CERIAS Tech Report 2001-14

APPLICATION OF RANDOMIZED RESPONSE STRATEGY IN PRIVACY-PRESERVING SURVEY (EXTENDED ABSTRACT)

Wenliang Du and Rajeev Gopalakrishna

Center for Education and Research in
Information Assurance and Security,
Purdue University, West Lafayette, IN 47907

Application of Randomized Response Strategy in Privacy-Preserving Survey (Extended Abstract)

Wenliang Du
CERIAS and
Department of Computer Sciences
Purdue University
1315 Recitation Building
West Lafayette, IN 47907-1315
Email: duw@cs.purdue.edu
Tel: (765) 496-6765
Fax: (765) 496-3181

Rajeev Gopalakrishna
CERIAS and
Department of Computer Sciences
Purdue University
1315 Recitation Building
West Lafayette, IN 47907-1315
Email: rgk@cs.purdue.edu
Tel: (765) 496-6765
Fax: (765) 496-3181

1 Purpose and Related Work

The proliferation of the Internet has not only made computer systems widely accessible but also open to attacks and intrusions. Intrusion detection has therefore become an important area of research.

As important as detecting intrusions may be, dissemination of information related to intrusions that have occurred is equally important. This helps others gear up and patch the vulnerabilities that resulted in those intrusions. In such a scenario, a survey of the vulnerabilities exploited and the number of intrusions experienced will give us valuable information about the trends in exploits and help direct effort towards preventing them.

Consider a large group of companies. Over a period of time every company will detect a certain number of vulnerabilities in the various pieces of software being used in the company. In order to find out how rampant those vulnerabilities are being exploited in the corporate world and direct concerted effort in patching them up, the companies might be interested in a survey of the vulnerabilities exploited in various companies and the number of exploits using them. All the companies would be interested in knowing the total number of exploits using a certain vulnerability detected in all the companies taking part in the survey. But none of the companies would want to reveal the actual number of exploits in their company as they might feel that this will hurt their image and that their customers might lose faith in dealing with them. Furthermore, they might be paranoid that this information could be used by malicious people to break into their system in the future.

In general, in the type of the surveys described above, the interviewer needs to ask interviewees sensitive questions whose answers are supposed to be confidential information. For example, questions like the following might be asked: "how many security break-in's does your company have in the last month", "please choose from the following the most common successful attacks your company is subject to, ...", and "please tell us the numbers of machines in your company that are running Windows NT, Linux, and Solaris, respectively". Companies want to keep the answer to these questions confidential, because, if falling to the wrong hands, the information could be used to attack the companies' computer systems. For example, telling other people that the number of break-ins is high is a sign of saying the company is easy to break in, and disclosing the most common successful attacks tells other people the right way to break into the company's computer systems.

This problem can be modeled as an a privacy preserving survey problem. Currently, in the survey situations where confidential questions are asked, two common strategies are adopted: The first approach is to assume the trustworthiness of the interviewer, or to assume the existence of a trusted third party. Such an assumption is quite risky in nowadays' dynamic and malicious environment. Furthermore, even for a trusted interviewer, if the confidential information collected from the interviewees is accidentally disclosed (perhaps by a disgruntled employee of the interviewer's company, or as a consequence of a system break-in), the interviewer might face expensive lawsuits from the interviewees. From this perspective, a trusted interviewer might even prefer not to know the actual answers from the interviewees if it can still conduct the normal statistical analysis. Therefore protocols that can support survey while protecting the participants' privacy are of growing importance.

The second commonly used approach is to use anonymous technique: each interviewee sends back their answers anonymously to the interviewer using a hard copy or using any of the anonymous communication protocol proposed in the literature [1, 5]. However, this straightforward use of anonymous reply does not guarantee that the results come from the intended interviewees; anyone else who knows the ongoing survey can make arbitrary answers and anonymously send it back to the interviewer. This renders the results more or less untrusted. Of course, more sophisticated approaches could be used to solve the above drawback, however we suspect that communications among the interviewees might be required, which is undesirable in the real life.

