

# K-Randomization

Maxim Zhilyaev

David Zeber

December 15, 2015

## 1 Outline of the procedure

•

## 2 Theoretical setup

In the following we work with data in the form of bit vectors. A **bit vector** is a vector  $v \in \{0, 1\}^L$ .

First we define the randomization procedure we will be applying.

**Definition.** The randomization procedure  $R$  with **lie probability**  $0 < q < 1/2$  flips a bit with probability  $q$ , and leaves it as-is with probability  $1 - q$ . In other words, for a bit  $b \in \{0, 1\}$ ,

$$R(b) = R(b; X) = (1 - b) \cdot X + b \cdot (1 - X) \quad \text{where } X \sim \text{Ber}(q).$$

When applied to a vector, each bit is randomized independently:

$$R(v) = R(v; (X_1, \dots, X_L)) = (R(v_1; X_1), \dots, R(v_L; X_L)) \quad \text{where } X \stackrel{\text{iid}}{\sim} \text{Ber}(q).$$

**Remark.** The randomization  $R$  reports the original bit value with probability  $1 - q > q$ , and lies with probability  $q$ . This is equivalent to the randomized response procedure where the value is reported as-is with probability  $1 - f$ , and with probability  $f$  the reported value is the outcome of the toss of a fair coin. In this case,  $q = f/2$ .

**Remark.** If  $q = 1/2$ , then  $R(0) \stackrel{d}{=} R(1)$ , and the reported value is “completely” randomly generated, i.e., independently of the original value.

Distribution of  $R(v)$ .

For a bit  $b$ , the randomization lies iff  $R(b) \neq b$ :

$$P[R(b) = s] = q^{\mathbf{1}_{\{b \neq s\}}} (1 - q)^{\mathbf{1}_{\{b = s\}}}$$

Hence, for a bit vector  $v$ ,

$$P[R(v) = s] = q^{\sum \mathbf{1}_{\{b_i \neq s_i\}}} (1 - q)^{\sum \mathbf{1}_{\{b_i = s_i\}}} = q^{L - m(v, s)} (1 - q)^{m(v, s)},$$

where  $m(v, s) = |\{i : v_i = s_i\}|$ . Note that this probability is maximized when  $m(v, s) = L$  (the reported vector  $s$  is identical to the original vector  $v$ ), and minimized when  $m(v, s) = 0$ . In other words, the most likely outcome of randomizing a bit vector is obtaining an identical vector.

For a collection  $T$ ,

$$P[s \in R(T)] = 1 - P[s \notin R(T)] = 1 - \prod_{v \in T} P[R(v) \neq s] = 1 - \prod_{v \in T} [1 - q^{L - m(v, s)} (1 - q)^{m(v, s)}].$$

### 3 Differential Privacy

The typical setting for differential privacy is the following. We consider a **database** as a collection of records. The records are elements of some space  $D$ , and a database  $\mathbf{x}$  is a vector of  $n$  records:  $\mathbf{x} \in D^n$ .

We wish to release information based on the database by applying a **query** to it. This is a function  $A$  mapping the database into another space:  $A : D^n \rightarrow \mathcal{S}$ . If the function  $A$  is random, i.e.,  $A(\mathbf{x}) = A(\mathbf{x}, X)$  for a random element  $X$ , then the output  $A(\mathbf{x})$  is a random element of  $\mathcal{S}$ .

In considering the differential privacy of  $A$ , we compare the result of applying  $A$  to two very similar databases  $\mathbf{x}, \mathbf{x}' \in D^n$ . We say the databases **differ in one row** if  $\sum_{i=1}^n \mathbf{1}_{\{x_i \neq x'_i\}} = 1$ . The random query  $A$  is said to be  $\epsilon$ -**differentially private** if, for any two databases  $\mathbf{x}, \mathbf{x}' \in D^n$  differing in one row,

$$P[A(\mathbf{x}) \in S] \leq \epsilon \cdot P[A(\mathbf{x}') \in S]$$

for all  $S \subset \mathcal{S}$  (measurable). An alternative notion of differing in one row that is sometimes used is that  $\mathbf{x} \in D^n$ ,  $\mathbf{x}' \in D^{n+1}$ , and  $x_i = x'_i$  for  $i = 1, \dots, n$ . In other words,  $\mathbf{x}'$  includes an additional record that is not in  $\mathbf{x}$ .

