K-Randomization

Maxim Zhilyaev David Zeber

March 19, 2016

0.1 Maximum and Local differential privacy

Assuming local privacy ratio reaches maximum when D consists of $(N-1)$ zeros and a single 1 bit, and the set original bit is switched to 0 to obtain D_m - modified collection of all zeros. Suppose S is the number of sets bits observed in the synthetic output. Then we can express privacy ratio for every value of S.

$$P(s=i|D_m) = \binom{N}{i} q^i p^{N-i} \tag{0.1}$$

$$P(s=i|D) = \binom{N-1}{i} q^{i+1} p^{N-1-i} + \binom{N-1}{i-1} q^{i-1} p^{N-i+1} \tag{0.2}$$

$$R_i = \frac{P(s=i|D_m)}{P(s=i|D)} = \frac{\binom{N}{i} q^i p^{N-i}}{\binom{N-1}{i} q^{i+1} p^{N-1-i} + \binom{N-1}{i-1} q^{i-1} p^{N-i+1}} \tag{0.3}$$

It's actually more convenient to work with $\frac{1}{R_i}$ as in:

$$\frac{1}{R_i} = \frac{P(s=i|D)}{P(s=i|D_m)} = \frac{N-i}{N}\frac{q}{p} + \frac{i}{N}\frac{p}{q} \tag{0.4}$$

When all $i=0$ - all synthetic bits are 0, the ratio reaches its maximum:

$$R_0 = \frac{1}{1/R_i} = \frac{p}{q}$$

When $i=N$ - the synthetic output consists of set bits entirely, the privacy ratio reaches minimum:

$$R_N = \frac{1}{1/R_i} = \frac{q}{p}$$

The ratio reduces as i increases, and becomes 1 when number of synthetic bits is equal to expected number of set synthetic bits after randomization:

$$\frac{1}{R_{qN}} = \frac{N-qN}{N}\frac{q}{p} + \frac{pN}{N}\frac{q}{p} = (1-q)\frac{q}{p} + q\frac{p}{q} = p+q = 1$$

The notion of a **local differential privacy**, considers the probabilistic ratio only for values of i that have realistic chance of being observed. The expected number of observed synthetic bits is qN, while the deviation of S random variable is $\sigma = \sqrt{pqN}$. Consider the interval $[qN - 3\sigma, qN + 3\sigma]$. Since the probabilistic ratio grows as i decreased, the maximum ratio will be attained when $i = qN - 3\sigma$. Hence, the local differential privacy reaches maximum at $i = qN - 3\sigma$, and we want to express analytically the relationship between the probabilistic privacy ratio λ, number of records N, and RRT parameters p and q:

$$i = qN - 3\sigma = qN - 3\sqrt{pqN} \tag{0.5}$$

$$R_i = \frac{P(s=i|D_m)}{P(s=i|D)} \leq \lambda \tag{0.6}$$

$$\frac{1}{R_i} = \frac{P(s=i|D)}{P(s=i|D_m)} \geq \frac{1}{\lambda} \tag{0.7}$$

$$\frac{N-i}{N}\frac{q}{p} + \frac{i}{N}\frac{p}{q} \geq \frac{1}{\lambda} \tag{0.8}$$

$$\frac{N - qN + 3\sqrt{pqN}}{N} \cdot \frac{q}{p} + \frac{qN - 3\sqrt{pqN}}{N} \cdot \frac{p}{q} \geq \frac{1}{\lambda} \tag{0.9}$$

From here:

$$p + q - 3\sqrt{\frac{pq}{N}}\left(\frac{p}{q} - \frac{q}{p}\right) \geq \frac{1}{\lambda} \tag{0.10}$$

$$1 - 3\sqrt{\frac{pq}{N}}\frac{p^2-q^2}{pq} \geq \frac{1}{\lambda} \tag{0.11}$$

$$3\sqrt{\frac{pq}{N}}\frac{p^2-q^2}{pq} \leq 1 - \frac{1}{\lambda} \tag{0.12}$$

$$3\sqrt{\frac{1}{N}} \cdot \frac{p-q}{\sqrt{pq}} \leq 1 - \frac{1}{\lambda} \tag{0.13}$$

$$\frac{pqN}{(p-q)^2} \geq \frac{9}{(1-\frac{1}{\lambda})^2} \tag{0.14}$$

This is an interesting result. Note that left side of inequality is the variance of estimate $\bar{T}$. The local differential privacy grantee simply places a lower bound on the variance of RRT estimates:

$$VAR(\bar{T}) = \frac{pqN}{(p-q)^2} \geq \frac{9}{(1-\frac{1}{\lambda})^2} \tag{0.15}$$

