K-Randomization

Maxim Zhilyaev David Zeber

April 4, 2016

1 Privacy ratio in multivariate case

Definitions

D - original collection
D_m - modified collection

N - size of collection in vectors
L - length of a vector
0 - 0-vector, a vector consisting of 0 buts
1 - unit-vector. a vector consisting of 1 bits
S - observed synthetic collection

Note that there are 2^L possible distinct synthetic vectors. We denote v_i a distinct synthetic vector, whereby i is ranging from 1 to 2^L.

$$v_1 = \underbrace{0000\ldots0000}_{L} \tag{1.1}$$

$$v_2 = \underbrace{1000\ldots0000}_{L} \tag{1.2}$$

$$v_3 = \underbrace{0100\ldots0000}_{L} \tag{1.3}$$

$$\ldots \tag{1.4}$$

$$v_{2^L} = \underbrace{1111\ldots1111}_{L} \tag{1.5}$$

The synthetic output S can be represented by a count of same v_i vectors found after randomization:

$$S = [s_1, s_2, \ldots, s_{2^L-1}, s_{2^L}]$$

Suppose that collection D contains a unit-vector that is modified into a zero-vector to receive a collection D_m. The remaining $N-1$ vectors form a collection called D'. Obviously:

$$D = D' + 1 \tag{1.6}$$

$$D_m = D' + 0 \tag{1.7}$$

The privacy ratio is a function of synthetic collection S and expressed as a ratio of probabilities of S given original and modified collections:

$$R(S) = \frac{P(S|D_m)}{P(S|D)} = \frac{P(S|D'+0)}{P(S|D'+1)} \tag{1.8}$$

2 properties of privacy ratio

The probability of generating S from collection D' and a single vector y is given by:

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

Given that a unit-vector is modified into a zero-vector, we may re-write the privacy ratio $R(S)$ as

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

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

2.1 privacy ratio of a homogenous vectors collection

An important class of collections is when all collection vectors are the same. In which case the distribution of randomized vectors generated by D' is multinomial because the probability of generating a particular synthetic v_i from any vector of D' remains constant. Hence, we consider D' to be a collection of $N-1$ identical vectors x of L bits long. Each x has k zero bits and r one bits in the exact same bit positions.

Since $P(s_1, s_2, \ldots, s_i-1, \ldots, s_{2^L}|D')$ is multinomial, we can express probabilistic ratio term in 2.3 as:

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

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

Using that result in the ratio expression 2.3 we have:

$$\frac{P(S|D'+0)}{P(S|D'+1)} = \frac{p(v_1|0) + \sum_{i=2}^{2^L} \frac{s_i}{s_1} \cdot \frac{p(v_1|x)}{p(v_i|x)} p(v_i|0)}{p(v_1|1) + \sum_{i=2}^{2^L} \frac{s_i}{s_1} \cdot \frac{p(v_1|x)}{p(v_i|x)} p(v_i|1)} = \tag{2.6}$$

$$\frac{s_1 \frac{p(v_1|0)}{p(v_1|x)} + \sum_{i=2}^{2^L} s_i \cdot \frac{p(v_i|0)}{p(v_i|x)}}{s_1 \frac{p(v_1|1)}{p(v_1|x)} + \sum_{i=2}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|x)}} = \tag{2.7}$$

$$\frac{\sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|0)}{p(v_i|x)}}{\sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|x)}} \tag{2.8}$$

We restate this important result as it will be used extensively below:

$$R(S) = \frac{\sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|0)}{p(v_i|x)}}{\sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|x)}} \tag{2.9}$$

3 privacy ratio of a zero-valued collection

Suppose that D' contains $N-1$ zero vectors, denote such collection as D'_0. Then, replacing $x=0$ in formula 2.9, we get:

$$R(S|D'_0) = \frac{\sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|0)}{p(v_i|0)}}{\sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|0)}} = \frac{\sum_{i=1}^{2^L} s_i}{\sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|0)}} = \frac{N}{\sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|0)}} \tag{3.1}$$

Note that if v_i and v_j have same number of set bits (call this number l), the corresponding probabilities ratios inside the sum are the same:

$$\frac{p(v_i|1)}{p(v_i|0)} = \frac{p(v_j|1)}{p(v_j|0)} = \frac{p^l q^{L-l}}{p^{L-l} q^l} = \left(\frac{q}{p}\right)^{L-2l} \tag{3.2}$$