The goal of this paper is to investigate another technique, the random response technique, and discuss how it can be used to conduct such type of privacy-preserving survey. Random response technique was initially proposed by Warner [6, 7, 2, 4] to compute the mean value of the sample data. We extend the random response technique also to various other standard statistical analysis operations, such as computing standard deviation, correlation coefficient, and linear regression line. Another goal of this paper is to apply this technique to the intrusion detection area for collection of information related to intrusions.

2 Privacy-Preserving Survey Problem

In this section, we will formally define the *Privacy-Preserving Survey Problem*.

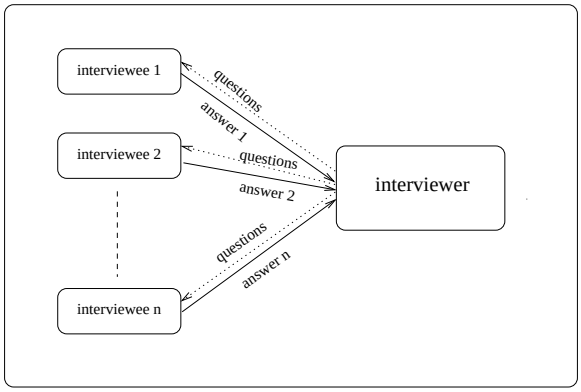

Figure 1: Models

Problem 1. *(Privacy-Preserving Survey Problem)* To conduct a survey, an interviewer sends out questions to many interviewees; each interviewee is supposed to send answers back to the interviewer (the answers could be quantitative answers or `yes/no` answers). The interviewer, after collecting all the answers, wants to conduct certain statistical analysis: if the answers are quantitative answers, it wants to calculate the sum,

mean, standard derivation, correlation and regression; if the answers are `yes/no` answers, it wants to count the number of `yes` answers and the number of `no` answers. Throughout the whole survey process, the following constraints should be satisfied:

1. The interviewer should not learn the exact answers of each interviewee.
2. No interviewee should learn the exact answers of other interviewees.
3. Interviewees are not supposed to communicate with each other.

The first two constraints guarantee the privacy of each interviewee's answer. The last constraint is necessary in the real world because it is undesirable to have the participants communicate with each other during a survey because of the scalability problem and the anonymity issue: in the anonymous survey situation, the participants' identities are not supposed to be revealed. Figure 1 describes such a survey process.

3 Solutions

3.1 The Encrypted Circuit Approach

This problem is a special case of the general secure multi-party computation problem (SMC) [8, 3]. Goldreich proposed a general approach, the encrypted circuit, to solve the general secure multi-party computation problem in [3].

Using this approach, the interviewer can apply for an encrypted circuit from a survey issuer (need not be a trusted party, however, the interviewer and the survey issuer cannot collude with each other). The encrypted circuit is designed to conduct n additions while not revealing to the interviewer the inputs from the interviewees and any meaningful intermediate results. The interviewer, cooperating with the interviewees, can use the encrypted circuit to get the sum of all inputs.

Although this approach can gain the accurate results while achieving privacy constraints, it comes with its cost: first of all, the communication cost between the interviewer and the survey issuer is $O(cn \log L)$, where n is the number of participants, and L is the largest number among the replies, and c is a non-negligible constant associated with the encrypted circuit. For example, for $n = 1000$, $L = 2^{20}$ the circuit size could be in the order of Mega-bytes. Secondly, each interviewee has to participate in multiple rounds of communication with interviewer. Therefore, if the number of interviewees are small, this technique tends to be a good approach because it can produce accurate results. We consider this technique as the complement to the one we are going to described in the following because the next one actually requires a larger number of interviewees in order to get more accurate results.

3.2 Randomized Response

Quantitative Answer

Observation 1. If every interviewee generates a random number according to certain pre-agreed random number generation parameters, the mean of these random numbers tends to be a pre-determined number if the number of interviewees is large enough. For example, if random numbers are generated uniformly in $[-m, m]$, the mean of these random numbers will be 0.

Based on this observation, each interviewee could add a random number to its actual answer, thus hiding the actual answer. Then, from the interviewer's point of view, although it does not know each actual answer, it can still estimate the mean value of the actual answers.

