Negative Survey with Manual Selection: A Case Study in Chinese Universities

Jianguo Wu[1], Jianwen Xiang[1], Dongdong Zhao[1*], Huanhuan Li[2], Qing Xie[1], and Xiaoyi Hu[1]

[1] School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China
[2] School of Computer Science, China University of Geosciences, Wuhan, China
{jgwu, jwxiang, zdd, felixxq, huxiaoyi}@whut.edu.cn, julylhh@gmail.com

Abstract. Negative survey is a promising method which can protect personal privacy while collecting sensitive data. Most of previous works focus on negative survey models with specific hypothesis, e.g., the probability of selecting negative categories follows the uniform distribution or Gaussian distribution. Moreover, as far as we know, negative survey is never conducted with manual selection in real world. In this paper, we carry out such a negative survey and find that the survey may not follow the previous hypothesis. And existing reconstruction methods like NStoPS and NStoPS-I perform poorly on the survey data. Therefore, we propose a method called NStoPS-MLE, which is based on the maximum likelihood estimation, for reconstructing useful information from the collected data. This method also uses background knowledge to enhance its performance. Experimental results show that our method can get more accurate aggregated results than previous methods.

Keywords: Privacy Protection, Negative Survey, Reconstruction Method.

1 Introduction

Nowadays, the rapid development of computer network and big data technologies brings great convenience to people, but it also increases the risk of disclosing sensitive data and personal privacy. Negative Survey [1, 2] is a promising privacy protection technique. In a negative survey, participants are asked to answer a question by selecting a category that they do NOT belong to (this kind of category is called negative category). When the number of categories in a question is larger than 2, the privacy of the participants can be protected because attackers cannot determine the real answer of a participant. After collecting negative survey results, statistical results about population distribution over different categories could be reconstructed by several methods.

Previous works about negative survey mainly focus on models with specific hypotheses, e.g., the probability that participants select negative categories follows the uniform distribution or Gaussian distribution. These models could be reasonable when negative categories are selected by electronic devices instead of humans. However, in

* Corresponding author: Dongdong Zhao

some applications with high security requirements, participants need/want to manually select a negative category as the answer, and they do not want to use electronic devices because the security cannot be guaranteed. Therefore, we conduct a negative survey in Wuhan University of Technology and China University of Geosciences, and the answers are manually selected by participants. Based on the survey results, we have several findings (as shown in section 3.3). Moreover, we propose a method called NStoPS-MLE to reconstruct useful aggregated results. To enhance the performance, the proposed method uses background knowledge about the overall probabilities of selecting negative categories. Experimental results show that NStoPS-MLE performs better than NStoPS [1, 2] and NStoPS-I [5] on most of the questions.

2 Related Work

Negative survey is first proposed by Esponda [1, 2] in 2006. For example, in a positive survey, the question is designed as follow:

What is the rank of your score in your class:

A. 1-5 B. 6-15 C. 16-25 D. >=26?

In negative survey, this question is designed as follow:

Which is **NOT** the rank of your score in your class:

A. 1-5 B. 6-15 C. 16-25 D. >=26?

If the rank of Alice's score is 3, in positive survey, she should select A. But in negative survey, she should select one answer among B, C and D at random.

Generally, assume that the number of categories in a question is c, the number of participants is n, and Q is the reconstructed matrix composed by q_{ij}, where q_{ij} denotes the probability that a participant, who actually belongs to the i^{th} category, selects the j^{th} category as negative category. The statistical results collected from negative survey is $\boldsymbol{r} = (r_1 \ldots r_c)$, where r_i is the number of participants that select the i^{th} category as negative category. Our goal is to reconstruct aggregated results $\boldsymbol{t} = (t_1 \ldots t_c)$ from negative survey results, where t_i denotes the number of participants that actually belong to the i^{th} category. A theoretical model called NStoPS for reconstructing $\boldsymbol{t}$ is: $\boldsymbol{t} = \boldsymbol{r}Q^{-1}$ [1, 2].

Presently, there are some researches about negative survey. Typically, Bao et al. pointed out in [5] that NStoPS would produce unreasonable negative values, and they proposed two algorithms called NStoPS-I and NStoPS-II to handle negative values. Xie et al. [4] proposed Gaussian negative survey, in which the probability that participants select negative categories follows Gaussian distribution. Zhao et al. [6] suggested to use background knowledge in reconstructing useful information from negative survey results. Recently, Esponda et al. [3] proposed a personalized negative survey model, which could meet different privacy requirements from users. Negative survey has been applied to several scenarios. For example, in [7], Horey et al. employed negative survey for collecting anonymous data in sensor networks. In 2012, Horey et al. [8] used negative survey in collecting the location information of users. In [9], Liu et al. applied negative survey to the privacy protection of cloud data. Overall, previous researches (e.g., [1, 2, 4-6]) mainly focus on uniform negative survey or Gaussian negative survey, in which the probability that participants select negative categories is assumed to follow

the uniform distribution or Gaussian distribution. In their experiments, they used electronic devices to simulate the negative selection.