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:

$$R(S) = \frac{P(S|D'+0)}{P(S|D'+1)} = \frac{N}{\sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|0)}} = \frac{N}{\sum_{l=0}^{L} s_l \cdot \left(\frac{q}{p}\right)^{L-2l}} \tag{3.3}$$

$$\frac{1}{R(S)} = \frac{1}{N} \sum_{l=0}^{L} s_l \cdot \left(\frac{q}{p}\right)^{L-2l} \tag{3.4}$$

An observer does not gain any more privacy insight by looking at counts of identical vectors than by looking at aggregated counts in a histogram buckets each collecting synthetic vectors with equal number of set bits. Hence, we can equivalently represent S as the count of vectors that have same number of set bits l, and there are $L+1$ such buckets, where $l=0$ corresponds to the bucket of synthetic zero-vectors and $l=L+1$ corresponds to the bucket of synthetic unit-vectors.

3.1 properties of the sum

Consider random variable X such that:

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

The privacy ratio depends only on X and the size of the collection. Hence, we study the properties of X - mainly it's expectation and variance. Note that probability of generating a synthetic vector containing l set bits from either unit or zero original is given by:

$$p(l|1) = \binom{L}{l} p^l q^{L-l} \tag{3.5}$$

$$p(l|0) = \binom{L}{l} q^l p^{L-l} \tag{3.6}$$

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 $L+1$ correlated variables X_l, where:

$$X_l = \left(\frac{q}{p}\right)^{L-2l} Binomial(p(l|0), N) \tag{3.7}$$

$$X = \sum_{l=0}^{L} X_l = \sum_{l=0}^{L} \left(\frac{q}{p}\right)^{L-2l} Binomial(p(l|0), N) \tag{3.8}$$

The expected values of X is given below:

$$E(X) = E(\sum_{l=0}^{L} X_l) = \sum_{l=0}^{L} E(X_l) \left(\frac{q}{p}\right)^{L-2l} = \sum_{l=0}^{L} N \cdot p(l|0) \left(\frac{q}{p}\right)^{L-2l} = \tag{3.9}$$

$$\sum_{l=0}^{L} N \cdot \binom{L}{l} q^l p^{L-l} \left(\frac{q}{p}\right)^{L-2l} = N \sum_{l=0}^{L} \binom{L}{l} \cdot q^{L-l} p^l = N(p+q)^L = N \tag{3.10}$$

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

$$VAR(X) = \sum_{l=0}^{L} VAR(X_l) + 2\sum_{i\leq j}\sum_{<j\leq L} COV(X_i, X_j) \tag{3.11}$$

$$VAR(X) = \sum_{l=0}^{L} N\left[\left(\frac{q}{p}\right)^{L-2l}\right]^2 p(l|0)(1-p(l|0)) - 2\sum_{j\neq j} Np(i|0)\left(\frac{q}{p}\right)^{L-2i} p(j|0)\left(\frac{q}{p}\right)^{L-2j} = \tag{3.12}$$

$$VAR(X) = N\left(\sum_{l=0}^{L}\left[\left(\frac{q}{p}\right)^{L-2l}\right]^2 p(l|0) - \sum_{l=0}^{L}\left[p(l|0)\left(\frac{q}{p}\right)^{L-2l}\right]^2 - 2\sum_{j\neq j} p(i|0)\left(\frac{q}{p}\right)^{L-2i} p(j|0)\left(\frac{q}{p}\right)^{L-2j}\right) \tag{3.13}$$

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

$$VAR(X) = N\left(\sum_{l=0}^{L}\left[\left(\frac{q}{p}\right)^{L-2l}\right]^2 p(l|0) - \left[\sum_{l=0}^{L} p(l|0)\left(\frac{q}{p}\right)^{L-2l}\right]^2\right) \tag{3.14}$$

$$VAR(X) = N\left(\sum_{l=0}^{L}\left[\left(\frac{q}{p}\right)^{L-2l}\right]^2 p(l|0) - 1\right) \tag{3.15}$$