For example, let x_i be the actual answer from the ith interviewee, for $i = 1, \ldots, n$, and r_i be the random number generated by the interviewee; let $\bar{x}$ be the mean value of x_i's and $\bar{r}$ be the mean value of r_i's. After the data collection, the interviewer will get:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i + r_i) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} r_i = \bar{x} + \bar{r}$$

Next we will show how to use this random response scheme to compute other statistical values, such as standard deviation σ, correlation coefficient r and linear regression b.

For the standard deviation

$$\begin{aligned} \sigma &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \\ &= \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} \end{aligned}$$

to compute σ, in addition to getting the mean value of x_i's, the interviewer needs also to get the value of $\sum_{i=1}^{n} x_i^2$. Using the same random response technique, interviewees make it possible for the interview to compute $\sum_{i=1}^{n} x_i^2$ by sending back their x_i^2 disguised by random numbers.

Similar methods could be used to compute the correlation coefficient and linear regression line if the interviewer asks two related questions, whose relationship is the subject of the survey. For example, let (x_i, y_i) be the actual answer for the ith interviewee, according to the following equations,

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^{n} x_i \cdot y_i - n * \bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}} \end{aligned}$$

the interviewer can compute the correlation coefficient r if it knows $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i^2$, $\sum_{i=1}^{n} x_i \cdot y_i$ and mean values $\bar{x}$ and $\bar{y}$ all of which can be obtained using the randomized response technique without disclosing the actual values of x_i and y_i. Moreover, knowing these numbers also allows the interviewer to compute the linear regression line $y = bx + (\bar{y} - b\bar{x})$, where

$$b = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

In what follows, we assume there are two questions asked by the interviewer, and each interviewee sends back a tuple (x, y) to the interviewer, where x and y are two numbers. The protocol works for one-question situation by ignoring y; it can be straightforwardly extended to an n-question scenario as well.

1. The interviewer sends the questions and parameters for the random number generation to the interviewees.

2. For the interviewee i, if the exact answers to the questions is (x_i, y_i), it generates five random numbers $r_{i,1}$, $r_{i,2}$, $r_{i,3}$, $r_{i,4}$ and $r_{i,5}$ according to the parameters from the interviewer; then the interviewee sends $x_i + r_{i,1}$, $x_i^2 + r_{i,2}$, $y_i + r_{i,3}$, $y_i^2 + r_{i,4}$, and $x_i \cdot y_i + r_{i,5}$ to the interviewer.

3. The interviewer can conduct the following statistical analysis:

(a) sum of x: $s_x = \sum_{i=1}^{n} x_i = \sum_{i=1}^{n}(x_i + r_{i,1}) - \sum_{i=1}^{n} r_{i,1}$.

(b) sum of y: $s_y = \sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(y_i + r_{i,3}) - \sum_{i=1}^{n} r_{i,3}$

(c) mean: $\bar{x} = \frac{s_x}{n}$ and $\bar{y} = \frac{s_y}{n}$.

(d) sum of x^2: $s_{x^2} = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n}(x_i^2 + r_{i,2}) - \sum_{i=1}^{n} r_{i,2}$.

(e) sum of y^2: $s_{y^2} = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n}(y_i^2 + r_{i,4}) - \sum_{i=1}^{n} r_{i,4}$.

(f) sum of xy: $s_{xy} = \sum_{i=1}^{n} x_i \cdot y_i = \sum_{i=1}^{n}(x_i \cdot y_i + r_{i,5}) - \sum_{i=1}^{n} r_{i,5}$.

(g) standard deviation: $\sigma = \sqrt{\frac{s_{x^2} - n\bar{x}}{n-1}}$.

(h) correlation coefficient: $r = \frac{s_{xy} - n * \bar{x}\bar{y}}{\sqrt{(s_{x^2} - n\bar{x}^2)(s_{y^2} - n\bar{y}^2)}}$

(i) regression line: $y = bx + (\bar{y} - b\bar{x})$, where $b = \frac{s_{xy} - \frac{s_x s_y}{n}}{s_{x^2} - \frac{s_x^2}{n}}$

Yes/No Answer

1. Ursula sends the questions and a parameter p to the interviewees.

2. Each interviewee prepares a biased "coin" with a head and a tail on each side. If flipping the coin, the interviewee has p possibility to get the head and $1 - p$ possibility to get the tail.