For a randomization algorithm applied independently to N bits to be ϵ-differentially private in local sense, means that estimate deviation is lower-bounded by:

$$\sigma(\bar{T}) \geq \frac{3}{1-\frac{1}{\lambda}} = \frac{3}{1-\frac{1}{e^{\epsilon}}} \tag{0.16}$$

From here, we can express RRT noise parameter q through N and λ:

$$\frac{pqN}{(p-q)^2} \geq \frac{9}{(1-\frac{1}{\lambda})^2} \tag{0.17}$$

$$\frac{(1-q)q}{(1-2q)^2} \geq \frac{9}{(1-\frac{1}{\lambda})^2 N} \tag{0.18}$$

$$q \geq \frac{1}{2}\left(1 - \frac{1}{\sqrt{1+4\frac{9}{(1-\frac{1}{\lambda})^2 N}}}\right) \tag{0.19}$$

Suppose $\lambda = 2$ and there are 1000 single bits records in D. The required noise is:

$$q = 0.032527$$

Compare that to the level of noise that absolute differential privacy bound would require for $\epsilon = ln(2)$.

$$\frac{p}{q} \leq 2 \tag{0.20}$$

$$q \geq \frac{1}{3} = 0.333 \tag{0.21}$$

The notion of local privacy allowed us to reduce RRT noise 10 times and enabled drastic improvement in estimation accuracy. In the classical case, the estimation deviation is $\sigma = 44.7$, while for the local privacy the deviation is $\sigma = 5.6$, meaning that precision of RRT estimates had grown 8 fold. It's worth reflecting on what's exactly going on and why such a drastic performance increase is achievable.

Consider confidence intervals for both an original collection D and modified collection D_m. D_m contains 1000 empty bits and D contains 999 empty bits. Corresponding means and deviation for sum of observed synthetic bits in each case is given below:

$$E(S) = q \cdot 1000 \tag{0.22}$$

$$\sigma(S) = \sqrt{pq \cdot 1000} \tag{0.23}$$

$$E(S_m) = q \cdot 1000 + p \tag{0.24}$$

$$\sigma(S_m) = \sqrt{pq \cdot 999 + pq} \tag{0.25}$$

Consider the confidence intervals for both S and S_m for RRT under classical and local differential privacy constrains. If $q = 0.333$ the confidence interval for S and S_m are:

$$S- > [198.9, 467.1] \tag{0.26}$$

$$S_m- > [198.6, 467.8] \tag{0.27}$$

Under local differential privacy, the noise level $q = 0.0.033$, and the confidence intervals become:

$$S- > [16.05, 49.94] \tag{0.28}$$

$$S_m- > [17.02, 50.91] \tag{0.29}$$

The intervals are nearly identical in either case. Which illustrates the point - we do not need the full power of the absolute differential privacy bound: the local privacy bound will guarantee privacy ratio for 99.98% of possible synthetic outcomes. Effectively, we exploit the noise of large collection to reduce the RRT noise required to randomize each individual record. Rephrasing this important idiom - hiding a record among other records needs less noise than obfuscating a single record.

1 K-randomization for a single bit case

We know consider an important technique for further increasing the estimation precision while providing same local privacy guarantees. Recall from previous example, that if collection D consists of $N = 1000$ records, the corresponding RRT noise at $\lambda = 2$ is $q = 0.0325$. The deviation of the estimate in this case is $\sigma = 6$. Hence our estimation error will be roughly 18 in either direction. We can increase the estimate precision by repeating randomization k times, hence the name **k-ranomization**.

