Problem Set 1: Data Privacy.

March 6, 2012

Solution to Ex 3. We will use the following (completely useless in real-world applications) deterministic mechanism M as a counter-example. For a database A, of the format described in the question, replace all quasi-identifier fields with *(star), completely "anonymizing" A except for the sensitive attribute "Disease". Clearly the changed database A' satisfies $k-$anonymity for $k = |A|$ and also t-closeness for $t = 0$ since the whole database is a fully anonymized equivalence class with the exact same sensitive data distribution as A.

There is a really large class of queries that M would be able to answer but let us assume the queries are of the form "Does there exist a person with "Disease I" in the the database?", (really M could just output the whole anonymized database instead of a particular query answer and still maintain k-anonymity and t-closeness) which clearly results in a non-constant mechanism . By construction M is a deterministic mechanism and as shown in (4), no non-constant deterministic mechanism can provide ϵ-differential privacy for $\epsilon < \infty$.

Solution to Ex 4.

1. Since M is non-constant there exist at least two databases X, Y such that $M(X) = c_1$, $M(X) = c_2$ and $c_1 \neq c_2$. Also since M is deterministic, its output is fully defined by its input with no randomness which means that $\Pr[M(X) = c_1] = 1$ and since $M(Y) = c_2$ it follows that $\Pr[M(Y) = c_1] = 0$. Therefore $\frac{\Pr[M(X)=c_1]}{\Pr[M(Y)=c_1]} \sim \infty$ and thus it cannot be bounded by $e^{\epsilon(X \oplus Y)}$ for no $\epsilon < \infty$.

2. Since x never appears as output of $M(A)$, it follows that $\Pr[M(A) = x] = 0$. Therefore $\Pr[M(B) = x]$ for any database B should also be equal to 0. Otherwise, in the exact same way as above, the ratio of the probabilities would be ∞ and therefore ϵ-differential privacy would be impossible for any $\epsilon < \infty$ which contradicts our assumption that M offers ϵ-differential privacy.

3. Let database A be of the format described in the question and set $Z = \{\text{zip code } z | \text{ at least one tuple with code } z^* \text{ exists in } A\}$. Without loss of generality assume value $z^* \in Z$ is associated with only one tuple in A. Then if we remove this tuple from A we create database B and corresponding set $Z' = Z \setminus z^*$ (since z^* was only associated with this particular tuple that was removed). From the above, it follows that $\Pr[M(A) = \{z^*, c\}] > 0$, $\forall c \in Range(M)$ since there exist a tuple with zip code z^* in A. Accoridngly, $\forall c \in Range(M)$, $\Pr[M(B) = \{z^*, c\}] = 0$ since there are no tuples with zip code z^* in B. Therefore as of the above analysis in (2) it follows that M cannot offer ϵ-differential privacy for any $\epsilon < \infty$.

4. The above problem results from the fact that we are outputting a pair of two things: a count (noisy) and a zip code (non-noisy) and what causes the problem is that the absence of particular zip code from the database mandates the absence of an output pair containing this code. There are many alternative approaches that can be used to overcome this. We acknowledge that adding "noise" to the zip code makes no sense since zip code is an alpharithmetic number.

   Therefore one way to solve the problem would be, to make M output not only pairs for codes in the database but also for codes not in it (a count for a code z' not in database would be $0+ \mathcal{L}(\frac{1}{\epsilon})$). This would mean that the answer size would roughly be 10^5 pairs, making such a mechanism unattractive for real-world applications but it would satisfy the ϵ-differential privacy definition for some ϵ.

   One alternative to that would be to "bucketize" the output, for example splitting the output for in 10 buckets over the whole domain of zip codes, and output only these ten values $\{zipcodespan, noisycount\}$, (implying $0+ \mathcal{L}(\frac{1}{\epsilon})$ as above for empty buckets). Such an approach yields a smaller answer size, however we cannot make any particular assumption for the bucket boundaries if we want to achieve differential privacy, thus leading to a poor alternative to the previous output-everything approach.