3 Overview of the Survey

3.1 Survey Goal and Questionnaire Design

The main goal of our work is carrying out a realistic negative survey and finding its characteristics. We conducted a survey in two universities. Our questionnaire has three parts. The first part is an anonymous positive survey, we assume the statistical results from this part is close to the truth, and we evaluate the reconstructed results from negative survey based on the results from this part. The second part is a real-name negative survey, and the third part is a real-name positive survey. The third part is simply used to construct some background knowledge about the probabilities of negative selection, and the background knowledge will be used in reconstructing results from negative survey. The respondents can answer all surveys or part of them.

Our questions are sensitive issues about university students. Each part has 15 questions, but the questions in the negative survey are designed in a different form. Several examples of questions are listed as follows: “how often do you skip class”, “what’s the rank of your scores in your class”, “how often do you watch xanthic films”. In our questionnaire, four questions have 3 categories, six questions have 4 categories, and five questions have 5 categories. To avoid that the order of categories would affect the choice of the respondents, we rearrange the order of categories in the second survey. Because we finally analyze the collected data based on the content of each category, for a convenience, we use “category A, B, C, D, E” in part 2 the same as in part 1 and part 3 in the rest of this paper.

3.2 Data Statistics

We collect data by surveys online and offline. For online surveys, we program the survey website, and participants are guided to the anonymous positive survey, real-name negative survey and real-name positive survey in turn. The answers of participants are automatically stored to the server database. For offline surveys, first, we conduct the anonymous positive survey, and then, we conduct the two real-name surveys. In the end, we collect 811 valid records (corresponds to 811 respondents) from the anonymous positive survey, 550 valid records from the real-name negative survey, and 528 valid records from the real-name positive survey.

The statistical results of each category for anonymous positive survey and real-name negative survey are shown in table 1. All these results are rounded to one decimal place. The statistical results of the real-name positive survey are not presented because we just use it to get Q in reconstruction.

3.3 Survey Findings

Based on the collected survey data, we have the following findings:

(1) The probability that participants select negative categories might not follow the

uniform distribution or Gaussian distribution. We extract all valid records that have the same identity (i.e., name) in the real-name negative survey and positive survey, and we compare the answers of each participant in the two surveys. Finally, we can obtain a matrix Q for each question, and $Q(i, j)$ denotes the percentage of participants who select the j^{th} category in real-name negative survey among those participants who select the i^{th} category in real-name positive survey. The $Q(i, j)$ could represent (at least approximate to) the probability that participants, who actually belong to the i^{th} category, select the j^{th} category in negative survey. We find that $Q(i, j)$ might not follow the uniform or Gaussian distribution, for example, as shown in fig. 1, the distributions of $Q(1, 1)$~$Q(1, 4)$ and $Q(2, 1)$~$Q(2, 4)$ are neither uniform nor Gaussian distribution.

Table 1. Statistical results of the anonymous positive survey and real-name negative survey.

	Anonymous positive survey					**Real-name negative survey**				
	A	B	C	D	E	A	B	C	D	E
1	45.5	49.3	3.2	2.0		12.9	8.5	40.4	38.2	
2	80.6	13.6	3.8	2.0		10.5	14.0	15.3	60.2	
3	23.1	33.3	31.6	12.1		35.5	8.5	15.5	40.5	
4	76.6	19.6	2.5	1.4		9.1	10.2	17.5	63.3	
5	13.4	37.1	39.8	5.3	4.3	18.7	10.5	5.3	25.8	39.6
6	32.3	42.8	18.7	3.1	3.1	17.1	3.6	12.5	18.9	47.8
7	29.6	55.0	11.0	4.4		16.9	5.8	13.8	63.5	
8	40.0	54.6	5.4			24.2	15.5	60.4		
9	91.9	6.3	1.8			9.8	25.6	64.5		
10	15.4	37.5	33.3	8.5	5.3	25.1	8.2	9.8	16.0	40.9
11	94.8	4.1	1.1			8.1	22.0	69.8		
12	94.8	3.8	1.3			6.9	48.5	44.5		
13	8.9	56.5	28.2	3.5	3.0	20.5	7.3	8.7	30.4	33.1
14	6.0	18.4	29.6	38.1	7.9	43.8	5.6	7.1	22.0	21.3
15	5.8	4.6	40.8	48.6		38.2	38.0	6.7	17.1	

