Private Data Collection Using Negative Surveys

WeiSheng, Chiu
School of Computing Augmented Intelligence
Arizona State University
Tempe, AZ, United States of America
wchiu6@asu.edu

ABSTRACT

When we do a survey, we must fill in our personal information. If the host leaks surveys, we have no chance to stop our personal information from being on the market. To prevent this, the paper *Enhancing Privacy in Participatory Sensing Applications with Multidimensional Data[1]* proposed that we can do negative surveys and reconstruct the original distribution from disguised data in the base station. This paper will implement the paper on estimating the prevalence of diseases in five cities in Arizona. This survey collects one million responses. I will answer two questions in this paper from those surveys. The first one is the prevalence of sexual diseases in this community. The second one is which city has the most diagnosed with herpes. In addition, I will analyze this method by calculating utility and privacy metrics.

1 Introduction

When doing surveys on some topic, we must first collect the data, carry it to a base station, analyze it, and store it. There is a high chance of data leakage during this process. For example, some people see the surveys during transportation to a base station, or a hacker hacks into the base station's database and releases it. Those may leak the participants' personal information. Therefore, fewer people will be willing to participate in a survey due to the privacy issue.

There is a way to solve it. The survey carrier could encrypt it before carrying it to the base station. This works, but it costs computing resources to encrypt the survey. Therefore, we must find a way to disguise the survey's answer with fewer computing resources. This is the situation in which negative surveys stand out.

Negative surveys do not ask participants to answer the question correctly; instead, they ask participants to choose the option that is not correct. Then, the base station reconstructs it and gets the original distribution. This prevents the consequence of the data leak since the survey content is not the participant's answer. The hacker could not extract any personal information from the survey.

In this paper, I am going to reconstruct the disguised survey result. This survey investigates the prevalence of sexual disease in a conservative community. The population of the surveys is one million. In part one, I will extract the prevalence of sexual disease in this community. In the second part, I will estimate the population of individuals diagnosed with herpes in each of the five cities.

2 Part 1

After all surveys are collected, they are returned to the base station. The disguised surveys would be reconstructed in the base station to disclose the original data distribution. In this section, I will reveal the prevalent disease in this community from those negative surveys. Formula (1) would be used to calculate the original data distribution. I used one million negative surveys to reconstruct the original data and analyze the results. In addition, I show my discovery when changing the probability of selecting a perturbed category.

$$\forall i | A_i = N - (\alpha - 1)Y_i \quad (1)$$

2.1 Estimating the Prevalence of Disease

I reconstructed one million collected surveys to estimate the prevalence of the disease. Table 1 shows the result and its 99% confidence interval. From the table, 49.82% ± 0.16% of individuals in this conservative community are healthy. Besides those individuals, the prevalent disease is herpes, which is 20.17% ± 0.13% of the individuals. Chart 1 shows the estimation of disease prevalence in the individual in a histogram.

Table 1. The result of disease estimation. The confidence interval is calculated under a 99% confidence level.

	Disguised Data	Reconstructed Data	Rate	Confidence Interval
Chlamydia	183677	81615	8.16%	8.07% ~ 8.24%
Gonorrhoea	186660	66700	6.67%	6.59% ~ 6.74%
Syphilis	180381	98095	9.81%	9.72% ~ 9.90%
Herpes	159655	201725	20.17%	20.05% ~ 20.30%
HIV	189268	53660	5.37%	5.30% ~ 5.44%
Healthy	100359	498205	49.82%	49.66% ~ 49.98%

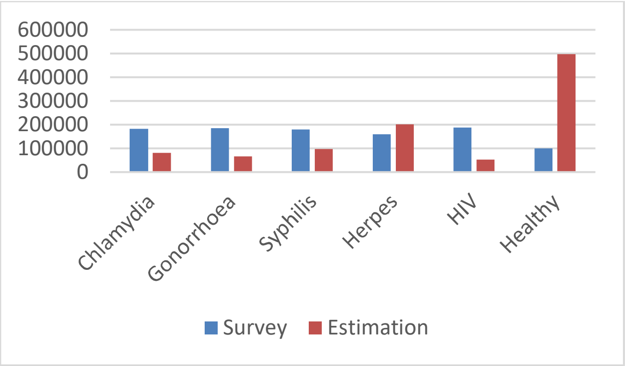

Chart 1 This histogram shows the number of people in each category. The x-axis represents disease. The y-axis is the population.