Solution to Ex 5. (Graded out of 5, with 3 bonus points for getting the last part correct.)

1. NoisyCount(A,3ϵ). This is just the Noisy Count mechanism described in class; its privacy budget is 3ϵ, and its standard deviation is $\frac{\sqrt{2}}{3\epsilon}$.

2. Repeat NoisyCount(A,ϵ), three times to obtain RV's Y_1, Y_2, Y_3, and output the average $\frac{1}{3}\sum_{i=1}^{3} Y_i$. Since we call NoisyCount(A,ϵ) three times, the privacy budget of this mechanism will be 3ϵ by composition of differentially private mechanisms. Then to calculate standard deviation, we first calculate variance, using both that $Var[aX + b] = a^2 Var[X]$, and that if X and Y are independant RVs, $Var[X + Y] = Var[X] + Var[Y]$. Then

$$\begin{aligned} Var[\frac{Y_1 + Y_2 + Y_3}{3}] &= \frac{1}{9}(Var[Y_1 + Y_2 + Y_3]) \\ &= \frac{1}{9}(Var[Y_1] + Var[Y_2] + Var[Y_3]) \\ &= \frac{1}{9}(\frac{2}{\epsilon^2} + \frac{2}{\epsilon^2} + \frac{2}{\epsilon^2}) \\ &= \frac{2}{3\epsilon^2} \end{aligned}$$

   Then taking the square root, we get that the standard deviation of this mechanism is $\frac{\sqrt{2}}{\sqrt{3}\epsilon}$.

3. Repeat NoisyCount(A,ϵ), three times to obtain RV's Y_1, Y_2, Y_3, and output the median of these values.

   By composition of differentially private mechanisms, this mechanism has privacy budget 3ϵ. To calculate standard deviation, we first work to define the PDF of the median function as described. For any x, the probability that $X = x$ is the probability that one of our three draws is x, one is less than or equal to x, and one is greater than x, drawing these three in

any order. Furthermore, there are $3! = 6$ ways to order of Y_1, Y_2, Y_3 such that the first RV is $< y$, the second is $= y$ and the third is $> y$. Therefore, for the median mechanism,

$$\begin{aligned} PDF(x) &= 6 * PDF(\mathcal{L}(\frac{1}{\epsilon})) * CDF(\mathcal{L}(\frac{1}{\epsilon})) * (1 - CDF(\mathcal{L}(\frac{1}{\epsilon}))) \\ &= 6 * \frac{\epsilon}{2} e^{-\epsilon|x|} * \frac{1}{2} e^{-\epsilon|x|} * (1 - \frac{1}{2} e^{-\epsilon|x|}) \end{aligned}$$

Then to compute the variance of a median random variable X from this mechanism, we have that $Var[X] = E[X^2] - E[X]$. By definition,

$$E[X] = \int_{-\infty}^{\infty} x * PDF(x)\, dx$$

Notice that the function $PDF(x)$ described above is even, since $PDF(x) = PDF(-x)$ for all $x \in \mathbb{R}$. Therefore, since the identity function is odd, we have that

$$E[X] = \int_{-\infty}^{\infty} x * PDF(x)\, dx = 0$$

since the integral over all of $\mathbb{R}$ of the product of an even function and an odd function is 0. Therefore,