We now simplify the first term of the sum:

$$\sum_{l=0}^{L}\left[\left(\frac{q}{p}\right)^{L-2l}\right]^2 p(l|0) = \tag{3.16}$$

$$\sum_{l=0}^{L}\binom{L}{l} q^l p^{L-l}\left(\frac{q}{p}\right)^{L-2l} \cdot \left(\frac{q}{p}\right)^{L-2l} = \tag{3.17}$$

$$\sum_{l=0}^{L}\binom{L}{l} p^l q^{L-l} \cdot \left(\frac{q}{p}\right)^{L-2l} \tag{3.18}$$

$$\sum_{l=0}^{L}\binom{L}{l}\frac{q^{2L-3l}}{p^{L-3l}} = \sum_{l=0}^{L}\binom{L}{l}\frac{q^{3L-3l}p^{3l}}{(pq)^L} = \tag{3.19}$$

$$\frac{1}{(pq)^L}\sum_{l=0}^{L}\binom{L}{l}(q^3)^{L-l}(p^3)^l = \left(\frac{p^3+q^3}{pq}\right)^L \tag{3.20}$$

Hence the variance of X has the final form of

$$VAR(X) = N\left(\left(\frac{p^3+q^3}{pq}\right)^L - 1\right) \tag{3.21}$$

3.2 local privacy ratio

The local differential privacy requires that $R(S)$ should have a reasonable probability of occurring. Which we express as a requirement that the X should not deviate more than certain number σ deviations away from expected value.

Remark: It's not immediately obvious why the last statement is valid. First, the probabilities of $R(S)$ do not necessarily equal to probabilities of X. Suppose that we know that $P(X < E(X)-3\sigma)$ is sufficiently small, does it mean that $P(R(S) > \frac{N}{E(X)-3\sigma})$ is as small. I actually think it does, because there should be just as many values of $R(S) > \frac{N}{E(X)-3\sigma}$ as there are values of $X < E(X) - 3\sigma$. Another issue, is why do we think that $P(X < E(X) - 3\sigma)$ is sufficiently small? I think it holds because X is essentially a sum of wighted binomials and it should have a bell shaped PDF, however we need a better proof of it.

Assuming 3σ away form the mean is sufficient and the privacy ratio limit equal λ, the local differential privacy ratio condition is met when:

$$R(S) == \frac{N}{N-3\sigma} = \frac{N}{E(X)-3\sqrt{VAR(X)}} \leq \lambda \tag{3.22}$$

$$\frac{1}{R(S)} = \frac{1}{N}(E(X)-3\sqrt{VAR(X)}) \geq \frac{1}{\lambda} \tag{3.23}$$

$$1-3\sqrt{\frac{1}{N}\left(\left(\frac{p^3+q^3}{pq}\right)^L-1\right)} \geq \frac{1}{\lambda} \tag{3.24}$$

$$3\sqrt{\frac{1}{N}\left(\left(\frac{p^3+q^3}{pq}\right)^L-1\right)} \geq \frac{\lambda-1}{\lambda} \tag{3.25}$$

4 privacy ratio of a unit-vector collection

We now suppose that D' contains $N-1$ unit vectors. Again, the distribution of randomized vectors generated by D' is multinomial, because the probability of generating a particular v_i from a unit vector remains constant. The privacy ratio is derived in a similar fashion and is equal to:

$$R(S) = \frac{P(S|D'+0)}{P(S|D'+1)} = \frac{\sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|0)}{p(v_i|1)}}{N} = \frac{\sum_{l=0}^{L} s_l \cdot \left(\frac{p}{q}\right)^{L-2l}}{N} \tag{4.1}$$

Considering the sum as a random variable Y:

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

Y is the sum of correlated random variables Y_l each accepting binomial distribution multiplied by a constant:

$$Y = \sum_{l=0}^{L} Y_l = \sum_{l=0}^{L} \left(\frac{p}{q}\right)^{L-2l} Binomial(p(l|1), N) \tag{4.2}$$

Using same approach as in zero-valued case, one can show that:

$$E(Y) = N \tag{4.3}$$

$$VAR(Y) = N\left(\left(\frac{p^3+q^3}{pq}\right)^L - 1\right) \tag{4.4}$$

$$R(S) = \frac{Y}{N} \tag{4.5}$$

Lemma 1

Denote a collection D' consisting of all zeros-vectors as D'_0, and a collection of only unit-vectors as D'_1. Then the privacy ratio for D'_0 is greater than that of D'_1 when they both deviate from the mean by the same amount.