If  $\mathcal{S}$  is countable, then we can write

$$P[A(\mathbf{x}) \in S] = \sum_{s \in S} P[A(\mathbf{x}) = s].$$

Hence,

$$\frac{P[A(\mathbf{x}) \in S]}{P[A(\mathbf{x}') \in S]} = \frac{\sum_{s \in S} P[A(\mathbf{x}) = s]}{\sum_{s \in S} P[A(\mathbf{x}') = s]} \leq \max_{s \in S} \frac{P[A(\mathbf{x}) = s]}{P[A(\mathbf{x}') = s]}$$

by the Lemma (need reference).

Furthermore, if  $A$  randomizes each record in the database independently, i.e.,  $A(\mathbf{x}) = A(\mathbf{x}, \mathbf{X}) := (A_0(x_1, X_1), \dots, A_0(x_n, X_n))$  where  $X_i$  are independent, then  $\mathcal{S} = \mathcal{S}_0^n$  and  $s = (s_1, \dots, s_n)$  with  $s_i \in \mathcal{S}_0$ . In this case  $P[A(\mathbf{x}) = s] = P[A_0(x_1) = s_1, \dots, A_0(x_n) = s_n] = \prod P[A_0(x_i) = s_i]$ . If  $\mathbf{x}$  and  $\mathbf{x}'$  differ in one row (wlog  $x_1 \neq x'_1$  and  $x_i = x'_i$  for  $i = 2, \dots, n$ ), then

$$\frac{P[A(\mathbf{x}) = s]}{P[A(\mathbf{x}') = s]} = \frac{P[A_0(x_1) = s_1]}{P[A_0(x'_1) = s_1]}.$$

Therefore, in this case, the query  $A$  will satisfy differential privacy if

$$P[A_0(x) = s] \leq \epsilon \cdot P[A_0(x') = s]$$

for all  $x, x' \in D$  and  $s \in \mathcal{S}_0$ . This is the formulation used in the RAPPOR paper that applies to differences between individual records rather than collections differing on a single element.

Consider a collection  $T$  of bit vectors, and write  $T_v = T \setminus \{v\}$ . The randomization procedure  $R$  is  $\epsilon$ -differentially private if

$$\log \left( \frac{P[R(T) \in S]}{P[R(T_v) \in S]} \right) \leq \epsilon$$

for any set of bit vectors  $S$ .

Anonymity:

$$A_p = \min_{v \in T, s \in \{0,1\}^L} \frac{P[s \in R(T_v)]}{P[s = R(v)]}$$

## 4 Single bit case

### 4.1 Estimating number of single bits

Suppose there are  $T$  set bits in the original collection of  $N$  single bit records. After randomization is performed the number of observed synthetic bits  $S$  is a random variable which we express as:

$$S = p \cdot T + q \cdot (N - T)$$

From here we can express an estimate for  $T$ , computed from observed value of  $S$ :

$$\bar{T} = \frac{S - qN}{p - q} \tag{4.1}$$

The expectation, variance and deviation of  $\bar{T}$  random variable are given by:

$$E(\bar{T}) = T \tag{4.2}$$

$$VAR(\bar{T}) = \frac{qpN}{(p - q)^2} \tag{4.3}$$

$$\sigma(\bar{T}) = \sqrt{\frac{qpN}{(p - q)^2}} \tag{4.4}$$

### 4.2 Local Differential Privacy

We now study how differential privacy ratio changes depending on the configuration of underlying database  $D$ . Assuming that  $D$  consists of  $N$  single bit records, we are interested in deriving

the expression of differential probability ratio as a function of observed number of set bits after randomization is performed.

#### 4.2.1 Choice of D

We are seeking collection  $D$  that maximizes differential privacy ratio for any number of observed bits in the randomized collection  $S$ . Since we initially consider  $D$  to consists of single bits only, the modified record switches the original bit to an opposite value. Without loss of generality, assume that the original record was 1 and it was modified to 0. Hence the original collection  $D$  contains at least one set bit, and the modified collection  $D_m$  contains one less set bits. Both collections generate synthetic collection  $S$ . Call the number of set bits in the synthetic collection a random variable  $s$ . Then, the differential privacy ratio when  $s$  is equal a particular number  $i$  of set bits is given by:

$$R_i = \frac{P(s = i|D_m)}{P(s = i|D)}$$

**Theorem 4.1.**  $R_i$  is maximized when  $D$  contains  $N$  set bits

*Proof.* Suppose there are  $m$  set bits in the original collection  $D_m$ . Consider generating function for the number of set bits  $s$  in  $S$ .