It will be shown that repeating randomization k times achieves increase in precision proportional to $\sqrt{k}$, it also causes slight increase in RRT noise necessary to maintain same differential privacy guarantee. However, the RRT noise increase is usually insignificant compared to the precision gain, which gives a nice dimension to the usual privacy vs. precision tradeoff. K-randomization enables precision increase at the same privacy level for the expense of increasing synthetic record volume k times. Instead of trading privacy for precision, k-randomization allows to trade infrastructure cost for precision while keeping privacy the same. This is especially apparent for long multivariate records, but we will lay mathematical grounds starting from a single bit case.

1.1 Estimating number of single bits under k-randomization

Suppose there are T set bits in the original collection of N single bit records. Each record is randomized k-times. The number of observed synthetic bits S is a random variable expressed as:

$$S = p \cdot kT + q \cdot (kN - kT)$$

The estimate for T, computed from observed value of S is:

$$\bar{T} = \frac{S - qkN}{k(p-q)} \tag{1.1}$$

The aggregator simply divides the estimate computed from kN records by k. The expectation, variance and deviation of $\bar{T}$ random variable are given by:

$$E(\bar{T}) = T \tag{1.2}$$

$$VAR(\bar{T}) = \frac{qpkN}{k^2 \cdot (p-q)^2} = \frac{qpN}{k \cdot (p-q)^2} \tag{1.3}$$

$$\sigma(\bar{T}) = \sqrt{\frac{qpN}{k \cdot (p-q)^2}} \tag{1.4}$$

Note that deviation of the estimate is reduced by $\sqrt{k}$ compared to a single randomization case.

1.2 Choice of D

WE NEED PROOF FOR MAXIMALITY UNDER K

1.3 Local differential privacy under k-randomization

Consider probabilities of seeing s set bits in the synthetic output for D_m and D respectively:

Since D_m consists of N empty bits, the probability of see s synthetic bits after randomization is binomial

$$P(S = s|D_m) = \binom{kN}{s} q^s p^{kN-s} \tag{1.5}$$

While the original collection D has a single set bit, and the probability of finding s set synthetic bits is:

$$P(S = s|D) = \sum_{i=0}^{k} \binom{k(N-1)}{s-i} q^{s-i} p^{k(N-1)-s+i} \cdot \binom{k}{i} p^i q^{k-i} \tag{1.6}$$

$$P(S = s|D_m) = \sum_{i=0}^{k} \binom{k(N-1)}{s-i} \binom{k}{i} q^{s+k-2i} p^{kN-s-(k-2i)} \tag{1.7}$$

Expressing the quotient of privacy ratio at given s, we have:

$$\frac{1}{R_s} = \sum_{i=0}^{k} \frac{\binom{k(N-1)}{s-i} \cdot \binom{k}{i}}{\binom{kN}{s}} \cdot \frac{q^{k-2i}}{p^{k-2i}} \tag{1.8}$$

Consider the binomial ratio in the sum:

$$\frac{\binom{k(N-1)}{s-i}}{\binom{kN}{s}} = \frac{(kN-k)!}{(kN)!} \cdot \frac{s!}{(s-i)!} \cdot \frac{(kN-s)!}{(kN-s-(k-i))!} = \frac{\prod_{j=0}^{i-1}(S-j) \cdot \prod_{j=0}^{k-i-1}(kN-S-j)}{\prod_{j=0}^{k-1}(kN-j)} \tag{1.9}$$

For positive B, A and e such that $A < B$ the following holds:

$$\frac{A-e}{B-e} < \frac{A}{B} \tag{1.10}$$

Hence the expression in 5.10 is upper bounded by:

$$\frac{\prod_{j=0}^{i-1}(s-j) \cdot \prod_{j=0}^{k-i-1}(kN-s-j)}{\prod_{j=0}^{k-1}(kN-j)} < \frac{\prod_{j=0}^{i-1} s \cdot \prod_{j=0}^{k-i-1}(kN-s)}{\prod_{j=0}^{k-1} kN} = \frac{s^i \cdot (kN-s)^{k-i}}{(kN)^k} \tag{1.11}$$