We can observe from Chart 1 that the result of my estimation is the opposite of the result of surveys. For example, in the HIV category, the number of individuals who have HIV is the highest in the results of surveys, but it is the lowest one in the results of my estimation. In the healthy category, the number of healthy individuals is less than others in the results of surveys, but it occupies almost half of the population in my estimation. This is reasonable because the negative survey makes participants select a disease they have not been diagnosed with, which means the more significant the disease population in the survey, the fewer individuals diagnosed with it.

2.2 Experient: How the number of surveys affect confidence interval

I used a 99% confidence level for this project to calculate the confidence interval. To observe the relationship between the number of surveys and the confidence interval, I plot the difference of an interval every 10,000 records for every disease. Chart 2 is the result for all six categories. As the number of surveys increases, the difference in intervals gets lower. This phenomenon matched that if the number of collected surveys is large enough, it could represent the population more. Chart 3 describes the execution time for each reconstruction of negative surveys.

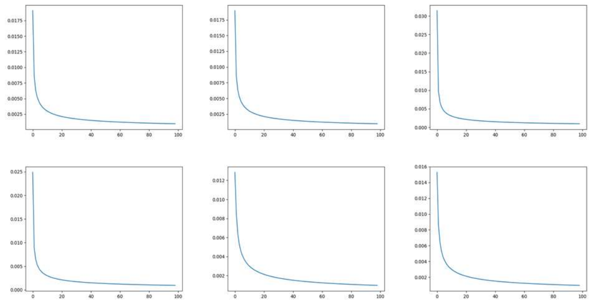

Chart 2. Plot the difference of an interval every 10,000 records for each disease. In top-to-down and left-to-right order, there are chlamydia, gonorrhoea, healthy, herpes, HIV, and syphilis.

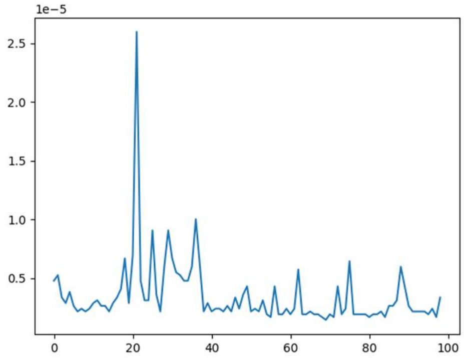

Chart 3. It shows the execution time of the reconstruction of negative surveys. The average is 3.45e-6 seconds. The maximum is 2.6e-5 seconds. The minimum is 1.43e-6 seconds.

2.3 Change the probability of selecting a perturbed category

From Formula 1, every disguised value of a disease times $\alpha - 1$ is because the paper assumes the probability of selecting a perturbed category is equal. Therefore, I tried to modify the possibility. I presume the first-time reconstruction is correct, so I use this to simulate a new result of negative surveys. Then, reconstruct it and compare it with the result of first-time reconstruction. The way I reconstruct it is that I time every disguised value of a disease to the reciprocal of its probability of being selected. However, I always got negative values in the healthy and the herpes categories. At first, I thought that it was because the number of the surveys was not large enough. After thoroughly digging into the paper, I found that my formula was too naïve. Since different categories would have different probabilities of choosing the same disguised category, I need to calculate the proportion of origin categories in the disguised category and times the reciprocal of the probability during reconstruction. The tricky thing is that I do not know the proportion of origin categories in the disguised category.

However, I discovered something interesting while tuning the probability. I set all the probability to 1 and simulated how every individual does when doing the survey. The individual's decision is based on the result of first-time reconstruction. Then, I execute Formula 1 again with the new result of the negative surveys. It turns out that I got a similar distribution compared to the first-time reconstruction. I showed the result in Table 2. This result means that Formula 1 and my implementation is correct.

Table 2. The comparison between first-time and second-time reconstruction and its difference. It shows that second-time reconstruction could derive a similar distribution to first-time reconstruction.