Proof

Recall that the probability ratio for D'_0 is given by:

$$R(S_0) = \frac{N}{X} \tag{4.6}$$

$$\text{where X is a random variable for the following sum:} \tag{4.7}$$

$$X = \sum_{l=0}^{L} \left(\frac{q}{p}\right)^{L-2l} Binomial(p(l|0), N) \tag{4.8}$$

Conversely, the probability ratio for D'_1 is given by:

$$R(S_1) = \frac{Y}{N} \tag{4.9}$$

$$\text{where Y is a random variable for the following sum:} \tag{4.10}$$

$$Y = \sum_{l=0}^{L} \left(\frac{p}{q}\right)^{L-2l} Binomial(p(l|1), N) \tag{4.11}$$

Note that $R(S_0)$ and $R(S_1)$ are both random variables depending entirely on X and Y. X and Y have same expectation N. Let X and Y deviate from the mean by the same distance d. The corresponding privacy ratios become:

$$R(S_0) = \frac{N}{E(X) - d} = \frac{N}{N - d} \tag{4.12}$$

$$R(S_1) = \frac{E(Y) + d}{N} = \frac{N + d}{N} \tag{4.13}$$

$$R(S_0) > R(S_1) \tag{4.14}$$

$$\frac{N}{N - d} > \frac{N + d}{N} \tag{4.15}$$

$$N^2 > N^2 - d^2 \tag{4.16}$$

Last formula is always true, which proves the lemma. Setting $d = 3\sigma$ we immediately prove that privacy ratio for D'_0 is greater than that of D'_1 at the end of local privacy range.

Remark: This way of reasoning assumes that probabilities of X and Y at 3 deviations of the mean are small and comparable (i.e. $P(X < E(X) - 3\sigma) \leq P(Y > E(Y) + 3\sigma)$. This obviously needs a better argumentation.

5 privacy ratio of a homogenous vectors collection

We now consider D' to be a collection of N-1 identical vectors x of L bits long. Each x has k zero bits and r one bits in the exact same bit positions.

Theorem 1

Privacy ratio of for a homogenous collection D' is the product of privacy ratios of two collections: $D'_{0,k}$ - a collection of zero-vectors of length k, and $D'_{1,r}$ - a collection of unit-vectors of length r. To rephrase the statement: If we partition D' into two collections - one containing only 0 bit columns and another containing only 1 bit columns, the privacy ratio of the original collection is the product of privacy ratios of each partition.

Proof

Recall that privacy ratio of homogenous collections is given by 2.9:

$$R(S) = \frac{\sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|0)}{p(v_i|x)}}{\sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|x)}} \tag{5.1}$$

Recall that each x contains k zero bits and r sets bits in the exact same positions. Call 0-columns the positions where x has 0 bit, and 1-columns the positions where x has 1 bit. Suppose a synthetic vector v_i has j sets bits in 0 columns and t set bits in 1 columns. Then the probability ratios for each sum is expressed as:

$$\frac{p(v_i|0)}{p(v_i|x)} = \frac{\overbrace{q^j p^{k-j}}^{0-bits} \cdot \overbrace{q^t p^{r-t}}^{0-bits}}{\underbrace{q^j p^{k-j}}_{0-bits} \cdot \underbrace{p^t q^{r-t}}_{1-bits}} = \left(\frac{p}{q}\right)^{r-2t} \tag{5.2}$$

$$\tag{5.3}$$

$$\frac{p(v_i|1)}{p(v_i|x)} = \frac{\overbrace{p^j q^{k-j}}^{1-bits} \cdot \overbrace{p^t q^{r-t}}^{1-bits}}{\underbrace{q^j p^{k-j}}_{0-bits} \cdot \underbrace{p^t q^{r-t}}_{1-bits}} = \left(\frac{q}{p}\right)^{k-2j} \tag{5.4}$$

Using this result in 5.1, we get

$$R(S) = \frac{\sum_{i=1}^{2^L} s_i \cdot \left(\frac{p}{q}\right)^{r-2t}}{\sum_{i=1}^{2^L} s_i \cdot \left(\frac{q}{p}\right)^{k-2j}} \tag{5.5}$$

By adding together s_i corresponding to the same number of set bits in the 1-columns for the numerator, and 0-columns for the denominator, we arrive to:

$$R(S|D') = \frac{\sum_{t=0}^{r} s_t \cdot \left(\frac{p}{q}\right)^{r-2t}}{\sum_{j=0}^{k} s_j \cdot \left(\frac{q}{p}\right)^{k-2j}} = \frac{\sum_{t=0}^{r} s_t \cdot \left(\frac{p}{q}\right)^{r-2t}}{N} \cdot \frac{N}{\sum_{j=0}^{k} s_j \cdot \left(\frac{q}{p}\right)^{k-2j}} \tag{5.6}$$

$$\tag{5.7}$$

$$R(S|D') = R(S|D'_{1,r}) \cdot R(S|D'_{0,k}) \tag{5.8}$$

Theorem 2

Privacy ratio of D'_0 at the end of the local privacy range is larger than that of any other homogenous collection.

