &= Var(c_i' + \nu_i) + (E(c_i' + \nu_i - c_i))^2 \\ &= Var(c_i') + Var(\nu_i) + (E(c_i') - E(c_i))^2 \ . \end{aligned} \tag{14}$$

Note that c_i' and ν_i are mutually independent.

Let p_i denote the popularity of item V_i, i.e. the probability of any record having $vID = V_i$. For simplicity, we assume that users are mutually independent, records are mutually independent, and every user has M records in the raw data set D. To obtain D_R, l records out of M are randomly chosen for each user in D. Thus for any item V_i, c_i' can be represented as the sum of independent random variables:

$$c_i' = \sum_{k=1}^{n} \sum_{r \in T_k} \delta_{r,i} \tag{15}$$

$$\delta_{r,i} = \begin{cases} 1 & \text{if } r.vID = V_i \ \& \ r \in D_R \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{16}$$

The event of $\delta_{r,i} = 1$ is equivalent to the event of record r is about V_i and r is sampled in D_R by chance:

$$Pr[\delta_{r,i} = 1] = Pr[r.vID = Vi \ \& \ r \in D_R] = p_i \frac{l}{M} \ . \tag{17}$$

Therefore, we can obtain the following expectation and variance for c_i':

$$\begin{aligned} E(c_i') &= \sum_{k=1}^{n} \sum_{r \in T_k} E(\delta_{r,i}) \\ &= \sum_{k=1}^{n} \sum_{r \in T_k} p_i \frac{l}{M} \\ &= nlp_i \end{aligned} \tag{18}$$

$$\begin{aligned} Var(c_i') &= \sum_{k=1}^{n} \sum_{r \in T_k} Var(\delta_{r,i}) \\ &= \sum_{k=1}^{n} \sum_{r \in T_k} p_i \frac{l}{M}(1 - p_i \frac{l}{M}) \\ &= nlp_i(1 - p_i \frac{l}{M}) \end{aligned} \tag{19}$$

Similarly, we can obtain the expectation of c_i:

$$E(c_i) = nMp_i \ . \tag{20}$$

From the above results, we can re-write Equation 14 as follows:

$$MSE(\tilde{c}_i) = nlp_i(1 - p_i \frac{l}{M}) + 2\frac{l^2}{\epsilon_l^2} + (nlp_i - nMp_i)^2 \tag{21}$$

and we can perform the standard least square method to minimize the MSE. The optimal l value is thus:

$$l = \frac{2n^2 p_i^2 M - np_i}{4/\epsilon_1^2 - 2np_i^2/M + 2n^2 p_i^2} \tag{22}$$

We conclude that the optimal l value is a monotonically increasing function of ϵ_1^2. □

B. PROOF OF LEMMA 6

PROOF. By definition of differential privacy, we are to prove that for any neighboring raw databases D_1 and D_2, $\mathcal{A} \circ \mathcal{S}$ satisfies the following inequality for $\widetilde{D} \in Range(\mathcal{A} \circ \mathcal{S})$:

$$Pr[\mathcal{A} \circ \mathcal{S}(D_1) = \widetilde{D}] \leq e^{\epsilon} Pr[\mathcal{A} \circ \mathcal{S}(D_2) = \widetilde{D}] \ . \tag{23}$$

Without loss of generality, we assume D_2 contains one more user than D_1. Let u denote the user that is contained in D_2 but not D_1 and T be user u's set of records in D_2. By definition of neighboring databases, we can rewrite $D_2 = D_1 \oplus T$ [3].

Let $\hat{D}_1$ denote any possible sampling output of $\mathcal{S}(D_1)$. We have:

$$\begin{aligned} &Pr[\mathcal{A} \circ \mathcal{S}(D_1) = \widetilde{D}] \\ &= \sum_{\hat{D}_1} Pr[\mathcal{A} \circ \mathcal{S}(D_1) = \widetilde{D} | \mathcal{S}(D_1) = \hat{D}_1] Pr[\mathcal{S}(D_1) = \hat{D}_1] \\ &= \sum_{\hat{D}_1} Pr[\mathcal{A}(\hat{D}_1) = \widetilde{D}] Pr[\mathcal{S}(D_1) = \hat{D}_1] \end{aligned} \tag{24}$$

Let $\hat{T}$ denote any possible sampling output of $\mathcal{S}(T)$. We note that $\hat{T}$ can take values from the entire domain, in general:

$$\sum_{\hat{T}} Pr[\mathcal{S}(T) = \hat{T}] = 1 \ . \tag{25}$$

Since $\mathcal{S}$ is performed independently on each user, we can derive:

$$\begin{aligned} &Pr[\mathcal{S}(D_1) = \hat{D}_1] \\ &= \sum_{\hat{T}} Pr[\mathcal{S}(D_1) = \hat{D}_1] Pr[\mathcal{S}(T) = \hat{T}] \\ &= \sum_{\hat{T}} Pr[\mathcal{S}(D_1 \oplus T) = \hat{D}_1 \oplus \hat{T}] \ . \end{aligned} \tag{26}$$

Note that since D_1 and T are disjoint, the sampling output on D_1 and T are also independent and disjoint. Therefore,

$$\begin{aligned} &Pr[\mathcal{A} \circ \mathcal{S}(D_1) = \widetilde{D}] \\ &= \sum_{\hat{D}_1} Pr[\mathcal{A}(\hat{D}_1) = \widetilde{D}] \sum_{\hat{T}} Pr[\mathcal{S}(D_1 \oplus T) = \hat{D}_1 \oplus \hat{T}] \\ &= \sum_{\hat{D}_1, \hat{T}} Pr[\mathcal{A}(\hat{D}_1) = \widetilde{D}] Pr[\mathcal{S}(D_1 \oplus T) = \hat{D}_1 \oplus \hat{T}] \\ &\leq \sum_{\hat{D}_1, \hat{T}} e^{\epsilon} Pr[\mathcal{A}(\hat{D}_1 \oplus \hat{T}) = \widetilde{D}] Pr[\mathcal{S}(D_1 \oplus T) = \hat{D}_1 \oplus \hat{T}] \end{aligned} \tag{27}$$

$$= e^{\epsilon} \sum_{\hat{D}_2} Pr[\mathcal{A}(\hat{D}_2) = \widetilde{D}] Pr[\mathcal{S}(D_2) = \hat{D}_2] \tag{28}$$

$$= e^{\epsilon} Pr[\mathcal{A} \circ \mathcal{S}(D_2) = \widetilde{D}] \ . \tag{29}$$

Line 27 is due to the fact that $\mathcal{A}$ is ϵ-differentially private and $\hat{D}_1$ and $\hat{D}_1 \oplus \hat{T}$ are neighboring databases. In line 28 we change notation and let $\hat{D}_2$ represent $\hat{D}_1 \oplus \hat{T}$. The proof is hence complete. □

[3] $\oplus$ is used to denote a co-product, or disjoint union of two databases.