$$G_m(x) = (q + px)^m(p + qx)^{N-m} = \sum_{i=0}^N a_i^m x^i$$

Note that coefficients  $a_i^m$  in the expansion of the generating function  $G_m$  represent probabilities of  $P(s = i|D)$ . We prove that for any  $i$ , the differential privacy ratio grows with  $m$ :

$$\frac{a_i^m}{a_i^{m+1}} > \frac{a_i^{m-1}}{a_i^m}$$

which holds when

$$(a_i^m)^2 > a_i^{m-1} a_i^{m+1} \tag{4.5}$$

$$(a_i^m)^2 - a_i^{m-1} a_i^{m+1} > 0 \tag{4.6}$$

Consider generating functions for  $m + 1$ ,  $m$  and  $m - 1$  respectively:

$$G_{m+1}(x) = (p + qx)^{m+1}(p + qx)^{N-m-1} \quad (4.7)$$

$$G_m(x) = (p + qx)^m(p + qx)^{N-m} \quad (4.8)$$

$$G_{m-1}(x) = (p + qx)^{m-1}(p + qx)^{N-m+1} \quad (4.9)$$

Define  $Q(x)$  as:

$$Q(x) = (p + qx)^{m-1}(p + qx)^{N-m-1} = \sum_{i=0}^{N-2} b_i x^i$$

Then generating functions above are expressed as:

$$G_{m+1}(x) = Q(x)(q + px)^2 = \sum_{i=0}^N [b_i q^2 + 2qpb_{i-1} + b_{i-2}p^2]x^i \quad (4.10)$$

$$G_m(x) = Q(x)(q + px)(p + qx) = \sum_{i=0}^N [b_i qp + (q^2 + p^2)b_{i-1} + b_{i-2}qp]x^i \quad (4.11)$$

$$G_{m-1}(x) = Q(x)(p + qx)^2 = \sum_{i=0}^N [b_i p^2 + 2qpb_{i-1} + b_{i-2}q^2]x^i \quad (4.12)$$

From here we can express coefficients of each generating function through coefficients of  $Q(x)$

$$a_i^{m+1} = b_i q^2 + 2qpb_{i-1} + b_{i-2}p^2 \quad (4.13)$$

$$a_i^m = b_i qp + (q^2 + p^2)b_{i-1} + b_{i-2}qp \quad (4.14)$$

$$a_i^{m-1} = b_i p^2 + 2qpb_{i-1} + b_{i-2}q^2 \quad (4.15)$$

Now, replace the coefficients  $a_i$  in the 3.2 with their expressions through  $b_i$ .

$$(a_i^m)^2 - a_i^{m-1}a_i^{m+1} = (b_i qp + (q^2 + p^2)b_{i-1} + b_{i-2}qp)^2 - (b_i q^2 + 2qpb_{i-1} + b_{i-2}p^2) \cdot (b_i p^2 + 2qpb_{i-1} + b_{i-2}q^2)$$

After trivial algebraic transformations the above expressions simplifies to:

$$(a_i^m)^2 - a_i^{m-1}a_i^{m+1} = (p^2 - q^2)^2 \cdot (b_i^2 - b_{i+1}b_{i-1}) \geq 0$$

Note that the first term of the product is always greater than 0, and we will show that the second term is greater or equal to zero as well.

**Lemma 1** If a polynomial has the form bellow

$$Q(x) = \sum_{i=0}^n a_i x^i = a_n \prod_i^n (r_i + x), \text{ where } r_i \geq 0$$

Then

$$(a_i^2 - a_{i+1}a_{i-1}) \geq 0$$

*Proof.* Assume polynomial is monic (e.g.  $a_n = 1$ ), and prove lemma by induction.

For  $n=2$ :

$$(r_1 + x)(r_2 + x) = r_1 * r_2 + (r_1 + r_2)x + x^2 \quad (4.16)$$

$$a_1^2 - a_0a_2 = (r_1 + r_2)^2 - r_1 * r_2 = r_1^2 + r_2^2 + r_1r_2 > 0, \text{ since } r_1 > 0 \text{ and } r_2 > 0 \quad (4.17)$$

