Privacy-Aware Negative Surveys with a Hidden Category in Mobile Sensing

Shoko Minagawa
Graduate School of Media and Governance
Keio University, Japan
choco@sfc.keio.ac.jp

Jin Nakazawa and Hideyuki Tokuda
Faculty of Environment and Information Studies
Keio University, Japan

ABSTRACT

The global spread of mobile phones creates a new vision in the world. It is called mobile sensing, in which human beings are regarded as sensors to produce aggregated models and knowledge. In this setting, it is likely that user privacy is violated. Therefore, we investigate privacy-preserving methods and propose a privacy-aware method to construct statistical data in mobile sensing.

1. INTRODUCTION

With the spread of mobile devices equipped with various sensors, much attention has been paid to mobile sensing applications. Almost all of these applications aim at constructing statistical data such as urban congestion levels, by analyzing location data and environmental data collected from many mobile devices. Data sensed by each mobile device is related to the user and his/her activities, and so user privacy is a problem that cannot be ignored in mobile sensing. Among the solutions of this privacy issue, encryption, k-anonymity and differential privacy are proposed. However, these methods still allow malicious administrators or system crackers to steal the private data. To solve this problem, we focus on Negative Surveys (NS) [1, 2] and propose NS with a hidden category. In NS, each mobile device sends a data different from sensing data, called a false data, to a server, and the server reconstructs statistical data from the false data set. User privacy can be protected since the sensing data of each mobile device is not sent to the outside of the mobile device. Furthermore, by adjusting the way to choose a false data, our method improves reconstruction accuracy of statistical data as compared to existing methods.

2. NEGATIVE SURVEYS

Negative Surveys (NS) [1, 2] has the node protocol (sensing, negation and transmission) and the base station protocol (counting and reconstruction). Suppose there is a sensing task consisting of m categories shown as $C = \{c_0, c_1, ..., c_{m-1}\}$. k of c_k is called a category ID.

MOBIQUITOUS 2014, December 02-05, London, Great Britain

DOI 10.4108/icst.mobiquitous.2014.257915

Sensing A user chooses c_i, or his/her mobile device senses data belongs to c_i. c_i is called a true data.

Negation The mobile device chooses c_j from $m-1$ categories other than c_i with the probability of $\frac{1}{m-1}$ at random. c_j is called a false data.

Transmission The mobile device sends c_j to a server.

Counting The server counts f_k which is the number of mobile devices having chosen c_k as a false data. Its set is called a false data set $F = \{f_0, f_1, ..., f_{m-1}\}$. The number of samples is $N = \sum_{k=0}^{m-1} f_k$.

Reconstruction The server computes t'_k which is an estimation number of mobile devices having got c_k as a true data, by using $\forall_k \mid t'_k = N-(m-1)\cdot f_k$. Its set is called a reconstructed data set $T' = \{t'_0, t'_1, ..., t'_{m-1}\}$.

In addition, t_k stands for a real number of mobile devices having got c_k as a true data. Its set is called a true data set $T = \{t_0, t_1, ..., t_{m-1}\}$. Note that T cannot be observed in NS, but is needed to calculate reconstruction accuracy (RA), which shows the difference between T and T'. When P and P' are discrete probability distribution of T and T' respectively, the Jensen-Shannon divergence of P' from P is shown as $D_{js}(P||P')$. The Jensen-Shannon divergence is symmetric, and takes any value between zero and one when it uses the base 2 logarithm. RA (%) is calculated by $\{1 - D_{js}(P||P')\} \times 100$. There are cases in which RA decreases: when the number of categories has increased for the same number of samples, or when the number of samples has decreased for the same number of categories.

3. APPROACH

Our goal is to solve the problem that RA decreases in NS, reducing the number of candidates for the false data of a true data. As an existing method having the same goal, there is Multidimensional Negative Surveys (MNS) [3]. Though MNS improves RA factorizing the number of categories, the ways of factorization are limited, especially when the number of categories is a prime number or a number including a large prime number (e.g. 26). Therefore, we introduce a hidden category to MNS. A hidden category is a category chosen as a false data, but not chosen as a true data. For example, if the number of categories m is 11, it cannot be factorized. If a hidden category is added to 11 categories, the number of categories m' becomes 12. Figure 1 shows a case of factorizing 12 into 3 and 4. Each category ID is shown as a vector. When a true data is $c_5 = c_{(1,1)}$, the false data is chosen applying Negation in NS to each element of

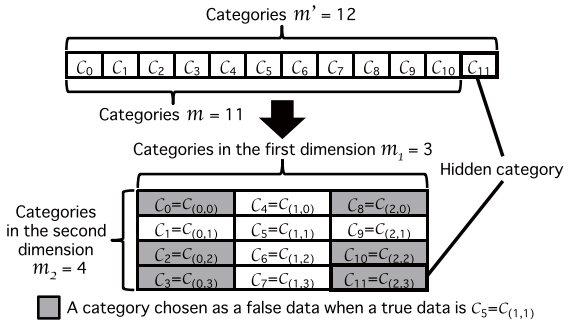

Figure 1: A hidden category and dimensions

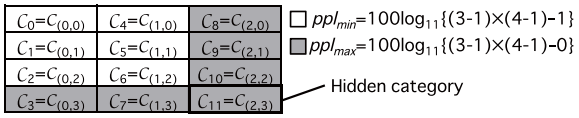

Figure 2: Privacy-preserving level

the vector. The number of candidates for the false data becomes $(3-1)\times(4-1)=6$, and the RA is the same as the case in which the number of categories is $6+1=7$ in NS. We generalize the above example. When d is the number of dimensions and m_l is the number of categories in the lth dimension, c_k is shown as $c_{(v1_k,v2_k,...,vd_k)}$, in which each vl_k takes on integer values between 0 and m_l-1. We obtain the following node protocol.