Proof

By **lemma 1** , for the deviation d from the mean, the following holds:

$$R(S|D'_{0,r}) > R(S|D'_{1,r}) \tag{5.9}$$

Therefore, at the end of the local privacy range, the ratio $R(S|D')$ will be less than:

$$R(S|D') = R(S|D'_{0,k}) \cdot R(S|D'_{1,r}) < R(S|D'_{0,k}) \cdot R(S|D'_{0,r}) \tag{5.10}$$

Not consider the product of privacy ratio for $D'_{0,k}$ and $D'_{0,r}$:

$$R(S|D'_{0,k}) \cdot R(S|D'_{0,r}) = \frac{N}{\sum_{t=0}^{r} s_t \cdot \left(\frac{q}{p}\right)^{r-2t}} \cdot \frac{N}{\sum_{j=0}^{k} s_j \cdot \left(\frac{q}{p}\right)^{k-2j}} = \frac{N^2}{X_k \cdot X_r} \tag{5.11}$$

We then need to prove that:

$$R(S|D_{0,L}) > R(S|D'_{0,k}) \cdot R(S|D'_{0,r}), \text{ which holds if} \tag{5.12}$$

$$\frac{N}{E(X_L) - 3\sqrt{VAR(X_L)}} > \frac{N^2}{E(X_k \cdot X_r) - 3\sqrt{VAR(X_r \cdot X_k)}} \tag{5.13}$$

$$E(X_L) - 3\sqrt{VAR(X_L)} < \frac{E(X_k \cdot X_r) - 3\sqrt{VAR(X_r \cdot X_k)}}{N} \tag{5.14}$$

$$E(X_L) - 3\sqrt{VAR(X_L)} < \frac{E(X_k \cdot X_r)}{N} - \frac{3\sqrt{VAR(X_r \cdot X_k)}}{N} \tag{5.15}$$

Remark: it is again questionable why we choose the end of the local privacy range for the right side of the inequality as 3σ away from the mean. $X_r \cdot X_k$ is a product of random variables, does it have same shape, how does probability cut off correspond to deviation cut off in this case? Nonetheless, proceeding with the proof.

First consider the expectations:

$$E(X_L) = N \tag{5.16}$$

$$E(X_k \cdot X_r) = E(X_k) \cdot E(X_r) = N^2 \tag{5.17}$$

$$\frac{E(X_k \cdot X_r)}{N} = N \tag{5.18}$$

Which reduces 5.15 to:

$$\sqrt{VAR(X_L)} > \frac{\sqrt{VAR(X_r \cdot X_k)}}{N} \tag{5.19}$$

$$VAR(X_L) > \frac{VAR(X_r \cdot X_k)}{N^2} \tag{5.20}$$

Recall that the variance of $VAR(X_L)$ is given by 3.21:

$$VAR(X_L) = N\left(\left(\frac{p^3 + q^3}{pq}\right)^L - 1\right) \tag{5.21}$$

Denote ϕ as:

$$\phi = \frac{p^3 + q^3}{pq}$$

Then $VAR(X_L)$ takes the form of:

$$VAR(X_L) = N(\phi^L - 1)$$

Since X_r and X_k are independent and $k + r = L$, we have the variance of the product $X_r \cdot X_k$ given by:

$$VAR(X_k) = N(\phi^k - 1) \tag{5.22}$$

$$VAR(X_r) = N(\phi^r - 1) \tag{5.23}$$

$$VAR(X_r \cdot X_k) = N^2 \cdot VAR(X_k) + N^2 \cdot VAR(X_r) + VAR(X_k)VAR(X_r) = \tag{5.24}$$

$$N^3(\phi^k + \phi^r - 2) + N^2(\phi^k \cdot \phi^r - \phi^k - \phi^r + 1) = N^3(\phi^k + \phi^r - 2) + N^2(\phi^L - (\phi^k + \phi^r - 1)) \tag{5.25}$$