Dividing each numerator term by kN we arrive to an upper bound of the privacy ratio:

$$\frac{1}{R_s} < \sum_{i=0}^{k} \left(\frac{s}{kN}\right)^i \left(1 - \frac{s}{kN}\right)^{k-i} \cdot \binom{k}{i} \cdot \frac{q^{k-2i}}{p^{k-2i}} \tag{1.12}$$

Again, under local privacy constrains we compute privacy ratio for s located 3σ bellow the mean:

$$s = qkN - 3\sqrt{pqkN}$$

Replacing s in formula 5.13, we get:

$$\sum_{i=0}^{k} \left(\frac{qkN - 3\sqrt{pqkN}}{kN}\right)^i \left(1 - \frac{qkN - 3\sqrt{pqkN}}{kN}\right)^{k-i} \cdot \binom{k}{i} \cdot \frac{q^{k-2i}}{p^{k-2i}} = \tag{1.13}$$

$$\sum_{i=0}^{k} \left(q - 3\sqrt{\frac{pq}{kN}}\right)^i \left(1 - q + 3\sqrt{\frac{pq}{kN}}\right)^{k-i} \cdot \binom{k}{i} \cdot \frac{q^{k-2i}}{p^{k-2i}} = \tag{1.14}$$

$$\sum_{i=0}^{k} \frac{p^i}{q^i} \left(q - 3\sqrt{\frac{pq}{kN}}\right)^i \cdot \frac{q^{k-i}}{p^{k-i}} \left(p + 3\sqrt{\frac{pq}{kN}}\right)^{k-i} \cdot \binom{k}{i} = \tag{1.15}$$

$$\sum_{i=0}^{k} \binom{k}{i} \left(p - 3p\sqrt{\frac{p}{qkN}}\right)^i \cdot \left(q + 3q\sqrt{\frac{q}{pkN}}\right)^{k-i} = \tag{1.16}$$

$$\left(q + 3q\sqrt{\frac{q}{pkN}} + p - 3p\sqrt{\frac{p}{qkN}}\right)^k = \tag{1.17}$$

$$\left(q + p - 3p\sqrt{\frac{p}{qkN}} + 3q\sqrt{\frac{q}{pkN}}\right)^k = \tag{1.18}$$

$$\left(1 - \frac{3(p-q)}{\sqrt{qpkN}}\right)^k \tag{1.19}$$

Should the differential privacy ratio limit be λ we have the lower bound below:

$$\left(1+\frac{3(p-q)}{\sqrt{qpkN}}\right)^k > \frac{1}{R_s} \geq \frac{1}{\lambda} \tag{1.20}$$

From here we have:

$$\left(1+\frac{3(p-q)}{\sqrt{qpkN}}\right)^k \geq \frac{1}{\lambda} \tag{1.21}$$

$$\frac{qpkN}{(p-q)^2} \geq \frac{9}{(1-\frac{1}{\sqrt[k]{\lambda}})^2} \tag{1.22}$$

From here we express required RRT noise through λ, N and k.

$$q \geq \frac{1}{2}\left(1-\frac{1}{\sqrt{1+4\frac{9}{(1-\frac{1}{\sqrt[k]{\lambda}})^2 kN}}}\right) \tag{1.23}$$

Using exact same example as before: $N = 1000$ records and $\lambda = 2$. Suppose the randomization is repeated 16 times, the corresponding RRT noise $q = 0.1668$. The noise increased to make the privacy stay at the same level, however the precision of the measurement actually decreased, because of $\frac{1}{\sqrt{k}}$ factor. The corresponding sigma is:

$$\sigma(\bar{T}) = \sqrt{\frac{qpN}{k \cdot (p-q)^2}} = 4.42 \tag{1.24}$$

So we gain 25% precision increase by repeating randomization 16 times. The k-randomization gain main not be very significant for a single bit reporting, but becomes very useful for high-dimentionality vectors, where privacy level are increased due to longer bits vectors reported.