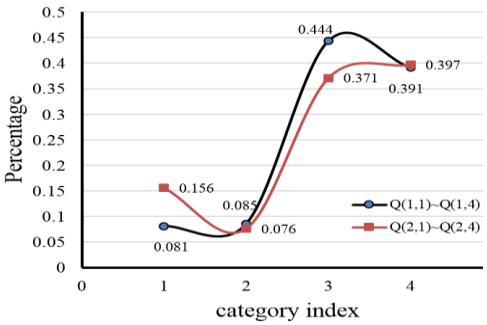

Fig. 1. The distributions of Q(1, 1)~Q(1, 4) and Q(2, 1)~Q(2, 4).

Moreover, we find participants prefer to select negative categories that have extreme values. For example, among the participants who select category A in the 3rd question (see the example in section 2) in real-name positive survey, about 57% of them select

D as negative category in negative survey while D has an extreme value. The percentages of selecting B and C are about 14% and 18%, respectively. Note that there are usually about 35 students in the classes where we conduct surveys.

(2) Typical reconstruction methods (i.e., NStoPS and NStoPS-I) perform poorly on the collected data. As shown in tables 3-5, the accuracy of NStoPS and NStoPS-I on several questions is low. For example, the errors of the results reconstructed by NStoPS on the 2~7th questions are larger than 0.60. The errors of the results reconstructed by NStoPS-I on the 2nd, 3rd, 6th, 14th questions are larger than 0.40, and especially, the error is about 0.78 for the 14th question. Those results are almost useless.

(3) Reconstructed results might not be better on the questions with less categories. For example, the result on the 13th question (with 5 categories) is better than several results on the questions with 3 categories or 4 categories (as shown in table 6).

(4) There might be more unreasonable answers in negative surveys when participants manually select negative categories. For example, we find some participants select the same option for all questions and some participants write fake names in real-name surveys. We regard these records as dirty data and remove them when reconstructing. Moreover, there are some non-negative values in Q matrix for most of the questions (e.g., the 1st, 2nd, 3rd questions, see table 2.) The method of getting Q is showed in section 4.1. It indicates that some participants might have not followed the rule of negative selection, i.e., they have selected a category they really belong to in negative survey, therefore $Q(i, i)$ in some matrices are not 0. Furthermore, we find that the results reconstructed by the theoretical model (i.e., NStoPS) contain unreasonable negative values for most of the questions.

Table 2. Matrices from the samples from the real-name negative and positive survey.

	1st Question	**2nd Question**	**3rd Question**
Q	$\begin{bmatrix} 0.13 & 0.09 & 0.41 & 0.37 \\ 0.11 & 0.1 & 0.3 & 0.49 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \end{bmatrix}$	$\begin{bmatrix} 0.07 & 0.12 & 0.17 & 0.64 \\ 0.24 & 0.08 & 0.04 & 0.64 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \end{bmatrix}$	$\begin{bmatrix} 0.11 & 0.12 & 0.18 & 0.59 \\ 0.25 & 0.07 & 0.2 & 0.48 \\ 0.53 & 0.11 & 0.06 & 0.3 \\ 0.7 & 0.06 & 0.12 & 0.12 \end{bmatrix}$

4 Reconstruction Algorithm

In this section, a method called NStoPS-MLE is proposed for reconstructing useful aggregated results from negative data.

4.1 Using Background Knowledge

In real world, we usually have some background knowledge. Therefore, we try to use part of Q as background knowledge to improve the accuracy of reconstructed results.

As we have conducted a real-name positive survey, we collect part information about Q by randomly sampling a number of (e.g., 100 or 50) participants and comparing their answers in the real-name negative survey and the real-name positive survey. For example, for the 1st question, $Q(1, 2)$ is set to the percentage of participants who select B in the negative survey among the participants who select A in the real-name positive survey. Except the part of Q we obtain, the remaining part is set as that in uniform negative

surveys, i.e., if $i = j$, then $Q(i, j) = 0$; otherwise, $Q(i, j) = 1/(c-1)$.