Replacing corresponding quantities in 5.20 with the above expressions , we have:

$$VAR(X_L) > \frac{VAR(X_r \cdot X_k)}{N^2} \tag{5.26}$$

$$N(\phi^L - 1) > \frac{N^3(\phi^k + \phi^r - 2) + N^2(\phi^L - (\phi^k + \phi^r - 1))}{N^2} \tag{5.27}$$

$$\phi^L - 1 > \phi^k + \phi^r - 2 + \frac{\phi^L - (\phi^k + \phi^r - 1)}{N} \tag{5.28}$$

$$\phi^L - (\phi^k + \phi^r - 1) > \frac{\phi^L - (\phi^k + \phi^r - 1)}{N} \tag{5.29}$$

Note that:

$$k + r = L \text{ , and} \tag{5.30}$$

$$\phi = \frac{p^3 + q^3}{pq} = \frac{(p+q)(p^2 - pq + q^2)}{pq} = \frac{(p-q)^2}{pq} + 1 > 1 \tag{5.31}$$

From that:

$$\phi^L - (\phi^k + \phi^r - 1) = \phi^{k+r} - \phi^k - (\phi^r - 1) = \phi^k(\phi^r - 1) - (\phi^r - 1) = (\phi^k - 1)(\phi^r - 1) > 0 \tag{5.32}$$

This proves inequality 5.29 and the **Theorem 2**.

Lemma 2.

When $N = 2$, and the D' consists only of a single vector, the privacy ratio at the end the local privacy range is alway maximized in D'_0. This follows immediately from **Theorem 2**, because a single vector collection is necessarily a homogenous collection.

6 Generic collections

Remark: Attempting to prove D'_0 maximality by induction

Theorem 3.

Privacy ratio of D'_0 at the end of the local privacy range is the largest compared to any other collection.

Step 1. For $N = 2$ **Theorem 3** holds due to **Lemma 2**.

Step 2. Suppose that for $N - 2$ local privacy ratio maximizes at D'_0. At $N - 1$ step we add a unit-vector to D'_0 and compare its privacy ratio to that of D' consisting of $N - 1$ zero vectors.

6.1 D' containing one extra unit-vector

D' has $N - 2$ zero-vectors and 1 unit vector. The privacy ratio in this case is given by:

$$R(S) = \frac{P(S|D'_0 + 1 + 0)}{P(S|D'_0 + 1 + 1)} = \frac{P(S|D'_0 + 0 + 1)}{P(S|D'_0 + 1 + 1)} \tag{6.1}$$

Consider the numerator of the above ratio, which has a familiar form of $P(S|D'_0 + 1)$:

$$P(s_1 - 1, s_2, \ldots, s_{2^L}|D') \cdot P(v_1|1) + \cdots + P(s_1, s_2 - 1, \ldots, s_{2^L}|D') \cdot P(v_{2^L}|1) = \tag{6.2}$$

$$\frac{p(v_1|0)P(s_1 - 1, s_2, \ldots, s_{2^L}|D')}{s_1} \sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|0)} = \frac{p^L P(s_1 - 1, s_2, \ldots, s_{2^L}|D')}{s_1} \sum_{l=0}^{L} s_l \cdot \left(\frac{q}{p}\right)^{L-2l} \tag{6.3}$$

The denominator requires double conditioning to take advantage of the multinomial distribution of D'_0. $P(S|D'_0 + 1 + 1)$ is given by:

$$\begin{gathered} p(v_1|1)\ [p(v_1|1)P(s_1 - 2, s_2, s_3 \ldots) + p(v_2|1)P(s_1 - 1, s_2 - 1, s_3, \ldots) + p(v_3|1)P(s_1 - 1, s_2, s_3 - 1, \ldots) \cdots + \\ p(v_2|1)\ [p(v_1|1)P(s_1 - 1, s_2 - 1, s_3 \ldots) + p(v_2|1)P(s_1, s_2 - 2, s_3, \ldots) + p(v_3|1)P(s_1, s_2 - 1, s_3 - 1, \ldots) \cdots] + \\ p(v_3|1)\ [p(v_1|1)P(s_1 - 1, s_2, s_3 - 1 \ldots) + p(v_2|1)P(s_1, s_2 - 1, s_3 - 1, \ldots) + p(v_3|1)P(s_1, s_2, s_3 - 2, \ldots) \cdots] + \\ \cdots \end{gathered} \tag{6.4}$$