Assume that for  $n$ , the statement holds for all  $i$ , then for  $n + 1$  we can express the polynomial as:

$$Q^{n+1}(x) = \sum_{i=0}^{n+1} a_i x^i = \prod_i^{n+1} (r_i + x) = Q^n(x) \cdot (r_{n+1} + x) = \left( \sum_{i=0}^n b_i x^i \right) \cdot (r_{n+1} + x) \quad (4.18)$$

$$\sum_{i=0}^{n+1} a_i x^i = \sum_{i=0}^{n+1} [b_i r_{n+1} + b_{i-1}] x^i \quad (4.19)$$

$$a_i = b_i r_{n+1} + b_{i-1} \quad (4.20)$$

The index of  $r_{n+1}$  is irrelevant for the proof, hence we drop it. We now express  $(a_i^2 - a_{i+1}a_{i-1})$  through coefficients of  $Q^n(x)$  and perform algebraic simplifications:

$$a_i^2 - a_{i+1}a_{i-1} = [b_i r + b_{i-1}]^2 - [b_{i+1}r + b_i] \cdot [b_{i-1}r + b_{i-2}] \quad (4.21)$$

$$a_i^2 - a_{i+1}a_{i-1} = r^2(b_i^2 - b_{i+1}b_{i-1}) + r(b_i b_{i-1} - b_{i+1}b_{i-1}) + (b_{i-1}^2 - b_i b_{i-2}) \quad (4.22)$$

$$b_i^2 - b_{i+1}b_{i-1} \geq 0 \text{ by induction hypothesis} \quad (4.23)$$

$$b_{i-1}^2 - b_i b_{i-2} \geq 0 \text{ by induction hypothesis} \quad (4.24)$$

$$(4.25)$$

$$b_i b_{i-1} - b_{i+1}b_{i-1} \geq 0 \text{ because all } b_i \text{ are positive and} \quad (4.26)$$

$$b_i^2 \geq b_{i+1}b_{i-1} \text{ and } b_{i-1}^2 > b_i b_{i-2} \quad (4.27)$$

$$b_i^2 \cdot b_{i-1}^2 \geq b_{i+1}b_{i-1}b_i b_{i-2} \quad (4.28)$$

$$b_i b_{i-1} \geq b_{i+1}b_{i-1} \quad (4.29)$$

This completes the proof of **Lemma 1** for monic polynomials. Same result is true for non-monic polynomials because if  $(a_i^2 - a_{i+1}a_{i-1}) \geq 0$ , then multiplying each coefficient by constant factor does not change the inequality.

We now ready to finish the proof of **Theorem 4.1**. Consider the generating function  $G_m(x)$  again:

$$G_m(x) = (q + px)^m(p + qx)^{N-m} = p^m q^{N-m} \left(\frac{q}{p} + x\right)^m \left(\frac{p}{q} + x\right)^{N-m}$$

Note that since  $p$  and  $q$  are probabilities, the expressions in parenthesis are of the form necessary for **Lemma 1** to hold. Which proves that:

$$\frac{a_i^m}{a_i^{m+1}} > \frac{a_i^{m-1}}{a_i^m}$$

Which in turn proves that the differential privacy ratio maximizes when  $m = N$  □

□

### 4.3 Maximum and Local differential privacy

Since we established a notion of a differential privacy ratio  $R_i$  to be a function of the observed number of set bits in the synthetic output, it's instructive to see how this ratio changes with  $i$ . Since  $D$  consists of set bits, we have for any  $i$

$$P(s = i|D) = \binom{N}{i} p^i q^{N-i} \tag{4.30}$$

$$P(s = i|D_m) = \binom{N-1}{i} p^{i+1} q^{N-i} + \binom{N-1}{i-1} p^{i-1} q^{N-i+1} \tag{4.31}$$

$$R_i = \frac{P(s = i|D_m)}{P(s = i|D)} = \frac{N-i}{N} \frac{p}{q} + \frac{i}{N} \frac{q}{p} \tag{4.32}$$

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

$$R_0 = \frac{p}{q}$$

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

$$R_N = \frac{q}{p}$$

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:

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

This observation raises a question of reducing absolute theoretical bound of classical differential privacy by considering realistic values of  $i$ , rather than all possible outcomes of randomization. Indeed, the probability of all  $N$  bits of  $D$  generating  $N$  zeros is very low. For example, assuming  $p = 0.7$ ,  $q = 0.3$  and  $N = 100$ , the probability of seeing no synthetic ones is  $q^{100} = 5e^{-53}$ , which is improbable for any realistic scenario. Instead, we should consider values of  $i$  that are realistic, which, in statistical sense, means those values of  $i$  that lie within certain number of  $\sigma$  away from the expected mean.