	First-time reconstruction	Second-time reconstruction	Variant
Chlamydia	8.16%	8.19%	+0.03%
Gonorrhoea	6.67%	6.38%	-0.29%
Syphilis	9.81%	10.07%	+0.26%
Herpes	20.17%	20.43%	+20.26%
HIV	5.37%	5.31%	-0.06%
Healthy	49.82%	49.60%	-0.22%

2.4 Utility and privacy metric

Utility and privacy metrics are two indicators that reveal the accuracy and security of this method. The privacy metric shows how well these negative surveys prevent the leak of original data to attackers. Formula 2 does this mission for one-dimensional surveys. The α is the number of categories. The privacy metric for this project is 0.5307. The utility metric displays the difference between the distribution of the original data and the reconstructed data's distribution, so the smaller, the better. Formula 3 shows it. The α is the number of categories, and the N is the number of participants. The utility metric for this disease prevalence investigation is 4e-6.

$$Privacy_{model} = \frac{2.5}{(\ln(\alpha))^2+1.5} \qquad (2)$$

$$Utility_{model} = \frac{\alpha-2}{N} \qquad (3)$$

3 Part 2

In part 2, I will extract the information that the disease herpes is prevalent in which city from disguised data. I will use Formula 4[1] to calculate for each city. I used one million surveys. I calculated the confidence interval with a 99% confidence level.

$$\forall \vec{x} \mid A(\vec{x}) = N + \sum_{k=1}^{D} (-1)^k \cdot \Gamma(\vec{x}, k), \qquad (4)$$

where $\Gamma(\vec{x}, k)$ is given as:

$$\Gamma(\vec{x}, k) = \sum_{\substack{d \in \\ B(\{1,\ldots,D\},k)}} \left(\left[\prod_{j \in d} (\alpha_j - 1) \right] \cdot \sum_{\substack{\vec{y} \text{ s.t.} \\ y_i \in \vec{x}, \\ \forall i \in d}} Y(\vec{y}) \right), \qquad (5)$$

3.1 Estimating the Prevalence of Disease

Table 3 shows my estimation. We can find that Tempe has the highest population diagnosed with herpes, which is 5.00% ± 0.07%. Chart 4 shows the prevalence in the histogram. If we sum up the column reconstructed data in Table 3, we would get 201725, the number of individuals diagnosed with herpes. This value is the same as my estimation in part 1.

Table 3. The prevalence of herpes in five cities in Arizona. The confidence interval is calculated under a 99% confidence level. The rate represents the proportion of individuals who satisfy the condition in the population, which is one million.

	Disguised Data	Reconstructed Data	Rate	Confidence Interval
Mesa	35265	43773	4.38 %	4.31% ~ 4.44%
Tempe	32366	50025	5.00%	4.93% ~ 5.07%
Scottsdale	35536	49545	4.95%	4.89% ~ 5.02%
Phoenix	23964	24777	2.48%	2.43% ~ 2.53%
Chandler	32524	33605	3.36%	3.30% ~ 3.42%

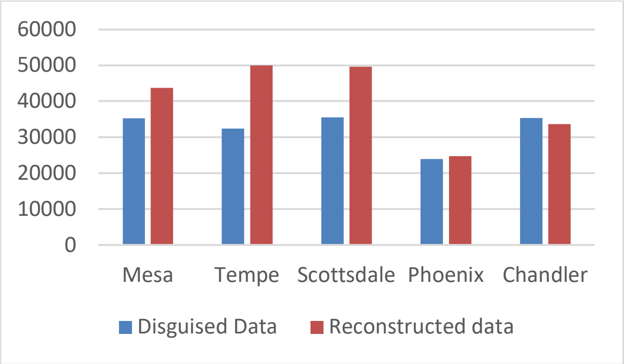

Chart 4. This histogram shows the location of the population diagnosed with herpes. The x-axis represents cities. The y-axis is the population.

ACKNOWLEDGMENTS

In this paper, I reconstruct the negative surveys and derive the distribution of the original data. In part one, I estimate the prevalence of the disease is herpes. The formula used is more straightforward in this part since there is one-dimensional data. In addition, I tried to change the probability of perturbed category selection. Even though I failed at it, I discovered something interesting. I constructed another negative survey using the same distribution derived from the reconstruction. Then, I reconstruct the new negative surveys and get a similar distribution as the first reconstruction. In part two, I will describe the population of individuals diagnosed with herpes in each of the five cities. The city of Tempe has the highest population.

REFERENCES

[1] Michael M. Groat, Benjamin Edwards, James Horey, Wenbo He, and Stephanie Forrest, 2012. Enhancing Privacy in Participatory Sensing Applications with Multidimensional Data. *2012 IEEE International Conference on Pervasive Computing and Communications, Lugano (19-23 March 2012)*