$\blacksquare$

Opening brackets and dividing by $P(s_1-2, s_2, s_3\dots)$ we receive the following expression:

$$\begin{array}{ccc}
p(v_1|1)^2+ & p(v_1|1)p(v_2|1)\frac{s_2}{s_1-1}\frac{p(v_1|0)}{p(v_2|0)}+ & p(v_1|1)p(v_3|1)\frac{s_3}{s_1-1}\frac{p(v_1|0)}{p(v_3|0)}+\dots \\
p(v_1|1)p(v_2|1)\frac{s_2}{s_1-1}\frac{p(v_1|0)}{p(v_2|0)}+ & p(v_2|1)^2\frac{s_2(s_2-1)}{s_1(s_1-1)}\left(\frac{p(v_1|0)}{p(v_2|0)}\right)^2+ & p(v_2|1)p(v_3|1)\frac{s_2\cdot s_3}{s_1(s_1-1)}\frac{p(v_1|0)^2}{p(v_2|0)p(v_3|0)}+\dots \\
p(v_1|1)p(v_3|1)\frac{s_3}{s_1-1}\frac{p(v_1|0)}{p(v_3|0)}+ & p(v_2|1)p(v_3|1)\frac{s_2\cdot s_3}{s_1(s_1-1)}\frac{p(v_1|0)^2}{p(v_2|0)p(v_3|0)}+ & p(v_3|1)^2\frac{s_3(s_3-1)}{s_1(s_1-1)}\left(\frac{p(v_1|0)}{p(v_3|0)}\right)^2+\dots \\
\dots & &
\end{array} \tag{6.5}$$

▮

Multiplying each term of the sum by $\frac{s1(s1-1)}{p(v1|0)^2}$ produces the following:

$$\begin{array}{ccc}
\frac{p(v_1|1)^2}{p(v_1|0)^2}s_1(s_1-1)+ & \frac{p(v_1|1)}{p(v_1|0)}\frac{p(v_2|1)}{p(v_2|0)}s_1s_2+ & \frac{p(v_1|1)}{p(v_1|0)}\frac{p(v_3|1)}{p(v_3|0)}s_1s_3+\dots \\
\frac{p(v_1|1)}{p(v_1|0)}\frac{p(v_2|1)}{p(v_2|0)}s_1s_2+ & \frac{p(v_2|1)^2}{p(v_2|0)^2}s_2(s_2-1)+ & \frac{p(v_2|1)}{p(v_2|0)}\frac{p(v_3|1)}{p(v_3|0)}s_2s_3+\dots \\
\frac{p(v_1|1)}{p(v_1|0)}\frac{p(v_3|1)}{p(v_3|0)}s_1s_3+ & \frac{p(v_2|1)}{p(v_2|0)}\frac{p(v_3|1)}{p(v_3|0)}s_2s_3+ & \frac{p(v_3|1)^2}{p(v_3|0)^2}s_3(s_3-1)+\dots \\
\dots & &
\end{array} \tag{6.6}$$

The sum 6.6 admits a very simple expression:

$$\left(\sum_{i=1}^{2^L} s_i\cdot\frac{p(v_i|1)}{p(v_i|0)}\right)^2-\sum_{i=1}^{2^L} s_i\cdot\left(\frac{p(v_i|1)}{p(v_i|0)}\right)^2 \tag{6.7}$$

$$\left(\sum_{l=0}^{L} s_l\cdot\left(\frac{q}{p}\right)^{L-2l}\right)^2-\sum_{l=0}^{L} s_l\cdot\left(\frac{q}{p}\right)^{2(L-2l)} \tag{6.8}$$

We now ready to come back to the privacy ratio:

$$P(S|D_0'+0+1)=\frac{p^LP(s_1-1,s_2,\dots,s_{2^L})}{s_1}\sum_{l=0}^{L}s_l\cdot\left(\frac{q}{p}\right)^{L-2l} \tag{6.9}$$

$$P(S|D_0'+1+1)=\frac{(p^L)^2P(s_1-2,s_2,\dots,s_{2^L})}{s_1(s_1-1)}\left[\left(\sum_{l=0}^{L}s_l\cdot\left(\frac{q}{p}\right)^{L-2l}\right)^2-\sum_{l=0}^{L}s_l\cdot\left(\frac{q}{p}\right)^{2(L-2l)}\right] \tag{6.10}$$

$$\frac{P(S|D_0'+0+1)}{P(S|D_0'+1+1)}=\frac{p^LP(s_1-1,s_2,\dots,s_{2^L})\cdot s1(s1-1)}{p^{2L}P(s_1-2,s_2,\dots,s_{2^L})\cdot s_1}\cdot\frac{\sum_{l=0}^{L}s_l\cdot\left(\frac{q}{p}\right)^{L-2l}}{\left(\sum_{l=0}^{L}s_l\cdot\left(\frac{q}{p}\right)^{L-2l}\right)^2-\sum_{l=0}^{L}s_l\cdot\left(\frac{q}{p}\right)^{2(L-2l)}} \tag{6.11}$$

