Problem 1: Post-processing

Recall the example discussed in class: given a database of student names and their grades for CSCI599, we would like to make public the number of As, Bs, Cs, Ds, and Fs received while preserving the privacy of each student. Recall the $\epsilon$-differentially private algorithm derived in class (for the purpose of the differential privacy definition, we consider two databases to differ in one student’s data if the student’s data is present in both but his/her grade may differ):

On input database $D$ consisting of grades of $n$ students, for each possible grade $g$ in $\{A, B, C, D, F\}$: compute $n_g$ - the number of students in $D$ who have received grade $g$. For each possible grade $g$, output $n'_g = n_g + T$ where $T$ is noise drawn independently for each grade from the discrete Laplace distribution with scale $\frac{2}{\epsilon}$.

Recall that the discrete Laplace distribution with scale $b$ is defined as follows: $Pr[T = x] = \frac{1}{Z} \exp\left(-\frac{|x|}{b}\right)$, $x$ is an integer and $Z = \sum_{x=-\infty}^{\infty} \exp\left(-\frac{|x|}{b}\right)$ is a normalization factor which ensures that the formula above describes a probability distribution over the integers.

1. The 5-tuple of grade counts returned by the above algorithm may contain counts that are negative or exceed $n$, and their total may not add up to $n$, leading to a possible confusion if the above algorithm was applied in practice. Propose a computation on $n'_A, n'_B, n'_C, n'_D, n'_F$ that would transform it into a 5-tuple whose properties are semantically consistent with those expected of the output and prove that even after such a computation $\epsilon$-differential privacy is preserved.

2. More generally, prove that differential privacy is immune to post-processing – any computation performed on an output of a differentially private algorithm does not make it less differentially private.

3. Consider an alternative proposal for a privacy-preserving algorithm suggested by several of you in class: compute $n'_A, n'_B, n'_C, n'_D$ as above, and compute $n'_F$ using semantic consistency constraints (the constraint mentioned in class was $n'_F = n - (n'_A + n'_B + n'_C + n'_D)$, but that’s not quite sufficient...). Complete the proposal to make it into a differentially private algorithm whose output is semantically consistent with expectations, and analyze its privacy guarantee.

---

1 Assume $n$ is publicly known
4. Suppose the algorithm of 1. or 3. was to be implemented in software to be used by Registrar’s offices in universities across the US. Reason and motivate (experimentally or theoretically) as to which one would be preferable.

Problem 2: Most Popular Item

Suppose we have a database that describes whether a student likes a particular subject:

<table>
<thead>
<tr>
<th>Likes?</th>
<th>Physics</th>
<th>Math</th>
<th>CS</th>
<th>...</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>...</td>
<td>yes</td>
</tr>
<tr>
<td>Bob</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>...</td>
<td>no</td>
</tr>
<tr>
<td>Eve</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>...</td>
<td>yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

1. Propose a differentially-private algorithm for answering the question: “What is the subject that is liked by the most number of students?” For a fixed level of desired privacy, \( \epsilon \), does the number of possible subjects affect your proposed algorithm’s utility?

2. (harder) Can you eliminate the dependence on the number of subjects?

For the purpose of the differential privacy definition, we consider two databases to differ in one student’s data if the student’s data is present in one but not in the other database.