$$\begin{aligned} Var[X] &= E[X^2] - (E[X])^2 \\ &= E[X^2] \\ &= \int_{-\infty}^{\infty} x^2 * PDF(x)\, dx \\ &= \int_{-\infty}^{\infty} x^2 * 6 * \frac{\epsilon}{2} e^{-\epsilon|x|} * \frac{1}{2} e^{-\epsilon|x|} * (1 - \frac{1}{2} e^{-\epsilon|x|})\, dx \\ &= \frac{3\epsilon}{2} \int_{-\infty}^{\infty} x^2 * e^{-2\epsilon|x|} * (1 - \frac{1}{2} e^{-\epsilon|x|}) dx \\ &= \frac{3\epsilon}{2} \int_{-\infty}^{\infty} x^2 * e^{-2\epsilon|x|} dx - \frac{3\epsilon}{4} \int_{-\infty}^{\infty} x^2 e^{-3\epsilon|x|} dx \\ &= \tfrac{3}{2}\text{VAR}[\mathcal{L}(\tfrac{1}{2\epsilon})] - \tfrac{3}{4}\tfrac{2}{3}\text{VAR}[\mathcal{L}(\tfrac{1}{3\epsilon})] \\ &= \tfrac{3}{2}\tfrac{2}{(2\epsilon)^2} - \tfrac{3}{4}\tfrac{2}{3}\tfrac{2}{(3\epsilon)^2} = (\tfrac{3}{4} - \tfrac{1}{27})\tfrac{1}{\epsilon^2} = \tfrac{77}{108\epsilon^2} \end{aligned}$$

Then taking the square root, we get that the standard deviation of this mechanism is $.844\frac{1}{\epsilon}$ (which is slightly larger than the standard deviation of the 'take the average of three' mechanism, which was $\sqrt{\frac{2}{3}}\frac{1}{\epsilon} = .81\frac{1}{\epsilon}$.

4. The described mechanism is 5ϵ-differentially private.

Proof. First, the Y_i that minimizes $(Y_i - |A|)^2$ also minimizes $|Y_i - |A||$.

Next, since each Y_i is a random variable $\mathcal{L}(\frac{1}{\epsilon}) + |A|$, then for each Y_i, $X_i = Y_i - |A|$ has distribution $\mathcal{L}(\frac{1}{\epsilon})$. Therefore, the absolute value $|X_i|$ has an exponential distribution with parameter ϵ; this follows because for $x \geq 0$

$$\Pr[|X_i| = x] = \Pr[X = x] + \Pr[X = -x] = \tfrac{\epsilon}{2} e^{-\epsilon x} + \tfrac{\epsilon}{2} e^{-\epsilon|-x|} = \epsilon e^{-\epsilon x}$$

There is a the theorem on p.198 of Mitzenmacher and Upfal that says that the minimum of indep drawn exponential random variables with parameter ϵ_i is an exponential random variable with parameter $\sum_i \epsilon_i$. Thus the distribution of $\min_i(|Y_i - |A|) = \min_i(|X_i|) = |X_k|$ is exponentially distributed with parameter 5ϵ.

Using the same argument as above, if $|X_k|$ is exponentially distributed with parameter 5ϵ we have that X_k is Laplace with parameter $\frac{1}{5\epsilon}$. Finally, $Y_k = |A| + X_k$ is just a count plus Laplace noise of parameter $\frac{1}{5\epsilon}$; this is exactly the noisy count mechanism described in class with parameter 5ϵ and the result follows.

Solution to Ex 6.

- $DPcount(A, f)$ stability 2

  Fix a dataset A, and consider a dataset B that is just A with one record removed (such that $|A \oplus B| = 1$). Then in the worst case choice of f, the record x removed from B will be such that there is a record $y \in B | f(y) = f(x)$. Then the ordered pairs $(f(y), c) \in DPcount(A, f)$ and $(f(y), c') \in DPcount(B, f)$ are such that $c \neq c'$. Therefore, $|DPcount(A, f) \oplus DPcount(B, f)| = 2$, and DPcount has stability constant 2.