2 multivariate vectors

2.1 Differential privacy ratio

2.1.1 Sufficient Statistics proof

The original collection D consists of $N-1$ zero vectors vectors and one unit vector of length L. Denote a zero vector as 0 and a unit vector as 1. The unit-vector 1 is modified into a zero-vector

0, hence the modified collection D_m consists of only 0 vectors. There are 2^L possible distinct synthetic vectors. Denote v_i a distinct synthetic vector. Denote a observed synthetic configuration S as $s_1, s2, \ldots, s_{2^L}$, whereby s_i represents a count of original vectors that mapped into specific synthetic vector v_i after randomization. Denote D' as a collection of $(N-1)$ zero vectors. Then the probability of generating S from collection D' and a single vector y is given by:

$$P(S|D'+y) = P(s_1-1, s_2, \ldots, s_{2^L}|D') \cdot P(v_1|y) + \cdots + P(s_1, s_2-1, \ldots, s_{2^L}|D') \cdot P(v_{2^L}|y) \tag{2.1}$$

Re-writing the ratio

$$\frac{P(S|D'+0)}{P(S|D'+1)} = \frac{P(s_1-1, s_2, \ldots, s_{2^L}|D') \cdot P(v_1|0) + \cdots + P(s_1, s_2-1, \ldots, s_{2^L}|D') \cdot P(v_{2^L}|0)}{P(s_1-1, s_2, \ldots, s_{2^L}|D') \cdot P(v_1|1) + \cdots + P(s_1, s_2-1, \ldots, s_{2^L}|D') \cdot P(v_{2^L}|1)} \tag{2.2}$$

$$\frac{P(S|D'+0)}{P(S|D'+1)} = \frac{P(v_1|0) + \sum_{i=2}^{2^L} \frac{P(s_1, s_2, \ldots, s_i-1, \ldots, s_{2^L}|D')}{P(s_1-1, s_2, \ldots, s_{2^L}|D')} p(v_i|0)}{P(v_1|1) + \sum_{i=2}^{2^L} \frac{P(s_1, s_2, \ldots, s_i-1, \ldots, s_{2^L}|D')}{P(s_1-1, s_2, \ldots, s_{2^L}|D')} p(v_i|1)} \tag{2.3}$$

Note that distribution of randomized vectors generated by D' is multinomial, since the probability of generating a particular v_i from a zero vector remains constant over all N trials.

$$\frac{P(s_1, s_2, \ldots, s_i-1, \ldots, s_{2^L}|D')}{P(s_1-1, s_2, \ldots, s_i, \ldots, s_{2^L}|D')} = \frac{\frac{(2^L)!}{s_1! \cdot s_2! \ldots (s_i-1)! \ldots} p(v_1|0)^{s_1} \ldots p(v_i|0)^{s_i-1} \ldots}{\frac{(2^L)!}{(s_1-1)! \cdot s_2! \ldots s_i! \ldots} p(v_1|0)^{s_1-1} \ldots p(v_i|0)^{s_i} \ldots} = \tag{2.4}$$

$$\frac{(s_1-1)! s_i!}{s_1!(s_i-1)!} \cdot \frac{p(v_1|0)^{s_1} p(v_i|0)^{s_i-1}}{p(v_1|0)^{s_1-1} p(v_i|0)^{s_i}} = \frac{s_i}{s_1} \cdot \frac{p(v_1|0)}{p(v_i|0)} = \frac{s_i}{s_1} \cdot \frac{p^L}{p(v_i|0)} \tag{2.5}$$

Using that result in the ratio expression we have:

$$\frac{P(S|D'+0)}{P(S|D'+1)} = \frac{p^L + \sum_{i=2}^{2^L} \frac{s_i}{s_1} \cdot \frac{p^L}{p(v_i|0)} p(v_i|0)}{q^L + \sum_{i=2}^{2^L} \frac{s_i}{s_1} \cdot \frac{p^L}{p(v_i|0)} p(v_i|1)} = \frac{s_1 + \sum_{i=2}^{2^L} s_i \cdot \frac{p(v_i|0)}{p(v_i|0)}}{s_1 \left(\frac{q}{p}\right)^L + \sum_{i=2}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|0)}} = \tag{2.6}$$

$$\frac{s_1 + \sum_{i=2}^{2^L} s_i}{s_1 \left(\frac{q}{p}\right)^L + \sum_{i=2}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|0)}} = \frac{N}{s_1 \left(\frac{q}{p}\right)^L + \sum_{i=2}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|0)}} = \frac{N}{\cdot \sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|0)}} \tag{2.7}$$