Sensing A user chooses $c_{(v1_i,v2_i,...,vd_i)}$, or his/her mobile device senses data belongs to $c_{(v1_i,v2_i,...,vd_i)}$.

Negation The mobile device chooses $c_{(v1_j,v2_j,...,vd_j)}$. The lth element vl_j is chosen from m_l-1 categories other than vl_i with the probability of $\frac{1}{m_l-1}$ at random.

Transmission The mobile device sends $c_{(v1_j,v2_j,...,vd_j)}$ to a server.

In the base station protocol, an equation that is based on the reconstruction equation in NS can be used [3]. The number of candidates for the false data of a true data is shown by $\alpha=\prod_{k=1}^{d}(m_k-1)$. The RA is the same as the case in which the number of categories is $\alpha+1$ in NS. Though RA is improved in this way, there is a trade-off between RA and privacy-preserving level (PPL). PPL represents the level that the true data can be inferred from a false data. It is measured by conditional entropy and maximum entropy. Suppose there are $m'=m+1$ categories shown as $C=\{c_0,c_1,...,c_{m-1},c_{m'-1}\}$, a false data c_j and the number of candidates for the true data of c_j, β. The entropy of C conditioned on c_j is given by $H(C|c_j)=-\log_2\frac{1}{\beta}=\log_2\beta$. Generally, the entropy becomes maximum when all events appear with the equal probability. The maximum entropy of C conditioned on c_j is shown as $H(C|c_j)_{max}=\log_2 m$. From the above equations, PPL (%) is defined as $\frac{H(C|c_j)}{H(C|c_j)_{max}}\times 100=100\log_m\beta$. As illustrated in Figure 2, when each white cell is a false data, β becomes $\alpha-1$, for a hidden category is included into the candidates for the true data. When each gray cell is a false data, β is equal to α, for a hidden category is not included. We adopt ppl_{min} as the standard PPL.

Table 1: Simulation results

Method	Factorization	RA (%)	PPL (%)
NS, MNS	[23]	83.63	98.58
Proposed	[4,6]	94.02	84.17
Proposed	[3,8]	94.63	81.80
Proposed	[2,12]	94.95	73.44
Proposed	[2,3,4]	99.10	51.33
Proposed	[2,2,6]	99.11	44.21

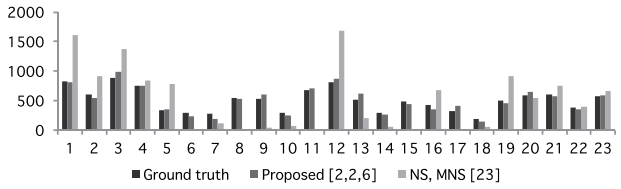

Figure 3: Population of the 23 special wards

4. SIMULATION

We estimate the population of the 23 special wards of Tokyo, Japan. The ground truth population data can be obtained from National Census 2010 in Japan. The total population in the area is about ten million. 0.1 (%) of the population data, that is, about ten thousand is used in this simulation. We apply Negation to each person of the population data, and calculate RA and PPL. The results of the simulation are shown in Table 1. In our approach, RA can be improved by creating a hidden category even though the number of categories is a prime number. It is useful for collecting data when a target area is divided into a prime number or a number including a large prime number. Though PPL decreases in accordance with the improvement of RA, system administrators can choose the way to factorize the number of categories while considering application types, the number of samples, RA and PPL. Figure 3 shows the true data set T, the reconstructed data set T' obtained by our approach, and T' obtained by NS and MNS.

5. CONCLUSION AND FUTURE WORK

We investigate privacy-preserving methods and propose a privacy-aware method to construct statistical data in mobile sensing. Our method improves reconstruction accuracy of statistical data regardless of the number of divisions of an attribute, that is, the number of categories. We plan to confirm how many hidden categories can be created while considering privacy-preserving level. If several hidden categories are created, our method can be used widely.

6. REFERENCES

[1] F. Esponda and V. M. Guerrero. Surveys with negative questions for sensitive items. *Statistics & Probability Letters*, Vol. 79, No. 24, pp. 2456–2461, 2009.

[2] J. Horey, M. M. Groat, S. Forrest, and F. Esponda. Anonymous data collection in sensor networks. In *MobiQuitous*, pp. 1–8, 2007.

[3] M. M. Groat, B. Edwards, J. Horey, W. He, and S. Forrest. Enhancing privacy in participatory sensing applications with multidimensional data. In *PerCom*, pp. 144–152, 2012.