4.2 NStoPS-MLE

The probability that a participant actually belongs to the i^{th} category and selects the j^{th} category in negative survey is $\frac{t_i}{n} \times q_{ij}$. Consequently, the probability that a participant selects the j^{th} category in negative survey is:

$$p_j = \sum_{i=1}^{c} \frac{t_i}{n} \times q_{ij}. \quad (1)$$

Let $\boldsymbol{p} = (p_1 \ldots p_c)$, and in an event of the negative selection on the question in negative survey: the probabilities that the $1^{st} \ldots c^{th}$ category is selected as a negative category are $p_1 \ldots p_c$, respectively. The negative selection event happens n times, and the probability that the $1^{st} \ldots c^{th}$ categories are selected as negative categories $r_1 \ldots r_c$ times respectively, can be calculated as:

$$\Pr(\boldsymbol{r}|\boldsymbol{p}) = \frac{n!}{r_1! \times \ldots \times r_c!} p_1^{r_1} \times \ldots \times p_c^{r_c}. \quad (2)$$

It subjects to multinomial distribution. When reconstructing, we have the observed results $\boldsymbol{r} = (r_1 \ldots r_c)$ but $\boldsymbol{p} = (p_1 \ldots p_c)$ remains unknown because $\boldsymbol{t}$ is unknown. The reconstruction can be formalized as:

$$\hat{\boldsymbol{p}}_{mle} = \arg\max_{\boldsymbol{p} \in P} \left\{ \frac{n!}{r_1! \times \ldots \times r_c!} p_1^{r_1} \times \ldots \times p_c^{r_c} \right\}. \quad (3)$$

Where P contains all feasible values of $\boldsymbol{p}$. Because when $\boldsymbol{t}$ is known, $\boldsymbol{p}$ can be calculated from $\boldsymbol{t}$, we have $\Pr(\boldsymbol{r}|\boldsymbol{t}) = \Pr(\boldsymbol{r}|\boldsymbol{p})$ and (3) can be converted to:

$$\begin{aligned} \hat{\boldsymbol{t}}_{mle} &= \arg\max_{\boldsymbol{t} \in T} \left\{ \frac{n!}{r_1! \times \ldots \times r_c!} \prod_{i=1}^{c} \left(\sum_{j=1}^{c} \frac{t_j}{n} \times q_{ji} \right)^{r_i} \right\} \\ &= \arg\max_{\boldsymbol{t} \in T} \left\{ \sum_{i=1}^{c} r_i \times \log\left(\sum_{j=1}^{c} t_j \times q_{ji} \right) \right\}. \end{aligned} \quad (4)$$

Where T contains all feasible values of $\boldsymbol{t}$. According to the definition of $\boldsymbol{t}$, it has the following constrains: $\sum_{i=1}^{c} t_i = n$ and $0 \le t_i \le n$. By the constrains, unreasonable negative values can be avoided in reconstruction.

The steps of NStoPS-MLE are shown as follows. Firstly, we counts the total number of participants by $n = r_1 + r_2 + \cdots + r_c$. Next, we revise the unreasonable values of q_{ii} in Q to 0, and scale the other values at the same row by $q_{ij} = q_{ij} / \sum_{j=1 \ldots c, j \neq i} q_{ij}$. Then we solves (4) with constrains $\sum_{i=1}^{c} t_i = n$ and $0 \le t_i \le n$. Finally, we can get $\hat{\boldsymbol{t}}$ out. Note that, there are many methods can efficiently solve (4) with constrains, like the interior point algorithm. When $c<4$ and $n<1000$, it is feasible in practice to enumerate every possible assignments of $\boldsymbol{t}$ to find the best one according to (4). In our experiments, we solve (4) by the built-in function called fmincon in Matlab.

5 Experimental Results

In this section, we carry out several experiments on reconstructing aggregated results by NStoPS-MLE, and compare it with NStoPS and NStoPS-I.

For each question, we make a list of categories, for which we will collect background knowledge. For each listed category, we randomly select 100 or 50 participants from those who finished the real-name negative survey and positive survey. Table 3 shows

the number of the sampled participants for each category in each question. Next, we collect the values in Q from the records of the sampled participants. The remaining part of Q is set according to that of uniform negative surveys. Finally, using the Q, we reconstruct aggregated results $\hat{\boldsymbol{t}}$ from negative survey results, and we evaluate $\hat{\boldsymbol{t}}$ by the error formula $\frac{1}{n}\sqrt{\sum_{i=1}^{c}(\hat{t}_i - t_i)^2}$ [5].

Table 3. The number of different categories we sample.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
A	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0
B	100	50	100	100	100	100	100	100	0	100	0	0	100	100	0
C	0	0	100	0	100	50	0	0	0	100	0	0	100	100	100
D	0	0	50	0	0	0	0			0			0	100	100
E					0	0				0			0	0	

We carry out the above experiment 30 times for each question, and the average value of error is presented in table 4, 5 and 6. Note that, NStoPS and NStoPS-I are executed with the matrix Q for uniform negative surveys. Table 4, 5, 6 shows the results for the questions which have three, four and five categories respectively.