▮

Using the expression below:

$$\frac{P(s_1-1, s_2, \dots, s_i-1, \dots, s_{2^L}|D'_0)}{P(s_1-2, s_2, \dots, s_i, \dots, s_{2^L}|D'_0)} = \frac{\frac{(N-1)!}{(s_1!-1)!\cdot s_2!,s_3!\dots}p(v_1|0)^{s_1-1}p(v_2|0)^{s_2}\dots}{\frac{(N-2)!}{(s_1-2)!s_2!s_3!\dots}p(v_1|0)^{s_1-2}p(v_2|0)^{s_2}\dots} = \frac{N-1}{s_1-1}p^L \tag{6.12}$$

We finally arrive to the privacy ratio of a collection with an extra unit vector:

$$R(S|D'_0+1) = (N-1)\cdot\frac{\sum_{l=0}^{L} s_l\cdot\left(\frac{q}{p}\right)^{L-2l}}{\left(\sum_{l=0}^{L} s_l\cdot\left(\frac{q}{p}\right)^{L-2l}\right)^2 - \sum_{l=0}^{L} s_l\cdot\left(\frac{q}{p}\right)^{2(L-2l)}} \tag{6.13}$$

Recall that for purely zero-vectored D'_0, the privacy ratio is given by:

$$R(S|D'_0+0) = \frac{N}{\sum_{l=0}^{L} s_l\cdot\left(\frac{q}{p}\right)^{L-2l}} \tag{6.14}$$

From here we could express maximality condition as:

$$\left(\sum_{l=0}^{L} s_l\cdot\left(\frac{q}{p}\right)^{L-2l}\right)^2 > N\sum_{l=0}^{L} s_l\cdot\left(\frac{q}{p}\right)^{2(L-2l)} \tag{6.15}$$

Remark: I believe there's an error in this derivation. Specifically, i do not trust the negative term in 6.13. When taking expectation of that expression it can get arbitrary high as in:

$$E(\sum_{l=0}^{L} s_l\cdot\left(\frac{q}{p}\right)^{2(L-2l)} = \sum_{l=0}^{L} E(s_l)\cdot\left(\frac{q}{p}\right)^{2(L-2l)} = N\sum_{l=0}^{L} p(l|0)\left(\frac{q}{p}\right)^{2(L-2l)} = \tag{6.16}$$

$$N\sum_{l=0}^{L}\binom{L}{l}q^l p^{L-l}\left(\frac{q}{p}\right)^{L-2l}\left(\frac{q}{p}\right)^{L-2l} = N\sum_{l=0}^{L}\binom{L}{l}q^{L-l}p^l\left(\frac{q}{p}\right)^{L-2l} = \tag{6.17}$$

$$N\sum_{l=0}^{L}\binom{L}{l}\frac{q^{2L-3l}}{p^{L-3l}} = N\sum_{l=0}^{L}\binom{L}{l}\frac{q^{2L-2l}}{p^{2L-2l}}\cdot\left(\frac{p}{q}\right)^l\cdot p^L = Np^L\sum_{l=0}^{L}\binom{L}{l}\left(\left(\frac{q}{p}\right)^2\right)^{L-l}\left(\frac{p}{q}\right)^l = \tag{6.18}$$

$$Np^L\left[\left(\frac{q}{p}\right)^2+\frac{p}{q}\right]^L = N\left(\frac{q^3+p^3}{pq}\right)^L \tag{6.19}$$

$\left(\frac{q^3+p^3}{pq}\right)^L$ could be arbitrary large, and the expectation for the denominator of 6.13 becomes negative:

$$E\left[\left(\sum_{l=0}^{L} s_l \cdot \left(\frac{q}{p}\right)^{L-2l}\right)^2 - \sum_{l=0}^{L} s_l \cdot \left(\frac{q}{p}\right)^{2(L-2l)}\right] = N^2 - N\left(\frac{q^3+p^3}{pq}\right)^L \qquad (6.20)$$

This smells fishy :(