Note that if v_i and v_j have same number of set bits, the ratio inside the sum is the same:

$$\text{if } v_i \text{ has same number of set bits as } v_j \text{, and this number is } l \text{ ,then} \tag{2.8}$$

$$\frac{p(v_i|1)}{p(v_i|0)} = \frac{p(v_j|1)}{p(v_j|0)} = \frac{p^l q^{L-l}}{p^{L-l} q^l} = \left(\frac{q}{p}\right)^{L-2l} \tag{2.9}$$

This allows us to express privacy ratio as function of synthetic output S through counts of synthetic vectors that have same number of set bits l:

$$R(S) = \frac{P(S|D'+0)}{P(S|D'+1)} = \frac{N}{\cdot \sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|0)}} = \frac{N}{\sum_{l=0}^{L} s_l \cdot \left(\frac{q}{p}\right)^{L-2l}} \tag{2.10}$$

$$\frac{1}{R(S)} = \frac{1}{N} \sum_{l=0}^{L} s_l \cdot \left(\frac{q}{p}\right)^{L-2l} \tag{2.11}$$

Hence, an observer does not gain any more privacy insight by looking at individual vectors than by looking at aggregated counts in a histogram buckets each collecting synthetic vectors with same bit count.

2.1.2 Local differential privacy

As mentioned above, we can equivalently represent collection S by set-bits-histogram counts. For vectors of length L, there are $L+1$ histogram buckets raging from $l=0$ to $l=L$. Let's consider the privacy ratio when the synthetic collection is in the expected state S_e and assume bucket l is sufficiently filled, that is $s_0 \geq 1$. We should represent state S as:

$$S = [s_0, s_1, \ldots, s_L]$$

For the expected synthetic state S_e we choose the state generated from modified collection D_m consisting of N zero vectors. The distribution S is a some of N independent random vectors of size L consisting of probabilities of finding 1 in a bucket l:

Note that probability of generating a synthetic vector containing l set bits from either unit or zero original is given by:

$$p(l|1) = \binom{L}{l} p^l q^{L-l} \tag{2.12}$$

$$p(l|0) = \binom{L}{l} q^l p^{L-l} \tag{2.13}$$

We now consider $\frac{1}{R(S)}$ to be a random variable X of its own.

$$X = \frac{1}{R(S)} = \frac{1}{N}\sum_{l=0}^{L} s_l \cdot \left(\frac{q}{p}\right)^{L-2l}$$

Note that bucket counts s_l assume multinomial distribution with bucket probabilities:

$$p(l|0) = \binom{L}{l} q^l p^{L-l}$$

Each count is multiplied by a constant factor $\left(\frac{q}{p}\right)^{L-2l}$, hence X is the sum of L correlated variables X_l, where:

$$X_l = \left(\frac{q}{p}\right)^{L-2l} Binomial(p(l|0), N) \tag{2.14}$$

The expected values of X is given below:

$$E(X) = \sum_{l=0}^{L} N \cdot p(l|0)\left(\frac{q}{p}\right)^{L-2l} = \sum_{l=0}^{L} N \cdot \binom{L}{l} q^l p^{L-l}\left(\frac{q}{p}\right)^{L-2l} = N\sum_{l=0}^{L}\binom{L}{l} \cdot q^{L-l}p^l = N(p+q)^L = N \tag{2.15}$$ ∎

The variance of X is expressed through variance-covariance of multinomial distribution:

$$VAR(X) = \sum_{l=0}^{L} VAR(X_l) + 2\sum_{i\leq j}\sum_{<j\leq L} COV(X_i, X_j) \tag{2.16}$$

$$VAR(X) = \sum_{l=0}^{L} N\left[\left(\frac{q}{p}\right)^{L-2l}\right]^2 p(l|0)(1-p(l|0)) - 2\sum_{j\neq j} Np(i|0)\left(\frac{q}{p}\right)^{L-2i} p(j|0)\left(\frac{q}{p}\right)^{L-2j} = \tag{2.17}$$

$$VAR(X) = N\left(\sum_{l=0}^{L}\left[\left(\frac{q}{p}\right)^{L-2l}\right]^2 p(l|0) - \sum_{l=0}^{L}\left[p(l|0)\left(\frac{q}{p}\right)^{L-2l}\right]^2 - 2\sum_{j\neq j} p(i|0)\left(\frac{q}{p}\right)^{L-2i} p(j|0)\left(\frac{q}{p}\right)^{L-2j}\right) \tag{2.18}$$ ∎