Table 4. *Errors* for the questions which have 3 categories.

	8	**9**	**11**	**12**
NStoPS	0.32089732	0.537624429	0.668872531	0.13349797
NStoPS-I	**0.083677434**	0.290930698	0.308991642	**0.129740317**
NStoPS-MLE	0.117258544	**0.163318938**	**0.155915267**	0.131531038

Table 5. *Errors* for the questions which have 4 categories.

	1	**2**	**3**	**4**	**7**	**15**
NStoPS	0.417023068	1.071013116	0.646269694	1.133704834	1.113415734	0.477458566
NStoPS-I	**0.126676882**	0.421710766	0.40721515	0.362863376	0.292960686	0.37690809
NStoPS-MLE	0.164332104	**0.261843698**	**0.238704758**	**0.185308482**	**0.241277358**	**0.23388729**

Table 6. *Errors* for the questions which have 5 categories.

	5	**6**	**10**	**13**	**14**
NStoPS	0.782076105	1.101897433	0.860045928	0.592824026	1.122504328
NStoPS-I	0.309181876	0.477232203	0.236083239	0.19853143	0.782297744
NStoPS-MLE	**0.229283031**	**0.275135786**	**0.180742101**	**0.155891588**	**0.231061221**

As shown in table 4, 5 and 6, NStoPS-MLE performs better than NStoPS over all questions, and it performs better than NStoPS-I over most questions. NStoPS-MLE performs worse than NStoPS-I only in the 1st, 8th and 12th questions, because NStoPS-I has already obtained very good results. Note that the effectiveness of the background knowledge about Q is related to the accuracy of the results in the real-name positive survey. However, the "real-name" rule may induce inexact background knowledge.

The results of NStoPS-MLE seem more stable than that of NStoPS-I, and all *error*s for NStoPS-MLE are less than 0.276, but the *error*s of NStoPS-I in the 2nd, 3rd, 6th and 14th questions are larger than 0.40. Specifically, NStoPS-I has an *error* larger than 0.78 on the 14th question, and that makes its result almost useless.

6 Conclusion and Future Work

In this paper, we present and analyze a real-world negative survey and obtain several findings. Existing reconstruction methods like NStoPS and NStoPS-I perform poorly on the data for several questions. Thus, we propose a method called NStoPS-MLE to effectively reconstruct aggregated results from negative survey results. Experimental results show that NStoPS-MLE using background knowledge performs better than NStoPS and NStoPS-I over most of questions. The method in this paper could be used in real-world negative survey.

In future work, we will try to carry out several surveys to investigate the influence of the privacy degree of different categories on the utility of the collected data. And we will try to use other background knowledge to enhance NStoPS-MLE.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (No. 61672398), the Key Natural Science Foundation of Hubei Province of China (No. 2015CFA069), the Applied Fundamental Research of Wuhan (No. 20160101010004), and the Fundamental Research Funds for the Central Universities (No. 173110002).

References

1. Esponda, F.: Negative surveys. arXiv:math/0608176 (2006)
2. Esponda, F., Guerrero, V.M.: Surveys with negative questions for sensitive items. Stat. Probab. Lett. 79, 2456-2461 (2009)
3. Esponda, F., Kael H., Victor M. G.: A Statistical approach to provide individualized privacy for surveys. PloS one 11.1: e0147314 (2016)
4. Xie, H., Kulik, L., Tanin, E.: Privacy-aware collection of aggregate spatial data. Data Knowl. Eng. 70, 576-595 (2011)
5. Bao, Y., Luo, W., Zhang, X.: Estimating positive surveys from negative surveys. Stat. Probab. Lett. 83, 551-558 (2013)
6. Zhao, D., Luo, W., Yue, L.: Reconstructing positive surveys from negative surveys with background knowledge. In: The 2016 International Conference on Data Mining and Big Data, 86-99 (2016)
7. Horey, J., Groat, M., Forrest, S., Esponda, F.: Anonymous data collection in sensor networks. In: The Fourth Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 1-8 (2007)
8. Horey, J., Forrest, S., Groat, M.M.: Reconstructing spatial distributions from anonymized locations. In: The 28th International Workshop on Data Engineering, 243-250 (2012)
9. Liu, R., Tang, S.: Negative survey-based privacy protection of cloud data. In: The 2015 International Conference in Swarm Intelligence, 151-159 (2015)