- $BoundedJoin(A_1, A_2, f_1, f_2)$ stability 10

  Fix datasets A_1 and A_2, and consider it one large dataset $A = (A_1, A_2)$. Then let $B = (B_1, B_2)$ such that it is A with one record removed. Then $|A \oplus B| = 1$, so either B_1 has a record removed, or B_2 has a record removed. In the worst case choice of f_1 and f_2, the record x removed from one of the datasets B_i was in a key selection group with more than 5 elements. Then an item y that was previously trimmed from that key selection group will now be present. Therefore, there will be at most 5 records in $BoundedJoin(A_1, A_2, f_1, f_2)$ containing x that will not be present in $BoundedJoin(B_1, B_2, f_1, f_2)$, and vice-versa with records containing y. This is because the transform outputs cartesian products between trimmed key selection groups in A_1 and A_2 with the same values assigned by f_1 and f_2. Therefore, $|BoundedJoin(A_1, A_2, f_1, f_2) \oplus BoundedJoin(B_1, B_2, f_1, f_2)| = 10$, so BoundedJoin has stability constant 10.

- $BoundedJoin(A, A, f_1, f_2)$ stability 20

  Fix dataset A, and consider the dataset B that is just A with one record removed. Then $|A \oplus B| = 1$. In the worst-case choice of f_1 and f_2, the x removed in B is such that $f_1(x) \neq f_2(x)$, and under each key selection function, x appeared in a key selection group of more than 5 elements. Then two items y_1 and y_2 that were previously trimmed will take the place of x in the f_1 and f_2 groups respectively. Therefore, there will be at most 5 elements containing y_1 and 5 elements containing y_2 in $BoundedJoin(B, B, f_1, f_2)$ that are not in $BoundedJoin(A, A, f_1, f_2)$, since the transform outputs cartesian products between trimmed key selection groups under f_1 and f_2 that have the same value. Similarly, there will be 10 records containing x (5 as a first coordinate, and 5 as a second) that are in $BoundedJoin(A, A, f_1, f_2)$ but not in $BoundedJoin(B, B, f_1, f_2)$. Therefore, $|BoundedJoin(A, A, f_1, f_2) \oplus BoundedJoin(B, B, f_1, f_2)| = 20$, so BoundedJoin has stability constant 20 when given the same dataset.

Solution to Ex 7. There we two ways to solve this one. The following solution assumes that the elements in the dataset can be any real valued number:

1. To compute the privacy budget of the mechanism, I will first compute the sensitivity $\Delta = max_{r,A,B||A\oplus B|=1}(|score(A,r) - score(B,r)|)$ of the scoring function. Since $|A \oplus B| = 1$ we can assume without loss of generality that B is dataset A with one record added. Notice that for any collection of values $a_1, \ldots a_n$, we can, by adding one value b, cause the new collection to average to any $r \in \mathbb{R}$ of our choosing. Then $\forall A\ \forall x$ either $score(A,x) = 0$, when $x = Avg(A)$, or $score(A,x) = -1$, when $x \neq Avg(A)$. Therefore, for any r, A, B, $|score(A,r) - score(B,r)|$ will be 1 if $r = Avg(A)$ or $r = Avg(B)$, but not both. Therefore, $\Delta = 1$. Then by the theorem we proved in class, the mechanism has privacy budget $\epsilon\Delta = \epsilon$.

2. Since $\forall A\ \forall x\ score(A,x) = 0 \vee score(A,x) = -1$, and $score(A,x) = 0$ only when $Avg(A) = x$, running $NoisyAvg(A,\epsilon)$ will output $Avg(A)$ with probability proportional to 1, and every other value is output with probability $e^{-\frac{1}{2}}$, since $score(A,x) = -1$ in this case. Then the mechanism has the same probability of outputting $Avg(A) + 1$ as it does $Avg(A) + 100000$. While the mechanism is not very accurate, its accuracy is independent of the dataset A.

Now, if you assumed that entries of the datasets are bounded in some way, for example, suppose that they may only be integers in the range $(0, ..., 1000)$, you will find that the accuracy depends on the database itself, using an argument similar to the one we used in class for the median mechanism where the score function $score(A,r)$ was $-\min_B |A \oplus B|$ where $med(B) = r$.