3. Each interviewee i will first flip the "coin" before answering the question. If he gets the head, he will tell the truth; if he gets the tail, he will tell a lie.

4. Suppose the number of `yes` and `no` the interviewer gets is u and v respectively, the estimate of the number of actual `yes` answers u' and the estimate of the number of actual `no` answers v' can be computed from the following:

$$u = u'p + (n - u')(1 - p)$$

$$v = v'p + (n - v')(1 - p)$$

Warner gave a detailed analysis of the choice of p in [6], and we summarize the result here. The choice of p is important in this protocol. If $p = 1/2$, the interviewer will gain nearly nothing from the survey because he will get about 50% `yes` answers and another 50% `no` answers. If $p = 1$ or $p = 0$, i.e., everybody tells the truth or tells a lie, the result the interviewer gets will always be accurate. For p between $1/2$ and 1 (or between $1/2$ and 0) the interviewee provides useful but not absolute information as to exactly what the actual answer is. p's being more close to $1/2$ means that a larger sample size is required to get more accurate results.

4 Applications

IRDB project, an incident response database for gathering cost and incidence information on types of security events, is an ongoing project using the traditional survey technique. The outcome is not satisfactory because of the privacy concerns of its participants. We believe the outcome could be improved if it takes advantage of our purposed survey technique.

The IRDB project attempts to provide a framework to record incident information. the incidents information includes a risk type and an attack type. The risk type expresses the consequences of the attack (e.g., root access). The attack type identifies kinds of attacks. One of the objective of this project is to collect statistical data from many different organizations, and assemble them into a coherent picture of incident

costs and frequencies on a national scale. In the current implementation, participants have to totally trusted the database maintainer in keeping their information confidential. Because of such a trust assumption, the number of participants are not significant enough for a meaningful statistical analysis.

Using our proposed randomized response technique, the participants do not need to trust the database maintainer. The database maintainer will not be able to derive the actual information from each single response, but as long as the number of participates are large enough, he can still achieve the statistical goal.

There are many other type of survey projects, whose tasks are to collect and analyze various kinds of intrusion information, such as vulnerabilities, exploits, damage of security breaches, situation of intrusion detection deployment, companies' budget devoted to the IT security, costs and the duration of the recovery and so on. We believe they could be able to benefit from applying the proposed randomized response technique if they can integrate this technique well with their projects.

5 Conclusion

We have presented a randomized response technique in this paper, and have discussed how it can be applied in intrusion detection for the purpose of understanding vulnerabilities, intrusions, recovery, and the global trend, pattern and situation of intrusions and vulnerabilities. The technique can also be used in other survey situations where participants have to answer sensitive questions that they are not willing to answer.

References

[1] P. F. Syverson, D. M. Goldschlag and M. G. Reed. Anonymous connections and onion routing. In *Proceedings of 1997 IEEE Symposium on Security and Privacy*, Oakland, California, USA, May 5-7 1997.

[2] M. S. Goodstadt and V.Gruson. The randomized response technique: A test on drug use. *Journal of the American Statistical Association*, 70(352):814–818, December 1975.

[3] O. Goldreich, S. Micali and A. Wigderson. How to play any mental game. In *Proceedings of the 19th annual ACM symposium on Theory of computing*, pages 218–229, 1987.

[4] K. H. Pollock and Y. Bek. A comparison of three randomized response models for quantitative data. *Journal of the American Statistical Association*, 71(356):994–886, December 1976.

[5] M. K. Reiter and A. D. Rubin. Crowds: anonymity for web transaction. *ACM Transactions on Information and System Security*, 1(1):Pages 66–92, 1998.

[6] S. L. Warner. Randomized response: A surey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, March 1965.

[7] S. L. Warner. Randomized response: A surey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 66(336):884–888, December 1971.

[8] A. Yao. Protocols for secure computations. In *Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science*, 1982.