Private Data Collection using Negative Surveys
BioComputing, CSE 598
Assignment 1, Spring Semester 2019
Due: March 21, 2024 (11:59 pm)

1 Introduction

In this assignment, you will use a form of negative selection, known as a negative survey, to privately collect data about the prevalence of certain sexually transmitted diseases. Let's suppose that you wish to conduct a study of individuals in a religious community, who might be reluctant to report such sensitive information or share it with a central authority. To avoid this problem, your survey will ask a sample of individuals to list one disease that they have not been diagnosed with, out of five diseases in the study: chlamydia, gonorrhea, syphilis, and herpes. (The dataset contains three additional features, cities, latitude and longitude, which are relevant only for Part II (cities) and the extra credit (longitude and latitude).)

The negative survey provides almost the same level of protection as encryption, but it doesn't require the individual to trust a central server and it avoids the high computational cost of computing over encrypted data.

Late Policy: You are allowed three "free" late days during the semester. Once you have used up those late days, 10% of your grade will be deducted for each day that it is late.

2 Part I: One-dimensional reconstruction

The process works as follows:

1. Questionnaires are distributed to individuals, who respond with one disease they have not contracted.

2. The negative questionnaires are returned to a base station.

3. The base station estimates the histogram of occcurrences across diseases using the following formula: $\forall j | A_j = P - Y_j(\alpha - 1)$ where j is a disease, P is the number of participants, α is the number of diseases, and Y_j is the number reported js.

This is a simplified form of the process described in the Groat et al. paper assigned for class in which we consider only one class of categories, and you will implement the functionality of the base station.

We have generated a database of 1 million samples to simulate the survey responses, which can accessed using the url `http://178.16.143.126:8000`. This points to a webserver that provides an API for retrieving survey responses from the dataset. If you go directly to the URL, you can use our web interface for downloading samples of the responses. Alternatively, you can use the following get request:

1. `http://178.16.143.126:8000/get_samples?num_samples=N` where you replace N with the number of responses you want between 1 and 1000.

2. `http://178.16.143.126:8000/ground_truth` which provides the ground truth data only for individuals in the dataset with syphilis. This may be used for computing utility and privacy metrics.

The responses are in CSV format with the headers: `negative_disease`, `negative_city`, `negative_lat`, and `negative_lon`. You will request a sample of such answers from the server and use them to reconstruct an estimate of the prevalence of each disease. This will be reported as a histogram with the different diseases listed on the x-axis and on the y-axis you will report a count of the sampled students you estimate have that particular disease.

3 Part II: Two-dimensional reconstruction

Next we will consider the case of two-dimensional data, asking whether or not the individual has contracted herpes (herpes simplex virus or HSV) and what city they live in. Your task is to estimate the prevalence of herpes in each city. The first dimension is the herpes field and the second dimension is the cities field. There are five different cities, each city indicated by an integer in $(1, 5)$. To calculate the prevalence of herpes infections in each city, use the following formula, where $R(u, v)$ is the number of reports that contain category u of dimension 1 and category v of dimension 2:

$$A(k_1, k_2) = P - (m_1 - 1)\sum_{y=1}^{m2} R(k_1, y) - (m_2 - 1)\sum_{x=1}^{m_1} R(x, k_2) + (m_1 - 1)(m_2 - 1)R(k_1, k_2)$$

To clarify, dimension 1 will have two categories: Does the student report not having that disease (1) or not report (0), and dimension 2 will have five categories, one for each city.

4 Extra Credit: Use latitude and longitude

The above approach is limited because the the formula for calculating the negative image over multiple dimensions scales exponentially with the number of dimensions. A more efficient approach is described in the paper "Enhancing privacy in participatory sensing applications with multidimensional data." Using that approach, use the longitude and latitude fields to make a map for each disease showing its prevalence by longitude and latitude. One approach to this problem is to bin the longitude and latitude into, for instance, 25 categories creating a 5x5 map and then treat this as a multidimensional negative survey with the three dimensions: disease, latitude bin, and longitude bin.

5 What to report?

For Part I, please report the following:

1. A histogram showing your estimate (and a confidence interval) of the prevalence of each disease.

2. Conduct an experiment using different numbers of responses and report both the computation time it takes and how the confidence in your result changes as you increase the number of

samples (to get more than 1000 samples, simply make multiple queries to the database. Note: It is unlikely that you will need to sample the database very many times. Consider that if you took 1000 samples of size 1000, you would have gotten a million responses, which is the population size of the survey. Report this in a separate table or graph.

3. Compute the privacy and utility values for syphilis using the method described in the paper "Enhancing privacy in participatory sensing applications with multidimensional data." For this disease, we provide ground truth `http://178.16.143.126:8000/ground_truth` for three different sample sizes.

For Part II, make a histogram showing your estimate of the prevalence of herpes in each different city.

For Part III, make a two-dimensional map, showing the prevalence of each disease by geographic location.

6 Details

6.1 Reporting results

Please hand in a 4-5 page report using the ACM format, written in appropriate academic style with proper citations. Please print a copy of your assignment paper and bring it to class on March 21 and submit a .pdf of the writeup and a .zip file with your code, instructions for running it through Canvas. Your Canvas submission time will document when you completed the assignment. If you miss class on the due date, then bring a printed copy the next time you come to class. Your paper should describe the following:

1. The problem you are trying to solve (in your own words) and a brief description of how you solved it;

2. The basic parameters for your runs (sample sizes you used);

3. Experimental results, as detailed above;

4. Discussion of your results, how well your project worked, and what you learned.