Note that negative terms is an expansion of the square of the sum, hence:

$$VAR(X) = N\left(\sum_{l=0}^{L}\left[\left(\frac{q}{p}\right)^{L-2l}\right]^2 p(l|0) - \left[\sum_{l=0}^{L} p(l|0)\left(\frac{q}{p}\right)^{L-2l}\right]^2\right) \tag{2.19}$$

$$VAR(X) = N\left(\sum_{l=0}^{L}\left[\left(\frac{q}{p}\right)^{L-2l}\right]^2 p(l|0) - 1\right) \tag{2.20}$$

We now simplify the first term of the sum:

$$\sum_{l=0}^{L}\left[\left(\frac{q}{p}\right)^{L-2l}\right]^2 p(l|0) = \sum_{l=0}^{L}\binom{L}{l}q^l p^{L-l}\left(\frac{q}{p}\right)^{L-2l}\cdot\left(\frac{q}{p}\right)^{L-2l} = \tag{2.21}$$

$$\sum_{l=0}^{L}\binom{L}{l}p^l q^{L-l}\cdot\left(\frac{q}{p}\right)^{L-2l} \tag{2.22}$$

$$\sum_{l=0}^{L}\binom{L}{l}\frac{q^{2L-3l}}{p^{L-3l}} = \sum_{l=0}^{L}\binom{L}{l}\frac{q^{3L-3l}p^{3l}}{(pq)^L} = \tag{2.23}$$

$$\frac{1}{(pq)^L}\sum_{l=0}^{L}\binom{L}{l}(q^3)^{L-l}(p^3)^l = \left(\frac{p^3+q^3}{pq}\right)^L \tag{2.24}$$

Hence the variance of X has the final form of

$$VAR(X) = N\left(\left(\frac{p^3+q^3}{pq}\right)^L - 1\right) \tag{2.25}$$

The local differential privacy requires that X should not be too far away from X_e. Which we express as a requirement that the X should not deviate more than certain number σ away from expected value. Hence the local privacy expression is given by:

$$\frac{1}{R(S)} = \frac{1}{N}(E(X) - 3\sqrt{VAR(X)}) \geq \frac{1}{\lambda} \tag{2.26}$$

$$1 - \frac{3}{N}\sqrt{N\left(\left(\frac{p^3+q^3}{pq}\right)^L - 1\right)} \geq \frac{1}{\lambda} \tag{2.27}$$

$$\left(\frac{p^3+q^3}{pq}\right)^L - 1 \leq (1-\frac{1}{\lambda})^2\frac{N}{9} \tag{2.28}$$

$$\left(\frac{p^3+q^3}{pq}\right)^L \leq 1 + (1-\frac{1}{\lambda})^2\frac{N}{9} \tag{2.29}$$

$$\frac{p^3+q^3}{pq} \leq \sqrt[L]{1 + (1-\frac{1}{\lambda})^2\frac{N}{9}} \tag{2.30}$$

From here we express q:

$$\frac{p^3+q^3}{pq} = \frac{1-3q+3q^2}{q(1-q)} \leq \sqrt[L]{1+(1-\frac{1}{\lambda})^2\frac{N}{9}} \tag{2.31}$$

$$q^2 - q + \frac{1}{3+\sqrt[L]{1+(1-\frac{1}{\lambda})^2\frac{N}{9}}} \leq 0 \tag{2.32}$$

$$q \geq \frac{1}{2}\left(1-\sqrt{1-\frac{4}{3+\sqrt[L]{1+(1-\frac{1}{\lambda})^2\frac{N}{9}}}}\right) \tag{2.33}$$