This brings about a notion of a **local differential privacy**, whereby the probabilistic ratio is considered only for values of  $i$  that have realistic chance of being observed. Consider the expression for  $R_i$  again.

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

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

$$i = pN - 3\sigma = pN - 3\sqrt{pqN} \quad (4.33)$$

$$R_i = \frac{P(s = i | D_m)}{P(s = i | D)} = \frac{N - i}{N} \frac{p}{q} + \frac{i}{N} \frac{q}{p} \leq \lambda \quad (4.34)$$

$$\text{Max}(R_i) = \frac{N - pN + 3\sqrt{pqN}}{N} \cdot \frac{p}{q} + \frac{pN + 3\sqrt{pqN}}{N} \cdot \frac{q}{p} \leq \lambda \quad (4.35)$$

From here:

$$\frac{N - pN + 3\sqrt{pqN}}{N} \cdot \frac{p}{q} + \frac{pN - 3\sqrt{pqN}}{N} \cdot \frac{q}{p} \leq \lambda \quad (4.36)$$

$$p + q + 3\sqrt{\frac{pq}{N}} \left( \frac{p}{q} - \frac{q}{p} \right) \leq \lambda \quad (4.37)$$

$$1 + 3\sqrt{\frac{pq}{N}} \frac{p^2 - q^2}{pq} \leq \lambda \quad (4.38)$$

$$1 + 3\sqrt{\frac{1}{N}} \cdot \frac{p - q}{\sqrt{pq}} \leq \lambda \quad (4.39)$$

$$\frac{pqN}{(p - q)^2} \geq \frac{9}{(\lambda - 1)^2} \quad (4.40)$$



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:

$$VAR(\bar{T}) = \frac{pqN}{(p-q)^2} \geq \frac{9}{(\lambda-1)^2} \quad (4.41)$$

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

$$\sigma(\bar{T}) \geq \frac{3}{\lambda-1} = \frac{3}{e^\epsilon - 1} \quad (4.42)$$

We can replace the local differential privacy bound with slightly less optimal, but more analytically convenient bound. Note that:

$$\frac{q}{p} < \frac{pq}{(p-q)^2} \quad (4.43)$$

$$(p-q)^2 < p^2 \quad (4.44)$$

$$(1-2q)^2 < (1-q)^2 \quad (4.45)$$

Hence, the local differential privacy holds when  $N$  is large enough to guarantee:

$$\frac{pqN}{(p-q)^2} > \frac{q}{p}N \geq \frac{9}{(\lambda-1)^2} \quad (4.46)$$

From here, we can express RRT noise parameter  $q$  through  $N$  and  $\lambda$ :

$$q \geq \frac{1}{1 + \frac{(\lambda-1)^2 N}{9}} \quad (4.47)$$

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

$$q = \frac{1}{1 + \frac{(2-1)^2 1000}{9}} = 0.009$$

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

$$\frac{p}{q} \leq 2 \quad (4.48)$$

$$q \geq \frac{1}{3} = 0.333 \quad (4.49)$$

The notion of local privacy allowed us to reduce RRT noise 37 times utilizing the noise created by 1000 records existing in the collection. This also 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  $\sigma = 3$ , meaning that precision of RRT estimates had grown 10 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  $D_m$ .  $D$  contains 1000 set bits and  $D_m$  contains 999 set bits. Corresponding means and deviation for sum of observed synthetic bits in each case is given below:

$$E(S) = p \cdot 1000 \quad (4.50)$$

$$\sigma(S) = \sqrt{pq \cdot 1000} \quad (4.51)$$

$$E(S_m) = p \cdot 999 + q \quad (4.52)$$

$$\sigma(S_m) = \sqrt{pq \cdot 999 + pq} \quad (4.53)$$

We now consider the confidence intervals for both  $S$  and  $S_m$  for RRT under classical and local differential privacy constraints. If  $q = 0.333$  the confidence interval for  $S$  and  $S_m$  are:

$$S - > [621.98, 711.42] \quad (4.54)$$

$$S_m - > [621.65, 711.09] \quad (4.55)$$

Under local differential privacy, the noise level  $q = 0.009$ , and the confidence intervals become:

$$S - > [982.04, 999.96] \quad (4.56)$$

$$S_m - > [981.06, 998.98] \quad (4.57)$$

Note that 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.

## 5 K-randomization for a single bit case

We now consider an important technique for further increasing the estimation precision while providing same local privacy guarantees.